Statistica: principi e metodi

Capitolo 5

Variabilità

Variabilità

Per variabilità si intende l'attitudine dei fenomeni, naturali e sociali, a manifestarsi in modi differenti.

- La variabilità è l'attitudine di un carattere a presentare modalità differenti nel collettivo in esame.
- La distribuzione di un carattere presenta variabilità nulla se su tutte le unità statistiche si rileva la stessa modalità. In tal caso tutti gli indici di variabilità assumono valore zero.

Variabilità

Costituisce la ragione stessa dell'esistenza della Statistica.

Tutta la metodologia statistica ha a che fare con la variabilità,

- per "**neutralizzarla**", con le medie, che mirano a far emergere "il costante nel variabile",
- per "**misurarla**", con gli indici di variabilità, oggetto di questo capitolo,
- per "**spiegarla**", con le tecniche della regressione e, in generale, con i metodi dell'inferenza statistica.

Variabilità

Fonti di variabilità

il fenomeno si manifesta su entità distinte

i caratteri osservati in unità distinte si manifestano con modalità diverse per una serie di circostanze e di cause

osservazione ripetuta di una stessa grandezza

i livelli rilevati sono in genere diversi in relazione a una molteplicità di elementi (l'ora della giornata, l'accuratezza della lettura dello strumento, l'eventuale assunzione di farmaci e così via)

Variabilità per caratteri di natura qualsiasi: Eterogeneità

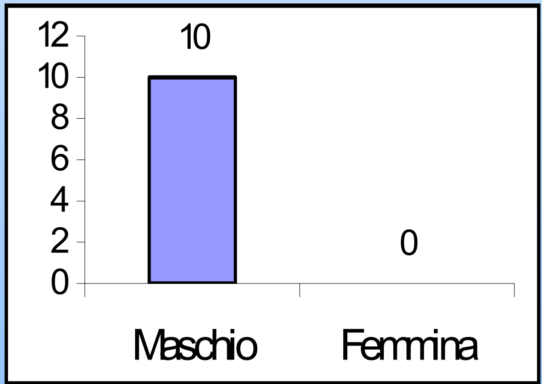
- Misura la variabilità nelle distribuzioni secondo caratteri qualitativi o quantitativi
- **minima eterogeneità (massima omogeneità):** tutte le unità del collettivo hanno la stessa modalità del carattere
- **massima eterogeneità** (**minima omogeneità**) le modalità presentano tutte la stessa frequenza.

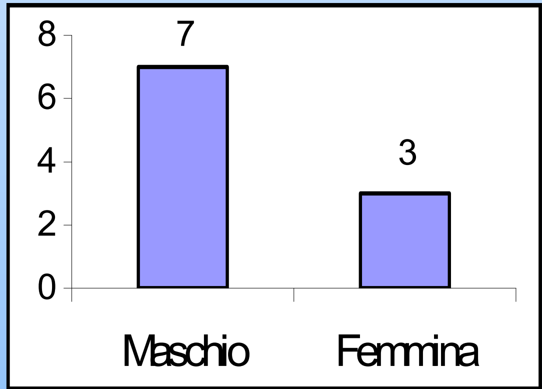
Eterogeneità: esempi

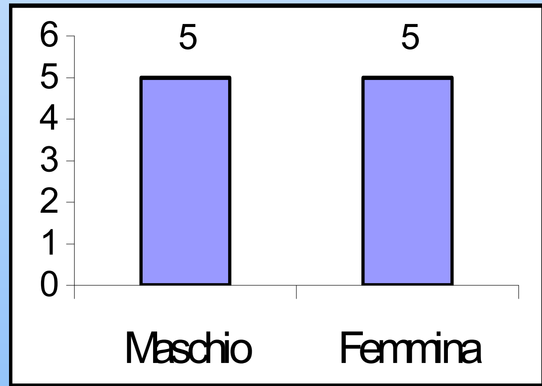
x_i	n_i
Maschio	10
Femmina	0
Totale	10

x_i	n_i
Maschio	7
Femmina	3
Totale	10

x_i	n_i
Maschio	5
Femmina	5
Totale	10







Eterogeneità nulla

Eterogeneità massima

Indice di eterogeneità

$$E = 1 - \sum_{i=1}^{k} f_i^2$$

valore minimo **0** (*eterogeneità nulla*)

valore massimo $\left(\frac{k-1}{k}\right)$ (*eterogeneità massima*)

x_i	n_i	f_i	f_i^2
Maschio	10	1	1
Femmina	0	0	0
Totale	**10**	**1**	**1**

$$E = 1 - \sum_{i=1}^{k} f_i^2 = 1 - 1 = 0$$

x_i	n_i	f_i	f_i^2
Maschio	7	0,7	0,49
Femmina	3	0,3	0,09
Totale	**10**	**1**	**0,58**

$$E = 1 - \sum_{i=1}^{k} f_i^2 = 1 - 0.58 = 0.42$$

x_i	n_i	f_i	f_i^2
Maschio	5	0,5	0,25
Femmina	5	0,5	0,25
Totale	**10**	**1**	**0,5**

$$E = 1 - \sum_{i=1}^{k} f_i^2 = 1 - 0.5 = 0.5$$

Indice di eterogeneità esempio

$$E = 1 - \sum_{i=1}^{k} f_i^2$$

	Frequenze	Frequenze relative	
	n_i	f_i	f_i^2
Licenza elementare o nessun titolo	19766	0.365	0.133
Licenza media	16554	0.306	0.093
Qualifica professionale	2541	0.047	0.002
Diploma di maturità	11749	0.217	0.047
Dottorato, laurea o diploma universitario	3547	0.065	0.004
Totale	**54157**	**1**	**0.280**

$$E = 1 - 0.280 = 0.720$$

$$\max(E) = \frac{k-1}{k} = \frac{5-1}{5} = 0.80$$

L'indice di eterogeneità relativo si calcola dividendo l'indice di eterogeneità per il suo massimo

$$e = \frac{E}{\max E} = \frac{0.72}{0.80} = \frac{5-1}{5} = 0.9$$

Variabilità per caratteri quantitativi

- Indici di dispersione rispetto alla media

- Indici di disuguaglianza.

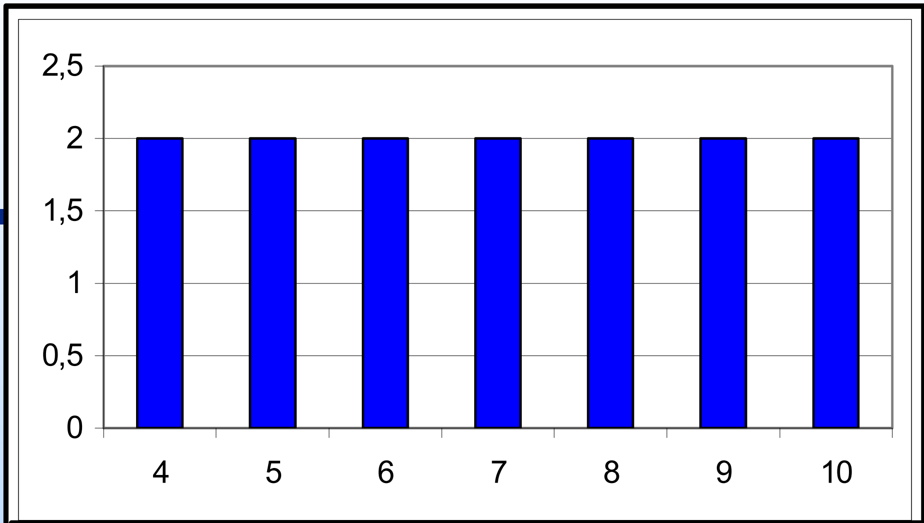
Dispersione rispetto alla media aritmetica

- La dispersione rappresenta la variabilità delle modalità presentata del carattere rispetto ad un valore di sintesi scelto, ad esempio la media aritmetica della distribuzione

- Gli indici che danno una misura della variabilità dei valori della distribuzione rispetto alla media aritmetica sono lo **scostamento semplice medio** e lo **scostamento quadratico medio**

Voto	n_i
4	2
5	2
6	2
7	2
8	2
9	2
10	2
Totale	**14**

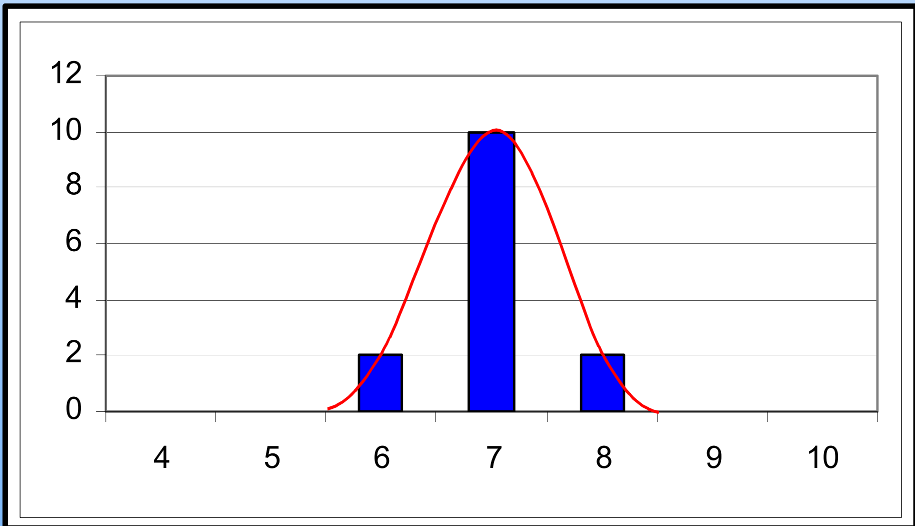
$\overline{x} = 7$



Alta DISPERSIONE attorno alla media

Voto	n_i
4	0
5	0
6	2
7	10
8	2
9	0
10	0
Totale	**14**

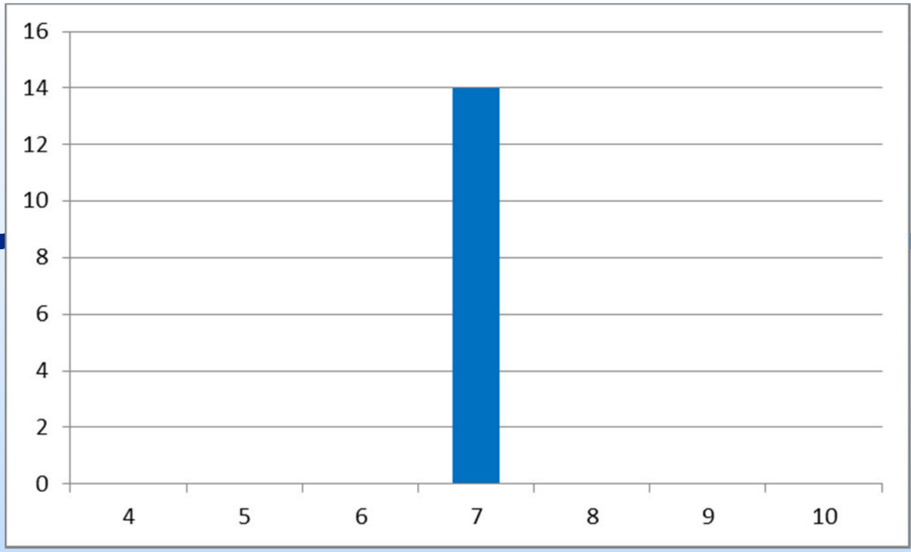
$\overline{x} = 7$



Valori più **ADDENSATI** attorno alla media

La media fornisce una misura di sintesi delle due distribuzioni ma non le rappresenta nello stesso modo. Occorre considerare anche gli indici di variabilità

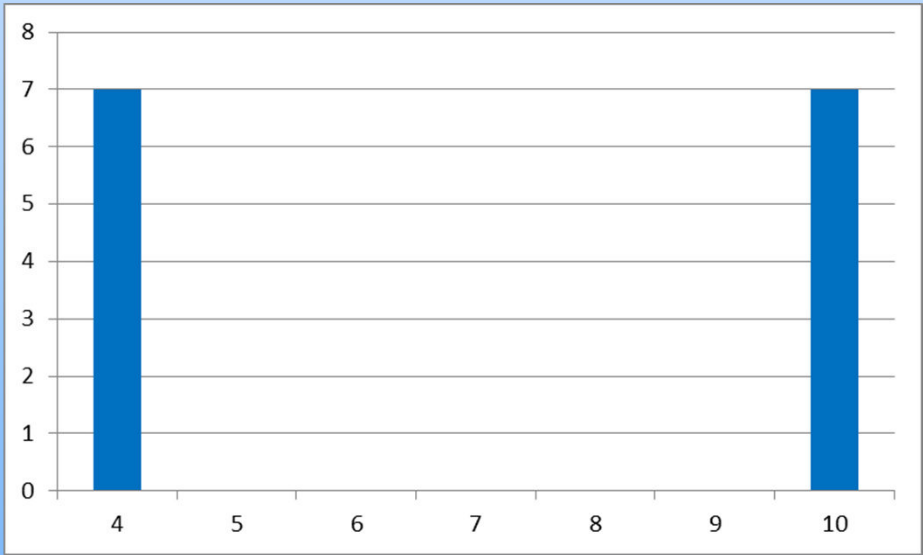
Voto	n_i
4	0
5	0
6	0
7	14
8	0
9	0
10	0
Totale	**14**



$$\bar{x} = 7$$

DISPERSIONE NULLA

Voto	n_i
4	7
5	0
6	0
7	0
8	0
9	0
10	7
Totale	**14**



$$\bar{x} = 7$$

DISPERSIONE MASSIMA

La media fornisce una misura di sintesi delle due distribuzioni ma non le rappresenta nello stesso modo. Occorre considerare anche gli indici di variabilità

Misura della variabilità: scostamento semplice medio

- Lo **scostamento semplice medio** è la **media aritmetica degli scarti dalla media presi in valore assoluto**

Distribuzione disaggregata

$$S_\mu = \frac{|x_1-\mu|+|x_2-\mu|+\cdots+|x_N-\mu|}{N}$$

$$=\frac{1}{N}\sum_{i=1}^{N}|x_i-\mu|.$$

Distribuzione di frequenza

$$S_\mu = \frac{|x_1-\mu|\cdot n_1+|x_2-\mu|\cdot n_2+\cdots+|x_k-\mu|\cdot n_k}{N}$$

$$=\frac{1}{N}\sum_{i=1}^{k}|x_i-\mu|\cdot n_i=\sum_{i=1}^{k}|x_i-\mu|\cdot f_i$$

Scostamento semplice medio per la distribuzione disaggregata: calcolo

$$S_\mu = \frac{1}{N}\sum_{i=1}^{N} |x_i - \mu|$$

x_i	$\lvert x_i - \mu \rvert$
0	7.14
5	2.14
6	1.14
8	0.86
9	1.86
10	2.86
12	4.86
Totale	20.86

$\mu = 7.14$ ***minuti***

Ritardi (in minuti) di un treno a lunga percorrenza alla stazione di Roma Termini, registrati in un campione di 7 osservazioni:

0, 5, 6, 8, 9, 10, 12

- *Scostamento semplice medio*:

$$S_\mu = \frac{|0-7.14| + |5-7.14| + \cdots + |12-7.14|}{7}$$

$$= \frac{20.86}{7} = 2.98 \quad \textbf{\textit{minuti}}$$

Scostamento semplice medio per una distribuzione di frequenze a modalità singole: calcolo

$$S_\mu = \frac{1}{N}\sum_{i=1}^{k} |x_i - \mu| \cdot n_i$$

Voti agli esami ottenuti da un laureando triennale

x_i	n_i	$x_i \cdot n_i$	$\lvert x_i - \mu \rvert$	$\lvert x_i - \mu \rvert \cdot n_i$
23	1	23	3.77	3.77
24	3	72	2.77	8.30
25	4	100	1.77	7.07
26	7	182	0.77	5.37
27	5	135	0.23	1.17
28	4	112	1.23	4.93
29	1	29	2.23	2.23
30	5	150	3.23	16.17
Totale	30	803		49.01

- Media aritmetica

$$\mu = \frac{23 \cdot 1 + 24 \cdot 3 + \cdots + 30 \cdot 5}{30} = \frac{803}{30} = 26.77$$

- *Scostamento semplice medio*

$$S_\mu = \frac{|23 - 26.77| \cdot 1 + \cdots + |30 - 26.77| \cdot 5}{30} = \frac{49.01}{30} = 1.63$$

Misura della variabilità: scostamento quadratico medio

Lo **scostamento quadratico medio** o **deviazione standard** è la **media quadratica degli scarti dalla media**

$$\sigma = \sqrt{\frac{(x_1-\mu)^2+(x_2-\mu)^2+\cdots+(x_N-\mu)^2}{N}} = \sqrt{\frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}}$$

Distribuzione disaggregata

$$\sigma = \sqrt{\frac{(x_1-\mu)^2 n_1+(x_2-\mu)^2 n_2+\cdots+(x_k-\mu)^2 n_k}{N}} = \sqrt{\frac{\sum_{i=1}^{k}(x_i-\mu)^2 n_i}{N}}$$

Distribuzione di frequenza

Formula alternativa

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

Scostamento quadratico medio per la distribuzione disaggregata: calcolo

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

x_i	$(x_i - \mu)^2$
0	51.02
5	4.59
6	1.31
8	0.73
9	3.45
10	8.16
12	23.59
Totale	92.86

$\mu = 7.14$ <u>*minuti*</u>

Ritardi (in minuti) di un treno a lunga percorrenza alla stazione di Roma Termini, registrati in un campione di 7:

0, 5, 6, 8, 9, 10, 12

- *Scostamento quadratico medio:*

$$\sigma = \sqrt{\frac{(0-7.14)^2 + (5-7.14)^2 + \cdots + (12-7.14)^2}{7}}$$

$$= \sqrt{\frac{92.86}{7}} = 3.64$$ <u>*minuti*</u>

σ per una distribuzione di frequenze a modalità singole: calcolo

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

Voti agli esami ottenuti da un laureando triennale

x_i	n_i	$x_i \cdot n_i$	$(x_i - \mu)^2$	$(x_i - \mu)^2 \cdot n_i$
23	1	23	14.19	14.19
24	3	72	7.65	22.96
25	4	100	3.12	12.48
26	7	182	0.59	4.11
27	5	135	0.05	0.27
28	4	112	1.52	6.08
29	1	29	4.99	4.99
30	5	150	10.45	52.27
Totale	30	803		117.35

- Media aritmetica

$$\mu = \frac{803}{30} = 26.77$$

- *Scostamento quadratico medio*

$$\sigma = \sqrt{\frac{(23-26.77)^2 \cdot 1 + \cdots + (30-26.77)^2 \cdot 5}{30}} = \sqrt{\frac{117.35}{30}} = 1.98$$

σ per una distribuzione di frequenze a modalità raggruppate in classi: calcolo

$$\sigma = \sqrt{\frac{\sum_{i=1}^{k}(\bar{x}_i - \mu)^2 n_i}{N}}$$

Classe	n_i	$\bar{x}_i$	$\bar{x}_i \cdot n_i$	$(\bar{x}_i - \mu)^2$	$(\bar{x}_i - \mu)^2 \cdot n_i$
19-21	31	20.0	620.0	3.28	101.67
21-24	45	22.5	1012.5	0.47	21.36
24-27	5	25.5	127.5	13.61	68.04
27-30	1	28.5	28.5	44.74	44.74
Totale	82		1788.5		235.81

Età degli studenti di un corso di Statistica

- Media aritmetica

$$\mu = \frac{1788.5}{82} = 21.81$$ ***anni***

- Scostamento quadratico medio

$$\sigma = \sqrt{\frac{(20.0-21.81)^2 \cdot 31 + (22.5-21.81)^2 \cdot 45 + \cdots + (28.5-21.81)^2 \cdot 1}{82}}$$

$$= \sqrt{\frac{235.81}{82}} = 1.70$$ ***anni***

Proprietà degli indici S_μ e σ

- Assumono il valore 0 nel caso di assenza di variabilità (*tutte le unità statistiche presentano la stessa modalità del carattere*)

- **Sono espressi nella stessa unità di misura del carattere considerato**

Proprietà degli indici S_μ e σ

- Non cambiano se a ciascun termine della distribuzione si aggiunge una quantità costante positiva o negativa
  - Se la distribuzione disaggregata $x_1, x_2, \ldots, x_N$, ha scostamento semplice medio S_μ e scostamento quadratico medio σ, la distribuzione $y_i=x_i+a$ ($y_1=x_1+a$, $y_2=x_2+a$, …, $y_N=x_N+a$) avrà scostamento semplice medio S_μ e scostamento quadratico medio σ

$$S_{\mu(y)} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \mu_y| = \frac{1}{N}\sum_{i=1}^{N}|x_i + a - (\mu_x + a)| = \frac{1}{N}\sum_{i=1}^{N}|x_i - \mu_x| = S_{\mu(x)}$$

$$\sigma_{(y)} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \mu_y)^2} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i + a - (\mu_x + a))^2} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_x)^2} = \sigma_{(x)}$$

Proprietà degli indici S_μ e σ

- La moltiplicazione di ciascun termine della distribuzione per una costante, positiva o negativa, ha come conseguenza la moltiplicazione degli indici per il valore assoluto della costante
  - Se la distribuzione disaggregata $x_1, x_2, \ldots, x_N$, ha scostamento semplice medio S_μ e scostamento quadratico medio σ, la distribuzione $y_i=bx_i$ ($y_1=bx_1$, $y_2=bx_2$, …, $y_N=bx_N$) avrà scostamento semplice medio $|b|S_\mu$ e scostamento quadratico medio $|b|\sigma$

$$S_{\mu(y)} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \mu_y| = \frac{1}{N}\sum_{i=1}^{N}|bx_i - b\mu_x| = \frac{|b|}{N}\sum_{i=1}^{N}|x_i - \mu_x| = |b|\,S_{\mu(x)}$$

$$\sigma_{(y)} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(y_i - \mu_y\right)^2} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(bx_i - b\mu_x\right)^2} = \sqrt{\frac{b^2}{N}\sum_{i=1}^{N}\left(x_i - \mu_x\right)^2} = |b|\,\sigma_{(x)}$$

Varianza

Il quadrato della deviazione standard si chiama varianza

(è espressa in unità di misura del carattere al quadrato).

$$\sigma^2 = \frac{(x_1-\mu)^2 + (x_2-\mu)^2 + \cdots + (x_N-\mu)^2}{N} = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2$$

Distribuzione disaggregata

$$\sigma^2 = \frac{(x_1-\mu)^2 n_1 + (x_2-\mu)^2 n_2 + \cdots + (x_k-\mu)^2 n_k}{N} = \frac{\sum_{i=1}^{k}(x_i-\mu)^2 n_i}{N}$$

Distribuzione di frequenza

Formula alternativa

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

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

Devianza

La somma dei quadrati degli scarti dalla media (numeratore della varianza) si chiama devianza
(*è espressa in* ***unità di misura del carattere al quadrato***)

$$D = (x_1 - \mu)^2 + (x_2 - \mu)^2 + \cdots + (x_N - \mu)^2 = \sum_{i=1}^{N} (x_i - \mu)^2$$

Distribuzione disaggregata

$$D = (x_1 - \mu)^2 n_1 + (x_2 - \mu)^2 n_2 + \cdots + (x_k - \mu)^2 n_k = \sum_{i=1}^{k} (x_i - \mu)^2 n_i$$

Distribuzione di frequenza

Varianza e devianza: calcolo per una distribuzione disaggregata

x_i	$(x_i - \mu)^2$
0	51.02
5	4.59
6	1.31
8	0.73
9	3.45
10	8.16
12	23.59
Totale	92.86

$\mu = 7.14$ ***minuti***

Ritardi (in minuti) di un treno a lunga percorrenza alla stazione di Roma Termini, registrati in un campione di 7:

0, 5, 6, 8, 9, 10, 12

$$\sigma = \sqrt{\frac{92.86}{7}} = 3.64$$ ***minuti***

$$\sigma^2 = \frac{92.86}{7} = 13.27$$ ***minuti²***

$$devianza = 92.86$$ ***minuti²***

Varianza e devianza per una distribuzione di frequenze a modalità singole: calcolo

Voti agli esami ottenuti da un laureando triennale

x_i	n_i	$xi \cdot ni$	$(x_i - \mu)^2$	$(xi - \mu)^2 \cdot n_i$
23	1	23	14.19	14.19
24	3	72	7.65	22.96
25	4	100	3.12	12.48
26	7	182	0.59	4.11
27	5	135	0.05	0.27
28	4	112	1.52	6.08
29	1	29	4.99	4.99
30	5	150	10.45	52.27
Totale	30	803		117.35

- Media aritmetica

$$\mu = \frac{803}{30} = 26.77$$

- Scostamento quadratico medio

$$\sigma = \sqrt{\frac{117.35}{30}} = 1.98$$

$$\sigma^2 = \frac{117.35}{30} = 3.91$$

$$devianza = 117.35$$

Indici di disuguaglianza

- La disuguaglianza evidenzia la diversità delle varie osservazioni tra loro
- **Gli indici di disuguaglianza** misurano la distanza fra tutte le modalità della distribuzione.
- Sono indici di disuguaglianza la **differenza semplice media** (*con o senza ripetizione*) e la **differenza quadratica media** (*con o senza ripetizione*)

Differenza semplice media per una distribuzione disaggregata

Si chiama **differenza semplice media** della distribuzione la **media aritmetica delle differenze in valore assoluto**, $|x_i - x_j|$, $i, j = 1, 2, \ldots, N$, tra le coppie di termini della distribuzione:

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

Senza ripetizione

$$\Delta_R = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} |x_i - x_j|$$

Con ripetizione

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

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

Differenza quadratica media per una distribuzione disaggregata

Si chiama **differenza quadratica media** della distribuzione la **media quadratica delle differenze**, $(x_i - x_j)$, $i, j = 1, 2, \ldots, N$, tra le coppie di termini della distribuzione:

$$\Delta^{(2)} = \sqrt{\frac{1}{N(N-1)} \sum_{\substack{i=1 \\ i \neq j}}^{N} \sum_{j=1}^{N} (x_i - x_j)^2}$$

Senza ripetizione

$$\Delta_R^{(2)} = \sqrt{\frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} (x_i - x_j)^2}$$

Con ripetizione

$$\Delta^{(2)} = \sqrt{\frac{N}{N-1}}\Delta_R^{(2)}$$

$$\Delta_R^{(2)} = \sqrt{\frac{N-1}{N}}\Delta^{(2)}$$

$$\Delta_R^{(2)} = \sqrt{2}\sigma$$

Differenze medie: formule operative

Le differenze sono a due a due uguali in quanto, per ogni coppia i, j, $|x_i - x_j| = |x_j - x_i|$, $(x_i - x_j)^2 = (x_j - x_i)^2$,
Ne segue che possiamo riscrivere la formule che semplificano i calcoli

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

$$\Delta_R = \frac{2}{N^2} \sum_{i=2}^{N} \sum_{j=1}^{i} |x_i - x_j|$$

$$\Delta^{(2)} = \sqrt{\frac{2}{N(N-1)} \sum_{i=2}^{N} \sum_{j=1}^{i-1} (x_i - x_j)^2}$$

$$\Delta_R^{(2)} = \sqrt{\frac{2}{N^2} \sum_{i=2}^{N} \sum_{j=1}^{i} (x_i - x_j)^2}$$

Differenza semplice media per una distribuzione disaggregata: calcolo

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

Ritardi (in minuti) di un treno a lunga percorrenza alla stazione di Roma Termini, registrati in un campione di 7 osservazioni (i valori vengono ordinati in senso crescente):

0, 5, 6, 8, 9, 10, 12

- Differenza semplice media mediante la formula operativa:

$$\Delta = \frac{2 \cdot 98}{7 \cdot 6} = \frac{196}{42} = 4.67 \text{ minuti}$$

$$\Delta_R = \frac{2 \cdot 98}{7 \cdot 7} = \frac{196}{49} = 4 \text{ minuti}$$

In forma tabellare

$|x_i - x_j|$

	0	5	6	8	9	10	12
0							
5	5						
6	6	1					
8	8	3	2				
9	9	4	3	1			
10	10	5	4	2	1		
12	12	7	6	4	3	2	

Totale 98

Differenza quadratica media per una distribuzione disaggregata : calcolo

$$\Delta^{(2)} = \sqrt{\frac{2}{N(N-1)} \sum_{i=2}^{N} \sum_{j=1}^{i-1} (x_i - x_j)^2}$$

Ritardi (in minuti) di un treno a lunga percorrenza alla stazione di Roma Termini, registrati in un campione di 7 osservazioni (i valori vengono ordinati in senso crescente):

0, 5, 6, 8, 9, 10, 12

- Differenza quadratica media mediante la formula operativa:

$$\Delta^{(2)} = \sqrt{\frac{2 \cdot 650}{7 \cdot 6}} = \sqrt{\frac{1300}{42}} = 5.56$$

minuti

$$\Delta_R^{(2)} = \sqrt{\frac{2 \cdot 650}{7 \cdot 7}} = \sqrt{\frac{1300}{49}} = 5.15$$

minuti

In forma tabellare

$(x_i - x_j)^2$

	0	5	6	8	9	10	12
0							
5	25						
6	36	1					
8	64	9	4				
9	81	16	9	1			
10	100	25	16	4	1		
12	144	49	36	16	9	4	

Totale 650

Differenze medie nel caso di distribuzioni di frequenze a modalità singole

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

$$\Delta_R = \frac{2}{N^2} \sum_{i=2}^{k} \sum_{j=1}^{i-1} |x_i - x_j| n_i n_j$$

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

$$\Delta_R^{(2)} = \sqrt{\frac{2}{N^2} \sum_{i=2}^{k} \sum_{j=1}^{i-1} (x_i - x_j)^2 n_i n_j}$$

Differenze medie nel caso di distribuzioni di frequenze in classi

- Quando le modalità sono raggruppate in classi, il calcolo delle differenze medie si effettua sulla base dei valori **centrali delle classi**

Differenza semplice media per una distribuzione di frequenze: calcolo

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

Voti agli esami ottenuti da un laureando triennale

Voto	N. di esami
23	1
24	3
25	4
26	7
27	5
28	4
29	1
30	5
Totale	**30**

In forma tabellare $|x_i - x_j| n_i n_j$

		23	24	25	26	27	28	29	30
		1	3	4	7	5	4	1	5
23	1								
24	3	3							
25	4	8	12						
26	7	21	42	28					
27	5	20	45	40	35				
28	4	20	48	48	56	20			
29	1	6	15	16	21	10	4		
30	5	35	90	100	140	75	40	5	
Totale	30								Totale

$\boxed{1003}$

- Differenza semplice media

$$\Delta = \frac{2}{30 \cdot 29} \cdot (|24-23| 3 \cdot 1 + \cdots + |30-29| 5 \cdot 1) = \frac{2 \cdot 1003}{30 \cdot 29} = 2.31$$

$$\Delta_R = \frac{2 \cdot 1003}{30 \cdot 30} = 2.23$$

Differenza quadratica media per una distribuzione di frequenze: calcolo

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

Voti agli esami ottenuti da un laureando triennale

Voto	N. di esami
23	1
24	3
25	4
26	7
27	5
28	4
29	1
30	5
Totale	**30**

In forma tabellare $(x_i - x_j)^2 n_i n_j$

		23	24	25	26	27	28	29	30
		1	3	4	7	5	4	1	5
23	1								
24	3	3							
25	4	16	12						
26	7	63	84	28					
27	5	80	135	80	35				
28	4	100	192	144	112	20			
29	1	36	75	64	63	20	4		
30	5	245	540	500	560	225	80	5	
Totale	30								3521

$$\Delta^{(2)} = \sqrt{\frac{2}{30 \cdot 29} \cdot \left[(24-23)^2 3 \cdot 1 + \cdots + (30-29)^2 5 \cdot 1) \right]} = \sqrt{\frac{2 \cdot 3521}{30 \cdot 29}} = 2.85$$

$$\Delta_R^{(2)} == \sqrt{\frac{2 \cdot 3521}{30^2}} = 2.80$$

Proprietà delle differenze medie

- Assumono il valore 0 nel caso di assenza di variabilità (tutte le unità statistiche presentano la stessa modalità del carattere)

- **Sono espresse nella stessa unità di misura del carattere considerato**

Proprietà delle differenze medie

- La differenza media non cambia se a ciascun termine della distribuzione si aggiunge una quantità costante positiva o negativa;
  - Se la distribuzione disaggregata $x_1, x_2, \ldots, x_N$, ha differenza semplice media Δ, la distribuzione $y_i=x_i+a$ ($y_1=x_1+a$, $y_2=x_2+a$, …, $y_N=x_N+a$) avrà differenza semplice medio Δ

$$\Delta_y = \frac{2}{N(N-1)} \sum_{i=2}^{N} \sum_{j=1}^{i-1} |y_i - y_j| = \frac{2}{N(N-1)} \sum_{i=2}^{N} \sum_{j=1}^{i-1} |x_i + a - x_j - a| =$$

$$\frac{2}{N(N-1)} \sum_{i=2}^{N} \sum_{j=1}^{i-1} |x_i - x_j| = \Delta_x$$

Lo stesso vale per le altre differenze medie

Proprietà delle differenze medie

- La moltiplicazione di ciascun termine della distribuzione per una costante, positiva o negativa, ha come conseguenza la moltiplicazione della differenza media per il valore assoluto della costante.
  - Se la distribuzione disaggregata $x_1, x_2, \ldots, x_N$, ha differenza semplice media Δ, la distribuzione $y_i=bx_i$ ($y_1=bx_1$, $y_2=bx_2$, …, $y_N=bx_N$) avrà scostamento semplice medio $|b|\,\Delta$

$$\Delta_y = \frac{2}{N(N-1)} \sum_{i=2}^{N} \sum_{j=1}^{i-1} |y_i - y_j| = \frac{2}{N(N-1)} \sum_{i=2}^{N} \sum_{j=1}^{i-1} |bx_i - bx_j| =$$

$$\frac{2|b|}{N(N-1)} \sum_{i=2}^{N} \sum_{j=1}^{i-1} |x_i - x_j| = |b|\,\Delta_x$$

Indici di variabilità relativa

Il problema si pone nel confrontare gli indici di variabilità di 2 o più distribuzioni *diverse*

In che senso diciamo "*diverse*"?

perché hanno **<u>diverse unità di misura</u>**

perché hanno **<u>diversa intensità media</u>**

Indice di variabilità relativa: numero puro a-dimensionale che elimina l'influenza dell'unità di misura e dell'intensità media

Indici di variabilità percentuali

- Si chiama **indice di variabilità percentuale** il rapporto, moltiplicato per 100, tra un indice di variabilità assoluto $(S_\mu, \sigma, \Delta, \Delta_c \text{ e } \Delta)$ e la media aritmetica.
- Particolare rilievo per le applicazioni ha **coefficiente di variazione**

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

Coefficiente di variazione: esempio

È maggiormente variabile il reddito medio annuo familiare o il numero di componenti della famiglia?

Reddito *(in migliaia)*	Media : 33364 euro Deviazione standard : 24636 euro
Numero di componenti	Media : 2.77 componenti Deviazione standard : 1.31 componenti

$$CV(reddito) = \frac{24636\ euro}{33364\ euro} 100 = 74\%$$

$$CV(componenti) = \frac{1.31\ componenti}{2.77\ componenti} 100 = 47\%$$

Coefficiente di variazione: esempio

Il reddito medio annuo è maggiormente variabile nell'insieme delle famiglie con 2 o con 4 componenti?

Reddito *(2 componenti)*	Media : 24451 euro Deviazione standard : 21218 euro
Reddito *(4 componenti)*	Media : 49260 euro Deviazione standard : 26050 euro

$$CV(reddito - 2componenti) = \frac{21218\ euro}{24451\ euro}100 = 86.4\%$$

$$CV(reddito - 4componenti) = \frac{26050\ euro}{49260\ euro}100 = 52.9\%$$

Campo di variazione e differenza interquartile

Sia $x_1, x_2, \ldots, x_N$ una distribuzione disaggregata. Sia $y_1, y_2, \ldots, y_N$ la stessa distribuzione con i termini disposti in ordine non decrescente.

campo di variazione

$$\Delta_c = y_N - y_1$$

Differenza tra massimo e minimo

Intervallo interquartile

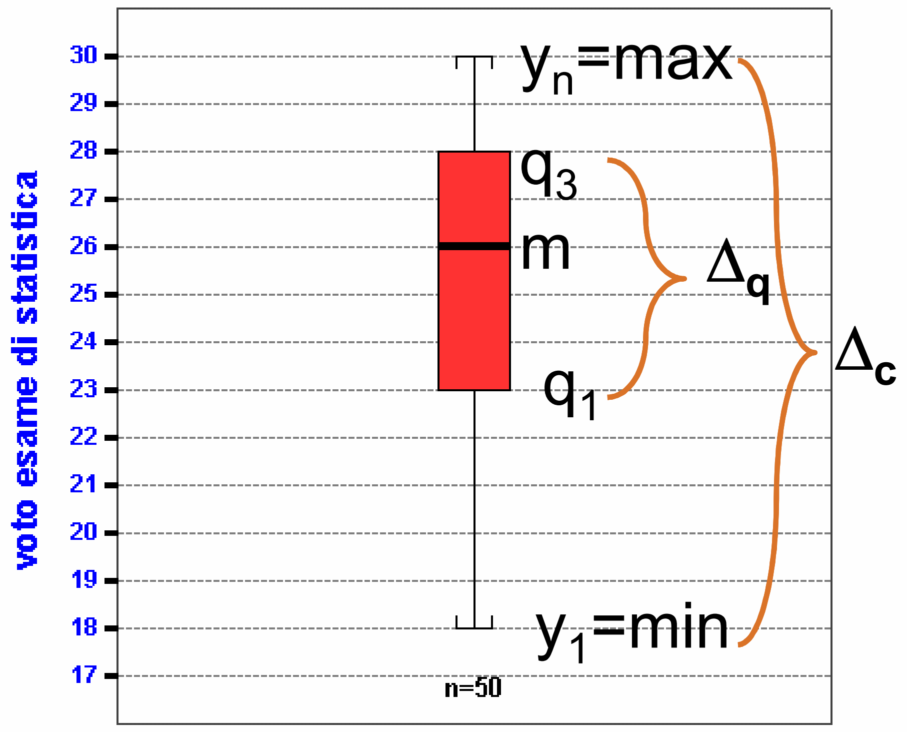
$$\Delta_q = q_3 - q_1$$

Differenza tra terzo e primo quartile

Box plot o Diagramma a scatola

Il box plot di una distribuzione è un grafico caratterizzato da tre elementi principali:

- una linea che indica la posizione della **mediana** della distribuzione
- un rettangolo (box) la cui altezza indica la variabilità dei valori prossimi alla **mediana**
- due segmenti che partono dal rettangolo e i cui estremi sono determinati in base ai **quartili** $\mathbf{Q_1}$ e $\mathbf{Q_3}$ della distribuzione

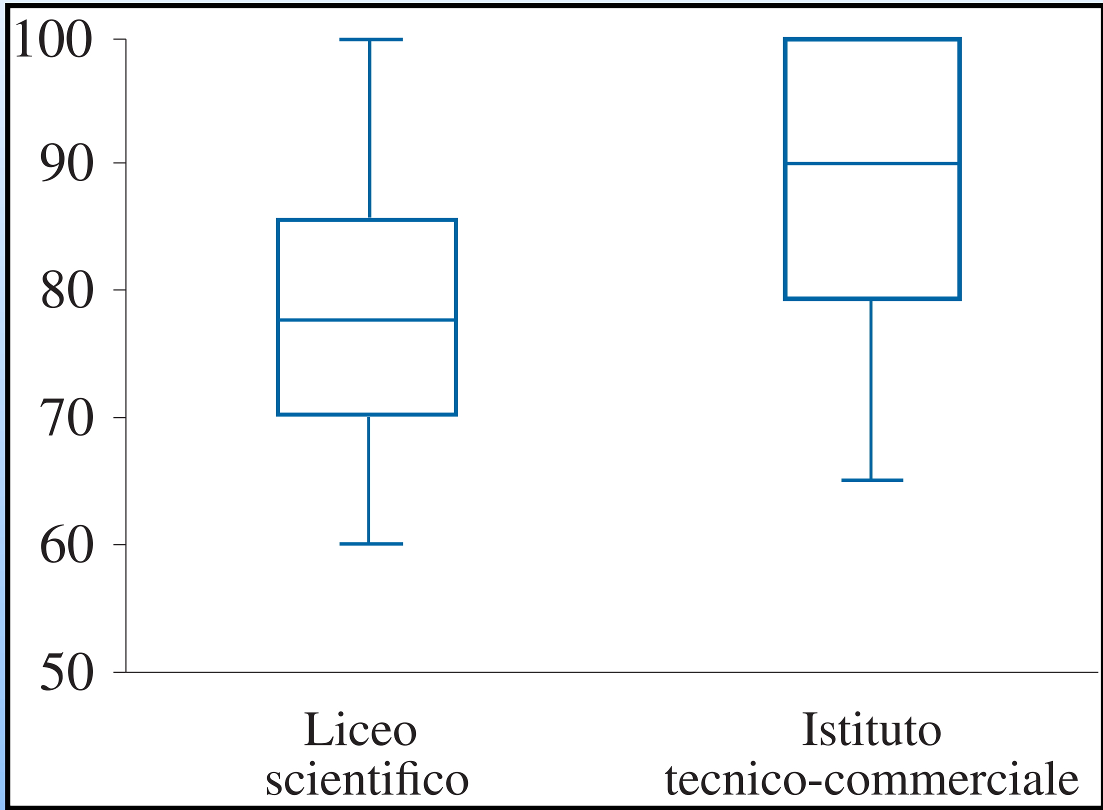


	y1	q1	m	q3	yn
voto esame di statistica	18	23	26	28	30

Diagramma a scatola: esempio

Aspetti salienti delle due distribuzioni

- **Intensità media** (*mediana*)
- La **misura della variabilità** (*campo di variazione* e *differenza interquartilica*)
- La presenza di **asimmetria** positiva o negativa che si desume dal raffronto delle distanze del primo e del terzo quartile dalla mediana



Voti riportati all'esame di Stato in un campione di studenti del corso di laurea in Economia aziendale distinti per scuola frequentata

Concentrazione

- È la tendenza di un **carattere trasferibile** a essere posseduto da un numero ristretto di unità
- **Carattere trasferibile:** un carattere è detto trasferibile se ha senso immaginare che un'unità statistica possa cedere tutto o parte del carattere posseduto a un'altra unità statistica.
  - **es. di carattere trasferibile: reddito**
  - **es. di carattere non trasferibile: età**

Concentrazione

- Una distribuzione statistica presenta una concentrazione tanto più elevata
  - quanto maggiore è la frazione (sul totale) del carattere posseduta dalle unità con le modalità più alte,
- ovvero
  - quanto minore è la frazione (sul totale) del carattere posseduta dalle unità con modalità più basse.

Equidistribuzione e massima concentrazione

- Un carattere trasferibile si dice **equidistribuito** tra le *N* unità di un collettivo se *l'ammontare complessivo del carattere è distribuito tra le N unità in parti uguali*, ossia se *ogni unità del collettivo possiede 1/N dell'ammontare totale del carattere*

- *La concentrazione è **massima** quando tutto l'ammontare del carattere è posseduto da una sola unità del collettivo*

Concentrazione

Distribuzione disaggregata per un carattere trasferibile $x_1, x_2, \ldots, x_N$

Distribuzione disaggregata corrispondente con <u>termini ordinati</u> $y_1, y_2, \ldots, y_N$ (con $y_1 \leq y_2 \leq \ldots \leq y_N$)

Il totale della distribuzione è

$$\sum_{i=1}^{N} x_i = \sum_{i=1}^{N} y_i = N \cdot \mu$$

Concentrazione

La domanda che ci poniamo è quale sia il **grado di disuguaglianza dei termini della distribuzione**.

La risposta a tale domanda è basata sul confronto della distribuzione data con le due situazioni estreme seguenti, **a parità di totale**:

equidistribuzione

Disuguaglianza minima $y_1 = y_2 = \ldots = y_N = \mu,$

Disuguaglianza massima $y_1 = y_2 = \ldots = y_{N-1} = 0$ e $y_N = N \cdot \mu$

massima concentrazione

Concentrazione: esempio

Fatturato, in milioni di euro, di 6 imprese appartenenti allo stesso gruppo societario:

2.3, 3.4, 37.3, 5.1, 74.8, 41.9

- Fase 1: ordinamento dei termini

2.3, 3.4, 5.1, 37.4, 41.9, 74.8

Il totale del carattere è 164.9.

Per stabilire il grado di concentrazione della distribuzione occorrerà confrontarla con le <u>situazioni estreme di equidistribuzione e massima concentrazione</u>

Concentrazione: esempio

Continuazione

Fatturato, in milioni di euro, di 6 imprese appartenenti allo stesso gruppo societario (dati ordinati):

2.3, 3.4, 5.1, 37.4, 41.9, 74.8

Situazioni di riferimento

- Equidistribuzione $y_1 = y_2 = \ldots = y_6 = 27.48$

I sei termini sono uguali alla media

- Massima concentrazione $y_1 = y_2 = \ldots = y_5 = 0$ e $y_6 = 164.90$

Cinque unità presentano il valore 0, una sola possiede tutto l'ammontare del carattere

Misura della concentrazione per le distribuzioni disaggregate

Indicato con $A_i = y_1 + y_2 + \cdots + y_i$ l'ammontare del carattere posseduto dalle *i* unità "più povere", i rapporti

$$Q_1 = \frac{A_1}{A_N}, Q_2 = \frac{A_2}{A_N}, Q_3 = \frac{A_3}{A_N}, \ldots, Q_{N-1} = \frac{A_{N-1}}{A_N},$$

rappresentano la **frazione del totale del carattere** posseduta dalla prima unità, dalle prime due unità, dalle prime tre unità, …, dalle prime N - 1 unità.

Misura della concentrazione per le distribuzioni disaggregate

Caso di equidistribuzione

$$\mathrm{y}_1 = \mathrm{y}_2 = \cdots = \mathrm{y}_\mathrm{i} = \cdots = \mathrm{y}_\mathrm{N} = \mu$$

$$\mathrm{A}_\mathrm{i} = \mathrm{y}_1 + \mathrm{y}_2 + \cdots + \mathrm{y}_\mathrm{i} = i\mu$$

$$\mathrm{A}_\mathrm{N} = N\mu$$

$$\mathrm{Q}_\mathrm{i} = \frac{\mathrm{A}_\mathrm{i}}{\mathrm{A}_N} = \frac{\mathrm{i} \cdot \mu}{\mathrm{N} \cdot \mu} = \frac{\mathrm{i}}{\mathrm{N}}$$

Misura della concentrazione per le distribuzioni disaggregate

Posto $P_i = \frac{i}{N}, i = 1, 2, \ldots, N-1,$

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

La quantità $$(P_1 - Q_1) + (P_2 - Q_2) + \cdots + (P_{N-1} - Q_{N-1}) = \sum_{i=1}^{N-1} (P_i - Q_i)$$

- **è nulla in caso di equidistribuzione** $P_i = Q_i = \frac{i}{N}$
- **è tanto maggiore quanto più la distribuzione data è lontana dalla equidistribuzione** *(All'aumentare della concentrazione aumentano le differenze: Pi –Qi)*
- **è massima quando la distribuzione osservata è di massima concentrazione** $Q_1 = Q_2 = \ldots = Q_{N-1} = 0$

Misura della concentrazione per le distribuzioni disaggregate

Rapporto di concentrazione di Gini

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

Massima concentrazione

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

- equidistribuzione : G=0
- G cresce al crescere della concentrazione
- massima concentrazione: G=1

Rapporto di concentrazione di Gini: esempi

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

X: reddito mensile netto; N=4

Equidistribuzione

unità	x_i	$P_i=i/N$	A_i	$Q_i=A_i/A_N$	P_i-Q_i
1	2500	0.25	2500	0.25	0
2	2500	0.50	5000	0.50	0
3	2500	0.75	7500	0.75	0
4	2500	1	10000	1	

$$\mathbf{G} = \frac{(0.25-0.25)+(0.5-0.5)+(0.75-0.75)}{0.25+0.5+0.75} = 0$$

Massima concentrazione

unità	x_i	$P_i=i/N$	A_i	$Q_i=A_i/A_N$	P_i-Q_i
1	0	0.25	0	0	0.25
2	0	0.50	0	0	0.50
3	0	0.75	0	0	0.75
4	10000	1	10000	1	

$$\mathbf{G} = \frac{(0.25-0)+(0.5-0)+(0.75-0)}{0.25+0.5+0.75} = 1$$

unità	x_i	$P_i=i/N$	A_i	$Q_i=A_i/A_N$	P_i-Q_i
1	1000	0.25	1000	0.10	0.15
2	2000	0.50	3000	0.30	0.20
3	3000	0.75	6000	0.60	0.15
4	4000	1	10000	1	

$$\mathbf{G} = \frac{(0.25-0.1)+(0.5-0.3)+(0.75-0.6)}{0.25+0.5+0.75} = 0.33$$

Proprietà dell'indice di Gini

- Assume valori nell'intervallo [0, 1]: è uguale a 0 nel caso di equidistribuzione, è uguale a 1 nel caso di massima concentrazione
- Diminuisce se si aggiunge a ogni termine della distribuzione una quantità positiva
- Non cambia se ogni termine della distribuzione è moltiplicato per una costante positiva
- Aumenta a seguito di un trasferimento di intensità da una data unità ad un'unità che presenta modalità non inferiore

Rapporto di concentrazione per una distribuzione disaggregata: calcolo

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

Popolazione nelle province della Toscana

Provincia	Popolazione	i	A_i	P_i	Q_i	$P_i - Q_i$
Massa-Carrara	197,652	1	197,652	0.10	0.06	**0.04**
Grosseto	211,086	2	408,738	0.20	0.12	**0.08**
Prato	227,886	3	636,624	0.30	0.18	**0.12**
Siena	252,288	4	888,912	0.40	0.25	**0.15**
Pistoia	268,503	5	1,157,415	0.50	0.33	**0.17**
Arezzo	323,288	6	1,480,703	0.60	0.42	**0.18**
Livorno	326,444	7	1,807,147	0.70	0.52	**0.18**
Lucca	372,244	8	2,179,391	0.80	0.62	**0.18**
Pisa	384,555	9	2,563,946	0.90	0.73	**0.17**
Firenze	933,860	10	3,497,806	1.00	1.00	0.00
Totale	3,497,806			4.50		1.26

$$G = \frac{1.26}{4.50} = 0.28$$

Misura della concentrazione per le distribuzioni disaggregate: formula operativa

- Formula operativa

$$G = 1 - \frac{2}{N-1}\sum_{i=1}^{N-1} Q_i$$

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

$$\sum_{i=1}^{N-1} P_i = \sum_{i=1}^{N-1} \frac{i}{N} = \frac{1}{N}\sum_{i=1}^{N-1} i = \\ = \frac{1}{N}\frac{(N-1)\cdot N}{2} = \frac{(N-1)}{2}$$

Regola: La somma dei primi n numeri naturali è n(n+1)/2

Rapporto di concentrazione per una distribuzione disaggregata: calcolo

$$G = 1 - \frac{2}{N-1} \sum_{i=1}^{N-1} Q_i$$

Popolazione nelle province della Toscana

Provincia	Popolazione	A_i	Q_i
Massa-Carrara	197,652	197,652	0.06
Grosseto	211,086	408,738	0.12
Prato	227,886	636,624	0.18
Siena	252,288	888,912	0.25
Pistoia	268,503	1,157,415	0.33
Arezzo	323,288	1,480,703	0.42
Livorno	326,444	1,807,147	0.52
Lucca	372,244	2,179,391	0.62
Pisa	384,555	2,563,946	0.73
Firenze	933,860	3,497,806	1.00
Totale	3,497,806		**3.23**

$$G = 1 - \frac{2}{9} \cdot 3.23$$

$$= 0.28$$

Curva di concentrazione

Si pongano sull'asse delle ascisse i valori di P_i e sull'asse delle ordinate le corrispondenti Q_i. Si individuino i punti di coordinate (P_i, Q_i), $i = 1, 2, \ldots, N$.

La spezzata che si ottiene unendo con segmenti di retta le coppie di punti contigui si chiama **curva di concentrazione** o **curva di Lorenz**

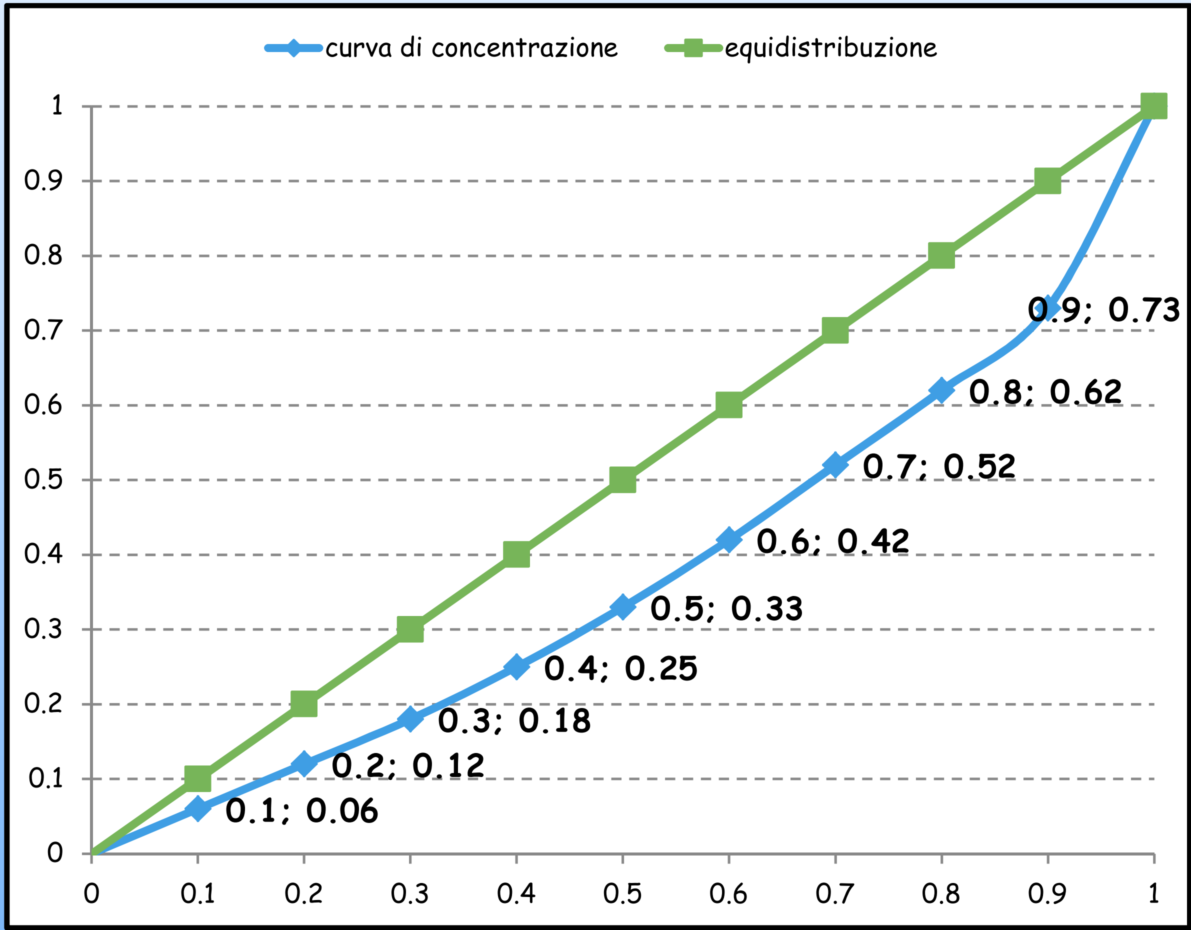
- il punto di coordinate (P_0, Q_0) è (0, 0)
- il punto di coordinate (P_N, Q_N) è (1, 1)

Curva di concentrazione: esempio

Popolazione nelle province della Toscana

Popolazione	P_i	Q_i
197,652	0.1	0.06
211,086	0.2	0.12
227,886	0.3	0.18
252,288	0.4	0.25
268,503	0.5	0.33
323,288	0.6	0.42
326,444	0.7	0.52
372,244	0.8	0.62
384,555	0.9	0.73
933,860	1	1

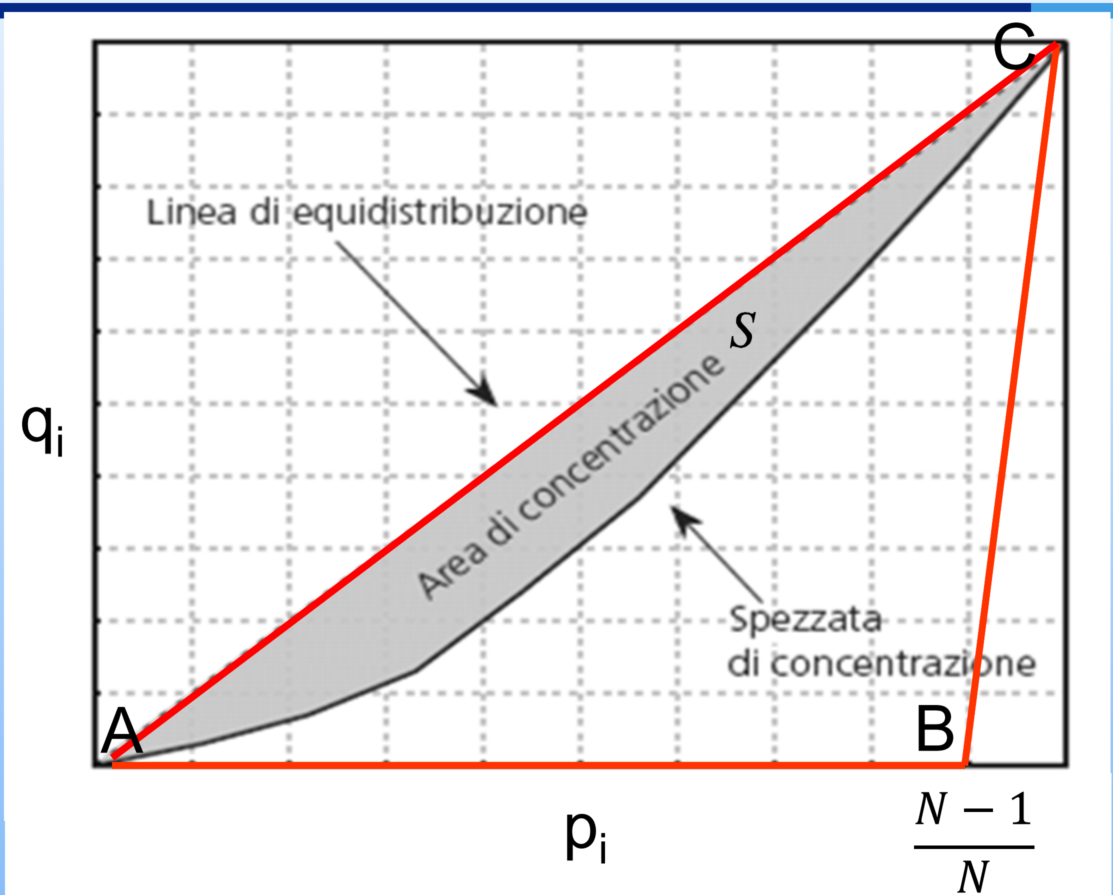
Curva di concentrazione



$$G = 0.28$$

diagramma di Lorenz

Interpretazione geometrica del rapporto di concentrazione



Il valore di R esprime l'area compresa tra la spezzata di concentrazione e la linea di equidistribuzione:

- più piccolo è R (fino a 0) più la spezzata si avvicina alla linea AC
- più grande è R (fino a 1) più la spezzata coincide con ABC (max concentrazione)

Nel caso di **equidistribuzione** la curva si riduce al segmento AC

Nel caso di **massima concentrazione** la curva coincide con ABC

Interpretazione geometrica del rapporto di concentrazione

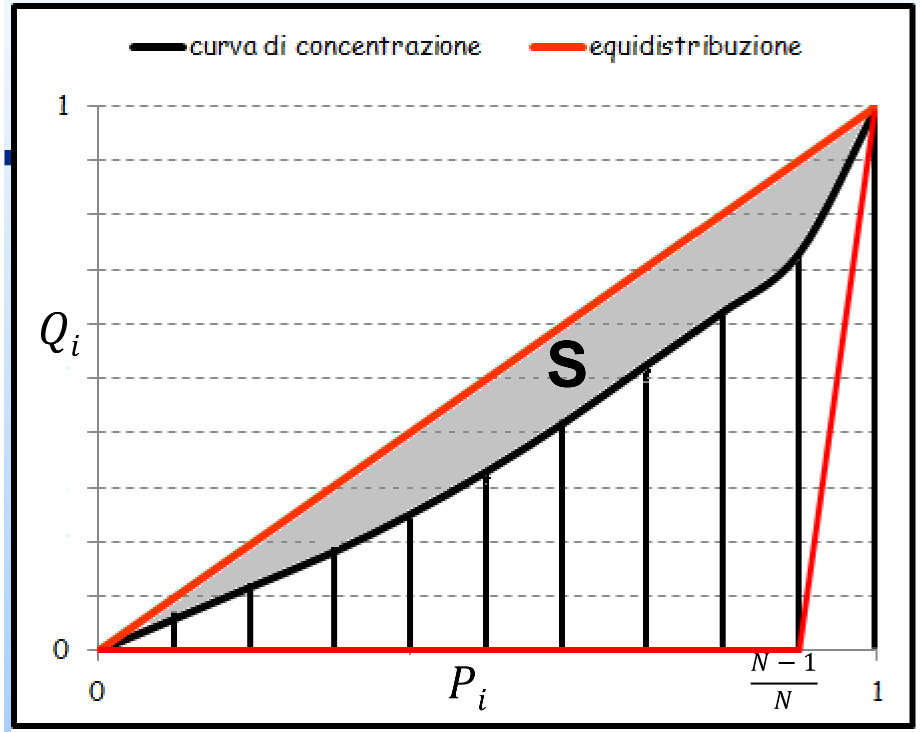
Sia S la superficie racchiusa tra il **segmento di equidistribuzione** e la curva di concentrazione.

Sia max(S) il massimo di S, dato dall'area racchiusa tra il segmento di equidistribuzione e la curva di massima concentrazione.

Allora, il rapporto di concentrazione G può essere espresso come

$$G = \frac{S}{\max S}$$

Interpretazione geometrica del rapporto di concentrazione



$$S=\frac{1}{2}-\sum_{i=1}^{N}\frac{(Q_i+Q_{i-1})(P_i-P_{i-1})}{2}$$

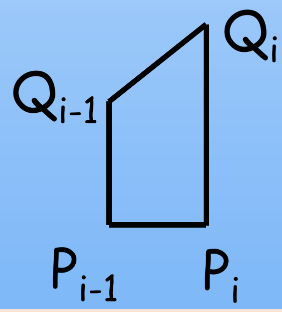
$$\max S=\frac{1}{2}-\frac{1}{2N}=\frac{N-1}{2N}$$

$$G=\frac{S}{\max S}=\left[\frac{1}{2}-\sum_{i=1}^{N}\frac{(Q_i+Q_{i-1})(P_i-P_{i-1})}{2}\right]\frac{2N}{N-1}$$

$$G=\frac{N}{N-1}\left[1-\sum_{i=1}^{N}(Q_i+Q_{i-1})(P_i-P_{i-1})\right]$$

S = area (triangolo di vertici (0,0), (1,0), (1,1)) – area sotto la curva di concentrazione

- area (triangolo)=1/2
- area (sotto la curva di concentrazione) =somma di aree di trapezi $\sum_{i=1}^{N}\frac{(Q_i+Q_{i-1})(P_i-P_{i-1})}{2}$



$P_0=0; Q_0=0$

Misura della concentrazione per le distribuzioni di frequenze

Sia data una distribuzione di frequenze secondo un carattere trasferibile. Sia N_i la generica frequenza cumulata e

$$A'_i = y_1 \cdot n_1 + y_2 \cdot n_2 + \ldots + y_i \cdot n_i$$

la quantità di carattere **posseduta complessivamente** dalle unità aventi modalità non superiori a x_i. Consideriamo i rapporti

$$P'_i = \frac{N_i}{N}, \quad Q'_i = \frac{A'_i}{A'_k}$$

che indicano, rispettivamente, la **frazione sul totale** delle unità e la **frazione sul totale del carattere relative alle unità che presentano modalità non superiori a x_i.**

$$G = \frac{N}{N-1}\left[1 - \sum_{i=1}^{k}(Q'_i + Q'_{i-1})(P'_i - P'_{i-1})\right]$$

Rapporto di concentrazione per una distribuzione di frequenze: calcolo

$$G=\frac{N}{N-1}\left[1-\sum_{i=1}^{k}(Q'_i+Q'_{i-1})(P'_i-P'_{i-1})\right]$$

Reddito in euro

Reddito (y_i)	individui (n_i)	y_in_i	N_i	A'_i	P'_i	Q'_i	$P'_i-P'_{i-1}$	$Q'_i+Q'_{i-1}$	$(P'_i-P'_{i-1})(Q'_i+Q'_{i-1})$
500	50	25000	50	25000	0.40	0.13	0.40	0.13	0.05
1000	25	25000	75	50000	0.60	0.27	0.20	0.40	0.08
1500	15	22500	90	72500	0.72	0.39	0.12	0.65	0.08
2000	20	40000	110	112500	0.88	0.60	0.16	0.99	0.16
5000	15	75000	125	187500	1.00	1.00	0.12	1.60	0.19
Totale	**125**	**187500**							0.56

$$G=\frac{125}{124}(1-0.56)=0.44$$

$$P'_i=\frac{N_i}{N},\quad Q'_i=\frac{A'_i}{A'_k}$$

Curva di concentrazione: esempio

Reddito

Reddito (y_i)	individui (n_i)	P'_i	Q'_i
500	50	0.40	0.13
1000	25	0.60	0.27
1500	15	0.72	0.39
2000	20	0.88	0.60
5000	15	1	1

$$G = 0.44$$

Curva di concentrazione

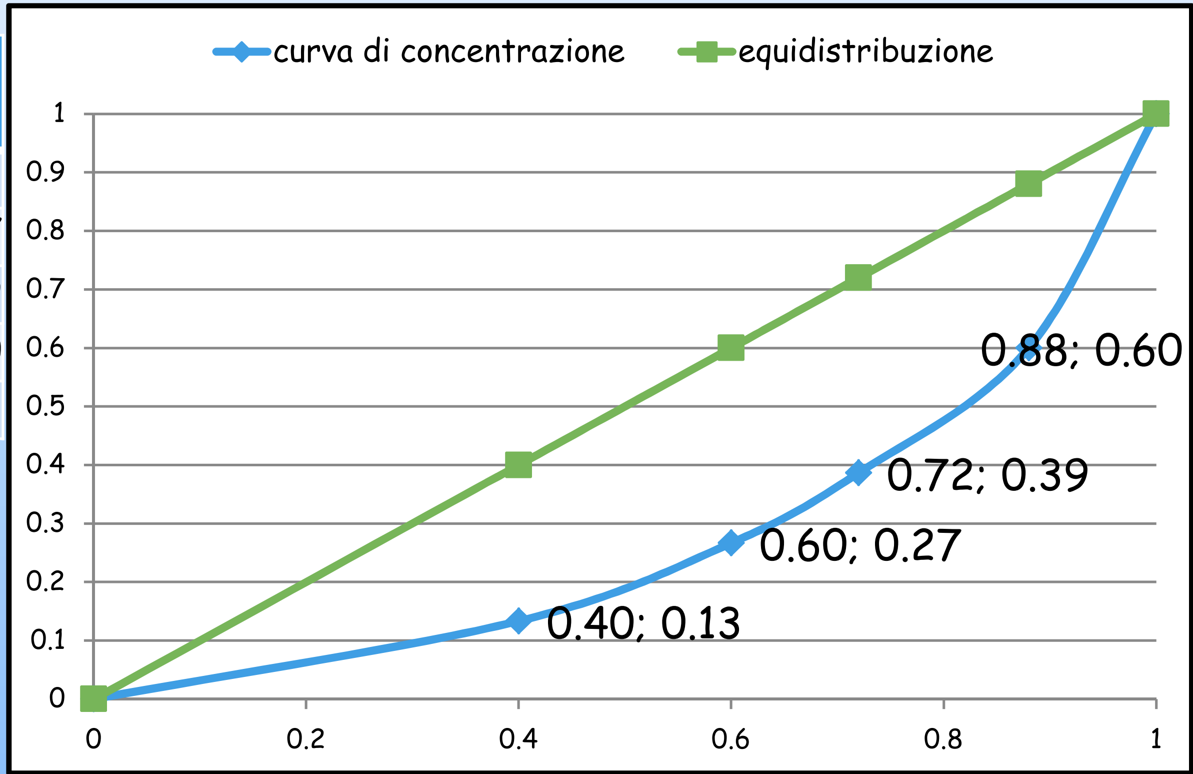


diagramma di Lorenz

Rapporto di concentrazione nel caso delle distribuzioni di frequenze (formula libro)

Il rapporto di concentrazione R assume la forma

$$R = \sum_{i=1}^{k} [(P'_{i-1} - Q'_{i-1}) + (P'_i - Q'_i)] \cdot \frac{n_i}{N}$$

Questa formula coincide con

$$\left[1 - \sum_{i=1}^{k} (Q'_i + Q'_{i-1})(P'_i - P'_{i-1})\right]$$

Rapporto di concentrazione per una distribuzione di frequenze: calcolo formula libro

$$R = \sum_{i=1}^{k} [(P'_{i-1} - Q'_{i-1}) + (P'_i - Q'_i)] \cdot \frac{n_i}{N}$$

Reddito in euro

$$\pi_i = \left(P'_{i-1} - Q'_{i-1}\right) + \left(P'_i - Q'_i\right) n_i$$

Reddito (y_i)	individui (n_i)	$y_i n_i$	N_i	A'_i	P'_i	Q'_i	$P'_i - Q'_i$	π_i
500	50	25000	50	25000	0.40	0.13	0.27	13.33
1000	25	25000	75	50000	0.60	0.27	0.33	15.00
1500	15	22500	90	72500	0.72	0.39	0.33	10.00
2000	20	40000	110	112500	0.88	0.60	0.28	12.27
5000	15	75000	125	187500	1.00	1.00	0.00	4.20
Totale	125	187500						54.80

$$R = \frac{54.8}{125} = 0.44$$

Misura della concentrazione per le distribuzioni in classi

Nel caso di distribuzioni in classi per studiare la concentrazione dobbiamo considerare alcune ipotesi iniziali:

- ***CASO A***: se **conosciamo l'ammontare di carattere** posseduto e il numero di unità si assume che ci sia equidistribuzione (ogni unità della classe possiede lo stesso ammontare di carattere)

- ***CASO B***: se **non conosciamo l'ammontare di carattere** posseduto dalle unità della classe allora possiamo stimarlo moltiplicando il valore centrale per il numero di unità statistiche della classe. **Il calcolo di A_i' si farà moltiplicando il valore centrale di classe per le frequenze**.

Distribuzione di frequenze con modalità raggruppate in classi <u>con ammontare di classe noto</u>: calcolo

$$G = \frac{N}{N-1}\left[1 - \sum_{i=1}^{k}(Q'_i + Q'_{i-1})(P'_i - P'_{i-1})\right]$$

Comuni della provincia di Perugia per ampiezza demografica

Ampiezza dem.	N. comuni	Popolazione	N_i	A'_i	P'_i	Q'_i	$P'_i-P'_{i-1}$	$Q'_i+Q'_{i-1}$	$(P'_i-P'_{i-1})(Q'_i+Q'_{i-1})$
fino a 5	33	74370	33	74370	0.56	0.11	0.56	0.11	0.06
5-10	11	76422	44	150792	0.75	0.22	0.19	0.34	0.06
10-15	2	26227	46	177019	0.78	0.26	0.03	0.49	0.02
15-20	5	84417	51	261436	0.86	0.39	0.08	0.65	0.06
20-30	3	70915	54	332351	0.92	0.49	0.05	0.88	0.04
30-40	2	72572	56	404923	0.95	0.60	0.03	1.10	0.04
oltre 40	3	266898	59	671821	1	1	0.05	1.60	0.08
Totale	59								**0.36**

$$G = \frac{59}{58}(1 - 0.36) = 0.65$$

Distribuzione di frequenze con modalità raggruppate in classi con ammontare di classe noto: calcolo

$$R = \sum_{i=1}^{k} [(P'_{i-1} - Q'_{i-1}) + (P'_i - Q'_i)] \cdot \frac{n_i}{N}$$

Comuni della provincia di Perugia per ampiezza demografica

Ampiezza dem.	N. comuni	Popolazione	N_i	A'_i	P'_i	Q'_i	$P'_i - Q'_i$	π_i
fino a 5	33	74370	33	74370	0.56	0.11	0.45	14.81
5-10	11	76422	44	150792	0.75	0.22	0.52	10.67
10-15	2	26227	46	177019	0.78	0.26	0.52	2.08
15-20	5	84417	51	261436	0.86	0.39	0.48	4.96
20-30	3	70915	54	332351	0.92	0.49	0.42	2.69
30-40	2	72572	56	404923	0.95	0.60	0.35	1.53
oltre 40	3	266898	59	671821	1	1	0	1.04
Totale	59							37.77

$$R = \frac{37.77}{59} = 0.64$$

$$\pi_i = \left(P'_{i-1} - Q'_{i-1}\right) + \left(P'_i - Q'_i\right) n_i$$

Distribuzione di frequenze con modalità raggruppate in classi con ammontare di classe non noto: calcolo

$$G = \frac{N}{N-1}\left[1 - \sum_{i=1}^{k}(Q'_i + Q'_{i-1})(P'_i - P'_{i-1})\right]$$

Reddito	Valore centrale	individui (n_i)	$y_i n_i$	N_i	A'_i	P'_i	Q'_i	$P'_i-P'_{i-1}$	$Q'_i+Q'_{i-1}$	$(P'_i-P'_{i-1})(Q'_i+Q'_{i-1})$
0-500	250	20	5000	20	5000	0.13	0.02	0.13	0.02	0.00
500-1000	750	35	26250	55	31250	0.35	0.12	0.23	0.14	0.03
1000-2000	1500	54	81000	109	112250	0.70	0.43	0.35	0.55	0.19
2000-3000	2500	30	75000	139	187250	0.90	0.72	0.19	1.16	0.22
3000-6000	4500	16	72000	155	259250	1	1	0.10	1.72	0.18
Totale		155	259250							0.63

$$R = \frac{155}{154}(1 - 0.63) = 0.37$$