Caso studio 6

Un investitore sta valutando il rendimento di due titoli alternativi. Sulla base del logaritmo dei rendimenti giornalieri della settimana passata vuole decidere quale titolo convenga acquistare. I dati sono riportati di seguito:

Data	Titolo 1	Titolo 2
01/02/2015	0,17	0,57
02/02/2015	1,22	0,64
03/02/2015	0,92	-0,01
04/02/2015	-0,53	0,71
05/02/2015	0,88	0,75

In entrambi i casi, la media aritmetica del logaritmo dei rendimenti è pari a 0,532. L'investitore può concludere che sia indifferente scegliere il primo o il secondo titolo? Oppure c'è qualche altra considerazione che possa essere fatta?

1

2

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.

3

Esempio

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

Popolazione 1

N. Figli (x_i)	Frequenze (n_i)
0	1
1	4
2	15
3	3
4	2
Totale	25

Popolazione 2

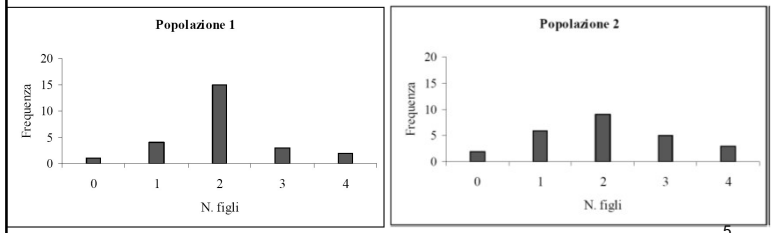
N. Figli (x_i)	Frequenze (n_i)
0	2
1	6
2	9
3	5
4	3
Totale	25

$$\bar{x}_1 = \frac{1}{n}\sum_{i=1}^{k} x_i n_i = \frac{1}{25}51 = 2{,}04$$

$$\bar{x}_2 = \frac{1}{n}\sum_{i=1}^{k} x_i n_i = \frac{1}{25}51 = 2{,}04$$

4

- 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

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

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

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

al posto delle modalità.

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

$$\sigma^2 = \frac{D}{n} = \frac{1}{n} \sum_{i=1}^{k} (x_i - \bar{x})^2 n_i = \sum_{i=1}^{k} (x_i - \bar{x})^2 f_i$$

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_i)	Frequenze (n_i)	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^2 n_i$
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

$$\bar{x}_1 = \frac{1}{n}\sum_{i=1}^{k} x_i n_i = \frac{1}{25}51 = 2{,}04$$

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

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

Popolazione 2

N. Figli (x_i)	Frequenze (n_i)	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^2 n_i$
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

$$\bar{x}_2 = \frac{1}{n}\sum_{i=1}^{k} x_i n_i = \frac{1}{25}51 = 2{,}04$$

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

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

Altezza (c_{i-1}--c_i)	Freq. (n_i)	Valori centrali (x_i)	$x_i n_i$	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^2 n_i$
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

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} x_i n_i = \frac{1}{50}8690 = 173{,}8$$

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{k}(x_i - \bar{x})^2 n_i = \frac{1}{50}2228 = 44{,}56$$

Esercizio

Si considerino i dati del Caso Studio 6.

Si calcoli la varianza del logaritmo dei rendimenti, separatamente per i due titoli.

Sulla base del risultato ottenuto, si determini in quale titolo sia meglio investire.

Titolo 1	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
0,17	-0,362	0,131
1,22	0,688	0,473
0,92	0,388	0,151
-0,53	-1,062	1,128
0,88	0,348	0,121
2,66	**0**	**2,004**

$\bar{x}$ =2,66/5=0,532
σ^2 =2,004/5=0,401

Titolo 2	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
0,57	0,038	0,001
0,64	0,108	0,012
-0,01	-0,542	0,294
0,71	0,178	0,032
0,75	0,218	0,048
2,66	**0**	**0,386**

$\bar{x}$ =2,66/5=0,532
σ^2 =0,386/5=0,077

Caso studio 7

E' stato rilevato il fatturato in due collettivi: uno di piccole imprese (P) e uno di imprese di grandi dimensioni (G).

I due fatturati medi sono risultati pari a:

$\bar{x}_p = 40$ (milioni di euro)
$\bar{x}_g = 110$ (milioni di euro)

I due scarti quadratici medi sono risultati pari a:

$\sigma_p = 6$ (milioni di euro)
$\sigma_g = 24$ (milioni di euro)

Si può concludere che nelle imprese grandi il fatturato è più variabile che nelle imprese piccole?

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:

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

- 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 $\sigma = \sqrt{0{,}758} = 0{,}871$ $CV = \frac{0{,}871}{2{,}04} 100 = 42{,}69\%$
  - Popolazione 2 $\sigma = \sqrt{1{,}238} = 1{,}113$ $CV = \frac{1{,}113}{2{,}04} 100 = 54{,}55\%$

- Per la distribuzione dell'altezza:

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

17

Esercizio

Si considerino i dati del Caso Studio 7.

Si confronti la variabilità del fatturato nei due gruppi di imprese utilizzando il coefficiente di variazione. Che cosa si può concludere?

$\bar{x}_p = 40$ (milioni di euro)
$\bar{x}_g = 110$ (milioni di euro)

$\sigma_p = 6$ (milioni di euro)
$\sigma_g = 24$ (milioni di euro)

$CV_p = \frac{6}{40} 100 = 15$

$CV_g = \frac{24}{110} 100 = 21{,}82$

18

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*)

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

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

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

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

19

- **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	-------->	$a\sigma$

Dim. : $D(aX + b) = \sum_{i=1}^{n}\left(ax_i + b - (a\bar{x} + b)\right)^2 = \sum_{i=1}^{n}(ax_i - a\bar{x})^2 = \sum_{i=1}^{n} a^2(x_i - \bar{x})^2 = a^2\sum_{i=1}^{n}(x_i - \bar{x})^2 = a^2 D(X)$

20

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

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

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

- 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_i)	Frequenze (n_i)	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^2 n_i$	x_i^2	$x_i^2 n_i$
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

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

$$\sigma_1^2 = \frac{D}{n} = \frac{1}{n}\sum_{i=1}^{k}x_i^2 n_i - \bar{x}^2 = \frac{1}{25}123 - 2{,}04^2 = 0{,}758$$

Esercizio

Si considerino i dati del Caso Studio 6.

Si calcoli la varianza del logaritmo dei rendimenti, separatamente per i due titoli, utilizzando la formula semplificata (proprietà 2).

Titolo 1	x_i^2
0,17	0,0289
1,22	1,4884
0,92	0,8464
-0,53	0,2809
0,88	0,7744
2,66	**3,419**

$\bar{x}$ =2,66/5=0,532

σ^2 =3,419/5-0,532^2=0,401

Titolo 2	x_i^2
0,57	0,3249
0,64	0,4096
-0,01	0,0001
0,71	0,5041
0,75	0,5625
2,66	**1,801**

$\bar{x}$ =2,66/5=0,532

σ^2 =1,801/5-0,532^2=0,077

Scostamenti semplici medi

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

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

- Per una *distribuzione di frequenza* non in classi

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

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

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

al posto delle modalità.

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

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

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

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_i)	Frequenze (n_i)	$\lvert x_i - \bar{x}\rvert$	$\lvert x_i - \bar{x}\rvert n_i$	$\lvert x_i - Me\rvert$	$\lvert x_i - Me\rvert n_i$
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

$$\bar{x}_1 = \frac{1}{25}51 = 2{,}04 \qquad S_{\bar{x}} = \frac{1}{n}\sum_{i=1}^{k}|x_i - \bar{x}|n_i = \frac{13{,}60}{25} = 0{,}544$$

$$Me = 2 \qquad S_{Me} = \frac{1}{n}\sum_{i=1}^{k}|x_i - Me|n_i = \frac{13}{25} = 0{,}52$$

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

$$W = Q_3 - Q_1$$

29

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:

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

30

Il teorema di Chebyshev

- (Teorema di Chebyshev) Data una distribuzione di valori, $x_1,\ldots,x_n$, dei quali si conoscono solo la media $\bar{x}$ e la deviazione standard σ, e dato un valore reale positivo k, possiamo affermare che

$$f(|x_i - \bar{x}| \geq k\sigma) \leq \frac{1}{k^2}$$

- (Teorema di Markov) Dato una distribuzione rispetto ad un carattere X che assume solo valori non negativi e per la quale sia nota la media $\bar{x}$, dato un qualsiasi valore $a > 0$, possiamo affermare che:

$$f(x_i \geq a) \leq \frac{\bar{x}}{a}$$

31

Esempi

- Il titolo ENI, nell'ultimo anno ha registrato un prezzo medio giornaliero di chiusura di 8,24 euro. Qual è la frequenza massima di giorni in cui il prezzo è salito al di sopra degli 11 euro?

$$f(x_i \geq a) \leq \frac{\bar{x}}{a} \rightarrow f(x_i \geq 11) \leq \frac{8{,}24}{11} = 0{,}75$$

- Sapendo inoltre che la deviazione standard è stata di 0,893 euro, qual è la frequenza massima di giorni in cui il titolo ha presentato un prezzo di chiusura con uno scarto dalla media superiore a 2 volte la deviazione standard?

$$f(|x_i - \bar{x}| \geq k\sigma) \leq \frac{1}{k^2} \rightarrow f(|x_i - \bar{x}| \geq 2 \cdot 0{,}893) \leq \frac{1}{4} \rightarrow f(|x_i - \bar{x}| \geq 1{,}78) \leq \frac{1}{4}$$

- E qual è la frequenza relativa di giorni in cui il prezzo si è allontanato dalla media meno di 2 euro?

$$f(|x_i - \bar{x}| < k\sigma) > 1 - \frac{1}{k^2} \rightarrow f(|x_i - \bar{x}| < 2) > 1 - \left(\frac{0{,}893}{2}\right)^2$$
$$\rightarrow f(|x_i - \bar{x}| < 2) > 0{,}80$$

32

La standardizzazione

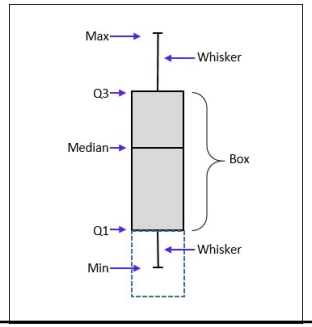
- Data una distribuzione di valori, $x_1,\ldots,x_n$, dei quali si conoscono la media $\bar{x}$ e la deviazione standard σ, è possibile considerare la trasformazione lineare

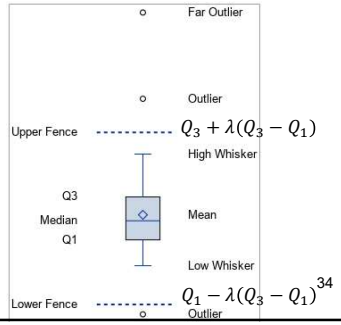
$$y_i = \frac{x_i - \bar{x}}{\sigma}$$

- La trasformazione precedente viene chiamata standardizzazione e permette di ottenere una distribuzione standardizzata, ossia avente media pari a 0 e deviazione standard pari a 1.

Il boxplot

- Il boxplot è un grafico che permette di riassumere alcune caratteristiche rilevanti di una distribuzione.
- E' caratterizzato da tre elementi:
  - Una linea o un punto in corrispondenza di un indice di posizione
  - Un rettangolo la cui altezza è pari ad un indice di dispersione
  - Due segmenti che partono dal rettangolo e i cui estremi sono determinati in base ai valori estremi della distribuzione





Esempi