La Variabilità statistica

- Una peculiarità dei caratteri rilevati nelle unità statistiche di un collettivo, è quella di presentare valori o attributi in tutto o in parte diversi.
- Si chiama *variabilità* (nel caso quantitativo) e *mutabilità* (nel caso qualitativo), l'attitudine dei caratteri o fenomeni ad assumere modalità differenti.
- La variabilità assume significati specifici a seconda del contesto in cui si manifesta, come evidenziato nei seguenti casi.
  1. variabilità come misura di imprecisione;
  2. variabilità come misura di incertezza;
  3. variabilità come concentrazione.

Due proprietà degli indici di variabilità

1. essere nulli se e solo se tutti i termini della distribuzione sono uguali fra loro, cioè se la popolazione è omogenea rispetto a X (e quindi il carattere risulta concentrato in una sola modalità).
2. crescere all'aumentare della distanza fra i termini, nel senso che la misura della variabilità cresca ogni volta che cresce almeno una delle quantità – differenze o rapporto – assunte per la misurazione della disuguaglianza fra i due termini, ferme restando le altre quantità.

Categorie di indici di variabilità 1/2

1. indici che misurano la variabilità del carattere mediante una sintesi delle *misure degli scarti in valore assoluto* $|a_i - M|$ *tra ogni termine della distribuzione ed una media* scelta come riferimento;
2. indici che misurano la variabilità del carattere mediante sintesi di *misure delle differenza in valore assoluto* $|a_i - a_j|$ *fra tutti i termini della distribuzione fra loro*;
3. indici che misurano la variabilità del carattere misurando la *diversità fra le modalità di due particolari unità della distribuzione o fra due quantili*.

Un indice di variabilità che è espresso nella stessa unità di misura con cui si misura il carattere, è detto *assoluto*.

Categorie di indici di variabilità 2/2

1. scostamenti medi: se sono ottenuti confrontando ogni termine della distribuzione con una media e che quindi sono basati sulla distribuzione:

$$|a_1 - M|, \quad |a_2 - M|, \quad \ldots, \quad |a_N - M| \, ;$$

2. differenze medie: se sono una sintesi dei confronti di tutti i termini fra di loro e che quindi sono basati sulla distribuzione:

$$|a_1 - a_1|, \quad |a_1 - a_2|, \quad \ldots, \quad |a_1 - a_N|$$
$$\ldots$$
$$|a_N - a_1|, \quad |a_N - a_2|, \quad \ldots, \quad |a_N - a_N|$$

3. campi di variazione: se sono il valore assoluto della differenza delle modalità presenti in due particolari unità della distribuzione o dei valori assunti da due quantità.

Scarto Quadratico Medio (s.q.m)

Sia X una v.s. su una popolazione Ω di media μ.

- distribuzione unitaria $a_1, \ldots, a_N$ di X:

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

- distribuzione di frequenze assolute $(x_1, n_1), \ldots, (x_k, n_k)$ di X:

$$\sigma := \sqrt{\frac{1}{N}\sum_{j=1}^{k}(x_j - \mu)^2 n_j}$$

- distribuzione di frequenze relative $(x_1, f_1), \ldots, (x_k, f_k)$ di X:

$$\sigma := \sqrt{\sum_{j=1}^{k}(x_j - \mu)^2 f_j}$$

Varianza, σ^2 o Var(X)

Sia X una v.s. su una popolazione Ω di media μ.

- distribuzione unitaria $a_1, \ldots, a_N$ di X:

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

- distribuzione di frequenze assolute $(x_1, n_1), \ldots, (x_k, n_k)$ di X:

$$\sigma^2 := \frac{1}{N} \sum_{j=1}^{k} (x_j - \mu)^2 n_j$$

- distribuzione di frequenze relative $(x_1, f_1), \ldots, (x_k, f_k)$ di X:

$$\sigma^2 := \sum_{j=1}^{k} (x_j - \mu)^2 f_j$$

Schema per il Calcolo della Varianza 1/3

Consideriamo la distribuzione dei voti riportati da uno studente universitario nei primi 10 esami: $30, 19, 24, 30, 30, 30, 24, 30, 28, 30$.

i	a_i	$a_i - \mu$	$(a_i - \mu)^2$
1	30	2,5	6,25
2	19	-8,5	72,25
3	24	-3,5	12,25
4	30	2,5	6,25
5	30	2,5	6,25
6	30	2,5	6,25
7	24	-3,5	12,25
8	30	2,5	6,25
9	28	0,5	0,25
10	30	2,5	6,25
somme	275	0	134,5
media	27,5		13,45

Schema per il Calcolo della Varianza 1/3

Consideriamo la distribuzione dei voti riportati da uno studente universitario nei primi 10 esami: $30, 19, 24, 30, 30, 30, 24, 30, 28, 30$.

i	a_i	$a_i - \mu$	$(a_i - \mu)^2$
1	30	2,5	6,25
2	19	-8,5	72,25
3	24	-3,5	12,25
4	30	2,5	6,25
5	30	2,5	6,25
6	30	2,5	6,25
7	24	-3,5	12,25
8	30	2,5	6,25
9	28	0,5	0,25
10	30	2,5	6,25
somme	275	0	134,5
media	27,5		13,45

Schema per il Calcolo della Varianza 2/3

Sia assegnata la seguente distribuzione dei voti riportati da 50 studenti in una sessione di esami:

x_i	18	20	21	22	24	25	26	27	28	30
n_i	2	7	2	3	11	9	4	5	2	5

Schema per il Calcolo della Varianza 2/3

x_i	n_i	$x_i n_i$	$x_i - \mu$	$(x_i - \mu)^2$	$(x_i - \mu)^2 n_i$
18	2	36	-6,36	40,45	80,90
20	7	140	-4,36	19,01	133,07
21	2	42	-3,36	11,29	22,58
22	3	66	-2,36	5,57	16,71
24	11	264	-0,36	0,13	1,43
25	9	225	0,64	0,41	3,69
26	4	104	1,64	2,69	10,76
27	5	135	2,64	6,97	34,85
28	2	56	3,64	13,25	26,50
30	5	150	5,64	31,81	159,05
Totale	50	1218			489,52
Medie		24,36			9,79

Schema per il Calcolo della Varianza 2/3

x_i	n_i	$x_i n_i$	$x_i - \mu$	$(x_i - \mu)^2$	$(x_i - \mu)^2 n_i$
18	2	36	-6,36	40,45	80,90
20	7	140	-4,36	19,01	133,07
21	2	42	-3,36	11,29	22,58
22	3	66	-2,36	5,57	16,71
24	11	264	-0,36	0,13	1,43
25	9	225	0,64	0,41	3,69
26	4	104	1,64	2,69	10,76
27	5	135	2,64	6,97	34,85
28	2	56	3,64	13,25	26,50
30	5	150	5,64	31,81	159,05
Totale	50	1218			489,52
Medie		24,36			9,79

Schema per il Calcolo della Varianza 3/3

Calcoliamo la varianza della seguente distribuzione dei residenti maschi in Calabria al 1 Gennaio 1995 per classe di età:

classi di età	n_i
meno di un anno	11.924
1 - 4	52.234
5 - 9	69.272
10 - 14	74.636
15 - 24	167.503
25 - 44	299.539
45 - 64	217.742
65 e più	129.033
Totale	1.021.883

Fonte: *ISTAT*

Schema per il Calcolo della Varianza 3/3

classi	Δ	x_i	n_i	$x_i n_i$	$x_i - \mu$	$(x_i - \mu)^2$	$(x_i - \mu)^2 n_i$
meno di 1	1	0,5	11.924	5.962	-35,81	1282,20	15.288.969
1 - 4	4	3	52.234	156.702	-33,31	1109,41	57.949.036
5 - 9	5	7,5	69.272	519.540	-28,81	829,89	57.488.252
10 - 14	5	12,5	74.636	932.950	-23,81	566,81	42.304.672
15 - 24	10	20	167.503	3.350.060	-16,31	265,95	44.546.691
25 - 44	20	35	299.539	10.483.865	-1,31	1,71	512.345
45 - 64	20	55	217.742	11.975.810	18,69	349,40	76.078.372
65 e più		75	129.033	9.677.475	38,69	1497,08	193.173.147
Totale			1.021.883	37.102.364			487.341.483
Medie				36,31			476,91

Schema per il Calcolo della Varianza 3/3

classi	Δ	x_i	n_i	$x_i n_i$	$x_i - \mu$	$(x_i - \mu)^2$	$(x_i - \mu)^2 n_i$
meno di 1	1	0,5	11.924	5.962	-35,81	1282,20	15.288.969
1 - 4	4	3	52.234	156.702	-33,31	1109,41	57.949.036
5 - 9	5	7,5	69.272	519.540	-28,81	829,89	57.488.252
10 - 14	5	12,5	74.636	932.950	-23,81	566,81	42.304.672
15 - 24	10	20	167.503	3.350.060	-16,31	265,95	44.546.691
25 - 44	20	35	299.539	10.483.865	-1,31	1,71	512.345
45 - 64	20	55	217.742	11.975.810	18,69	349,40	76.078.372
65 e più		75	129.033	9.677.475	38,69	1497,08	193.173.147
Totale			1.021.883	37.102.364			487.341.483
Medie				36,31			476,91

La distribuzione presenta inoltre scarto quadratico medio

$\sigma = \sqrt{476,91} = 21,84.$

Proprietà della Varianza

Sia X una v.s. con media $M(X) = \mu$. Assegnati $\alpha, c \in \mathbb{R}$ si ha:

1. $\mathsf{Var}(c) = 0$;
2. $\mathsf{Var}(\alpha X) = \alpha^2 \mathsf{Var}(X)$.
3. $\mathsf{Var}(X + \beta) = \mathsf{Var}(X)$.
4. $\mathsf{Var}(\alpha X + \beta) = \alpha^2 \mathsf{Var}(X)$.

Un'importante relazione di carattere pratico è la seguente:

$$\mathsf{Var}(X) = M(X^2) - \mu^2 \ .$$

Schema per il Calcolo della Varianza 1/3 b

Consideriamo la distribuzione unitaria precedente:

i	a_i	a_i^2
1	30	900
2	19	361
3	24	576
4	30	900
5	30	900
6	30	900
7	24	576
8	30	900
9	28	784
10	30	900
somme	275	7.697
medie	27,5	769,7

$$\sigma^2 = M(X^2) - \mu^2 = 769,7 - 27,5^2 = 769,7 - 756,25 = 13,45$$

Schema per il Calcolo della Varianza 2/3 b

Consideriamo la distribuzione per valori singoli precedente:

x_i	n_i	$x_i n_i$	x_i^2	$x_i^2 n_i$
18	2	36	324	648
20	7	140	400	2.800
21	2	42	441	882
22	3	66	484	1.452
24	11	264	576	6.336
25	9	225	625	5.625
26	4	104	676	2.704
27	5	135	729	3.645
28	2	56	784	1.568
30	5	150	900	4.500
Totale	50	1218	5.939	30.160
Medie		24,36		603,2

$$\sigma^2 = M(X^2) - \mu^2 = 603,2 - 24,36^2 = 603,2 - 593,41 = 9,79$$

Schema per il Calcolo della Varianza 3/3 b

Consideriamo la distribuzione per valori singoli precedente:

classi di età	Δ	x_i	n_i	$x_i n_i$	x_i^2	$x_i^2 n_i$
meno di 1	1	0,5	11.924	5.962	0,25	2.981
1 - 4	4	3	52.234	156.702	9,00	470.106
5 - 9	5	7,5	69.272	519.540	56,25	3.896.550
10 - 14	5	12,5	74.636	932.950	156,25	11.661.875
15 - 24	10	20	167.503	3.350.060	400,00	67.001.200
25 - 44	20	35	299.539	10.483.865	1225,00	366.935.275
45 - 64	20	55	217.742	11.975.810	3025,00	658.669.550
65 e più		75	129.033	9.677.475	5625,00	725.810.625
Totale			1.021.883	37.102.364		1.834.448.162
Medie				36,31		1795,16

$$\sigma^2 = M(X^2) - \mu^2 = 1795,16 - 36,31^2 = 1795,16 - 1318,25 = 476,91$$

Scarto Quadratico Medio

L'importanza dello *scarto quadratico medio* o *deviazione standard* come misura di dispersione si evince anche dalle seguenti relazioni proprie di distribuzioni unimodali ed approssimativamente simmetriche:

- L'intervallo $[\mu - \sigma, \mu + \sigma]$
  comprende circa il 67% delle osservazioni;
- L'intervallo $[\mu - 2\sigma, \mu + 2\sigma]$
  comprende circa il 95% delle osservazioni;
- L'intervallo $[\mu - 3\sigma, \mu + 3\sigma]$
  comprende approssimativamente tutte le osservazioni.

Scarti Standardizzati

Lo scarto quadratico medio è anche utilizzato per calcolare gli *scarti ridotti* o *standardizzati* espressi da:

$$z_i := \frac{a_i - \mu}{\sigma} \ .$$

La distribuzione degli scarti standardizzati origina una variabile statistica Z per cui risulta:

- $M(Z) = 0$ per la proprietà della media;
- Var$(Z) = 1$ perchè la media dei quadrati dei numeratori delle z_i è uguale a σ^2

Gli scarti ridotti sono numeri puri poichè esprimono lo scarto assoluto in unità di scarto quadratico medio.

Scarti Standardizzati – Esempio di Calcolo

Consideriamo la distribuzione per valori singoli precedente.

Essendo per tale distribuzione $\mu = 24.36$ e $\sigma = 3.13$, si ha:

	x_i	n_i	$x_i - \mu$	$z_i = \frac{x_i - \mu}{\sigma}$	$z_i n_i$	z_i^2	$z_i^2 n_i$
	18	2	-6,36	-2,03	-4,07	4,13	8,26
	20	7	-4,36	-1,39	-9,75	1,94	13,59
	21	2	-3,36	-1,07	-2,15	1,15	2,31
	22	3	-2,36	-0,75	-2,26	0,57	1,71
	24	11	-0,36	-0,12	-1,27	0,01	0,14
	25	9	0,64	0,20	1,84	0,04	0,38
	26	4	1,64	0,52	2,10	0,27	1,10
	27	5	2,64	0,84	4,22	0,71	3,56
	28	2	3,64	1,16	2,33	1,35	2,71
	30	5	5,64	1,80	9,01	3,25	16,24
Totale		50			0,00		50,00
Medie					0,00		1,00

Varianza per dati raggruppati

Sia X una v.s. suun collettivo Ω di N u.s. diviso in s gruppi. Per ciascuno di tali gruppi, indichiamo con:

μ_i : media dei valori dell'i-esimo gruppo,

σ_i^2 : varianza dei valori dell'i-esimo gruppo,

n_i : numero elementi dell'i-esimo gruppo.

Per la proprietà associativa, la media aritmetica μ di Ω data da:

$$\mu = \frac{\mu_1 n_1 + \cdots + \mu_s n_s}{n_1 + \cdots + n_s} .$$

Si può dimostrare che la varianza σ^2 dell'intero collettivo è data da:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{s} \sigma_i^2 n_i + \frac{1}{N}\sum_{i=1}^{s} (\mu_i - \mu)^2 n_i .$$

La quantità $\frac{1}{N}\sum_{i=1}^{s}(\mu_i - \mu)^2 n_i$ viene chiamata *varianza delle medie* dei singoli gruppi.

Differenze medie

Le *differenze medie* sono indici di variabilità in cui non si considera la dispersione dei termini rispetto ad una media, ma la disuguaglianza dei termini fra loro. Considerata pertanto la distribuzione $a_1, a_2, \ldots, a_N$ si considerano misure di diversità di ogni termine con tutti gli altri.

Le quantità da considerare sono:

$$\begin{array}{ccccc} |a_1 - a_1| & |a_1 - a_2| & |a_1 - a_3| & \cdots & |a_1 - a_N| \\ |a_2 - a_1| & |a_2 - a_2| & |a_2 - a_3| & \cdots & |a_2 - a_N| \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ |a_N - a_1| & |a_N - a_2| & |a_N - a_3| & \cdots & |a_N - a_N| \end{array}$$

- Si hanno N^2 termini se si considerano tutte le differenze fra ciascun termine e tutti i termini della distribuzione;
- Si hanno $N(N-1)$ termini se si considerano tutte le differenze fra ciascun termine e tutti i restanti termini della distribuzione.

Differenze semplici medie

La *differenza semplice media* e la *differenza semplice media con ripetizione* sono date rispettivamente da:

$$\Delta := \frac{\sum_{i=1}^{N}\sum_{j=1}^{N}|a_i - a_j|}{N(N-1)} \qquad \text{e} \qquad \Delta' := \frac{\sum_{i=1}^{N}\sum_{j=1}^{N}|a_i - a_j|}{N^2}\ ,$$

tra le quali sussiste la relazione:

$$\Delta' = \frac{N-1}{N}\Delta \qquad \text{o} \qquad \Delta = \frac{N}{N-1}\Delta'\ .$$

Nel caso in cui si abbia una distribuzione $(x_i, n_i)_{i=1,..,k}$, tali indici sono dati rispettivamente da:

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

Esistono delle formule semplificate per il calcolo delle differenze semplici medie.

Differenze quadratiche medie

La *differenza quadratica media* e la *differenza quadratica media con ripetizione* sono date rispettivamente da:

$$\Delta_2 := \sqrt{\frac{\sum_{i=1}^{N}\sum_{j=1}^{N}(a_i - a_j)^2}{N(N-1)}} \qquad \text{e} \qquad \Delta_2' := \sqrt{\frac{\sum_{i=1}^{N}\sum_{j=1}^{N}(a_i - a_j)^2}{N^2}} \ ,$$

o, nel caso di distribuzioni di frequenze $X = (x_i, n_i)_{i=1,..,k}$:

$$\Delta_2 = \sqrt{\frac{\sum_{i=1}^{k}\sum_{j=1}^{k}|(x_i - x_j)^2 n_i n_j}{N(N-1)}} \quad \text{e} \quad \Delta_2' = \sqrt{\frac{\sum_{i=1}^{k}\sum_{j=1}^{k}(x_i - x_j)^2 n_i}{N^2}}$$

Il calcolo della differenza quadratica media non presenta alcuna difficoltà in quanto si dimostrano le relazioni:

$$\Delta_2' = \sqrt{2}\sigma \qquad \text{e} \qquad \Delta_2 = \sqrt{2\frac{N}{N-1}}\sigma \ .$$

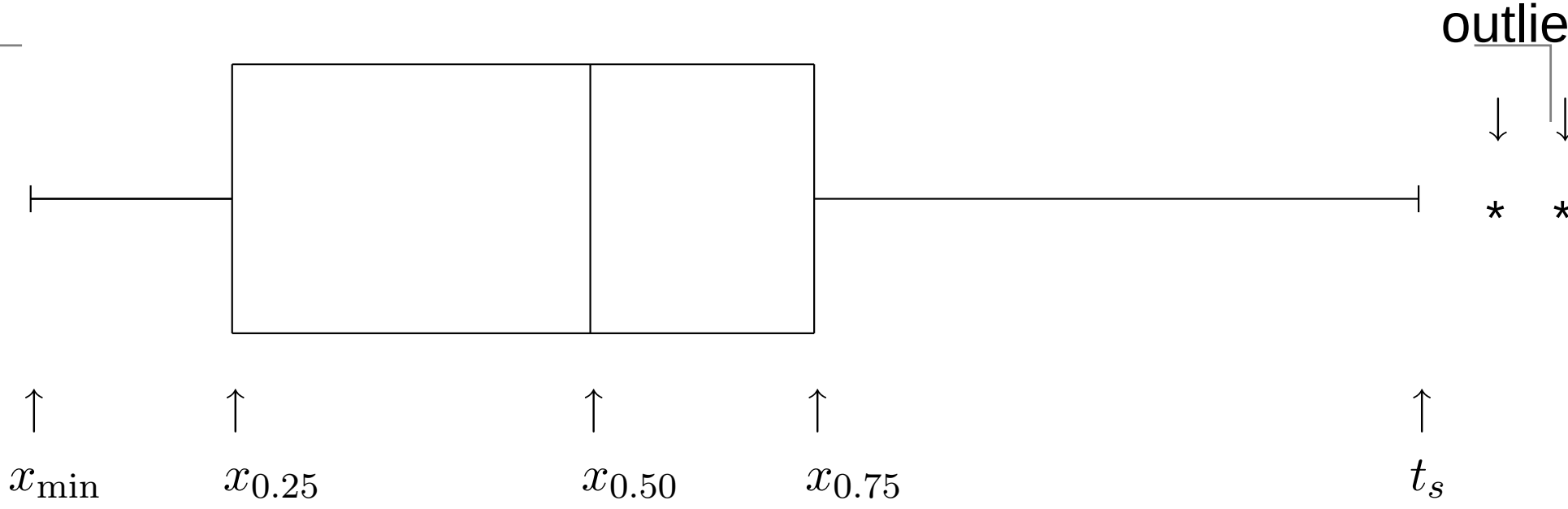
Intervalli di Variazione

Per *intervallo di variazione* si intende il valore assoluto, o una sua parte, della differenza tra le modalità presenti in due specifiche unità della distribuzione o tra i valori assunti da due quantili.
Si tratta di indici di facile e rapido calcolo che sono però grossolani perchè sono basati sulla conoscenza di due termini o di due quantili. Esempi:

- Il *campo di variazione* o *range* $a_{(N)} - a_{(1)}$
  dove $a_{(N)}$ e $a_{(1)}$ indicano rispettivamente il valore massimo e quello minimo della distribuzione.
- La *distanza interquartilica* $Q_3 - Q_1$
  dove Q_1 e Q_3 sono, rispettivamente il primo ed il terzo quartile della distribuzione.

Si possono generalizzare a differenze di opportuni percentili.

Il diagramma a scatola (box-plot)



Posto $DI = x_{0.75} - x_{0.25}$, si individuano quindi il "punto di troncamento inferiore" $t_i := \max\{x_{\min}, x_{0.25} - 1.5DI\}$ e il "punto di troncamento superiore" $t_s := \min\{x_{\max}, x_{0.75} + 1.5DI\}$.
Gli eventuali valori esterni rispetto ai punti di troncamento vengono considerati come possibili valori anomali e sono indicati con asterischi (o altro simbolo) sulla retta in prosecuzione del rispettivo segmento.

Variabilità relativa 1/2

Gli indici di variabilità visti finora sono chiamati *indici di variabilità assoluta* in quanto sono espressi nella stessa unità di misura dei termini della distribuzione. Essi non sono sempre sufficiente per poter confrontare variabilità di distribuzioni differenti. Infatti possiamo distinguere i seguenti casi:

1. le modalità delle distribuzioni a confronto sono espresse in unità di misura diverse, fra le quali vi è un rapporto costante (ad esempio: cambio ufficiale fra due monete in un dato giorno, etc.);
2. le modalità delle distribuzioni a confronto sono espresse nella stessa unità di misura, ma le loro intensità medie sono differenti (ad esempio: quantità di importazioni di beni diversi, prezzi di merci ordinarie e prezzi di merci pregiate, etc.).

Variabilità relativa 2/2

Nel primo caso, basta esprimere tutte le quantità nella stessa unità di misura o moltiplicare l'indice assoluto di variabilità per il rapporto che intercorre fra le rispettive unità di misura; nel secondo caso si ricorre agli *indici di variabilità relativa*.

Sia V_a il generico indice di variabilità assoluta, allora il corrispondente indice V_r di variabilità relativa può essere definito in base a due diversi criteri:

1. *indici percentuali di variabilità* o *di variabilità relativi alla media*:

$$\mathsf{V}_r := \frac{\mathsf{V}_a}{\mu} \ ;$$

2. *indici relativi di variabilità* o *di variabilità ralativi al massimo*:

$$\mathsf{V}_r = \frac{\mathsf{V}_a}{\max \mathsf{V}_a} \ .$$

Gli indici così ottenuti sono numeri puri.

Coefficiente di Variazione

E' l'indice più utilizzato, proposto da K. Pearson nel 1895:

$$CV = \frac{\sigma}{\mu} \, .$$

Si ha:

- $\max CV = 100\sqrt{N-1}$;
- è privo di significato se la media è uguale a zero;
- se la media è negativa, se ne considera il valore assoluto affinchè il CV risulti positivo;
- se la media, in valore assoluto, risulta prossima a zero (per effetto di parziali compensazioni fra valori positivi e negativi), il CV può segnalare, in maniera errata, una variabilità molto elevata del fenomeno.

Indici relativi al massimo

Gli indici relativi al massimo hanno il vantaggio di variare fra 0 e 1, ma lo svantaggio di essere strettamente dipendenti dalle ipotesi in base alle quali si determina il massimo dell'indice. Questo valore dipende dall'insieme dei valori possibili che l'indice può assumere, cioè dalla *distribuzione massimante*.

1. Carattere sempre positivo e non ha limite superiore: possiamo porre $a_{(1)} = 0$ e $a_{(N)} = N\mu$ e quindi la distribuzione massimante si presenta dunque formata da $N - 1$ termini uguali a zero ed un solo termine uguale a $N\mu$:
   $\max \sigma = \mu\sqrt{N-1}$

2. Distribuzioni simmetriche cioè i valori di $a_{(1)}$ e $a_{(N)}$ siano equidistanti da μ. Se inoltre poniamo $a_{(1)} = 0$, per la simmetria si avrà $a_{(N)} = 2\mu$ e quindi $\frac{1}{2}[a_{(N)} - a_{(1)}] = \mu$. In questo caso si ha: $\max \sigma = \mu$.

La concentrazione

- Un aspetto rilevante della variabilità di un carattere quantitativo è quello di *concentrazione*.
- La definizione ed il concetto comune di concentrazione (si parla di concentrazione urbana, finanziaria, etc.) richiede che il carattere quantitativo sia *trasferibile*.
- Un carattere trasferibile è *equidistribuito* fra le N unità del collettivo se l'ammontare complessivo A del carattere è distribuito in parti uguali fra le N unità, cioè se ogni unità possiede la quantità $\frac{A}{N}$.
- Se un carattere quantitativo trasferibile non è equidistribuito, diciamo che è *concentrato*.
- La situazione di *concentrazione massima* si ha quando una sola unità possiede tutto il carattere e tutte le altre unità statistiche non possiedono niente.

Misura della concentrazione 1/3

Un *indice di concentrazione* deve quindi assumere il valore zero nel caso di equidistribuzione, ed aumentano fino ad un valore massimo che viene assunto nel caso in cui una sola unità possiede l'ammontare totale del carattere.

Considerato un collettivo di N elementi, bisogna preliminarmente ordinare le unità statistiche secondo la loro modalità, in ordine non decrescente, del carattere quantitativo X; sia a_i l'ammontare del carattere posseduto dalla i-esima unità del collettivo (con $a_i \geq 0$). Risulta pertanto:

$$a_1 \leq a_2 \leq \cdots \leq a_N$$

La somma $A_i = a_1 + \cdots + a_i$ indica l'ammontare complessivo del carattere o *intensità* posseduta dalle i unità più povere del carattere stesso, per $i = 1, \ldots, N$.

Misura della concentrazione 2/3

Consideriamo le seguenti distribuzioni:

- La distribuzione $q_1, \ldots, q_N$ delle *intensità relative cumulate*

$$q_i = \frac{A_i}{A_N}$$

- La distribuzione $p_1, \ldots, p_N$ delle *frequenze relative cumulate*

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

Ovviamente p_i e q_i sono funzioni non decrescenti con i, cioè $p_i \leq p_{i+1}$ e $q_i \leq q_{i+1}$; risulta inoltre sempre per ogni $i = 1, \ldots, N$:

$$q_i \leq p_i \ .$$

Misura della concentrazione 3/3

Possono verificarsi tre casi:

1. caso di equidistribuzione, cioè nel caso in cui si ha:

$$a_1 = a_2 = \cdots = a_N = A/N \qquad \text{da cui :}$$

$$p_i = q_i \qquad \text{per ogni } i = 1, \ldots, N.$$

2. caso di massima concentrazione, cioè nel caso in cui si ha

$$a_1 = a_2 = \cdots = a_{N-1} = 0 \qquad \text{e} \qquad a_N = A\ . \qquad \text{da cui :}$$

$$p_i = i/N \qquad \text{per } i = 1, \ldots, N$$

$$q_1 = q_2 = \cdots = q_{N-1} = 0 \qquad \text{e} \qquad q_N = 1\ .$$

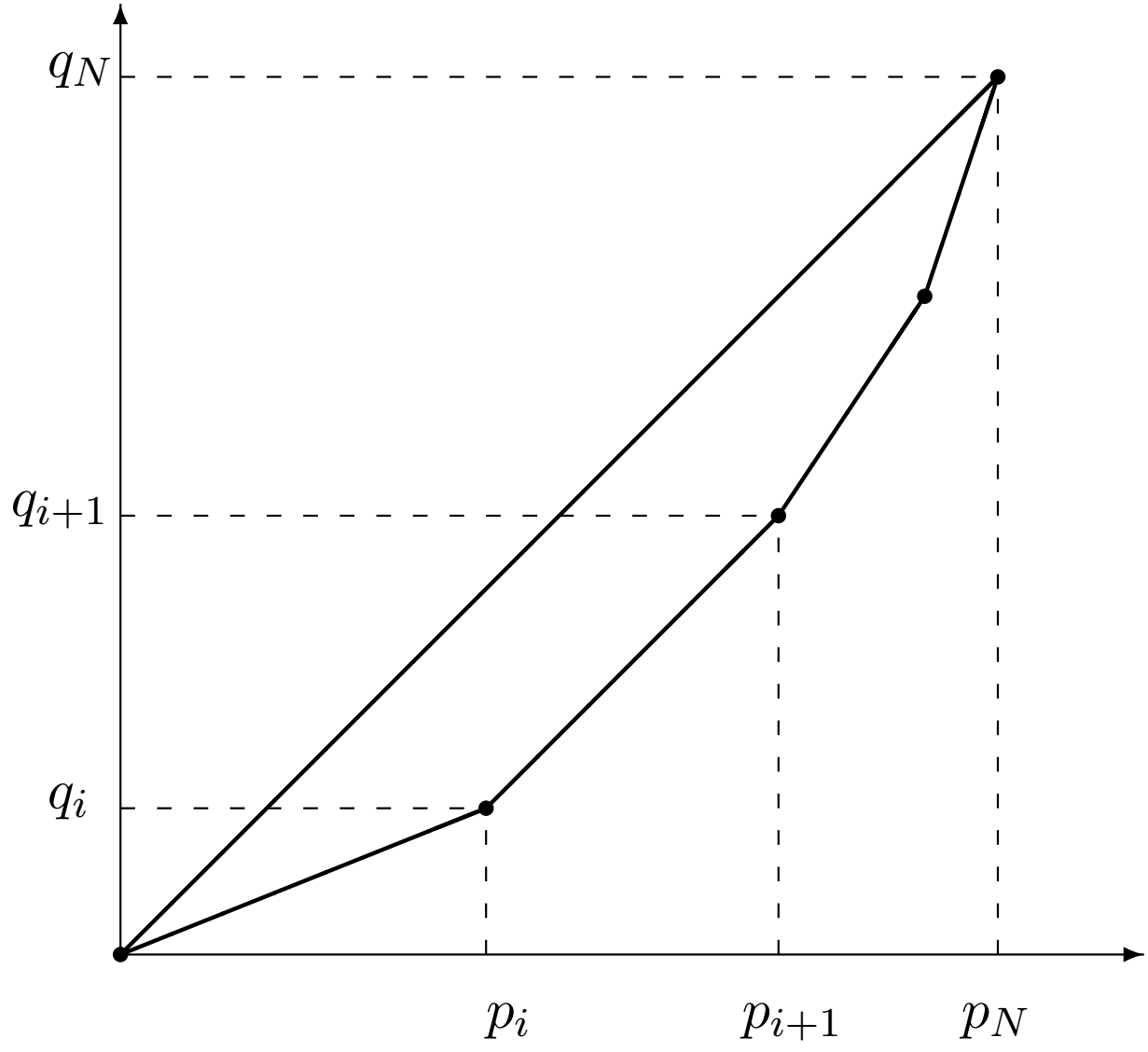
3. nei casi intermedi il carattere è tanto più concentrato quanto maggiore è la differenza $p_i - q_i$.

La curva di concentrazione 1/3

E' possibile introdurre un indice di concentrazione a partire dalla rappresenzazione grafica dei punti (p_i, q_i), per $i = 1, \ldots, N$. In un piano cartesiano, riportiamo in ascisse i valori di p_i e nelle ordinate i valori di q_i.

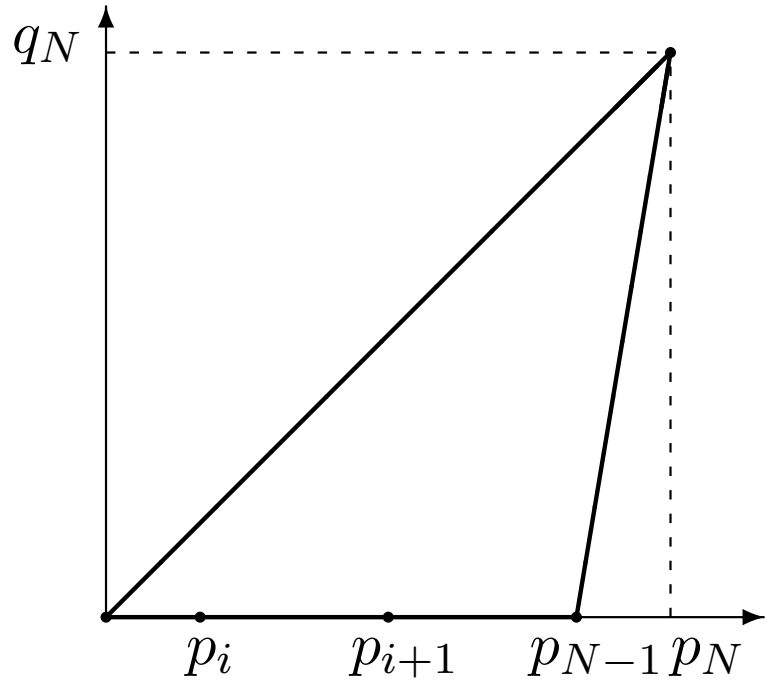
- Nel caso di equidistribuzione si ha $p_i = q_i$, e quindi i punti si dispongono sulla bisettrice del I quadrante: il segmento di estremi $(0, 0)$ e $(1, 1)$ viene chiamato *segmento di equidistribuzione*.
- Se non vi è equidistribuzione, i punti di coordinate (p_i, q_i) si trovano nel triangolo di vertici $(0, 0)$, $(1, 0)$ e $(1, 1)$. Unendo tali punti si ottiene una linea che viene chiamata *spezzata di concentrazione* o *curva di concentrazione* o *curva di Lorenz*.

La curva di concentrazione 2/3



La curva di concentrazione 2/3

L'area della superficie compresa fra il segmento di equidistribuzione e la spezzata di concentrazione viene scelta quale misura di concentrazione $A_G = \frac{1}{2} - \frac{1}{2}\sum_{i=0}^{N-1}(p_{i+1} - p_i)(q_{i+1} + q_i)$



Per avere un indice relativo bisogna dividere A_G per il suo valore massimo $A_{Gmax} = \frac{1}{2}\frac{N-1}{N}$, che si ottiene considerando che nel caso di massima concentrazione.

Il rapporto di concentrazione 1/2

Dal rapporto fra l'area di concentrazione ed il suo valor massimo si ottiene un indice chiamato *rapporto di concentrazione*:

$$\begin{aligned} R &= \frac{\frac{1}{2} - \frac{1}{2}\sum_{i=0}^{N-1}(p_{i+1} - p_i)(q_{i+1} + q_i)}{\frac{1}{2}\frac{N-1}{N}} \\ &= \frac{N}{N-1}\left(1 - \sum_{i=0}^{N-1}(p_{i+1} - p_i)(q_{i+1} + q_i)\right) \ . \end{aligned}$$

dove si pone $p_0 = q_0 = 0$. Nei casi pratici, il valore di N è molto grande e pertanto si ha:

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

Il rapporto di concentrazione 2/2

Si dimostra che la formula precedente può essere utilizzata nel calcolo del rapporto di concentrazione anche nel caso in cui viene assegnata una distribuzione di frequenza con k modalità (o classi di valori, assumendo l'ipotesi di equidistribuzione all'interno di ciascuna classe). In tal caso, si ha:

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

dove:

- p_i è frequenza relativa cumulata della i-esima modalità (o classe);
- q_i è intensità relativa cumulata della i-esima modalità (o classe).

dove, come in precedenza, si assume $p_0 = q_0 = 0$.

Rapporto di concentrazione - SdC a1

Consideriamo la seguente distribuzione del numero di dipendenti nel 1995 di alcune banche italiane. (Fonte *Il Sole 24-Ore*):

Banca	n. dipendenti
Banca Agricola Etnea	393
Banca Credito Siciliano	238
Popolare Sant'Angelo	675
Citibank	327
Credito Valtellinese	671
Inabanca	261

Rapporto di concentrazione - SdC a2

Indichiamo con X il "numero di dipendenti per banca", e con a_i la modalità di X, per $i = 1, 2, \ldots, 6$. Dopo aver ordinato in ordine non decrescente le modalità, possiamo calcolare l'insieme delle coppie (p_i, q_i), e quindi il rapporto di concentrazione R, essendo:

$$p_i = \frac{i}{6} \qquad A_i = \sum_{j=1}^{i} a_j \qquad q_i = \frac{A_i}{A_6}\ . \qquad i = 1, \ldots, 6.$$

i	a_i	A_i	p_i	q_i	$p_{i+1} - p_i$	$q_{i+1} + q_i$	$(p_{i+1} - p_i)(q_{i+1} + q_i)$
0			0,000	0,000	0,167	0,093	0,01546
1	238	238	0,167	0,093	0,167	0,287	0,04789
2	261	499	0,333	0,195	0,167	0,517	0,08609
3	327	826	0,500	0,322	0,167	0,797	0,13288
4	393	1.219	0,667	0,475	0,167	1,212	0,20201
5	671	1.890	0,833	0,737	0,167	1,737	0,28947
6	675	2.565	1,000	1,000			
Totale	2.565						0,77381

Il rapporto di concentrazione vale pertanto:

$$R \quad = \quad \frac{6}{5}\left(1 - \sum_{i=0}^{5}(p_{i+1} - p_i)(q_{i+1} + q_i)\right) \quad = \quad \frac{6}{5}(1 - 0,77381) \quad = \quad 0,27\ .$$

Rapporto di concentrazione - SdC b1

Consideriamo la seguente distribuzione di salari (in migliaia di euro) di $N = 85$ lavoratori:

Salari	n. lavoratori
25 - 50	15
50 - 75	25
75 - 100	20
100 - 125	15
125 - 150	10

Per calcolare il rapporto di concentrazione, costruiamo la tabella dei calcoli a partire dalla seguenti quantità:

$$N_i = \sum_{j=1}^{i} n_j \qquad p_i = \frac{N_i}{N} \qquad A_i = \sum_{j=1}^{i} x_j n_j \qquad q_i = \frac{A_i}{A_5} \ .$$

Rapporto di concentrazione - SdC a2

i	classi	n_i	x_i	$x_i n_i$	N_i	A_i	p_i	q_i	$p_{i+1} - p_i$	q_i
0							0,0000	0,0000	0,1765	
1	25 - 50	15	37,5	562,5	15	562,5	0,1765	0,0811	0,2941	
2	50 - 75	25	62,5	1.562,5	40	2.125,0	0,4706	0,3063	0,2353	
3	75 - 100	20	87,5	1.750,0	60	3.875,0	0,7059	0,5586	0,1765	
4	100 - 125	15	112,5	1.687,5	75	5.562,5	0,8824	0,8018	0,1176	
5	125 - 150	10	137,5	1.375,0	85	6.937,5	1,0000	1,000		
	Totale	85		6.937,5						

Il rapporto di concentrazione vale pertanto:

$$R \;=\; 1 - \sum_{i=0}^{4} (p_{i+1} - p_i)(q_{i+1} + q_i) \;=\; 1 - 0,7838 \;=\; 0,2162\ .$$

Il Rapporto di Concentrazione di Gini

Si può dimostrare che il rapporto di concentrazione può essere ricavato come rapporto fra la differenza semplice media (senza ripetizione) ed il doppio della media aritmetica della distribuzione:

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

dove Δ e μ sono rispettivamente la differenza semplice media senza ripetizione e la media aritmetica della distribuzione assegnata. Tale espressione prende il nome di *rapporto di concentrazione di Gini*.

Nei casi pratici, il rapporto di concentrazione di Gini può essere calcolato mediante opportune formule

Il Rapporto di Concentrazione di Gini

Nel caso di distribuzioni unitarieIl rapporto di concentrazione di Gini è dato dal rapporto:

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

Tale formula può essere giustificata anche come rapporto fra la somme delle differenze

$$\sum_{i=1}^{N-1} p_i - q_i$$

rilevata dai dati ed il massimo valore che tale somma può assumere (nel caso di massima concentrazione) cioè

$$\sum_{i=1}^{N-1} p_i = \frac{N-1}{2} .$$

Rapporto di Concentrazione di Gini - Sd

I valori possono essere sintetizzati nella tabella seguente:

i	a_i	A_i	p_i	q_i	$p_i - q_i$
1	238	238	0,167	0,093	0,074
2	261	499	0,333	0,195	0,138
3	327	826	0,500	0,322	0,178
4	393	1.219	0,667	0,475	0,191
5	671	1.890	0,833	0,737	0,096
6	675	2.565	1,000	1,000	0,000
Totale	2.565				0,677

Il rapporto di concentrazione vale pertanto:

$$R = \frac{2}{6-1}\sum_{i=1}^{5}(p_i - q_i) = 2 \cdot \frac{0,677}{5} = 0,27\ .$$