Concetto e interpretazione di indici di variabilità

Riassunto

Dopo aver presentato i concetti di base per la costruzione dei più usati indici di variabilità, sono esposti gli strumenti per la loro interpretazione, sia nella statistica descrittiva che nell'inferenza statistica.

Parole chiave. *Devianza, varianza, scarto quadratico medio, coefficiente di variazione, deviazione standard, errore standard.*

Summary

Concepts and interpretations of indices of variability

Methods for the construction of the most used variability indices are shown, and their interpretation is also given, both in descriptive and in inferential statistics.

Key words. *Sum of squares, variance, root mean square, coefficient of variation, standard deviation, standard error.*

La Statistica si occupa dei fenomeni che presentano variabilità nelle loro manifestazioni individuali. Centrale è dunque la misura della variabilità dei fenomeni sia per descrivere una distribuzione statistica, sia per confrontarne la variabilità con quella di altre distribuzioni; inoltre, com'è noto, una misura della variabilità delle stime è fondamentale per affrontare i problemi di inferenza statistica (v. "Statistica per concetti", in CASCO 5)

Indici assoluti di variabilità

La variabilità si definisce come attitudine di un carattere ad assumere diverse modalità quantitative.

La media esprime la caratteristica più importante, ma non l'unica, di una distribuzione. Infatti, fissando la media (aritmetica), possiamo immaginare innumerevoli (o infinite, nel caso di una misura) distribuzioni che soddisfino a tale vincolo.

Esempio 1. La distribuzione secondo il reddito di 200 famiglie ha per media 80. Vi sono innumerevoli distribuzioni di 200 unità che hanno tutte media 80; ad esempio, le seguenti A, B, C, D nella tabella sottostante.

La caratteristica più importante che differenzia le 4 distribuzioni è la variabilità. Nella distribuzione A non c'è variabilità, in quanto tutti i 200 termini sono uguali tra loro. Nella distribuzione B c'è variabilità, in quanto i redditi delle famiglie non sono tutti uguali tra loro, ma tale variabilità è esigua perché i redditi sono abbastanza simili. Nella distribuzione C la variabilità è elevata perché vi sono redditi molto differenti tra loro. Nella distribuzione D c'è la massima variabilità che avremmo potuto osservare, in quanto una sola famiglia detiene tutto il reddito e le altre hanno reddito nullo. Gli economisti indicano la distribuzione D come caso di *massima concentrazione del reddito* e la A come caso di *equidistribuzione (concentrazione nulla) del reddito*.

Nel descrivere le caratteristiche essenziali di una popolazione, oltre la media va sempre riportata una misura (o *indice*) di variabilità in modo da poter comprendere all'incirca di fronte a quale delle innumerevoli situazioni possibili ci si trova.

Osservazione 1. Niente come la sola media è servita agli ingenui detrattori della Statistica per suscitare ilarità e sarcasmo. Citiamo, per tutti, il celebre sonetto dei polli di Trilussa:

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

A

Reddito	N. fam.
80	200
Totale	*200*

B

Reddito	N. fam.
78	100
82	100
Totale	*200*

C

Reddito	N. fam.
10	50
60	50
100	50
150	50
Totale	*200*

D

Reddito	N. fam.
0	199
16 000	1
Totale	*200*

Se, accanto alla media, Trilussa avesse calcolato anche un indice di variabilità questo gioiellino di sonetto non sarebbe mai stato scritto.

La costruzione di una misura di variabilità non può che essere eseguita facendo delle operazioni sui termini di una distribuzione. Ma anche le medie, così come anche altri indici, si calcolano facendo delle operazioni sui termini della distribuzione. Pertanto è lecito chiedersi cosa caratterizzi un indice di variabilità rispetto ad altri indici statistici.

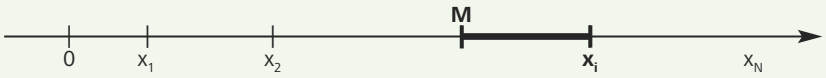
Gli indici di variabilità sono delle operazioni eseguite sui termini di una distribuzione il cui risultato gode delle seguenti due proprietà:

a. un indice di variabilità deve valere zero **quando e solo quando** tutti i termini sono uguali tra loro; in altre parole, se un indice di variabilità è nullo, allora tutti i termini della distribuzione sono uguali tra loro e, viceversa, quando i termini sono uguali tra loro, allora l'indice di variabilità vale zero;
b. un indice di variabilità deve assumere valori crescenti all'aumentare della diversità tra i termini.

Esempio 2. Considerando le 4 distribuzioni riportate nell'es. 1, un indice di variabilità deve valere zero nella distribuzione A, nella distribuzione B deve assumere un valore che non è nullo (in quanto c'è variabilità), ma inferiore a quello assunto nella distribuzione C. Infine, il valore assunto dall'indice in C deve essere inferiore a quello assunto dallo stesso indice in D.

Vi sono diversi metodi per la costruzione di un indice di variabilità, ma nella presente nota ne considereremo solo uno.

Riportiamo una distribuzione secondo un carattere quantitativo, X, con la sua media M, su un sistema di ascisse (si ricordi che un sistema di ascisse consiste in una retta orientata, con una origine ed in cui sia fissata un'unità di misura):



La distanza[1] del termine generico della distribuzione da M è data dal valore assoluto[2] della differenza tra xi e M: $|x_i - M|$; essa prende il nome di *valore assoluto dello scarto del termine x_i da M* ed è la lunghezza del segmento che va da M a x_i (nel precedente grafico è evidenziata in grassetto).

Il più usato indice di variabilità consiste nel calcolare la media quadratica di tali distanze; esso prende il nome di **scarto quadratico medio** *(root mean square)* ed è indicato con il simbolo σ (sigma minuscolo dell'alfabeto greco):

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - M)^2}{N}} \qquad (1)$$

Come si può osservare, non c'è più bisogno di indicare il valore assoluto dello scarto, in quanto ciascuno scarto, essendo considerato al quadrato, è sempre nullo o positivo.

Lo scarto quadratico medio misura, quindi, quanto ciascun termine della distribuzione è distante (diverso), in media (quadratica), dalla media aritmetica.

Il quadrato dello scarto quadratico medio

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - M)^2}{N} \qquad (2)$$

prende il nome di **varianza** (*variance*).

Il numeratore della varianza, cioè la somma dei quadrati degli scarti dei termini della distribuzione dalla media aritmetica, si chiama **devianza** *(sum of squares)*:

$$Dev(X) = \sum_{i=1}^{N}(x_i - M)^2 \qquad (3)$$

Esempio 3 (calcolo ed interpretazione di σ). Consideriamo la seguente distribuzione per unità dei 5 reparti di un ospedale secondo il numero dei posti letto:

12, 60, 42, 54, 32.

La media aritmetica è M = (12 + 60 + 42 + 54 + 32) / 5 = 200 / 5 = 40 (se tutti i reparti avessero lo stesso numero di posti letto, avrebbero 40 posti letto ciascuno).

Calcoliamo dapprima la devianza:

$Dev(X) = (12 - 40)^2 + (60 - 40)^2 + (42 - 40)^2 + (54 - 40)^2 + (32 - 40)^2 = 1448$.

La varianza è quindi $\sigma^2 = 1448/5 = 289{,}6$ e lo scarto quadratico medio $\sigma = \sqrt{289{,}6} = 17$.

Interpretazione: ciascun reparto ha un numero di posti letto che è diverso da quello medio (40) in media di 17 posti letto.

Osservazione 2. Devianza, varianza e scarto quadratico medio sono tutti indici di variabilità, in quanto soddisfano le due proprietà sopra esposte. Infatti, si può dimostrare che se i termini della distribuzione sono uguali tra loro qualunque media coincide con essi; in tal caso, ciascuno scarto dalla media è uguale a zero e, pertanto, la devianza è nulla. Viceversa, se la devianza è uguale a zero, allora è nullo ogni addendo, in quanto basterebbe un solo addendo positivo per negare l'ipotesi. In altre parole, se la devianza è nulla ogni scarto dalla media è uguale a zero e, pertanto, tutti i termini, coincidendo con la media, sono uguali tra loro. Inoltre, poiché la devianza è la somma dei quadrati delle distanze di ciascun termine dalla media,

1. In un sistema di ascisse, la distanza tra due punti, A e B, è data dalla lunghezza del segmento che li congiunge e, quindi, è calcolata come differenza tra le rispettive ascisse: $b - a$. Infatti, b indica la distanza del punto B dall'origine ed a la distanza del punto A dall'origine; la differenza tra le due ascisse rappresenta la lunghezza del segmento AB.

2. Il valore assoluto di una differenza è dato dal suo valore aritmetico, cioè senza segno; a fini pratici ciò equivale a considerarlo sempre positivo.

quanto più i termini sono diversi tra loro, tanto più si allontanano dalla media e, quindi, tanto più grandi sono gli scarti e tanto maggiore risulta la devianza. Con gli stessi ragionamenti si può dimostrare che anche varianza e scarto quadratico medio sono indici di variabilità. Però, lo scarto quadratico medio ha una proprietà aggiuntiva rispetto agli altri due indici: è espresso nella stessa unità di misura dei termini della distribuzione.

Gli indici di variabilità che sono espressi nella stessa unità di misura dei termini della distribuzione si dicono *assoluti*. In conclusione, devianza, varianza e scarto quadratico medio sono indici di variabilità, ma, tra essi, solo lo scarto quadratico medio è un **indice assoluto di variabilità**.

Esempio 4. Si consideri una scolaresca di 20 maschi sedicenni italiani secondo la statura. Supponiamo che sia M = 170 e $\sigma = 12$. Com'è noto, ciò vuol dire che se tutti i 20 soggetti avessero la stessa statura, sarebbero tutti alti cm 170 e che **ogni soggetto ha una statura che è diversa da quella media, in media, di cm 12**. La varianza è allora pari a 144 cm^2 e la devianza è cm^2 2880. Si osservi che i cm^2 non sono più un'unità di misura lineare (come la statura), ma di superficie; pur essendo indici di variabilità anche questi ultimi due, la loro interpretazione non è così immediata come quella dello scarto quadratico medio.

Confronto tra la variabilità di più distribuzioni

Quando si debba eseguire il confronto tra la variabilità di due o più distribuzioni, spesso un indice assoluto risulta inadeguato o perché le distribuzioni hanno termini espressi in diversa unità di misura (l'indice assoluto di variabilità è espresso nella stessa unità di misura dei termini di una distribuzione) o perché le medie sono diverse (l'indice assoluto di variabilità risente del valore della media).

Esempio 5. Consideriamo i dati dell'es. 4 (statura media: 170 cm con scarto quadratico medio 12 cm) ed aggiungiamo le corrispondenti informazioni sul peso: media 58 Kg, scarto quadratico medio 10,6 kg [interpretazione: se tutti i ragazzi avessero lo stesso peso, ciascuno peserebbe 58 Kg; ogni ragazzo ha un peso diverso da quello medio, in media, di 10,6 Kg]). Nel confrontare la variabilità delle due distribuzioni ci si pone il quesito: i ragazzi osservati sono più diversi tra loro rispetto al peso o rispetto alla statura? È evidente che l'indice assoluto non può essere usato per rispondere al quesito perché non si possono confrontare centimetri con chilogrammi.

Un altro esempio: dato un gruppo di donne che hanno partorito, è più diverso il peso delle mamme o il peso dei neonati? Stavolta l'unità di misura è la stessa, ma le medie sono molto diverse. Supponiamo che lo scarto quadratico medio, in entrambi i gruppi, sia pari a 1 Kg. Ciò vuol dire che ogni neonato ha un peso diverso da quello medio (supponiamo 3 Kg) in media di 1 Kg, il che denota una forte variabilità perché vuol dire che ci sono neonati che pesano poco e neonati che pesano molto. Per le madri che hanno un peso medio, poniamo, di 60 Kg, invece, $\sigma = 1$ vuol dire che hanno tutte all'incirca lo steso peso, in quanto ciascuna madre ha un peso diverso da quello medio (60 Kg), in media, di un chilo. Quindi anche quando l'unità di misura dei termini delle distribuzioni di cui si vuole confrontare la variabilità è la stessa, ma le medie sono alquanto diverse, l'indice assoluto di variabilità è inadeguato per eseguire il confronto.

Riepilogando, un indice assoluto di variabilità può essere usato per eseguire il confronto tra la variabilità di due o più distribuzioni solo quando

a. i termini delle distribuzioni siano espressi nella stessa unità di misura;
b. le medie delle distribuzioni siano all'incirca uguali tra loro.

Nel caso in cui una delle due condizioni non sia verificata, per eseguire il confronto è necessario ricorrere ad un altro tipo di indice di variabilità.

Ricordando la nozione di rapporto, esposta anche in "Statistica per concetti" in CASCO 13, si perviene facilmente alla costruzione del più usato indice idoneo ad eseguire il confronto tra la variabilità di due o più distribuzioni. Esso prende il nome di **coefficiente di variazione** (CV, *coefficient of variation*) e si ottiene rapportando lo scarto quadratico medio alla media aritmetica e moltiplicando per 100:

$$CV = (\sigma / M) \times 100.$$

Il risultato di CV è un numero puro in quanto rapporto tra quantità omogenee (cioè espresse nella stessa unità di misura). CV inoltre elimina dal valore di σ l'influenza della media. Il suo significato indica la misura della variabilità (media) per ogni 100 unità di media aritmetica, cioè, qual è il valore di σ posta uguale a 100 la media.

Esempio 6. Riprendiamo i dati dell'esempio 5. (Statura e peso). Il coefficiente di variazione, calcolato nella distribuzione secondo la statura è CV = (12 / 170) x 100 = 7,1%; in quella secondo il peso è CV = (10,6 / 58) x 100 = 20%: i ragazzi sono più diversi tra loro per il peso che per la statura. Infatti, considerando la statura, il coefficiente di variazione indica che lo scarto quadratico medio è il 7,1% della media; invece, per il peso è il 20% della media.
(Peso delle madri e peso dei neonati). Per i neonati il coefficiente di variazione è CV = (1 / 3) x 100 = 33,3%; per le madri è CV = (1 / 60) x 100 = 1,7%: pur essendo uguale il valore di σ (un chilo) per madri e neonati, CV indica che la variabilità misurata da σ è per i neonati il 33,3% della media e per le madri l'1,7% della media, denotando così che, mentre le madri hanno all'incirca lo stesso peso, i neonati sono hanno un peso assai diverso tra loro.

Variabilità nell'inferenza statistica

a. Deviazione standard

Quanto esposto finora riguarda un'intera popolazione (o più

popolazioni di cui vogliamo confrontare la variabilità). Spesso però si ha a disposizione un campione estratto a sorte da una popolazione e lo scopo principale è quello di riferire i risultati delle analisi statistiche eseguite sul campione alla popolazione da cui esso proviene (v. Statistica per concetti, CASCO 5).

Nel caso presente, lo scarto quadratico medio della popolazione, σ, è sconosciuto e vogliamo stimarlo a partire dalle osservazioni eseguite su un campione. Si può dimostrare che la stima corretta di σ è

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - m)^2}{n-1}} \qquad (4)$$

dove n è la numerosità del campione ed m è la media aritmetica calcolata sul campione.

Tale indice prende il nome di **Deviazione Standard** (*Standard Deviation, SD*).

Confrontandolo con la (1), si può osservare che la struttura è la stessa, ma il denominatore è diverso e, pertanto, non è più la media quadratica degli scarti e, quindi, a rigore, non avrebbe più la stessa interpretazione di σ. Però, poiché **s** è una stima di σ, possiamo continuare ad interpretarlo come stima della distanza che, nella popolazione, c'è in media tra ciascun termine e la media aritmetica.

b. Variabilità dello stimatore: errore standard

Come si è visto (è importante rivedere quanto esposto in "Statistica per concetti", CASCO 5), elemento cardine dell'inferenza statistica è la variabilità dello stimatore. Al variare del campione nell'universo dei campioni la stima di un parametro varia (da campione a campione) e descrive una variabile casuale detta stimatore. **Lo scarto quadratico medio dello stimatore prende il nome di errore standard** (dello stimatore) e misura la diversità che c'è, in media, tra ciascuna stima e la media di tutte le stime. Se la stima è *corretta* la media delle stime coincide con il valore del parametro della popolazione; in tal caso, quindi, l'errore standard dello stimatore indica quant'è diversa ciascuna stima, in media, dal valore del parametro.

A titolo di esempio, consideriamo come parametro la media M calcolata nella popolazione. Immaginiamo l'insieme di tutti i campioni diversi che possono essere estratti a sorte dalla popolazione considerata (universo dei campioni) e, in ciascun campione, calcoliamo la media aritmetica **m** che è una stima corretta di M; per definizione di stima corretta, la media di tutte le medie calcolate sui campioni dell'universo coincide con M. Lo scarto quadratico medio di tutte le stime prende il nome di **errore standard** (**ES**, nel nostro caso, della media: ES(M)*) e misura quanto ciascuna media, calcolata su un campione dell'universo, sia diversa, in media, dalla media M della popolazione.

Nell'espressione dell'errore standard compare il valore del parametro M che non è noto. Poiché operiamo su un unico campione, possiamo solo darne una stima, *m*. Quindi non possiamo conoscere l'esatto valore dell'errore standard, ma solo quello di una sua stima. La stima dell'errore standard della media è ES(M) = $s/\sqrt{n}$, dove s è la deviazione standard (4) ed *n* è la dimensione del campione.
Interpretazione: ES(M) stima quanto sia diversa ciascuna media calcolata sui campioni dell'universo, in media, dalla media aritmetica della popolazione. In concreto, si stima M con *m*: se l'errore standard è "piccolo" siamo quasi certi di non aver sbagliato molto, ma se è "grande" la media stimata potrebbe essere anche molto diversa dal parametro. Come si vede, l'errore standard dipende anche da *n* (numerosità del campione), nel senso che quanto più *n* è grande tanto minore è l'errore standard. Se la stima dell'errore standard è così grande da mettere in dubbio l'attendibilità della stima della media, *m*, è sufficiente aumentare *n* per avere una stima migliore.

Esempio 7 (riepilogativo). I 5 pazienti di un campione casuale di pazienti sottoposti alla stessa chemioterapia di moderato potere emetogeno e trattati con la stessa profilassi antiemetica, sono stati valutati in relazione alla massima nausea ritardata (quella che si presenta nei gg. 2-5 dopo la somministrazione della chemioterapia) registrata con un analogo visivo lineare lungo 100 mm.

Risultati: 25, 0, 10, 15, 10. Media, m = 12; deviazione standard, s = [$\sqrt{}$(169 + 144 + 4 + 9 + 4) / 4] = 9,1.

Interpretazione: la stima corretta della media M (sconosciuta) della popolazione è 12 mm; cioè, se tutti i pazienti nelle stesse condizioni di quelli osservati avessero la stessa massima intensità di nausea ritardata, la sua stima sarebbe pari a 12 mm. Si stima, inoltre, che la massima nausea ritardata di ciascun paziente della popolazione si discosti, in media, dalla media di 9,1 mm (deviazione standard, s).

Coefficiente di variazione (stimato): CV = (9,1/12) x 100 = 75,8%: la deviazione standard (stimata) è pari al 75,8% della media (stimata).

La stima della media della popolazione, *m*, varia in ciascun campione dell'universo (perché diversi sono i pazienti che ne fanno parte). Considerando tutti i valori di *m* (uno per ciascun campione), la stima dello scarto quadratico medio di tali valori (errore standard) è pari a ES(M) = 9,1/$\sqrt{}$5 = 4,1: considerando tutte le possibili stime di m (una in ogni campione dell'universo), si stima che ciascuna di esse si discosti dal vero valore del parametro M (media nella popolazione), in media, di 4,1 mm. Si osservi che se i dati si riferissero ad un campione più grande (poniamo di 25 anziché di soli 5 pazienti), a parità del valore della deviazione standard, la stima dell'errore standard sarebbe stata pari a ES(M) = 9,1/5 = 1,82, cioè, meno della metà di quello riferito a 5 pazienti.

Enzo Ballatori