4 La variabilità

Come si è detto, ogni categoria di indici sintetici descrive particolari aspetti di una distribuzione.

Gli indici di variabilità misurano l'attitudine che hanno i dati ad assumere valori diversi.

In quanto costanti di sintesi, gli indici di variabilità si distinguono in relazione:

- agli scopi che ci si propone;
- al tipo di dati in esame;
- al livello di informazione che si vuole ottenere.

In particolare, nell'ambito degli indici di variabilità assoluta, distinguiamo:

- gli indici di dispersione;
- gli indici di variazione;
- gli indici di diversita'.

4.1 Gli indici di variabilità assoluta

Gli indici di variabilità assoluta soddisfano le seguenti proprietà:

- risultano nulli se tutti i valori x_i sono uguali fra loro, cioè se non c'è variabilità fra i dati;
- assumono valori positivi se i valori x_i sono diversi fra loro e sono tanto più elevati quanto più è elevata la variabilità fra le x_i;
- sono invarianti per traslazione;
- sono espressi nella stessa unità di misura dei dati.

4.1.1 Gli indici di dispersione

Gli indici di dispersione fanno riferimento a dati omogenei, che derivano da misure ripetute di una medesima grandezza incognita, riguardante uno stesso soggetto/oggetto o anche soggetti diversi, ma rigorosamente selezionati dal punto di vista genetico.

Tali misure si suppone siano affette da errori accidentali, che non consentono di conoscere con esattezza il vero valore della grandezza investigata.

Compito della Statistica è individuare la migliore combinazione delle osservazioni ai fini di ridurre l'influenza degli errori.

La migliore combinazione delle osservazioni, cioè la combinazione che meglio rappresenta il vero valore, sotto l'ipotesi di additività degli errori, è la media in senso "decisionale" M_p.

In tale contesto, ha significato individuare un indice di dispersione che indichi di quanto il valore rilevato si discosta dal vero valore.

Se non ci fosse l'influenza degli errori accidentali, tutti i dati rilevati sarebbero uguali fra loro e uguali a M_p, quindi la dispersione sarebbe nulla, perché nulli sarebbero tutti gli scarti $(x_i - M_p)$, $i=1, 2, \ldots, n$.

Ma ciò, in realtà, non si verifica e la variabilità sarà tanto più elevata quanto più grandi sono gli scostamenti dei valori x_i da M_p.

Sotto questi presupposti, una buona misura della variabilità dei dati è rappresentata dall'indice di dispersione:

$$\sigma_p = \left[\frac{\sum_{i=1}^{n} \left| x_i - M_p \right|^p}{n} \right]^{1/p},$$

che risulta invariante se aggiungiamo a ciascun valore x_i una costante α, cioè se cambiamo sistema di riferimento.

In una distribuzione di frequenze, le osservazioni vanno ponderate, per cui:

$$\sigma_p = \left[\frac{\sum_{i=1}^{k} \left| x_i - M_p \right|^p n_i}{n} \right]^{1/p} = \left[\sum_{i=1}^{k} \left| x_i - M_p \right|^p f_i \right]^{1/p}.$$

Il parametro $p \in (0,\infty)$ dipende dalla particolare struttura degli errori che influenzano i dati:

- se $p=1$, dunque $\varepsilon_i \sim LAPLACE$, si ha lo "scostamento semplice medio dalla mediana"

$$\sigma_1 = \frac{\sum_{i=1}^{n} |x_i - M_1|}{n}$$

che misura, in media, di quanto i valori osservati x_i si discostano dalla mediana M_1;

- se $p=2$, cioè se $\varepsilon_i \sim GAUSS$, si ha lo "scarto quadratico medio"

$$\sigma_2 = \sqrt{\frac{\sum_{i=1}^{n} (x_i - M_2)^2}{n}},$$

che misura, in media quadratica, di quanto i valori osservati x_i si discostano dalla media aritmetica M_2.

Il quadrato di σ_2 è noto con il nome di "VARIANZA", il cui calcolo si può effettuare con facilità, evitando gli scarti; infatti è:

$$\frac{\sum_{i=1}^{n} (x_i - M_2)^2}{n} = \frac{\sum_{i=1}^{n} (x_i^2 - 2x_i M_2 + M_2^2)}{n} = \frac{\sum_{i=1}^{n} x_i^2}{n} - 2M_2 \frac{\sum_{i=1}^{n} x_i}{n} + \frac{nM_2^2}{n} = \frac{\sum_{i=1}^{n} x_i^2}{n} - M_2^2$$

,

ossia la varianza di una serie di valori è uguale al quadrato della media quadratica meno il quadrato della media aritmetica.

- Quando $p \to \infty$, cioè quando $\varepsilon_i \sim UNIFORME$, si dimostra che σ_p è il semi-intervallo di variazione:

$$\sigma_\infty = \frac{x_{(n)} - x_{(1)}}{2}.$$

σ_p è espresso nella stessa unità di misura dei valori osservati x_i.

4.1.2 Gli indici di variazione

Gli indici di variazione trovano applicazione quando la variabilità di una serie osservata non è dovuta all'influenza di errori accidentali, ma ciascun valore x_i differisce dagli altri e dal valore medio per l'effetto sistematico di una legge di dipendenza *g(.),* che descrive l'evolversi degli stessi valori x_i.

In questo contesto, le medie di riferimento traggono origine dalla definizione del Chisini. Tale definizione è legata alla natura del fenomeno, che può essere di tipo additivo o moltiplicativo e all'esistenza di una legge di dipendenza, che considera i valori osservati x_i funzioni di altre variabili y: $x=g(y)$.

La variabilità di una serie di valori, in questo caso, può ancora essere misurata in termini di valore medio degli scarti di ciascun valore dalla media di riferimento, che rappresenta il baricentro della serie, ma ogni scarto non può, in alcun modo, essere assimilato al concetto di errore accidentale.

La nuova famiglia di indici di variabilità è rappresentata dall'espressione:

$$V_m = \left[\frac{\sum_{i=1}^{n} \left| x_i - M \right|^m}{n} \right]^{1/m}$$

che al variare di *m,* dove *m=1, 2, 3,…,* fornisce i cosiddetti "indici di variazione".

Per una distribuzione di frequenze, bisogna ponderare le osservazioni, per cui:

$$V_m = \left[\frac{\sum_{i=1}^{k} \left| x_i - M \right|^m n_i}{n} \right]^{1/m} = \left[\sum_{i=1}^{k} \left| x_i - M \right|^m f_i \right]^{1/m}.$$

4.1.3 Gli indici di diversità

Se consideriamo una variabile quantitativa di tipo discreto, che deriva da enumerazioni o conteggi di uno stesso oggetto/soggetto, non ha senso logico calcolare le differenze di ciascun valore x_i da una media decisionale o da una

media secondo il Chisini, che assumerebbe, in questo caso, valore puramente indicativo-formale.

Acquistano, invece, pienezza di significato, in questo contesto, le medie di posizione e gli indici di diversità.

Gli indici di diversità, quali indici di variabilità, si fondano sulla eterogeneità dei valori di una serie, pertanto si possono ottenere come "media potenziata di ordine *m*" di tutte le possibili differenze in coppia dei valori osservati:

$$ {}_R\Delta_m = \left[\frac{\sum_{i,j} \left| x_i - x_j \right|^m}{n^2} \right]^{1/m} $$

dove $n^2 = {}_RD_{n,2}$ sono tutte le possibili differenze. Ad esempio, se $n=3$, la matrice delle differenze è:

$$ \begin{matrix} d_{11} & d_{12} & d_{13} \\ d_{21} & d_{22} & d_{23} \\ d_{31} & d_{32} & d_{33} \end{matrix} $$

Poiché la differenza fra ciascun valore e sé stesso risulta nulla: $d_{ii}=0$ (differenze sulla diagonale principale della matrice), si possono considerare solo le differenze fra valori diversi, che sono $n(n-1)=D_{n,2}$:

$$ \Delta_m = \left[\frac{\sum_{i \neq j} \left| x_i - x_j \right|^m}{n(n-1)} \right]^{1/m} $$

Le due espressioni ${}_R\Delta_m$ e Δ_m, differiscono solo per il denominatore e sono definite, rispettivamente, "*differenze medie potenziate di ordine m con e senza ripetizione*".

Gli indici di diversità più impiegati, in pratica, sono le "*differenze semplici medie*", con e senza ripetizione, che si ottengono per $m=1$.

Poiché la matrice delle differenze è una matrice simmetrica, l'espressione a numeratore

$$\sum_{i,j}\left|x_i - x_j\right| = \sum_{i\neq j}\left|x_i - x_j\right|$$

può essere sostituita dall'espressione

$$2\sum_{i<j}\left|x_i - x_j\right|,$$

che consente di dimezzare i calcoli.

In poche parole, basta calcolare $\frac{n(n-1)}{2}$ differenze, anziché *n(n-1).*

Quando il valore x_i si ripete n_i volte e il valore x_j si ripete n_j volte, le singole differenze $|x_i-x_j|$ andranno moltiplicate per il fattore n_in_j, che rappresenta il numero delle volte che si verificano tali differenze:

$$_R\Delta_1 = \frac{\sum_{i,j}\left|x_i - x_j\right| n_i n_j}{n^2} \qquad\qquad i, j=1,2,\ldots,n$$

e se non si considerano le differenze ripetute:

$$\Delta_1 = \frac{\sum_{i\neq j}\left|x_i - x_j\right| n_i n_j}{n(n-1)} \qquad i, j=1,2,\ldots,n \qquad i\neq j.$$

Se *n* è grande si ricorre alla formula di De Finetti-Paciello (cfr.par. 4.3.3).

4.2 Indici di variabilità relativa

Gli indici di variabilità, così come i valori medi, visti finora, sono espressi nella stessa unità di misura dei valori x_i, ossia sono "indici di variabilità assoluta".

Tuttavia, se vogliamo confrontare due o più serie di valori, espressi in unità di misura diverse o aventi diverso ordine medio di grandezza, ovvero aventi un diverso intervallo di variazione, è necessario considerare gli "indici di variabilità relativa". Gli indici di variabilità relativa sono "numeri puri", ossia numeri senza alcuna unità di misura, e si distinguono in:

- *coefficienti di dispersione*;
- *coefficienti di variazione*;
- *coefficienti di diversità*.

4.2.1 Coefficienti di dispersione

Si ottengono rapportando gli indici di dispersione assoluta alla propria media decisionale:

$$\frac{\sigma_p}{M_p}.$$

4.2.3 Coefficienti di variazione

Si ottengono rapportando gli indici di variazione assoluta sempre e soltanto alla media aritmetica:

$$\frac{V_m}{M}.$$

Le unità di misura sono eliminate dal rapporto.

Tali indici possono essere espressi anche in termini percentuali; in tal caso, basta moltiplicarli per 100.

I coefficienti di dispersione e di variazione variano tra 0 e un valore massimo, che dipende dalla particolare forma della distribuzione statistica.

Questi indici presentano inconvenienti se la media cui sono rapportati tende a 0.

4.2.3 Coefficienti di diversità

Gli indici di diversità ha più senso rapportarli ad una quantità simile, che misura lo stesso aspetto dei dati, nella stessa unità di misura, ma in una situazione diversa. Per fare questo, occorre definire il concetto di massima variabilità teorica, che fa riferimento alle “distribuzioni massimanti” della variabilità, in cui si ipotizza che la massa delle osservazioni sia concentrata in un unico valore, mentre gli altri valori assumono valore zero.

Si dimostra che il massimo valore teorico della differenza semplice media è *2M* (dove *M* è la media aritmetica), per cui un indice di variabilità relativa si può ottenere, in questo caso, dal rapporto:

$$\frac{\Delta_1}{\max \Delta_1} = \frac{\Delta_1}{2M}.$$

Consideriamo, ad esempio, la seguente distribuzione massimante:

x_1	x_2	x_3	x_4
L	0	0	0

e calcoliamo Δ_1:

$$\Delta_1 = \frac{2\sum_{i<j} \left| x_i - x_j \right|}{n(n-1)} = \frac{2(n-1)L}{n(n-1)} = \frac{2\sum x_i}{n} = 2M .$$

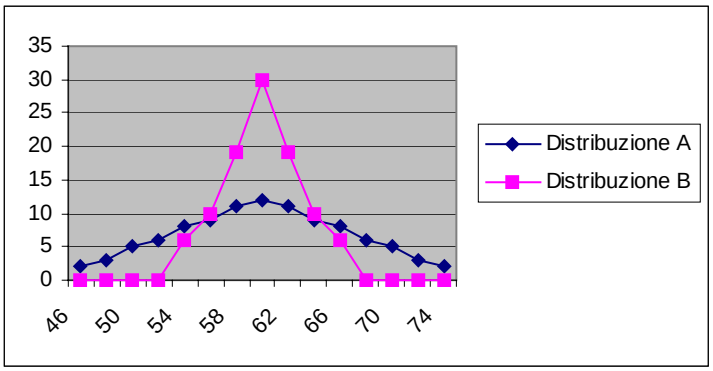
4.3 Esempi sugli indici di variabilità assoluta

Le misure di tendenza centrale da sole non sono sufficienti a caratterizzare una distribuzione.

Si consideri, ad esempio, la distribuzione del peso di due gruppi di persone:

peso	$\mathbf{n_A}$	$\mathbf{n_B}$
45-47	2	0
47-49	3	0
49-51	5	0
51-53	6	0
53-55	8	6
55-57	9	10
57-59	11	19
59-61	12	30
61-63	11	19
63-65	9	10
65-67	8	6
67-69	6	0
69-71	5	0
71-73	3	0
73-75	2	0
totale	100	100

Le due distribuzioni, pur avendo stessa media, mediana e moda, presentano una minore (distribuzione A) o una maggiore (distribuzione B) concentrazione dei valori intorno a tali medie:



$$M=M_e=M_o=60$$

Alle misure di tendenza centrale vanno, pertanto, affiancati gli indici di variabilità, che indicano appunto quanto i valori osservati sono più o meno dispersi rispetto alla media considerata.

4.3.1 Esempi sugli indici di variazione

Nella seguente tabella sono riportati i valori delle precipitazioni in *mm* rilevati in una stazione meteorologica di Roma nel periodo 1981-1987:

anni	$\mathbf{x_i}$
1	608,6
2	694
3	726,4
4	760,9
5	887,6
6	904,6
7	1128,6
TOTALE	**5710,7**

Calcoliamo gli indici di variazione V_1 e V_2; occorre effettuare tutti i possibili scarti dalla media:

$\mathbf{x_i}$	$\mathbf{x_i - M}$	$\mathbf{\lvert x_i - M \rvert}$	$\mathbf{(x_i - M)^2}$
608,6	-207,2	207,2	42931,84
694	-121,8	121,8	14835,24
726,4	-89,4	89,4	7992,36
760,9	-54,9	54,9	3014,01
887,6	71,8	71,8	5155,24
904,6	88,8	88,8	7885,44
1128,6	312,8	312,8	97843,84
5710,7		**946,7**	**179658**

$$M = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{5710,7}{7} = 815,8143 \text{ mm}$$

$$V_1 = \frac{\sum_{i=1}^{n} |x_i - M|}{n} = \frac{946,7}{7} = 135,2429 \text{ mm}$$

$$V_2 = \sqrt{\frac{\sum_{i=1}^{n} (x_i - M)^2}{n}} = \sqrt{\frac{179658}{7}} = 160,2043 \text{ mm}$$

Si noti che $V_1 < V_2$, in quanto V_1 altro non è che una media aritmetica di scarti, mentre V_2 ne è una media quadratica e, ricordando la relazione che lega le medie potenziate di ordine m (cfr.par. 3.1), la media aritmetica risulta inferiore alla media quadratica.

Il quadrato di V_2 è la varianza:

$\sigma^2 = V_2^2 = 25665,42\ mm^2$.

Volendo evitare di calcolare tutti gli scarti, si può calcolare σ^2 con la formula ridotta; occorrono, in tal caso, solo le due colonne:

x_i	x_i^2
608,6	370394
694	481636
726,4	527657
760,9	578968,8
887,6	787833,8
904,6	818301,2
1128,6	1273738
5710,7	**4838529**

$$\sigma^2 = \frac{\sum_{i=1}^{n} x_i^2}{n} - M^2 = \frac{4838529}{7} - (815{,}8143)^2 = 25665{,}42\,.$$

Consideriamo adesso la distribuzione relativa dei redditi familiari in Italia nel 1983 (in milioni di £):

x_i-x_{i+1}	n_i
0-4	18
4-6	41
6-8	52
8-10	84
10-12	98
12-14	89
14-16	90
16-18	76
18-20	66
20-22	55
22-25	69
25-30	97
30-35	50
35-40	45
40-45	27
45-50	14
50-100	29
totale	**1000**

Calcoliamo V_1 e V_2:

$_cx_i$	n_i	$_cx_i\ n_i$	$_cx_i - M$	$\lvert {}_cx_i - M \rvert$	$\lvert {}_cx_i - M \rvert\ n_i$	$({}_cx_i - M)^2$	$({}_cx_i - M)^2\ n_i$
2	18	36	-18,236	18,236	328,248	332,5517	5985,931
5	41	205	-15,236	15,236	624,676	232,1357	9517,564
7	52	364	-13,236	13,236	688,272	175,1917	9109,968
9	84	756	-11,236	11,236	943,824	126,2477	10604,81
11	98	1078	-9,236	9,236	905,128	85,3037	8359,762
13	89	1157	-7,236	7,236	644,004	52,3597	4660,013
15	90	1350	-5,236	5,236	471,24	27,4157	2467,413
17	76	1292	-3,236	3,236	245,936	10,4717	795,8489
19	66	1254	-1,236	1,236	81,576	1,527696	100,8279
21	55	1155	0,764	0,764	42,02	0,583696	32,10328
23,5	69	1621,5	3,264	3,264	225,216	10,6537	735,105
27,5	97	2667,5	7,264	7,264	704,608	52,7657	5118,273
32,5	50	1625	12,264	12,264	613,2	150,4057	7520,285
37,5	45	1687,5	17,264	17,264	776,88	298,0457	13412,06
42,5	27	1147,5	22,264	22,264	601,128	495,6857	13383,51
47,5	14	665	27,264	27,264	381,696	743,3257	10406,56
75	29	2175	54,764	54,764	1588,156	2999,096	86973,78
	1000	**20236**			**9865,808**		**189183,8**

$$M = \frac{\sum_{i=1}^{n} {}_c x_i n_i}{n} = \frac{20236}{1000} = 20{,}236 \text{ milioni di £}$$

$$V_1 = \frac{\sum_{i=1}^{n} \left| {}_c x_i - M \right| n_i}{n} = \frac{9865{,}808}{1000} = 9{,}8658 \text{ milioni di £}$$

$$V_2 = \sqrt{\frac{\sum_{i=1}^{n} \left({}_c x_i - M \right)^2 n_i}{n}} = \sqrt{\frac{189183{,}8}{1000}} = 13{,}7544 \text{ milioni di £.}$$

Volendo utilizzare la formula ridotta per il calcolo di V_2:

$_cx_i$	n_i	$_cx_i\, n_i$	$_cx_i^{\,2}$	$_cx_i^{\,2}\, n_i$
2	18	36	4	72
5	41	205	25	1025
7	52	364	49	2548
9	84	756	81	6804
11	98	1078	121	11858
13	89	1157	169	15041
15	90	1350	225	20250
17	76	1292	289	21964
19	66	1254	361	23826
21	55	1155	441	24255
23,5	69	1621,5	552,25	38105,25
27,5	97	2667,5	756,25	73356,25
32,5	50	1625	1056,25	52812,5
37,5	45	1687,5	1406,25	63281,25
42,5	27	1147,5	1806,25	48768,75
47,5	14	665	2256,25	31587,5
75	29	2175	5625	163125
	1000	**20236**		**598679,5**

$$V_2 = \sqrt{\frac{\sum_{i=1}^{n} {}_c x_i^2 n_i}{n} - M^2} = \sqrt{\frac{598679,5}{1000} - (20,236)^2} = 13,7544 \text{ milioni di £.}$$

Si noti, anche in questo caso, che $V_1 < V_2$.

4.3.2 Esempi sugli indici di dispersione

Riprendiamo l'esempio relativo ai valori di tensione misurati in un punto preciso di un circuito (cfr.par.3.3). Si tratta di misure ripetute della stessa grandezza, dunque gli indici di variabilità assoluta più idonei sono in tal caso gli indici di dispersione.

Se supponiamo che i dati provengano da una distribuzione normale ($p=2$), l'indice di dispersione più appropriato è lo scarto quadratico medio σ_2:

$_cx_i$	n_i	$_cx_i\ n_i$	$_cx_i$-M_1	$(_cx_i$-$M_1)^2$	$(_cx_i$-$M_1)^2\ n_i$	$_cx_i\ ^2$	$_cx_i\ ^2 n_i$
5,113	1	5,113	-0,0289	0,000835	0,0008352	26,138	26,138
5,118	2	10,235	-0,0239	0,000571	0,0011424	26,189	52,378
5,123	5	25,613	-0,0189	0,000357	0,0017860	26,240	131,200
5,128	13	66,658	-0,0139	0,000193	0,0025117	26,291	341,786
5,133	25	128,313	-0,0089	0,000079	0,0019802	26,343	658,564
5,138	26	133,575	-0,0039	0,000015	0,0003955	26,394	686,242
5,143	22	113,135	0,0011	0,000001	0,0000266	26,445	581,797
5,148	20	102,950	0,0061	0,000037	0,0007442	26,497	529,935
5,153	18	92,745	0,0111	0,000123	0,0022178	26,548	477,869
5,158	12	61,890	0,0161	0,000259	0,0031105	26,600	319,198
5,163	4	20,650	0,0211	0,000445	0,0017808	26,651	106,606
5,168	2	10,335	0,0261	0,000681	0,0013624	26,703	53,406
totale	150	771,210			0,0178935		3965,117

$$M_1 = \frac{\sum_{i=1}^{n} {}_c x_i n_i}{n} = \frac{771{,}21}{150} = 5{,}1414$$

$$\sigma_2 = \sqrt{\frac{\sum_{i=1}^{n} \left({}_c x_i - M_1 \right)^2 n_i}{n}} = \sqrt{\frac{0{,}017893}{150}} = 0{,}010922$$

Se avessimo considerato la mediana M_0, anziché la media aritmetica M_1, avremmo ottenuto un valore più alto per σ_2, essendo M_1 la media decisionale che minimizza la funzione di perdita quando $p=2$.

Con la formula ridotta:

$$\sigma_2 = \sqrt{\frac{\sum_{i=1}^{n} {}_c x_i^2 n_i}{n} - M_1^2} = \sqrt{\frac{3965{,}117}{150} - \left(5{,}1414\right)^2} = 0{,}010922\,.$$

Supponiamo adesso che i dati provengano da una distribuzione di Laplace ($p=1$); l'indice di dispersione più appropriato è lo scostamento semplice medio dalla mediana σ_1:

$_cx_i$	n_i	N_i	$_cx_i$-M_0	$\lvert_cx_i$-$M_0\rvert$	$\lvert_cx_i$-$M_0\rvert\, n_i$
5,113	1	1	-0,0282	0,0282	0,0282
5,118	2	3	-0,0232	0,0232	0,0464
5,123	5	8	-0,0182	0,0182	0,0910
5,128	13	21	-0,0132	0,0132	0,1716
5,133	25	46	-0,0082	0,0082	0,2050
5,138	26	72	-0,0032	0,0032	0,0832
5,143	22	94	0,0018	0,0018	0,0396
5,148	20	114	0,0068	0,0068	0,1360
5,153	18	132	0,0118	0,0118	0,2124
5,158	12	144	0,0168	0,0168	0,2016
5,163	4	148	0,0218	0,0218	0,0872
5,168	2	150	0,0268	0,0268	0,0536
totale	150			0,1800	1,3558

Calcolata la mediana M_0=5,1407 con il metodo dell'interpolazione (cfr.par. 3.2), si ha :

$$\sigma_1 = \frac{\sum_{i=1}^{n} \left|{}_c x_i - M_0\right| n_i}{n} = \frac{1{,}3558}{150} = 0{,}009039\,.$$

Se avessimo considerato la media aritmetica M_1 anziché la mediana M_0, avremmo ottenuto un valore più alto per σ_1, essendo M_0 la media decisionale che minimizza la funzione di perdita quando $p=1$.

4.3.3 Esempi sugli indici di diversità

Supponiamo di aver rilevato il numero di figli di 5 famiglie:

x_1	x_2	x_3	x_4	x_5
4	*2*	*1*	*3*	*2*

Calcoliamo la differenza semplice media con ripetizione ${}_R\Delta_1$ e senza ripetizione Δ_1.

Tutte le possibili differenze $d_{ij}=|x_i-x_j|$ sono ${}_RD_{n,2}=n^2=5^2=25$:

d_{11}	d_{12}	d_{13}	d_{14}	d_{15}
d_{21}	d_{22}	d_{23}	d_{24}	d_{25}
d_{31}	d_{32}	d_{33}	d_{34}	d_{35}
d_{41}	d_{42}	d_{43}	d_{44}	d_{45}
d_{51}	d_{52}	d_{53}	d_{54}	d_{55}

mentre le differenze senza ripetizione sono $D_{n,2}=n(n-1)=5 \cdot 4=20$.

Essendo la matrice delle differenze simmetrica, basta calcolare solo le differenze per cui $i<j$, ossia tutte le differenze al di sopra della diagonale principale:

|4-2|	|4-1|	|4-3|	|4-2|	2	3	1	2
	|2-1|	|2-3|	|2-2|		1	1	0
		|1-3|	|1-2|			2	1
			|3-2|				1

Quindi è:

$$_R\Delta_1 = \frac{2\sum\left|x_i - x_j\right|}{n^2} = \frac{2(2+3+1+2+1+1+0+2+1+1)}{25} = \frac{28}{25} = 1{,}12$$

e se non si considerano le differenze ripetute:

$$\Delta_1 = \frac{2\sum\left|x_i - x_j\right|}{n(n-1)} = \frac{28}{20} = 1{,}4 \ .$$

Consideriamo adesso la seguente distribuzione di frequenza; dobbiamo tener conto che ciascuna differenza si ripete n_in_j volte:

x_i	n_i	$\lvert x_i\text{-}x_j\rvert$				$\lvert x_i\text{-}x_j\rvert n_in_j$				$\lvert x_i\text{-}x_j\rvert n_in_j$			
2	3	|2-6|	|2-7|	|2-9|	|2-12|	4·3·5	5·3·6	7·3·4	10·3·2	60	90	84	60
6	5		|6-7|	|6-9|	|6-12|		1·5·6	3·5·4	6·5·2		30	60	60
7	6			|7-9|	|7-12|			2·6·4	5·6·2			48	60
9	4				|9-12|				3·4·2				24
12	2												
totale	20												576

Dunque è:

$$_R\Delta_1 = \frac{2\sum_{i<j}\left|x_i - x_j\right| n_i n_j}{n^2} = \frac{2\cdot 576}{20^2} = 2{,}88$$

e se non si considerano le differenze ripetute:

$$\Delta_1 = \frac{2\sum_{i<j}\left|x_i - x_j\right| n_i n_j}{n(n-1)} = \frac{2\cdot 576}{20\cdot 19} = 3{,}0316\,.$$

Se il totale delle osservazioni *n* è molto grande, per evitare di calcolare tutte le differenze, si può ricorrere alla formula di De Finetti-Paciello:

x_i	n_i	N_i	$n-N_i$	$N_i(n-N_i)$	$x_{i+1}-x_i$	$N_i(n-N_i)(x_{i+1}-x_i)$
2	3	3	17	51	4	204
6	5	8	12	96	1	96
7	6	14	6	84	2	168
9	4	18	2	36	3	108
12	2	20				
totale	20					576

$$\Delta_1 = \frac{2\sum_{i} N_i(n-N_i)(x_{i+1}-x_i)}{n(n-1)} = \frac{2\cdot 576}{20\cdot 19} = 3{,}0316\,.$$

Se le x_i sono in progressione aritmetica di ragione *h*, il numeratore si riduce ulteriormente:

$$\Delta_1 = \frac{2h\sum_{i} N_i(n-N_i)}{n(n-1)}\,.$$

4.4 Esempi sugli indici di variabilità relativa

L'indice di variabilità relativa più utilizzato è il "coefficiente di variazione":

$$CV = \frac{\sigma}{M}100\,.$$

Vediamo due esempi in cui è necessaria la sua applicazione.

Esempio 1

Le distribuzioni dei pesi e delle stature di un gruppo di studenti hanno presentato media e scarto quadratico medio come dal seguente prospetto; verificare se risulta maggiore la variabilità della distribuzione dei pesi o delle stature.

	M	σ
Peso (kg)	59,4	7,5
Statura (cm)	173,2	8,2

Si tratta di confrontare la variabilità di due distribuzioni espresse con diversa unità di misura, per cui non ha senso confrontare i due scarti quadratici medi, che sono espressi l'uno in kg e l'altro in cm.

Calcoliamo pertanto i due coefficienti di variazione, che sono numeri puri o adimensionali:

PESO $$CV = \frac{\sigma}{M}100 = \frac{7,5}{59,4}100 = 12,6\%$$

STATURA $$CV = \frac{\sigma}{M}100 = \frac{8,2}{173,2}100 = 4,7\%$$

Dal confronto dei due coefficienti di variazione risulta maggiore la variabilità della distribuzione dei pesi.

Esempio 2

In una regione si hanno 9 industrie che hanno installato un dispositivo anti-inquinante di tipo A ed altre 9 che hanno installato un dispositivo anti-inquinante di tipo B. Di seguito vengono riportate le quantità (in grammi al minuto) di pulviscolo eliminate giornalmente dalle industrie con i dispositivi A e B:

Industrie	x_A	x_B	x_A^2	x_B^2
1	69	35	4761	1225
2	80	62	6400	3844
3	44	43	1936	1849
4	52	23	2704	529
5	54	30	2916	900
6	54	28	2916	784
7	86	22	7396	484
8	77	40	5929	1600
9	66	25	4356	625
totale	582	308	39314	11840

$$M_A = \frac{\sum x_A}{n} = \frac{582}{9} = 64{,}67 \ \ gr/min$$

$$M_B = \frac{\sum x_B}{n} = \frac{308}{9} = 34{,}22 \ \ gr/min.$$

Si tratta di due distribuzioni che, pur essendo espresse nella stessa unità di misura, presentano un ordine medio di grandezza diverso dunque, per confrontarne la variabilità, calcoliamo i coefficienti di variazione, da cui risulta più variabile la distribuzione B:

$$\sigma_A = \sqrt{\frac{\sum x_A^2}{n} - M_A^2} = \sqrt{\frac{39314}{9} - (64{,}67)^2} = 13{,}64 \ \text{gr/min}$$

$$\sigma_B = \sqrt{\frac{\sum x_B^2}{n} - M_B^2} = \sqrt{\frac{11840}{9} - (34{,}22)^2} = 12{,}02 \ \text{gr/min}$$

$$CV_A = \frac{\sigma_A}{M_A} 100 = \frac{13{,}64}{64{,}67} 100 = 21\%$$

$$CV_B = \frac{\sigma_B}{M_B} 100 = \frac{12{,}02}{34{,}22} 100 = 35\% .$$

4.5 Proprietà della varianza

Sia X una variabile statistica con media M_X e varianza σ_X^2.

Sia Y una trasformazione lineare di X: $Y = \alpha + \beta X$; dimostriamo che:

$$\sigma_Y^2 = \beta^2 \sigma_X^2$$

dimostrazione

$M_Y = \alpha + \beta M_X$ (cfr.par.3.4)

$$\sigma_Y^2 = \frac{\sum_{i=1}^{n}(y_i - M_Y)^2}{n} = \frac{\sum_{i=1}^{n}(\alpha + \beta x_i - \alpha - \beta M_X)^2}{n} = \frac{\sum_{i=1}^{n}(\beta x_i - \beta M_X)^2}{n} = \beta^2 \sigma_X^2$$

Se $\beta=1$, $Y=\alpha+X$ ed è:

$\sigma_Y^2 = \sigma_X^2$

Ovvero, la varianza è invariante per traslazione.

Se $\alpha=0$, $Y=\beta X$ ed è ancora:

$\sigma_Y^2 = \beta^2 \sigma_X^2$.

Consideriamo adesso una variabile statistica X con media M_X e varianza σ_X^2 e una variabile statistica Y con media M_Y e varianza σ_Y^2. Sia $Z=X+Y$. Dimostriamo che:

$\sigma_Z^2=\sigma_X^2+\sigma_Y^2+2\sigma_{XY}$

dimostrazione

$M_Z=M_X+M_Y$ (cfr.par. 3.4)

$$\sigma_Z^2 = \frac{\sum_{i=1}^{n}(z_i - M_Z)^2}{n} = \frac{\sum_{i=1}^{n}(x_i + y_i - M_X - M_Y)^2}{n} = \frac{\sum_{i=1}^{n}[(x_i - M_X) + (y_i - M_Y)]^2}{n} =$$

$$= \frac{\sum_{i=1}^{n}(x_i - M_X)^2}{n} + \frac{\sum_{i=1}^{n}(y_i - M_Y)^2}{n} + 2\frac{\sum_{i=1}^{n}(x_i - M_X)(y_i - M_Y)}{n} = \sigma_X^2+\sigma_Y^2+2\sigma_{XY}$$

L'espressione $\sigma_{XY} = \frac{\sum_{i=1}^{n}(x_i - M_X)(y_i - M_Y)}{n}$ viene definita covarianza (cfr.par. 7.6); se X ed Y sono indipendenti in distribuzione (cfr.par. 7.2), allora $\sigma_{XY} = 0$ e $\sigma_Z^2=\sigma_X^2+\sigma_Y^2$.

4.6 Indici di eterogeneità

Gli indici di variabilità fin qui presentati possono essere utilizzati solo per variabili quantitative.

Con il termine “eterogeneità” si indica, in genere, la diversità fra le modalità di un carattere qualitativo.

Se tutte le unità statistiche rilevate presentano la stessa modalità del carattere, si dice che l’eterogeneità è nulla o che la concentrazione è massima:

x_i	n_i	f_i	f_i^2
x_1	n	1	1
x_2	0	0	0
x_3	0	0	0
…	…	…	…
x_k	0	0	0
totale	n	1	1

Se tutte le unità statistiche sono ripartite uniformemente fra le k modalità del carattere, allora l’eterogeneità è massima:

x_i	n_i	f_i	f_i^2
x_1	n/k	$1/k$	$1/k^2$
x_2	n/k	$1/k$	$1/k^2$
x_3	n/k	$1/k$	$1/k^2$
…	…	…	…
x_k	n/k	$1/k$	$1/k^2$
totale	n	1	$1/k$

Per valutare l’eterogeneità di una distribuzione, Gini ha proposto il seguente indice:

$$G = 1 - \sum_{i=1}^{k} f_i^{\,2} .$$

In caso di eterogeneità nulla è $G = 1 - \sum_{i=1}^{k} f_i^{\,2} = 1 - 1 = 0$.

In caso di eterogeneità massima è $G = 1 - \sum_{i=1}^{k} f_i^{\,2} = 1 - 1/k = \frac{k-1}{k}$.

Un indice di eterogeneità relativo è dunque dato da:

$$G' = \frac{G}{(k-1)/k}.$$

Supponiamo di avere rilevato su un gruppo di soggetti, separatamente per i due sessi, il titolo di studio. Si vuole confrontare l'eterogeneità delle due distribuzioni:

Femmine

x_i	n_i	f_i	f_i^2
Lic. media	2	0,17	0,03
Maturità	7	0,58	0,34
Laurea	3	0,25	0,06
totale	12	1,0	0,43

Maschi

x_i	n_i	f_i	f_i^2
Lic. elem.	3	0,17	0,03
Lic. media	6	0,33	0,11
Maturità	8	0,44	0,20
Laurea	1	0,06	0,00
totale	18	1,0	0,34

$$G_F' = \frac{G_F}{(k-1)/k} = \frac{0{,}57}{(3-1)/3} = 0{,}85$$

$$G_M' = \frac{G_M}{(k-1)/k} = \frac{0{,}66}{(4-1)/4} = 0{,}88\,.$$

In entrambi i casi *G'* risulta più vicino ad 1 che a 0, dunque c'è eterogeneità.

Inoltre, risulta più eterogenea la distribuzione dei maschi.