Concetto ed interpretazione di indici statistici: medie e asimmetria

Riassunto

Sono stati sviluppati i concetti e l'interpretazione della media aritmetica e della mediana senza introdurre formule matematiche. Inoltre, dal confronto tra tali indici statistici si è ricavato un semplice indicatore di asimmetria, utile per acquisire ulteriori conoscenze sulla distribuzione statistica considerata.

Parole chiave. *Media, mediana, simmetria, asimmetria.*

Summary

Concepts and interpretations of statistical indices: means and asymmetry

Concept and interpretation of both mean and median have been developed without introducing mathematical formulas. Moreover, a useful index of asymmetry was derived from the relationship between mean and median.

Key words. *Mean, median, symmetry, asymmetry.*

Negli articoli scientifici che riportano i risultati di studi clinici compaiono i valori di indici statistici che, in assenza di una preparazione specifica del Lettore, sono interpretati su base intuitiva. Inoltre, spesso non sono sfruttate tutte le caratteristiche e le relazioni tra gli indici che consentono di acquisire informazioni importanti ai fini della comprensione della struttura dei fenomeni oggetto di studio. Scopo della presente nota è dare conoscenze abbastanza rigorose, pur senza fornire l'alibi dell'impedimento a comprendere per via di una insufficiente preparazione matematica.

La Statistica è stata anche definita "Scienza del confronto" perché consente di comparare differenti popolazioni rispetto ad importanti caratteristiche: è anzitutto dal confronto che scaturisce la conoscenza.

Ad esempio, sapere che, per un certo tipo e stadio di tumore, curato con un determinato trattamento, A, la sopravvivenza mediana è di 8 mesi nulla ci dice sull'efficacia di A finché non la confrontiamo con quella riscontrata con un'altra terapia, B (con cui si ottiene, poniamo, una sopravvivenza mediana di 10 mesi). Supponendo che tali risultati siano stati acquisiti con uno studio clinico dal disegno rigoroso e da una conduzione ineccepibile, se una differenza continua a persistere anche una volta che è stata aggiustata per la variabilità accidentale, possiamo concludere che B è più efficace di A.

La Statistica si occupa di fenomeni collettivi, cioè di quei fenomeni che si presentano in modo diverso nelle unità della popolazione. La maggior complessità dei fenomeni collettivi rispetto a quelli deterministici, cioè a quelli che si manifestano sempre allo stesso modo in tutte le unità, discende appunto dalla variabilità.

Esempio di fenomeno deterministico: "Non c'è bisogno della Statistica per sapere quante volte il taglio del nervo sciatico produce la paralisi dei muscoli da questo innervati" scriveva Claude Bernard nella sua "Introduzione alla Medicina Sperimentale".

Il confronto è uno dei due pilastri su cui si fonda il metodo scientifico, essendo l'altro la relazione di causalità. Tali strumenti sono irrinunciabili per comprendere e vanno estesi al campo dei fenomeni collettivi, dove però si incontra una complicazione in più – la variabilità – che ne indebolisce la struttura.

Nel campo dei fenomeni deterministici è assai semplice eseguire un confronto. Ad esempio, dati due soggetti, il primo alto cm 182 e il secondo cm 176, sappiamo che il primo è più alto del secondo e che la sua statura è di 6 cm superiore. Se ci ponessimo la stessa domanda per due collettivi di soggetti, ad esempio, se considerassimo i ragazzi e le ragazze di 18 anni di età di una certa scuola, e volessimo sapere quale dei due gruppi è "più alto", la domanda resterebbe priva di risposta perché vi sono alcuni maschi che sono più alti di alcune ragazze, ma ci sono anche alcune ragazze più alte di alcuni maschi. La domanda corretta è allora: sono *complessivamente* più alti i ragazzi o le ragazze? Il che indurrebbe, in prima battuta, a sommare le stature dei ragazzi e a confrontare il risultato con la somma delle stature delle ragazze.

Quando si confrontano le intensità o le frequenze di due (o più) fenomeni collettivi esistono uno o più fattori di disturbo il cui effetto va eliminato per rendere corretto il confronto. L'effetto del fattore di disturbo più importante si elimina con il rapporto R = A/B, dove A è il fenomeno la cui intensità (o frequenza) si intende confrontare e B è l'intensità (o la frequenza) del fattore di disturbo.

Nell'esempio, il più importante fattore di disturbo è la diversa numerosità dei due gruppi, che va eliminata con un rapporto per rendere corretto il confronto (altrimenti è ovvio che solo il diverso numero di soggetti che compongono i due gruppi potrebbe essere responsabile della maggiore o minore statura complessiva). Si tratta, quindi, di dividere la somma delle stature dei

ragazzi per il numero dei ragazzi e la somma della stature delle ragazze per il numero delle ragazze. Ad esempio, se la somma delle stature dei ragazzi fosse pari a cm 21.120 e quella della ragazze cm 33.600, dividendo tali somme di stature, rispettivamente, per il numero dei ragazzi (120) e per quello delle ragazze (200), si ha, per i ragazzi, cm 176 e, per le ragazze, cm 168.

Il risultato del rapporto R = A/B esprime quanta parte di A spetta (compete), in media, ad ogni unità di B, dove **"in media"** significa che, se tutte le unità di B fossero ugualmente dotate dell'intensità (o della frequenza) di A, allora R rappresenta quanto competerebbe a ciascuna di loro.

Considerando i dati dell'esempio, se tutti i ragazzi fossero ugualmente alti, ciascuno di loro sarebbe alto cm 176; se le ragazze avessero la stessa statura, ciascuna di loro sarebbe alta cm 168. È come se avessimo "costruito" un ragazzo e una ragazza "tipo", ossia un rappresentante di ciascun gruppo, riconducendoci così alla soluzione deterministica del confronto tra due stature. Possiamo quindi concludere che i ragazzi sono **mediamente** più alti delle ragazze di cm 8.

Quella calcolata nell'esempio è la statura media. In generale la media aritmetica, o più semplicemente "media" (per antonomasia, data la sua importanza), si ottiene rapportando l'ammontare totale del carattere (ossia la somma delle intensità individuali) al numero delle unità della popolazione.

Altri tipi di rapporti statistici

Ci sono moltissimi altri tipi di rapporti statistici che possono interpretarsi come descritto sopra. A titolo di esempio, ne esponiamo due.

a. **Frequenze relative.** Sia dato un gruppo di pazienti, affetti tutti dalla stessa neoplasia allo stesso stadio. Considerando i sopravviventi a due anni, in uno studio clinico randomizzato di confronto tra due trattamenti, A e B, sono stati registrati 150 sopravviventi con il trattamento A e 240 con B. Il confronto tra questi due dati (conteggi o frequenze assolute) sarebbe corretto se i due gruppi fossero ugualmente numerosi. Se così non è, allora per poter eseguire un confronto corretto delle frequenze assolute bisogna eliminare l'effetto del più importante fattore di disturbo (diversa numerosità). Se i pazienti trattati con A sono complessivamente 300 e quelli trattati con B 600 (ad esempio nel caso di una randomizzazione 1:2), i corrispondenti rapporti valgono R(A) = 150/300 = 0,5 e R(B) = 240/600 = 0,4. Tali valori si chiamano frequenze relative (*proportions*) e moltiplicati per 100 danno le più note percentuali. In conclusione, se lo studio è stato adeguatamente programmato e ben condotto, e se una differenza permane dopo aver aggiustato per la variabilità accidentale, si può concludere che il trattamento A è più efficace di B.

b. **Quoziente di natalità.** La domanda è: c'è una tendenza a nascere di più nella regione A o nella regione B? La prima idea è confrontare il numero dei nati delle due regioni in un certo anno, ma ci si accorge subito che tale confronto è affetto da vari fattori di disturbo il più importante dei quali è la dimensione della popolazione, in quanto da essa provengono i nati. Si costruiscono allora i rapporti Q(A) = [N(A)/P(A)] x 1000 e Q(B) = [N(B)/(P(B)] x 1000, dove N(A) e N(B) sono i nati nelle due regioni in un certo anno e P(A) e P(B) il numero degli abitanti. I valori così ottenuti si chiamano quozienti generici grezzi di natalità ed esprimono quanti sono, in media, i nati nelle due regioni per ogni 1000 abitanti.

Warning: nella costruzione di un rapporto occorre prestare molta attenzione a che il fenomeno, la cui intensità (o frequenza) è posta a denominatore, sia proprio il fattore di disturbo più importante, altrimenti si può giungere a conclusioni aberranti. Ad esempio, a qualcuno è venuto in mente di confrontare l'abilità e la prudenza alla guida degli uomini e delle donne considerando il numero degli incidenti in cui sono stati coinvolti come guidatori. È evidente che non è sufficiente confrontare il solo numero di incidenti ma che occorra tener conto di fattori di disturbo, per cui gli analisti che hanno pubblicato i dati hanno diviso, in ciascuno dei due gruppi, il numero degli incidenti per il numero dei patentati. Ne è risultato che il numero medio di incidenti per patente era più alto per gli uomini che per le donne; da ciò conclusero che le donne guidano meglio degli uomini. Appena fu pubblicata, la notizia rimbalzò in tutti i telegiornali delle varie reti televisive (ed ancora oggi, seppur raramente, ancora se ne parla). In realtà l'esposizione al rischio di incidente è misurata soprattutto dalla percorrenza chilometrica; quindi un corretto rapporto avrebbe dovuto avere a denominatore non già il numero di patentati, ma la distanza media realmente percorsa in un anno da ciascuno dei due generi. Il confronto eseguito sarebbe stato corretto solo sotto l'inverosimile ipotesi che uomini e donne percorrano mediamente un'analoga distanza; ma si può asserire che così non è, conoscendo il differente uso dell'auto da parte dei due generi.

La mediana

Possono essere costruite infinite altre medie che, come la media aritmetica, si calcolano eseguendo operazioni su tutti i termini della distribuzione; tali medie sono dette *analitiche*. Vi sono però anche infinite medie *lasche* che sono calcolate considerando la posizione che certi valori hanno all'interno della graduatoria (ossia della distribuzione ordinata, ad es., in senso crescente) o per il significato che certi termini

esprimono. La più importante tra queste è la **mediana** che viene definita come **il termine che bipartisce la graduatoria lasciando a sinistra lo stesso numero di termini che lascia a destra**.

Si chiama graduatoria una distribuzione ordinata, ad es., in senso crescente (o meglio non decrescente, v. CASCO 9, Statistica per concetti 2).

Per calcolare la mediana, come prima cosa, occorre trasformare la distribuzione in graduatoria. Poi si va a vedere qual è il termine che la divide in due parti ugualmente numerose.

Esempio: calcolo della mediana. Siano state eseguite 5 determinazioni di glicemia su altrettanti pazienti diabetici alla diagnosi: 140, 180, 320, 240, 210. La corrispondente graduatoria (crescente) è: 140, 180, 210, 240, 320. Il valore mediano di glicemia tra i 5 pazienti è 210. Nel caso di un numero di termini pari, non c'è più una sola mediana, ma qualunque valore dell'intervallo più interno (detto intervallo mediano) può essere assunto come mediana. Solo convenzionalmente si assume per mediana la semisomma degli estremi dell'intervallo mediano.

Graduatoria: 140, 180, 210, 250, 280, 320. Intervallo mediano: 210 --- 250; mediana è qualunque valore tra 210 e 250 (estremi inclusi). Mediana convenzionale = (210 + 250)/2 = 230.

Esistono infinite altre medie simili alla mediana, dette "quantili". In una graduatoria, ogni quantile lascia alla sua sinistra una certa quantità di termini ed alla sua destra la restante quantità. Tra i quantili, i più importanti nelle applicazioni mediche sono i centili (o percentili, molto usati in Pediatria e in Auxologia per individuare il normale accrescimento, definito in Italia come l'intervallo compreso tra il 3° e il 97° percentile). I percentili sono 99: il primo lascia alla sua sinistra l'1% dei termini ed alla sua destra il restante 99%; il secondo percentile lascia a sinistra il 2% dei termini ed alla sua destra il restante 98%; … ; il 99° percentile lascia a sinistra il 99% dei termini ed a destra il restante 1%. Il 100° percentile lascerebbe a sinistra il 100% dei termini: quindi, non essendo univocamente determinato, nel senso che di "centesimi centili" ce ne sono infiniti, non ha interesse. Si osservi che la mediana coincide con il 50° percentile.

Applicazioni della mediana

Quando la media aritmetica non può essere calcolata, è opportuno determinare la mediana.

Esempi:

- Scale ordinali: se il fenomeno oggetto di interesse è collocabile su una scala ordinale (v. CASCO 9, Statistica per concetti 2) la media aritmetica non può essere calcolata perché i termini della distribuzione non sono intensità – esprimibili con numeri – ma attributi (cioè aggettivi) e, quindi, non è possibile sommarli. In tal caso si determina la mediana.
- Sopravvivenza: al momento della chiusura di uno studio clinico avente come endpoint la sopravvivenza, ci possono essere pazienti ancora in vita, ma non si sa per quanto tempo ancora. In tal caso, non è possibile calcolare la sopravvivenza media (perché, per farlo sono richieste le durate di vita di tutti i pazienti che vanno sommate), ma spesso è possibile determinare la sopravvivenza mediana.

Interpretazioni: esempi

- In un gruppo di pazienti ipertesi alla diagnosi, la pressione diastolica media è pari a 104 mmHg e la mediana a 98 mmHg. Ciò vuol dire che, se tutti i pazienti considerati avessero la stessa pressione diastolica, questa sarebbe pari a 104 mm/Hg. Inoltre, 98 mm/Hg è quel valore di pressione diastolica tale che la metà dei pazienti considerati ne ha uno più basso e l'altra metà uno più alto.
- Vita media e vita mediana. L'ISTAT (Istituto Centrale di Statistica) ha calcolato che, per le donne italiane, la vita media alla nascita (*expectation of life*, speranza di vita) è pari a 84 anni e la vita mediana pari a 82 anni. Quindi, se tutte le bambine che nascono oggi avessero la stessa durata di vita, questa sarebbe pari a 84 anni; invece, 82 anni è quell'età tale che la metà delle bambine che nascono oggi morirà prima e l'altra metà dopo.

Asimmetria di una distribuzione

Una distribuzione statistica, trasformata in graduatoria, si dice simmetrica se, stabilito un centro di simmetria (che si assume essere la mediana, proprio per la sua definizione), per ogni termine a sinistra della mediana, ne esiste uno ed uno solo a destra avente la stessa distanza dalla mediana.

Esempi

1. Nella distribuzione: 5, 7, 9, la mediana è 7. La distanza di 5 da 7 è $|5 - 7| = 2$. La distanza di 9 da 7 è ancora 2: la distribuzione è simmetrica.
2. La distribuzione 5, 7, 21 non è simmetrica in quanto la distanza di 5 dalla mediana 7 è 2, mentre quella di 21 da 7 è 14.

Nell'esempio (1) si può osservare che la media aritmetica è uguale a 7 [= (5 + 7 + 9)/3], cioè alla mediana, e non è un caso, potendosi dimostrare che, se la distribuzione è simmetrica, la media coincide sempre con la mediana.

Osservando i dati dell'esempio (2), senza fare calcoli, si può asserire che la media è maggiore della mediana, in quanto il termine "21" conta "uno" (è un termine solo, come 5) ai fini della determinazione della mediana, ma ha un peso ben maggiore di 5 quando si calcola la media. Infatti, svolgendo i calcoli, si ha mediana = 7, media = 11.

3. Confrontiamo la distribuzione 1, 7, 9 con quella dell'esempio (1). Si può osservare che anch'essa è asimmetrica, ma per un ragionamento analogo a quello fatto sopra, ci si attende che la media sia inferiore alla mediana. Infatti, la mediana è sempre 7, ma la media, stavolta, è 17/3 = 5,67, minore della mediana.

In sintesi, se la distribuzione è simmetrica allora la media coincide con la mediana. Attenzione: non vale il viceversa; non è detto, cioè, che se la media coincide con la mediana la distribuzione sia simmetrica: potrebbe esserlo o non esserlo. Però, se la media non è uguale alla mediana la distribuzione è certamente asimmetrica. In questo caso, se la media è maggiore della mediana [come nell'esempio (2)], l'**asimmetria** si dice **positiva**, se la media è inferiore alla mediana, si parla di **asimmetria negativa**.

Quindi, un pratico indicatore di asimmetria scaturisce dal confronto tra media e mediana: A_S = media – mediana. Se A_S è maggiore di zero, l'asimmetria si dice positiva ed è caratterizzata dalla presenza di pochi termini molto più grandi degli altri.

Invece, quando A_S è minore di zero, si parla di asimmetria negativa che è caratterizzata dalla presenza, nella distribuzione, di pochi termini molto più piccoli degli altri.

Infine, se è $A_S = 0$, si hanno poche informazioni: la distribuzione potrebbe essere simmetrica o asimmetrica.

Interpretazione

In entrambi gli esempi riportati in precedenza (pressione diastolica e vita media) la media è superiore alla mediana. Si tratta quindi di due casi di asimmetria positiva che può essere interpretata come dovuta, nel primo esempio, alla presenza di pochi pazienti con una pressione molto più alta di quella misurata negli altri, nel secondo esempio, a poche donne che sopravvivono molto più a lungo delle altre (ad es. le ultracentenarie).

In conclusione, poiché media e mediana sono calcolate da tutti i programmi statistici (ed anche da EXCEL), il loro confronto offre, a pochissimo prezzo, un'utile informazione su un rilevante aspetto della distribuzione, l'asimmetria, che contribuisce a descrivere ancor più precisamente la distribuzione.

Enzo Ballatori