LEZIONE 6

VARIABILITÀ

Si è detto che gli indici di tendenza centrale (medie di posizione e medie analitiche) sintetizzano l'insieme delle *n* osservazioni mediante un'unica determinazione che fornisce l'ordine di grandezza della variabile.
Questa informazione è così utile da essere usata per sintetizzare i risultati di una qualsiasi indagine statistica, ma va tenuto presente che gruppi diversi di unità statistiche possono avere una stessa moda, mediana o media, ma essere molto diversi fra loro per quanto riguarda i valori assunti dalla variabile.

Per esempio, date le sequenze dei rendimenti medi mensili di due diversi titoli (A e B) in borsa

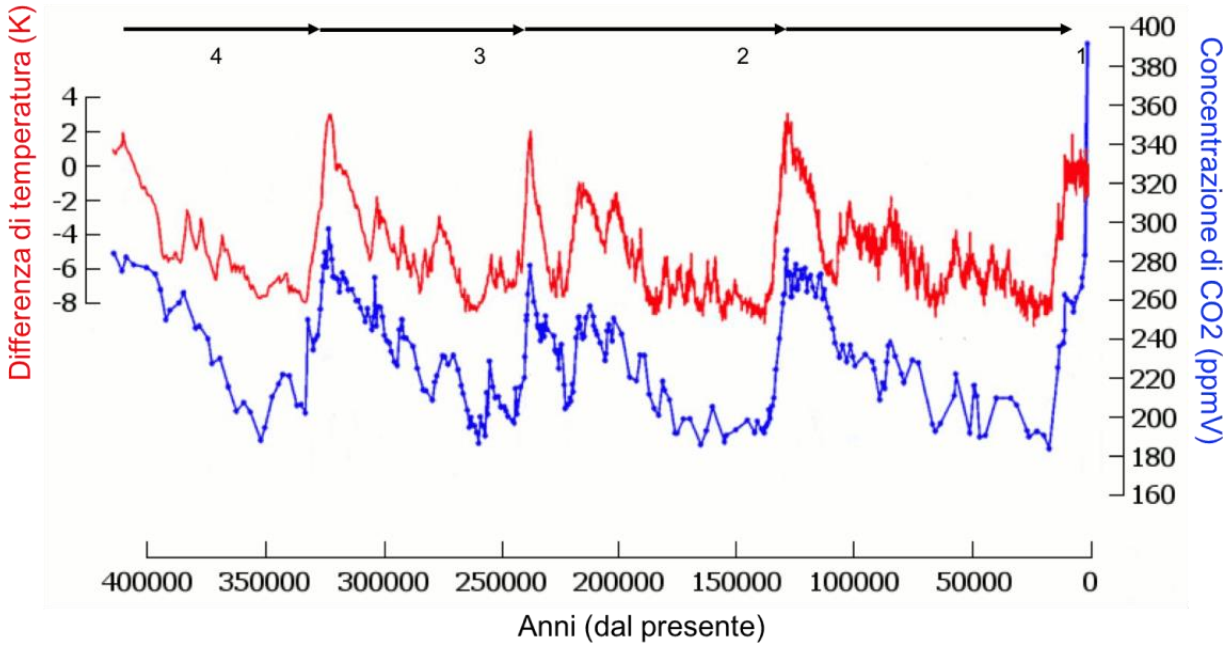
Titolo A:	1.52	0.48	1.20	1.41	0.39
Titolo B:	4.30	-2.42	-3.20	6.48	-0.16

La media dei rendimenti medi risulta in entrambi i casi pari a 1, ma il titolo A ha assunto valori stabili nel tempo, mentre il titolo B ha subito forti variazioni. A parità di questo valore medio, la scelta fra quale titolo acquistare dipenderà dal grado di rischio che un investitore è disposto a correre.

Allo stesso modo, a parità di reddito medio, due collettività possono essere composte da individui con redditi molto simili oppure da alcuni individui molto ricchi e da altri molto poveri.

Il grafico successivo, preso da https://ingvambiente.com/2019/04/09/cambiamenti-climatici-e-riscaldamento-globale/ riporta la "serie storica" (a partire da circa 400mila anni fa

fino al 1995) della variazione di temperatura (linea rossa) e della concentrazione di anidride carbonica in atmosfera (linea blu) ricostruite da dati Antartici.



Questo tipo di rappresentazione grafica, se anche non è stata esaminata in precedenza, dovrebbe essere di facile lettura: gli anni sono riportati in ascissa, mentre i valori delle due variabili, riportati in ordinata, sono stati uniti fra loro da segmenti di retta, così da evidenziarne l'andamento temporale.

Si nota immediatamente l'estrema variabilità che caratterizza entrambe le variabili rilevate, così come si nota anche l'evidente legame esistente tra loro: ad alte concentrazioni di CO2 sono associate alte temperature, e viceversa.

Lo studio delle eventuali relazioni fra variabili, con l'identificazione della loro natura e la misurazione della loro entità, verrà effettuato fra qualche lezione.

In questa lezione si esamineranno i più comuni indici che vengono utilizzati in statistica per misurare la variabilità di una variabile.

Se pure esistono moltissimi indici in grado di misurare la variabilità di una variabile X, anche a seconda della natura della X, tutti questi indici devono necessariamente rispettare le seguenti proprietà:
- assumere il valore minimo in assenza di variabilità
- assumere valori crescenti all'aumentare del grado di variabilità

Nelle pagine successive si esamineranno i soli indici utilizzabili per variabili di tipo quantitativo, ma è possibile valutare la variabilità anche di variabili qualitative (in questi casi si parla spesso di eterogeneità, anziché di variabilità, ma il concetto è sempre il medesimo).

INDICI DI VARIABILITÀ

1) AMPIEZZA DEL CAMPO DI VARIAZIONE

Questo indice, estremamente semplice, corrisponde alla differenza fra la massima intensità rilevata e la minima.

Considerata la sequenza ordinata delle n osservazioni $x_{(i)}$ relative a una variabile quantitativa X, l'ampiezza del campo (o intervallo) di variazione, corrisponde a

$$\omega_x = x_{(n)} - x_{(1)}$$

Esempio

Data la seguente sequenza di osservazioni

0 2 9 5 -1 7

L'ampiezza del campo di variazione è

$$\omega_x = 9 - (-1) = 10$$

Le caratteristiche di questo indice sono:

- Risulta uguale a zero se e solo se tutti gli n valori sono uguali fra loro
- Assume valori crescenti al crescere della variabilità della variabile
- Presenta il difetto di dipendere solo dai valori della più piccola e della più grande intensità rilevate, per cui è fortemente influenzato dalla presenza di eventuali valori anomali (anche detti *outliers*)
- Se la distribuzione è in classi e la prima e/o l'ultima classe sono aperte, il risultato ottenuto presenta il difetto di dipendere dai valori utilizzati per chiudere le classi aperte
-Per questi difetti è in realtà poco usato

2) DIFFERENZA INTERQUARTILE

Come il nome stesso suggerisce, questo indice è pari alla differenza fra il terzo e il primo quartile.

Considerata una variabile quantitativa X, la differenza interquartile corrisponde quindi a

$$W_x = x_{0.75} - x_{0.25}$$

Esempio

Considerata la sequenza ordinata dei valori

-1 0 2 5 7 9

per calcolare la differenza interquartile è necessario determinare il valore del primo e del terzo quartile.

Il posto occupato da $x_{0.25}$ è $\lceil 6 \times 0.25 \rceil = \lceil 1.5 \rceil = 2$, per cui $x_{0.25}=0$

Il posto occupato da $x_{0.75}$ è $\lceil 6 \times 0.75 \rceil = \lceil 4.5 \rceil = 5$, per cui $x_{0.75}=7$

La differenza interquartile è quindi $W_x = 7 - 0 = 7$

Le caratteristiche di questo indice sono:

- Assume valori crescenti al crescere della variabilità della variabile
- Presenta il difetto di dipendere solo da due valori caratteristici della variabile
- Ha il pregio di non risentire dell'eventuale presenza di valori anomali.

Oltre agli indici precedenti, esiste un'importante famiglia di indici di variabilità, detti **indici di dispersione**, che si basano sulle differenze $x_i - \bar{x}$, ossia sui valori assunti dalla variabile scarto.

In caso di assenza di variabilità, le intensità rilevate sulle n unità sono tutte uguali fra loro e, quindi, sono anche uguali alla media della variabile. In questo caso i valori assunti dalla variabile scarto sono tutti uguali a zero.
Al crescere delle differenze fra le intensità rilevate, invece, gli scarti dalla media della variabile tenderanno a risultare sempre più grandi in valore assoluto.

Gli indici di dispersione sono stati formulati con l'obiettivo di valutare l'**ordine di grandezza degli scarti** $(x_i - \bar{x})$ e si è visto come una valutazione dell'ordine di grandezza si ottenga tramite il calcolo della media.
Si è visto però che gli scarti di segno positivo e di segno negativo si compensano fra loro, così che la loro media (o la loro somma) è sempre pari a zero.

Sono stati quindi formulati degli indici di dispersione che si basano sul calcolo della media degli scarti presi in valore assoluto oppure elevati ad una potenza pari, in modo che tutti i termini risultino maggiori o uguali a zero (se l'intensità è uguale alla media).
Questa scelta è giustificata anche dalla considerazione che nella valutazione della variabilità uno scarto di segno negativo ha la stessa importanza di uno scarto di segno positivo.

3) VARIANZA

Questo indice di variabilità, fra i più frequentemente utilizzati, corrisponde alla media dei quadrati degli scarti della variabile dalla sua media.

Da questa definizione risulta che, in realtà, la varianza corrisponde al secondo momento centrale, che è stato studiato nella lezione precedente.
L'importanza che assume nel valutare la variabilità di una variabile fa sì che lo si indichi con il nome specifico di varianza, anziché con quello, più generico, di secondo momento centrale.
Dalle formule utilizzate per calcolare i momenti centrali si ottengono le formule di calcolo della varianza che, a seconda di come sono organizzati i dati, risultano

- Per una **sequenza di *n* osservazioni** x_i di una variabile quantitativa X, la varianza, indicata con i simboli s_x^2 o semplicemente s^2, assume la forma

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

- Per una **distribuzione di frequenza** che si riferisce a una **variabile discreta** X, si ha

$$s_x^2 = s^2 = \frac{1}{n}\sum_{j=1}^{k}\left(c_j - \bar{x}\right)^2 \times n_j$$

$$s_x^2 = s^2 = \sum_{j=1}^{k}\left(c_j - \bar{x}\right)^2 \times f_j$$

- Per una **distribuzione di frequenza in classi** che si riferisce a una **variabile continua** X, infine, risulta

$$s_x^2 = s^2 = \frac{1}{n}\sum_{j=1}^{k}\left(\bar{c}_j - \bar{x}\right)^2 \times n_j$$

$$s_x^2 = s^2 = \sum_{j=1}^{k}\left(\bar{c}_j - \bar{x}\right)^2 \times f_j$$

Le caratteristiche di questo indice sono:

- È uguale a zero se e solo se tutte le osservazioni sono uguali fra di loro (e uguali alla loro media)
- Assume valori crescenti all'aumentare dell'ampiezza degli scarti, ossia all'aumentare della variabilità della variabile
- Ha il difetto di essere espresso nell'unità di misura utilizzata nella rilevazione elevata, però, al quadrato

Esempio

Data la seguente sequenza

1.71 1.59 1.67 1.73 1.82 1.72 1.56 1.80

che riporta i valori assunti dalla variabile X "prezzo medio al litro della benzina" rilevato per i distributori di alcune marche, si vuole determinare il valore della varianza dei prezzi medi.

La media delle 8 osservazioni risulta pari a

$$\bar{x} = \frac{1}{8}(1.71 + 1.59 + \cdots + 1.80) = 1.70$$

per cui si può affermare che la media dei prezzi medi praticati dalle marche rilevate è pari a 1.70 euro al litro.

La varianza risulta invece

$$s_x^2 = \frac{1}{8}[(1.71 - 1.70)^2 + (1.59 - 1.70)^2 + \cdots + (1.80 - 1.70)^2] = 0.0073$$

per cui la varianza dei prezzi medi praticati dalle marche rilevate risulta uguale a 0.0073 euro al quadrato.

Per quanto appena visto, si comprende il motivo per cui molto spesso la variabilità venga misurata non mediante la varianza, ma utilizzando la sua radice quadrata. Questo indice di variabilità contiene infatti le stesse informazioni della varianza, ma ha il vantaggio di essere espresso nella stessa unità di misura utilizzata nella rilevazione.

Prima di passare a esaminare altri indici di variabilità, conviene però concludere lo studio della varianza, esaminandone le proprietà principali.

PROPRIETÀ DELLA VARIANZA

La varianza può essere calcolata con una formula alternativa, molto più semplice di quella utilizzata in precedenza. Si dimostra infatti che

PRIMA PROPRIETÀ

la varianza corrisponde alla differenza fra il secondo momento ordinario e la media elevata al quadrato o, in altri termini, che la varianza è uguale alla media dei quadrati meno il quadrato della media

Per questa dimostrazione è sufficiente partire dalla formula della varianza utilizzata per una sequenza di n valori e sviluppare il quadrato che compare nella formula.

Dimostrazione

Data la sequenza di n osservazioni x_i di una variabile quantitativa X, la varianza assume la forma

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i^2 - 2x_i\bar{x} + \bar{x}^2\right)$$

A questo punto si va a distribuire la somma e, portando fuori dal segno di sommatoria le quantità costanti, si ottiene

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - 2\bar{x}\frac{1}{n}\sum_{i=1}^{n} x_i + \frac{1}{n}\sum_{i=1}^{n} \bar{x}^2$$

Andando a esaminare i diversi termini così ottenuti si nota che:

- la quantità che compare nel rettangolo azzurro corrisponde al momento ordinario di ordine 2 della X, ossia a m_2
- la quantità che compare nel rettangolo giallo corrisponde al prodotto della costante $2\bar{x}$ per la media della variabile X, per cui risulta pari a $2\bar{x}^2$
- la quantità riportata nel rettangolo rosso corrisponde alla media della costante $\bar{x}^2$, per cui è pari a $\bar{x}^2$.

Risulta quindi

$$s_x^2 = m_2 - 2\bar{x}^2 + \bar{x}^2 = m_2 - \bar{x}^2$$

Esercizi

1) Riprendendo la sequenza dei prezzi della benzina, per la quale la media risultava pari a 1.7, il momento ordinario di ordine 2 è dato da

$$m_2 = \frac{1}{8}(1.71^2 + 1.59^2 + \cdots + 1.80^2) = 2.8973$$

per cui la varianza risulta pari a

$$s_x^2 = m_2 - \bar{x}^2 = 2.8973 - 1.7^2 = 0.0073$$

2) Considerata la seguente distribuzione di frequenza, se ne calcoli la varianza

X	Frequenza relativa
1	0.4
2	0.3
3	0.2
4	0.1
	1.0

Si ottengono i seguenti risultati

$\bar{x} = 1\times0.4+2\times0.3+3\times0.2+4\times0.1=2$
$m_2 = 1\times0.4+4\times0.3+9\times0.2+16\times0.1=5$
$\bar{m}_2=5-2^2=1$

SECONDA PROPRIETÀ

La varianza è un **minimo**. Questo significa che la media dei quadrati degli scarti delle osservazioni calcolati da un qualsiasi valore c diverso dalla media aritmetica risulta sempre maggiore della varianza.

Questa proprietà deriva dall'analoga proprietà già dimostrata a proposito della media aritmetica e ovviamente la dimostrazione si effettua in modo analogo.
È sufficiente riprendere quella dimostrazione e moltiplicare i diversi termini per la costante 1/n.

Dimostrazione

Considerata la sequenza delle n osservazioni x_i relative a una variabile quantitativa X, si consideri un valore $c \neq \bar{x}$. Si deve dimostrare che vale la seguente disuguaglianza

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

La quantità che compare a destra del segno di disuguaglianza può essere posta nella forma

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - c)^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x} + \bar{x} - c)^2 = \frac{1}{n}\sum_{i=1}^{n}[(x_i - \bar{x}) + (\bar{x} - c)]^2$$

Sviluppando il quadrato del binomio si ottiene

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - c)^2 = \frac{1}{n}\sum_{i=1}^{n}[(x_i - \bar{x})^2 + (\bar{x} - c)^2 + 2(x_i - \bar{x})(\bar{x} - c)]$$

e, distribuendo la somma, risulta

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - c)^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 + \frac{1}{n}\sum_{i=1}^{n}(\bar{x} - c)^2 + 2(\bar{x} - c) \times \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})$$

dove

- il primo termine a destra del segno di uguaglianza corrisponde alla varianza della X
- il secondo termine è la media della costante $(\bar{x} - c)^2$ che è sempre ≥ 0
- l'ultimo termine corrisponde al prodotto della costante $2(\bar{x} - c)$ per la media della variabile scarto che, come si è visto in precedenza, è sempre pari a zero.

La disuguaglianza da cui si è partiti può essere quindi scritta come indicato di seguito

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \leq \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 + (\bar{x} - c)^2$$

e risulta chiaro che il termine a sinistra è sempre minore del termine a destra, a meno che la costante c sia esattamente uguale alla media $\bar{x}$

TERZA PROPRIETÀ

Considerata la sequenza delle *n* osservazioni

$$x_1 \ , x_2 \ , \dots , x_n$$

relativa a una variabile quantitativa X ed indicato con s_x^2 la sua varianza, si dimostra che la varianza di una trasformazione lineare del tipo

$$\mathrm{Y} = a + b\mathrm{X}$$

risulta

$$s_y^2 = b^2 s_x^2$$

Questa dimostrazione, in realtà, è semplicemente un caso particolare della proprietà dei momenti centrali di ordine r. Si può quindi ottenere da quella dimostrazione ponendo l'uguaglianza r=2

Dimostrazione

Una volta effettuata la dimostrazione generale per il momento centrale *r*-esimo, si pone

$$r=2$$

ottenendo quindi

$$\bar{m}_{2y} = s_y^2 = b^2 s_x^2 = b^2 \bar{m}_{2x}$$

Questo risultato indica che un cambiamento dell'origine della scala di misura non modifica la varianza, che invece risente del cambiamento della scala di misura.

Per esempio, considerato il peso di una certa merce, la sua varianza resta la stessa sia se si fa riferimento al suo peso lordo oppure al suo peso netto.

Se invece si misurasse la statura di un gruppo di persone in millimetri si otterrebbe una varianza che sarebbe 100 volte più grande della varianza della statura delle medesime persone qualora l'unità di misura utilizzata fosse stato il centimetro.

Esercizi

1) Data una variabile X con media $\bar{x} = 4$ varianza $s_x^2 = 8$, si determini media e varianza della variabile trasformata

$$Y = -2 + \frac{1}{4}X$$

In base alle proprietà della media e della varianza di una trasformazione lineare si ottiene

$$\bar{y} = -2 + \frac{1}{4}\bar{x} = -2 + \frac{1}{4}4 = -1$$

$$s_y^2 = \left(\frac{1}{4}\right)^2 s_x^2 = \frac{1}{16} \times 8 = \frac{1}{2}$$

QUARTA PROPRIETÀ (SCOMPOSIZIONE DELLA VARIANZA)

Se le n unità statistiche sono suddivise in g gruppi distinti, la varianza complessiva di una variabile quantitativa X è data dalla somma della media delle varianze ponderata con le numerosità dei gruppi più la varianza delle medie dei gruppi.

Questa proprietà è molto importante in statistica, in quanto risulta utile in molti casi e verrà utilizzata anche in lezioni successive.

Per semplicità non se ne dà qui la dimostrazione (che è comunque riportata nelle dispense), ma ne viene chiarito il significato, anche mediante un esempio numerico.

Innanzitutto è opportuno precisare cosa si intende con i diversi termini utilizzati nell'enunciazione della quarta proprietà della varianza, in base alla quale la varianza complessiva della X, indicata con s_x^2 , corrisponde alla somma di queste due quantità:

A. la media delle varianze ponderata con le numerosità dei gruppi
B. la varianza delle medie dei gruppi

Considerati i g gruppi, sia h l'indice che individua i gruppi, per cui $h = 1, 2, \ldots g$.
Considerato il gruppo h-esimo, si utilizzano i seguenti simboli:

- n_h corrisponde alla sua numerosità, con $\sum_{h=1}^{g} n_h = n$
- $\bar{x}_h$ è la media della X nel gruppo
- s_h^2 è la varianza della X nel gruppo

La quantità A., che corrisponde a

$$\frac{1}{n}\sum_{h=1}^{g} s_h^2 \, n_h$$

viene chiamata **varianza within** (o **varianza all'interno dei gruppi**) e, per semplicità, verrà in seguito indicata con s_w^2

Per comprendere cosa si intende con la quantità B. bisogna innanzitutto ricordare che la media generale della X, $\bar{x}$, corrisponde alla media delle medie

della X nei gruppi ponderata con le numerosità dei gruppi stessi (in base all'ultima proprietà della media aritmetica dimostrata nella lezione 4).

Dato che le medie all'interno dei gruppi hanno una media pari a $\bar{x}$, la loro varianza è data dalla quantità B.

La quantità B., che corrisponde a

$$\frac{1}{n}\sum_{h=1}^{g}(\bar{x}_h - \bar{x})^2\, n_h$$

viene chiamata **varianza between** (o **varianza fra i gruppi**) e, per semplicità, verrà in seguito indicata con s_b^2

La quarta proprietà della varianza, quindi, stabilisce che, considerato un insieme di n unità suddivise in g gruppi distinti, la varianza complessiva della X corrisponde alla somma seguente

$$s_x^2 = \frac{1}{n}\sum_{h=1}^{g} s_h^2\, n_h + \frac{1}{n}\sum_{h=1}^{g}(\bar{x}_h - \bar{x})^2\, n_h = s_w^2 + s_b^2$$

Esempio numerico

Si consideri una popolazione naturalmente divisa in tre gruppi (a, b e c). Nella tabella successiva sono indicati i valori di una variabile quantitative discreta X rilevati all'interno di ciascun gruppo

a	1	3		
b	1	2	1	4
c	2	2	4	4

Considerando le 10 osservazioni totali, la X ha la distribuzione di frequenza

X	Frequenze assolute
1	3
2	3
3	1
4	3
	10

per cui si ottengono i seguenti risultati: $\bar{x} = 2.4$ e $s_x^2 = 1.44$

Passando ora ad analizzare i risultati per i tre gruppi (*a*, *b* e *c*) si ha

$n_a = 2, \qquad n_b = 4, \qquad n_c = 4$

$\bar{x}_a = \frac{1+3}{2} = 2, \qquad \bar{x}_b = \frac{1+2+1+4}{4} = 2, \qquad \bar{x}_c = \frac{2+2+4+4}{4} = 3$

$s_a^2 = 1, \qquad s_b^2 = 1.5, \qquad s_c^2 = 1$

La varianza within corrisponde quindi a

$$s_w^2 = \frac{1}{10}(1 \times 2 + 1.5 \times 4 + 1 \times 4) = 1.2$$

mentre la varianza between è pari a

$$s_b^2 = \frac{1}{10}[(2-2.4)^2 \times 2 + (2-2.4)^2 \times 4 + (3-2.4)^2 \times 4] = 0.24$$

La media generale di X corrisponde quindi alla media delle medie nei gruppi ponderata con le loro numerosità

$$\bar{x} = 2.4 = \frac{1}{10}(2 \times 2 + 2 \times 4 + 3 \times 4)$$

mentre la varianza complessiva della X corrisponde alla somma della varianza within e della varianza between

$$s_x^2 = 1.44 = 0.24 + 1.2$$