La variabilità

- L'utilizzo di una ***media*** permette di sintetizzare efficacemente l'informazione contenuta in una distribuzione statistica dal punto di vista dell'*intensità* del carattere. Tuttavia la sintesi può essere eccessiva, nel senso si possono perdere informazioni su altre caratteristiche fondamentali come la *variabilità*.

- La ***variabilità*** è definibile come la tendenza delle unità di un collettivo ad assumere modalità diverse tra loro.

- La Statistica (Trilussa)

Sai ched'è la statistica? È na' cosa
che serve pe fà un conto in generale
de la gente che nasce, che sta male,
che more, che va in carcere e che spósa.
Ma pè me la statistica curiosa
è dove c'entra la percentuale,
pè via che, lì, la media è sempre eguale
puro co' la persona bisognosa.
Me spiego: da li conti che se fanno
seconno le statistiche d'adesso
risurta che te tocca un pollo all'anno:
e, se nun entra nelle spese tue,
t'entra ne la statistica lo stesso
perch'è c'è un antro che ne magna due.

Esempio

- Consideriamo le distribuzioni secondo il numero di figli in due collettivi diversi di 25 famiglie ciascuno.

Popolazione 1

N. Figli (x_j)	Frequenze (n_j)
0	1
1	4
2	15
3	3
4	2
Totale	25

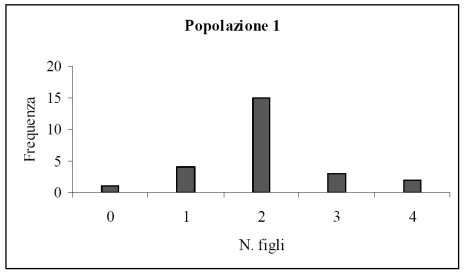
Popolazione 2

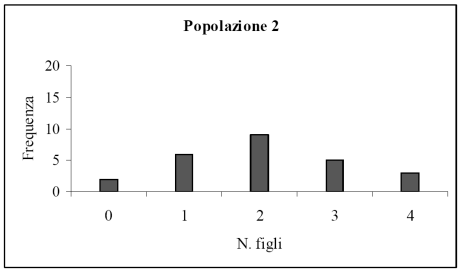
N. Figli (x_j)	Frequenze (n_j)
0	2
1	6
2	9
3	5
4	3
Totale	25

$$\bar{x}_1 = \frac{1}{n}\sum_{j=1}^{k} x_j n_j = \frac{1}{25}51 = 2{,}04$$

$$\bar{x}_2 = \frac{1}{n}\sum_{j=1}^{k} x_j n_j = \frac{1}{25}51 = 2{,}04$$

- Entrambe le distribuzioni hanno media 2,04 ma, come è possibile dedurre dai grafici, sono molto diverse: la prima assume delle modalità molto più concentrate attorno alla media e quindi ha minore variabilità.





Indici di variabilità

- Per avere una descrizione più completa della distribuzione è quindi opportuno utilizzare, oltre a una media, un ***indice*** che misuri la variabilità della distribuzione.
- Un indice di variabilità deve:
  - assumere il valore minimo (tipicamente 0) se e solo se tutte le unità della distribuzione presentano la stessa modalità;
  - aumentare all'aumentare della diversità tra le modalità del carattere assunte dalle varie unità.

- Gli indici di variabilità possono essere basati:
  - sullo *scostamento da una media*;
  - sulla *differenze tra statistiche d'ordine*.

Scostamenti da una media	**Differenze tra statistiche d'ordine**
Devianza	Campo di variazione
Varianza	Differenza interquartilica
Deviazione standard	
Coefficiente di variazione	
Scostamento semplice medio dalla mediana	

Devianza e varianza

- Per una *distribuzione unitaria* di un carattere *quantitativo*, la ***devianza*** è definita come

$$D = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

- Per una *distribuzione di frequenza* non in classi

$$D = \sum_{j=1}^{k} (x_j - \bar{x})^2 n_j$$

- Se il *carattere è in classi* si utilizzano i *valori centrali*

$$x_j = \frac{c_{j-1} + c_j}{2}$$

al posto delle modalità.

- La ***varianza*** è normalmente preferita alla *devianza* e si ottiene come:

$$\sigma^2 = \frac{D}{n} = \frac{1}{n}\sum_{j=1}^{k} (x_j - \bar{x})^2 n_j = \sum_{j=1}^{k} (x_j - \bar{x})^2 f_j$$

Esempi

- Per la distribuzione unitaria dei voti si ha:

Unità (i)	Voto (x_i)	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
1	25	0	0
2	27	2	4
3	22	-3	9
4	26	1	1
5	30	5	25
6	20	-5	25
Totale	150	0	64

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{6}150 = 25 \qquad \sigma^2 = \frac{D}{n} = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{1}{6}64 = 10{,}667$$

- Per la prima distribuzione del numero di figli per un collettivo di 25 famiglie, che ha media 2,04, si ha:

Popolazione 1

N. Figli (x_j)	Frequenze (n_j)	$x_j - \bar{x}$	$(x_j - \bar{x})^2$	$(x_j - \bar{x})^2 n_j$
0	1	-2,04	4,162	4,162
1	4	-1,04	1,082	4,326
2	15	-0,04	0,002	0,024
3	3	0,96	0,922	2,765
4	2	1,96	3,842	7,683
Totale	25	--	--	18,960

$$\bar{x}_1 = \frac{1}{n}\sum_{j=1}^{k} x_j n_j = \frac{1}{25}51 = 2{,}04$$

$$\sigma_1^2 = \frac{D}{n} = \frac{1}{n}\sum_{j=1}^{k}(x_j - \bar{x}_1)^2 n_j = \frac{1}{25}18{,}960 = 0{,}758$$

- Per la seconda distribuzione, che ha sempre media 2,04, si ha:

Popolazione 2

N. Figli (x_j)	Frequenze (n_j)	$x_j - \bar{x}$	$(x_j - \bar{x})^2$	$(x_j - \bar{x})^2 n_j$
0	2	-2,04	4,162	8,323
1	6	-1,04	1,082	6,490
2	9	-0,04	0,002	0,014
3	5	0,96	0,922	4,608
4	3	1,96	3,842	11,525
Totale	25	--	--	30,960

$$\bar{x}_2 = \frac{1}{n}\sum_{j=1}^{k} x_j n_j = \frac{1}{25}51 = 2{,}04$$

$$\sigma_2^2 = \frac{D}{n} = \frac{1}{n}\sum_{j=1}^{k}(x_j - \bar{x}_2)^2 n_j = \frac{1}{25}30{,}960 = 1{,}238$$

- Per la distribuzione dell'altezza per un collettivo di 50 soggetti si ha:

Altezza (c_{j-1}--c_j)	Freq. (n_j)	Valori centrali (x_j)	$x_j n_j$	$x_j - \bar{x}$	$(x_j - \bar{x})^2$	$(x_j - \bar{x})^2 n_j$
150-160	1	155	155	-18,8	353,44	353,44
160-170	10	165	1650	-8,8	77,44	774,4
170-180	35	175	6125	1,2	1,44	50,4
180-200	4	190	760	16,2	262,44	1049,76
Totale	50	--	8690	--	--	2228

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{k} x_j n_j = \frac{1}{50}8690 = 173{,}8$$

$$\sigma^2 = \frac{1}{n}\sum_{j=1}^{k}(x_j - \bar{x})^2 n_j = \frac{1}{50}2228 = 44{,}56$$

Deviazione standard e coefficiente di variazione

- La ***deviazione standard*** (o ***scostamento quadratico medio***) è l'indice di variabilità più utilizzato in quanto è espresso nella stessa unità di misura del carattere. Si ottiene come:

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{1}{n}\sum_{j=1}^{k}(x_j - \bar{x})^2 n_j}$$

- Nel caso in cui la distribuzione abbia media aritmetica positiva, il ***coefficiente di variazione*** si calcola come (normalmente in percentuale):

$$CV = \frac{\sigma}{\bar{x}}100$$

Esempi

- Per la distribuzione dei voti

$$\sigma=\sqrt{10{,}667}=3{,}27 \qquad CV=\frac{3{,}27}{25}100=13{,}1\%$$

- Per le distribuzioni del numero di figli:
  - Popolazione 1 $\sigma=\sqrt{0{,}758}=0{,}871$ $\quad CV=\frac{0{,}871}{2{,}04}100=42{,}69\%$
  - Popolazione 2 $\sigma=\sqrt{1{,}238}=1{,}113$ $\quad CV=\frac{1{,}113}{2{,}04}100=54{,}55\%$

- Per la distribuzione dell'altezza:

$$\sigma=\sqrt{44{,}56}=6{,}675 \qquad CV=\frac{6{,}675}{173{,}8}100=3{,}84\%$$

Proprietà

- **Proprietà 1**: gli indici D, σ^2 e σ sono sempre non negativi e assumono il valore minimo (0) se e solo se tutte le modalità della distribuzione sono uguali tra loro.

- **Proprietà 2**: la devianza può essere calcolata come (*formula semplificata*)

$$D=\sum_{i=1}^{n}x_i^2-n\bar{x}^2 \qquad \text{(distribuzione unitaria)}$$

$$D=\sum_{j=1}^{k}x_j^2 n_j-n\bar{x}^2 \qquad \text{(distribuzione di frequenze)}$$

che ha vantaggi nel calcolo anche della varianza e della deviazione standard

- **Proprietà 3**: se a ogni termine della distribuzione viene applicata la trasformazione *aX* + *b*, allora gli indici di variabilità cambieranno nel modo seguente:

Devianza	-------->	a^2D
Varianza	-------->	$a^2\sigma^2$
Deviazione standard	-------->	$a\sigma$

Esempi

- Per la distribuzione unitaria dei voti si ha:

Unità (i)	Voto (x_i)	$x_i-\bar{x}$	$(x_i-\bar{x})^2$	x_i^2
1	25	0	0	625
2	27	2	4	729
3	22	-3	9	484
4	26	1	1	676
5	30	5	25	900
6	20	-5	25	400
Totale	150	0	64	3814

$$\bar{x}=\frac{1}{6}150=25 \qquad \sigma^2=\frac{D}{n}=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2=\frac{1}{6}64=10{,}667$$

$$\sigma^2=\frac{D}{n}=\frac{1}{n}\sum_{i=1}^{n}x_i^2-\bar{x}^2=\frac{1}{6}3814-25^2=10{,}667$$

- Per la prima distribuzione del numero di figli per un collettivo di 25 famiglie, che ha media 2,04, si ha:

Popolazione 1

N. Figli (x_j)	Frequenze (n_j)	$x_j - \bar{x}$	$(x_j - \bar{x})^2$	$(x_j - \bar{x})^2 n_j$	x_j^2	$x_j^2 n_j$
0	1	-2,04	4,162	4,162	0	0
1	4	-1,04	1,082	4,326	1	4
2	15	-0,04	0,002	0,024	4	60
3	3	0,96	0,922	2,765	9	27
4	2	1,96	3,842	7,683	16	32
Totale	25	--	--	18,960	--	123

$$\bar{x}_1 = \frac{1}{25}51 = 2{,}04 \qquad \sigma_1^2 = \frac{D}{n} = \frac{1}{n}\sum_{j=1}^{k}(x_j - \bar{x}_1)^2 n_j = \frac{1}{25}18{,}960 = 0{,}758$$

$$\sigma_1^2 = \frac{D}{n} = \frac{1}{n}\sum_{j=1}^{k} x_j^2 n_j - \bar{x}^2 = \frac{1}{25}123 - 2{,}04^2 = 0{,}758$$

Scostamenti semplici medi

- Per una *distribuzione unitaria* di un *carattere quantitativo*, lo ***scostamento semplice medio dalla media aritmetica*** è definito come

$$S_{\bar{x}} = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|$$

- Per una *distribuzione di frequenza* non in classi

$$S_{\bar{x}} = \frac{1}{n}\sum_{j=1}^{k}|x_j - \bar{x}|n_j$$

- Se il *carattere è in classi* si utilizzano i valori centrali

$$x_j = \frac{c_{j-1} + c_j}{2}$$

al posto delle modalità.

- Lo ***scostamento semplice medio dalla mediana*** si ottiene sostituendo la mediana alla media aritmetica:

$$S_{Me} = \frac{1}{n}\sum_{i=1}^{n}|x_i - Me| \qquad \text{(\textit{distribuzione unitaria})}$$

$$S_{Me} = \frac{1}{n}\sum_{j=1}^{k}|x_j - Me|n_j \qquad \text{(\textit{distribuzione di frequenza})}$$

Esempi

- Per la distribuzione unitaria dei voti si ha:

Unità (i)	Voto (x_i)	$\lvert x_i - \bar{x}\rvert$	$\lvert x_i - Me\rvert$
1	25	0	0,5
2	27	2	1,5
3	22	3	3,5
4	26	1	0,5
5	30	5	4,5
6	20	5	5,5
Totale	150	16	16

$$\bar{x} = \frac{1}{6}150 = 25 \qquad S_{\bar{x}} = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}| = 16/6 = 2{,}667$$

$$Me = 25{,}5 \qquad S_{Me} = \frac{1}{n}\sum_{i=1}^{n}|x_i - Me| = 16/6 = 2{,}667$$

- Per la prima distribuzione del numero di figli per un collettivo di 25 famiglie, che ha media 2,04, si ha:

Popolazione 1

N. Figli (x_j)	Frequenze (n_j)	$\lvert x_j - \bar{x} \rvert$	$\lvert x_j - \bar{x} \rvert n_j$	$\lvert x_j - Me \rvert$	$\lvert x_j - Me \rvert n_j$
0	1	2,04	2,04	2	2
1	4	1,04	4,16	1	4
2	15	0,04	0,6	0	0
3	3	0,96	2,88	1	3
4	2	1,96	3,92	2	4
Totale	25	--	13,6	--	13

$$\bar{x}_1 = \frac{1}{25} 51 = 2{,}04 \qquad S_{\bar{x}} = \frac{1}{n} \sum_{j=1}^{k} \lvert x_j - \bar{x} \rvert n_j = \frac{13{,}60}{25} = 0{,}544$$

$$Me = 2 \qquad S_{Me} = \frac{1}{n} \sum_{j=1}^{k} \lvert x_j - Me \rvert n_j = \frac{13}{25} = 0{,}52$$

Altri indici di variabilità

- Per una distribuzione con modalità ordinate, $x_1, \ldots, x_k$, il ***campo di variazione*** è definito come

$$R = x_n - x_1$$

- E' l'indice di variabilità più semplice da calcolare, ma non è molto efficace nel misurare la variabilità della distribuzione.

- La differenza interquartilica si basa sul primo quartile (Q_1) e il terzo quartile (Q_3) ed è definita come

$$W = Q_3 - Q_1$$

Esempi

- Per la distribuzione unitaria dei voti si ha:

$$R = x_n - x_1 = 30 - 20 = 10$$
$$W = Q_3 - Q_1 = 27 - 22 = 5$$

- Per entrambe le distribuzioni del numero di figli si ha:

$$R = x_n - x_1 = 4 - 0 = 4$$

- Per la distribuzione delle altezze si ha:

$$W = 177{,}57 - 170{,}43 = 7{,}14$$