Epidemiologia & Statistica per tutti

Sintetizzare i dati: la statistica descrittiva

M. Nichelatti, M. Nordio, U. Maggiore, A. Limido

a nome del Comitato Scientifico SIN-RIDT

Dati

Chiameremo *dati* l'insieme di un certo numero di variabili misurate in un certo numero di casi (di pazienti, di soggetti ecc.). In pratica, immagineremo che i dati con cui avremo a che fare, siano raccolti in un database (come ad esempio Access) o in un foglio elettronico (come ad esempio Excel), ed assumeremo che i casi siano le righe orizzontali (ogni riga rappresenta un paziente) mentre le variabili saranno le colonne (ogni singola colonna rappresenta una variabile).

I dati vengono raccolti misurando le variabili in ogni singolo caso, che costituisce così un individuo del campione di cui stiamo raccogliendo i dati.

Variabili

Una variabile è per quanto di nostro interesse un *qualcosa di misurabile* che può assumere determinati valori e solo quelli.

Possiamo distinguerle essenzialmente in:

1. *Variabili continue*, che possono assumere qualsiasi valore possibile all'interno di un dato intervallo prestabilito (o almeno ragionevole); esempi di variabili continue sono l'altezza, il peso, l'età, il glucosio ematico, la conta degli eritrociti, eccetera.
2. *Variabili discrete*, concettualmente simili alle variabili continue, ma che possono assumere solo dei valori interi; di fatto per motivi pratici, le nostre variabili continue sono rappresentate da variabili discrete, perché misuriamo l'altezza in cm, e il peso in kg (non sarebbe granché utile misurare l'altezza o il peso con strumenti che discriminano oltre questo grado di risoluzione).
3. *Variabili categoriche ordinali* (cioè con gerarchia di valori), che possono assumere pochi valori discreti, disposti in una scala gerarchica, e in cui tra un valore e l'altro vi sono differenze non necessariamente uguali; esempi di variabili categoriche ordinali sono il livello di scolarità, la fascia di reddito, lo staging dei tumori, eccetera.
4. *Variabili categoriche pure*, che assumono valori discreti, ma in cui non esiste un ordine gerarchico in cui tali valori possono essere collocati; esempi per queste variabili sono il colore dei capelli, lo stato civile, l'etnia, eccetera.
5. *Variabili binarie*, che sono variabili categoriche ordinali o pure, che possono assumere solo due valori; esempi di queste variabili sono il sesso, lo stato in vita (vivo o morto), eccetera.

Campione

Ipotizziamo sempre che il campione sia rappresentativo della popolazione (o universo) da cui è estratto, cioè che i valori che assumono le variabili di interesse per la nostra ricerca non siano differenti dai valori che queste variabili assumerebbero se misurate in tutta la popolazione. Ad esempio, se la popolazione che vogliamo studiare è costituita dai pazienti trattati con un certo tipo di farmaco, e se il farmaco viene utilizzato in pazienti con malattie degenerative, occorrerà ad esempio che il campione che andiamo a selezionare (dai ricoveri in alcuni ospedali) abbia delle caratteristiche congrue a quelle generali dell'universo. Se ad esempio sappiamo dalla letteratura che l'età media dei pazienti trattati è pari a 70 anni, e se il campione che abbiamo selezionato è formato da pazienti con età media di 50 anni, è difficile sperare che il campione sia realmente rappresentativo dell'universo che vogliamo esplorare.

Infatti, l'ipotesi che sorregge la nostra idea è che il campione selezionato rappresenti, in modo adeguato e in tutte le sue caratteristiche, l'universo: se questa ipotesi non trova una realizzazione, quello che analizzeremo sarà un campione di soggetti che non consentirà di trarre conclusioni utili rispetto alla nostra indagine.

Universo

Per definizione, chiamiamo *universo* la popolazione di nostro interesse relativamente a una certa ipotesi di lavoro. Potremmo ad esempio ipotizzare che la percentuale dei due sessi colpita da una certa malattia sia identica, cioè che la malattia colpisca indifferentemente uomini e donne, oppure potremmo ipotizzare che a parità di percentuale, i maschi siano colpiti prima delle donne, ovvero che i pazienti maschi siano mediamente più giovani delle pazienti femmina. La *popolazione obiettivo*, il nostro universo, in questo caso è costituito da tutti i soggetti colpiti dalla malattia di nostro interesse, ma evidentemente condurre l'analisi su tutti i soggetti si tradurrebbe in un grandissimo sforzo organizzativo e in costi elevatissimi che di fatto impedirebbero lo svolgimento dell'analisi e quindi la verifica della nostra ipotesi.

Per questo motivo si seleziona un campione all'interno della popolazione di interesse utilizzando una serie di criteri che vedremo in uno dei moduli successivi. Di fatto, è indispensabile che il campione che selezioniamo per il nostro studio rappresenti in modo credibile l'universo di riferimento.

L'universo, cioè l'insieme di tutti i pazienti affetti dalla malattia di nostro interesse è una popolazione a tutti gli effetti e quindi è caratterizzata da variabili come il peso, l'altezza, l'età alla diagnosi e così via: per ciascuna di queste variabili, la popolazione obiettivo sarà caratterizzata da un certo parametro, ad esempio, se la malattia colpisce i soggetti sottopeso, ci aspettiamo che il peso medio dei pazienti che costituiscono l'universo sia al di sotto del peso medio della popolazione generale, e ci aspettiamo che il peso medio del campione che andremo ad analizzare per il nostro studio sia abbastanza simile al peso medio osservato atteso per l'universo.

In pratica, il valore atteso per il peso dei pazienti (considerando tutto l'universo dei pazienti) è un parametro, mentre il valore medio del peso dei pazienti che costituiscono il campione che abbiamo selezionato dovrebbe rappresentare un valore il più possibile vicino al valore del parametro, e dovrebbe stimarlo con una buona esattezza.

In altre parole, se selezioniamo un campione di dimensioni adeguate, ci aspettiamo che le caratteristiche di una certa variabile si avvicinino il più possibile alle caratteristiche assunte dalla stessa variabile nella totalità della popolazione.

- *Utilizzare un campione anziché l'intera popolazione consente sicuramente di facilitare le analisi risparmiando tempo e riducendo i costi: possiamo sicuramente stimare l'altezza media degli italiani calcolando l'altezza media di un campione, ma questo dovrebbe spingerci a fare qualche riflessione. Secondo voi, sarebbe possibile misurare effettivamente l'altezza media di tutti gli italiani?*
- *E l'altezza media di tutta la popolazione, secondo voi sarebbe un valore costante, o potrebbe cambiare – seppure di poco – anche da un giorno all'altro?*

Gli indici di posizione per variabili continue

Cominceremo ad analizzare le variabili continue, definendo gli indici di posizione più utilizzati per descrivere le caratteristiche per un dato campione: tali indici sono la media (campionaria) aritmetica, la media geometrica, la mediana e la moda.

Media aritmetica

La definizione di *media aritmetica* (o *valore medio*, o – più semplicemente – *media*) di un campione è intuitiva: la media aritmetica di una variabile è data dalla somma dei valori misurati divisa per il numero di misurazioni effettuate, secondo la formula

$$\overline{x} = \frac{\sum_{k=1}^{n} x_k}{n} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n},$$

dove, per un numero n di soggetti, x_1, x_2, x_3, ..., x_n, rappresentano le misure della variabile ottenute per il soggetto 1, 2, 3, fino a quella del soggetto n-esimo. Quindi, ad esempio, se misuriamo il peso w in kg di 5 soggetti, trovando $w_1 = 78$, $w_2 = 84$, $w_3 = 78$, $w_4 = 81$, $w_5 = 79$, il peso medio sarà

$$\overline{w} = \frac{\sum_{k=1}^{5} w_k}{5} = \frac{w_1 + w_2 + w_3 + w_4 + w_5}{5}$$
$$= \frac{78+84+78+81+79}{5} = \frac{400}{5} = 80.$$

La definizione di media aritmetica è così familiare che non perderemo tempo nel commentarla ulteriormente. Vale tuttavia la pena soffermarsi su una sua proprietà importantissima, ovvero sul fatto che il prodotto della media per il numero di misurazioni dà come risultato la somma di tutte le misurazioni, ovvero che

$$n\overline{x} = \sum_{k=1}^{n} x_k = x_1 + x_2 + x_3 + \cdots + x_n$$

questa proprietà deriva direttamente dalla definizione di media, e nel nostro esempio precedente si riassume semplicemente facendo notare che la media ottenuta (80 kg) moltiplicata per il numero delle misurazioni del peso (5 misurazioni) è uguale alla somma di tutti i 5 pesi (400 kg), e pertanto sembrerebbe una cosa da dare quasi per scontata, ma la sua importanza è fondamentale.

La media aritmetica del campione (se il campione è rappresentativo della popolazione) è una stima affidabile della media dell'intera popolazione.

- *La formula per il calcolo della media aritmetica non vi ricorda forse qualcosa che avete già studiato nei corsi di fisica delle scuole superiori? Che cosa, di preciso?*

Media geometrica

Per definizione, la *media geometrica* che si ottiene da n misure è la radice n-esima del prodotto delle misure, ovvero:

$$\overline{x}_g = \sqrt[n]{\prod_{k=1}^{n} x_k} = \sqrt[n]{x_1 \times x_2 \times x_3 \times \cdots \times x_n},$$

Utilizzando i valori del peso in kg misurato per i 5 soggetti precedenti avremo quindi:

$$\overline{w}_g = \sqrt[5]{\prod_{k=1}^{5} w_k} = \sqrt[5]{w_1 \times w_2 \times w_3 \times w_4 \times w_5}$$
$$= \sqrt[5]{78 \times 84 \times 78 \times 81 \times 79} = \sqrt[5]{3270247344} = 79.97$$

Osserviamo una cosa molto interessante: calcolando la media aritmetica e la media geometrica dei cinque pesi utilizzati come esempio, verifichiamo che la media aritmetica è maggiore della media geometrica; più in generale, possiamo dire che, dato un certo numero di misure per una variabile continua, la media aritmetica è *quasi sempre* maggiore (e mai minore) della media geometrica.

- *In effetti, c'è un'unica situazione in cui la media aritmetica e la media geometrica risultano perfettamente uguali: qual è questa situazione?*
- *C'è un'altra cosa che rende differenti la media aritmetica e la media geometrica: consideriamo i due valori 5 e 7: qual è la loro media aritmetica, e qual è la loro media geometrica? Consideriamo ora la nuova coppia di valori 4 e 8: qual è la media aritmetica e qual è la media geometrica? Cosa è cambiato nei risultati rispetto alla prima coppia di valori?*
- *Provate ora a considerare i due valori 6 e 6: ancora una volta quali sono la media aritmetica e la media geometrica? Tutte queste considerazioni non vi hanno fatto riportare alla mente qualcosa che avete sicuramente studiato nei corsi di geometria del liceo, e relativo al calcolo della superficie di una figura piana? Quale?*

Mediana

Dato un certo numero di valori misurati in un campione, la *mediana* è quel valore che divide i dati in due parti uguali: il 50% delle misure ottenute sarà maggiore della mediana e il 50% sarà minore della mediana. Per quello che riguarda il nostro livello di approfondimento, non facciamo riferimento ad alcuna specifica formula per il calcolo della mediana, ma allo stesso calcolo è di per sé molto semplice. Basta infatti disporre i valori misurati in ordine crescente o decrescente (operazione molto facile con un qualsiasi foglio elettronico) e quindi cancellare il valore maggiore e minore, poi il valore maggiore rimanente e quello minore rimanente, e così via: se il numero di misure è dispari, la mediana sarà il valore posizionato esattamente a metà una volta disposti i dati in ordine crescente; se invece il numero di misurazioni è pari, la mediana sarà il valore intermedio esistente tra le due misure posizionate a metà una volta disposti i dati in ordine crescente.

Ad esempio, se avessimo misurato l'altezza in centimetri di sei soggetti ottenendo 162, 183, 162, 174, 180 e 179, disponendo i dati in ordine crescente otterremmo nell'ordine: 162, 162, 174, 179, 180, 183, per cui la mediana sarebbe il valore intermedio tra 174 e 179, cioè 176.5.

La mediana viene utilizzata come indice di posizione quando la distribuzione della variabile che si sta studiando ha una forma irregolare, oppure in determinati casi per descrivere il comportamento di specifiche variabili: ad esempio, negli studi epidemiologici, è abbastanza comune utilizzare il valore mediano anziché medio per descrivere l'età di un certo campione.

- *La mediana potrebbe mai avere un valore maggiore di quello della media aritmetica?*
  Cercate di giustificare la risposta con un ragionamento di tipo generale.

Moda

Per definizione, la moda di una serie di misure è il valore più frequente che si riscontra in tali misure: se ritornassimo alle sei misure delle altezze in centimetri utilizzate nei paragrafi precedenti, troveremo che la moda è pari a 162. Per calcolare la moda è sufficiente fare una tabella con i dati e cercarne il valore più frequente, oppure ancora più semplicemente, può bastare un istogramma delle misure.

- *In un campione di numerosità sufficiente, la media aritmetica, la mediana e la moda possono avere esattamente lo stesso valore?*

Gli indici di dispersione per variabili continue

Gli indici di posizione per le variabili continue fissano un valore che in un modo o nell'altro può essere considerato il "centro" dei valori misurati in un campione; gli indici di dispersione, invece, servono per dare una misura di quanto tali valori siano "raggruppati" attorno al centro. Per i nostri scopi, qui ci occuperemo del range, della somma degli scarti dalla media, della varianza e della deviazione standard.

Range

Il *range* (o intervallo) di un certo numero di misure è l'intervallo che separa la misura maggiore da quella minore: in generale, si indica semplicemente scrivendo la misura minima e la misura massima separate da un trattino, o – in taluni testi – con il valore della differenza. Rifacendoci al caso delle sedi altezze misurate, il range e può essere espresso con la scrittura 162-183.

Si tratta di una misura abbastanza grezza, ma comunque molto utilizzata, in particolare per verificare se tutte le misure ottenute sono verosimili, oppure se vi sono delle misure "fuori scala".

Scarti dalla media

Gli *scarti dalla media* SM sono per definizione le differenze che esistono tra ogni singola misura e il loro valore medio. Dato che alcune misure sono maggiori della media e altre sono minori, gli scarti dalla media sono caratterizzati da un segno positivo o negativo, e la loro somma (*somma degli scarti dalla media,* SSM), proprio per le proprietà della media è sempre nulla. Lo scarto dalla media, quindi, a meno che non venga calcolato nel suo valore assoluto, è un indice di dispersione di poca o nessuna utilità, ma ne parliamo ugualmente per verificare l'importante proprietà della media aritmetica cui abbiamo già accennato nei paragrafi precedenti.

Rifacendoci all'esempio dei cinque soggetti cui era stato misurato il peso (ed in cui era stato ottenuto un peso medio di 80 kg), calcoliamo facilmente la somma degli scarti dalla media come

$$\begin{aligned} SSM &= (78 - 80) + (84 - 80) + (78 - 80) + (81 - 80) + (79 - 80) \\ &= -2 + 4 - 2 + 1 - 1 = 0. \end{aligned}$$

Il fatto che la somma degli scarti dalla media sia sempre uguale a zero, può essere dimostrato in modo del tutto generale facendo ricorso al fatto che la media moltiplicata per il numero delle misure è uguale

alla somma di tutte le misure, ottenendo:

$$\begin{aligned}\text{SSM} &= \sum_{k=1}^{n}(x_k-\bar{x}) = (x_1-\bar{x})+(x_2-\bar{x})+(x_3-\bar{x})+\cdots+(x_n-\bar{x})\\ &= x_1-\bar{x}+x_2-\bar{x}+x_3-\bar{x}+\cdots+x_n-\bar{x}\\ &= (x_1+x_2+x_3+\cdots+x_n)-n\bar{x}\\ &= (x_1+x_2+x_3+\cdots+x_n)-(x_1+x_2+x_3+\cdots+x_n)=0.\end{aligned}$$

Varianza

Utilizzare la somma degli scarti dalla media non ha molta utilità, pertanto per descrivere la dispersione dei dati attorno alla media si preferisce utilizzare la *varianza* (identificata dal simbolo s^2), che è definita come la somma dei quadrati degli scarti divisa per il numero di soggetti che compongono il campione, diminuito di una unità. Per la precisione questa definizione riguarda la varianza campionaria ovvero la varianza calcolata sui soggetti selezionati per lo studio; altra cosa è invece la varianza della popolazione (di cui parleremo in uno dei moduli successivi), che è invece riferita a tutto l'universo da cui il campione è stato estratto. In particolare la varianza del campione è una stima della varianza della popolazione.

La formula generale per la varianza è

$$s^2 = \frac{\sum_{k=1}^{n}(x_k-\bar{x})^2}{n-1} = \frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+(x_3-\bar{x})^2+\cdots+(x_n-\bar{x})^2}{n-1}$$

Possiamo calcolare il valore della varianza del peso misurato nei 5 soggetti, riportato precedentemente. Ricordando che il peso medio era 80 kg, otteniamo

$$\begin{aligned}s^2 &= \frac{\sum_{k=1}^{5}(w_k-\bar{w})^2}{5-1} = \frac{(w_1-\bar{w})^2+(w_2-\bar{w})^2+(w_3-\bar{w})^2+(w_4-\bar{w})^2+(w_5-\bar{w})^2}{5-1}\\ &= \frac{(78-80)^2+(84-80)^2+(78-80)^2+(81-80)^2+(79-80)^2}{4}\\ &= \frac{(-2)^2+(4)^2+(-2)^2+(1)^2+(-1)^2}{4}\\ &= \frac{4+16+4+1+1}{4} = \frac{26}{4} = 6.5.\end{aligned}$$

La varianza del campione (se il campione è rappresentativo della popolazione) è una stima affidabile della varianza dell'intera popolazione.

- *Qual è il principale e più evidente vantaggio che si ottiene elevando al quadrato gli scarti dalla media nella formula della varianza rispetto alla somma degli scarti?*
- *Se il peso medio dei cinque pazienti si misura in kg, quale sarà l'unità di misura della varianza del peso?*

Deviazione standard

La varianza è un'eccellente misura della dispersione, ma viene espressa con una unità di misura differente da quella della media, per la precisione con il suo quadrato; per ovviare a questo inconveniente si utilizza la *deviazione standard s*, che è la radice quadrata del valore della varianza.

Utilizzando il calcolo della varianza ottenuto precedentemente per i pesi, per il calcolo della deviazione standard sarà sufficiente ottenere la radice quadrata del valore della varianza: nel nostro caso quindi avremo

$$s = \sqrt{\frac{\sum_{k=1}^{5} (w_k - \bar{w})^2}{5-1}} = \sqrt{\frac{(78-80)^2 + (84-80)^2 + (78-80)^2 + (81-80)^2 + (79-80)^2}{4}}$$
$$= \sqrt{\frac{4+16+4+1+1}{4}} = \sqrt{\frac{26}{4}} = 2.55.$$

Le caratteristiche, le proprietà e l'importanza fondamentale che la deviazione standard riveste in statistica saranno descritte nei moduli che parleranno delle distribuzioni dei dati e della statistica inferenziale. Per il momento, accontentiamoci di capire da cosa derivi la deviazione standard e come si possa calcolare a partire da un certo numero di dati.

- *Se invece di considerare il quadrato degli scarti per calcolare la varianza e poi ottenerne la radice quadrata per calcolare la deviazione standard, avessimo calcolato direttamente la somma dei valori assoluti degli scarti, che valore numerico avremmo ottenuto?*

Statistiche descrittive per le variabili categoriche

Non esistono delle metodiche elementari equivalenti a quelle utilizzate con le variabili continue per descrivere le variabili categoriche. Il metodo migliore, e anche il più efficace, è quello di costruire delle tabelle di frequenza per valutare quante volte una determinata realizzazione di una data variabile categorica sia presente nel campione.

Facendo un esempio molto semplice, e riprendendo il caso delle altezze in centimetri raccolte per sei soggetti citate precedentemente, ipotizzando di avere raccolto anche la variabile sesso (variabile binaria), e di avere visto che due soggetti erano di sesso maschile, e quattro di sesso femminile, la descrizione della variabile binaria sesso viene facilmente espressa dalla ripartizione percentuale che vede il 33. 3% di maschi e il 66. 7% di femmine.

Esercizio di verifica

Sono riportate le altezze misurate in un campione complessivo di 18 soggetti (maschi e femmine).

Si calcoli la media aritmetica, la mediana, la moda e la deviazione standard delle altezze di tutti i soggetti; si rifaccia poi il calcolo per i soli maschi e le sole femmine, come se fossero due campioni separati. Si faccia una tabella con la frequenza dei due sessi.

Excel ha delle funzioni che consentono di fare il calcolo diretto di tutto quanto richiesto: tuttavia, sempre utilizzando Excel, si consiglia di effettuare il calcolo "a mano".

Altezza	Sesso
177	M
191	M
177	F
181	M
164	F
175	F
180	F
176	M
179	M
175	F
169	M
174	F
177	M
182	F
177	F
172	F
170	F
183	M

Le risposte alle domande presenti nell'articolo saranno pubblicate sul prossimo numero del *Giornale di Tecniche Nefrologiche & Dialitiche* **Vol. 24, No. 3**