Giornale di Tecniche Nefrologiche & Dialitiche 2013; 25 (4): 335-338
DOI: 10.5301/GTND.2014.11749

L'intervallo di confidenza

Michele Nichelatti, Maurizio Nordio, Umberto Maggiore, Maurizio Postorino, Aurelio Limido, Anteo Di Napoli

a nome del Comitato Scientifico SIN-RIDT

The confidence interval

Abstract. The paper presents a very simple introduction to the concept of confidence interval for both a proportion and a mean. The confidence interval is useful when analyzing data that refer to a sample extracted from a target population. The confidence interval defines the upper and lower values of the interval of the variable we are studying; in other words, it defines the interval that one should expect, with a predefined level of confidence, to contain the true value of the proportion or the mean value of the population.

Key words: Confidence interval, Mean, Proportion, Sample, Population

Conflict of interest: None.

Financial support: None.

Ricevuto: 25 Novembre 2013; Accettato: 28 Novembre 2013

Nell'articolo "Sintetizzare i dati: la statistica descrittiva" (*GTN&D 2012; 24 (2): 53-60*), si introducevano le nozioni di "dati" e di "campione" e si trattavano gli indici di posizione e quelli di dispersione per variabili continue.
Tra gli indici di dispersione descritti vi era la deviazione *standard*, un indice direttamente derivato dalla varianza (ne è la radice quadrata), utile per descrivere quanto siano variabili intorno al valore medio le singole unità osservate.
È noto che, attraverso la media e la deviazione *standard*, è possibile descrivere una distribuzione normale (gaussiana), con la sua caratteristica forma a campana. In particolare, si dimostra che il 68% delle osservazioni cade in un intervallo definito dalla media più o meno 1 deviazione *standard* e che il 95% delle osservazioni è compreso in un intervallo definito dalla media più o meno 1.96 deviazioni *standard*.
È molto raro che si possano studiare tutte le unità che costituiscono una popolazione di interesse. In genere, si studia solo una parte della popolazione, detta campione, generalizzando, poi, i risultati ottenuti attraverso un processo noto come inferenza.
In qualsiasi modo il campione venga estratto, non potrà mai essere identico alla popolazione da cui deriva, a causa del cosiddetto errore di campionamento, che potrà solo essere più o meno contenuto adottando metodi appropriati. Infatti, l'errore di campionamento è inevitabile quando si utilizza un campione per stimare una caratteristica della popolazione.
La differenza tra i risultati ottenuti dal campione e quelli della popolazione che vogliamo stimare non potrà essere determinata con esattezza, in quanto il valore vero della caratteristica nella popolazione è sostanzialmente ignoto. Tuttavia, con adatti metodi statistici l'entità dell'errore di campionamento può essere stimato.
L'accuratezza della stima del valore di una caratteristica di una popolazione, ottenuta attraverso lo studio di un campione, dipende dalla numerosità del campione stesso. All'aumentare della numerosità del campione, la media del valore della caratteristica studiata tende ad avvicinarsi a quella della popolazione; infatti, si riduce l'errore campionario derivante dallo stimare la media della popolazione attraverso quella del campione.
Supponiamo di disporre di un campione di individui con una misura della creatininemia. Sarà possibile calcolare un valore medio della creatininemia. Supponiamo, ora, di disporre di una serie di campioni, ognuno costituito da un certo numero di individui, provenienti dalla stessa popolazione generale studiata. Ovviamente, per ciascun campione si potranno determinare altrettante medie. Queste medie campionarie, se trattate come singole osservazioni, avranno una loro distribuzione di frequenza.
Il teorema del limite centrale afferma che, per un campione ragionevolmente grande (almeno 30 osservazioni), la distribuzione delle medie campionarie del valore di una variabile è una normale (gaussiana), indipendentemente dalla distribuzione della variabile studiata nella popolazione generale. La varianza di tale distribuzione di medie campionarie è una stima non distorta della varianza della popolazione generale solo se, al denominatore della devianza, vengono posti i gradi di libertà (n-1), poiché la media è stata stimata dal campione. La deviazione standard (radice quadrata della varianza) della distribuzione delle medie campionarie è detta errore standard. In sostanza, l'errore standard è la deviazione standard delle medie campionarie e non di osservazioni individuali. Inoltre, per campioni di qualsiasi numerosità, l'errore standard di tutte le

possibili medie campionarie è uguale alla deviazione standard della popolazione diviso la radice quadrata della numerosità.
Più è piccola la numerosità campionaria, più è grande l'errore standard, vale a dire la variabilità nella distribuzione del parametro stimato (media). Pertanto, se l'errore standard è grande, è più verosimile che, estraendo un altro campione della stessa dimensione, si possa ottenere un valore della media stimata anche molto diverso. Viceversa, tanto più grande è la numerosità campionaria, tanto più piccolo sarà l'errore standard. In questo caso, i valori della media saranno più vicini alla media vera della popolazione.
Poiché il ricercatore tratta quasi sempre i dati di un campione casuale di una qualsiasi popolazione, e non quelli dell'intera popolazione, il calcolo della media (ma anche di una proporzione, di un tasso) è una stima puntuale del valore nella popolazione da cui il campione è estratto.
Ma il ricercatore e il lettore di una ricerca sono interessati a capire quanto la media misurata nel campione sia affidabile, e quanto la stima della media della popolazione che ne deriva, possa essere diversa dalla media effettiva. È possibile costruire un intervallo che, con una probabilità prefissata, contenga la vera media nella popolazione. Questo è il cosiddetto intervallo di confidenza.
L'intervallo di confidenza di un parametro (una media, una proporzione, un tasso), descrive il range di valori nel quale dovrebbe trovarsi il valore vero del parametro della popolazione. Tale stima, per definizione, avrà un carattere probabilistico, misurando intervalli di confidenza per ogni prefissato livello di probabilità. Per esempio, i due valori di un intervallo di confidenza calcolato a un livello di probabilità pari al 95% ci dicono che, con una probabilità del 95%, il valore vero del parametro della popolazione cade in quell'intervallo. Questo significa che, estraendo 100 campioni, in 95 casi essi conterranno il valore vero della media della popolazione studiata.
In sostanza, un intervallo di confidenza ci informa sul grado di precisione di una nostra stima, dipendente, come si è detto, dalla dimensione campionaria e dalla variabilità della distribuzione del fenomeno studiato nella popolazione. È possibile costruire gli intervalli di confidenza di una stima a partire dal calcolo dell'errore standard, indice che tiene conto dell'incertezza determinata da questi due fattori.
Un esempio ben noto di intervallo di confidenza è rappresentato dalla cosiddetta forbice degli *exit poll* elettorali, che descrive la possibile oscillazione della percentuale di voti stimata per ciascun partito. Nel caso degli *exit poll*, noi possiamo selezionare un campione di 100 votanti all'uscita da un seggio e, assumendo che le risposte siano veritiere, ottenere, per esempio, un risultato che dice che 45 elettori (il 45% del campione) hanno votato per il partito X. Cosa si può dire, allora, della percentuale di voti per lo stesso partito ottenuta tra tutti gli elettori e non solo tra i 100 selezionati all'uscita del seggio? Si può affermare che, in tutta Italia, il partito X abbia ottenuto esattamente il 45% dei voti basandosi solo su quel campione? Certamente no: infatti, oltre che il problema della risposta sincera dei 100 elettori del campione del seggio, ci si deve interrogare su quanto quegli stessi 100 elettori siano rappresentativi dell'intero universo di interesse (tutti gli elettori di tutti i seggi di tutta Italia). È lecito attendersi che, per essere rappresentativo, il campione dovrebbe avere delle caratteristiche (sesso, età, occupazione, opinioni politiche, scolarità, reddito, ecc.) quanto più possibile simili a quelle dell'intera popolazione di tutti gli elettori e, quindi, è necessario che la numerosità del campione venga aumentata: un campione ben selezionato di 1000 elettori è certamente più affidabile di un campione ben selezionato di 100 elettori. Per questo motivo, la variabilità delle proiezioni (la forbice) si restringe man mano che il numero di sezioni elettorali scrutinate aumenta, cioè man mano che cresce la dimensione del campione, con la riduzione, così, dell'errore *standard* e con l'aumento, quindi, della precisione della stima.
Si ribadisce che, mentre la deviazione *standard* misura la variabilità delle osservazioni individuali nella popolazione, l'errore *standard* misura la variabilità delle medie campionarie.
Pertanto, così come la media più o meno 1.96 deviazioni *standard* stima il *range* in cui dovrebbe cadere il 95% delle osservazioni individuali, la media più o meno 1.96 errori standard stima il *range* in cui dovrebbe cadere il 95% delle medie campionarie.
Conoscendo il valore della media più o meno 1.96 errori *standard* è possibile calcolare l'intervallo di confidenza al 95%, vale a dire l'intervallo di valori in cui, al 95%, cade la media reale della popolazione generale.
A parità di dimensione campionaria e di variabilità del fenomeno in esame, un intervallo di confidenza calcolato, per esempio, con probabilità del 99% sarà più ampio rispetto a quello calcolato con probabilità del 95%.
Un utilizzo fondamentale degli intervalli di confidenza è quello che consente di valutare se, per esempio, una media, una proporzione o un tasso differiscano in maniera statisticamente significativa da un altro valore puntuale calcolato o rispetto a un valore fissato.
Se gli intervalli di confidenza che comprendono i valori puntuali calcolati non si sovrappongono, tali valori potranno essere considerati significativamente diversi. Se, viceversa, gli intervalli di confidenza si sovrappongono, non si potrà affermare che i due valori puntuali siano diversi in maniera statisticamente significativa.
Un esempio comune del confronto tra un valore puntuale calcolato e un riferimento è la verifica della significatività statistica di un rischio relativo, di un *odds ratio* e di un *hazard ratio*. Si tratta di misure di rapporto e, pertanto, l'indifferenza è rappresentata dal valore 1, essendo uguali due numeri il cui rapporto è pari a 1. In tal caso, se l'intervallo di confidenza della misura racchiude il valore 1, vorrà dire che il rischio relativo, l'*odds ratio* e l'*hazard ratio* non differiscono da esso in maniera statisticamente significativa. Viceversa, se l'intervallo di confidenza della misura di interesse non racchiude il valore 1, questo indicherà che tale valore differisce in maniera statisticamente significativa da quello che rappresenta l'indifferenza, appunto l'1 nel caso di misure di un rapporto. Supponiamo, per esempio, che il rischio relativo di una patologia nei maschi rispetto alle donne abbia valore 1.6, con un intervallo di confidenza al 95% di 0.8-2.1: ciò indicherebbe che il rischio stimato per i maschi non è diverso in maniera statisticamente significativa rispetto alle donne. Un rischio relativo di 1.6, con un intervallo di confidenza al 95% di 1.2-2.1, indicherebbe, invece, che il rischio stimato è più elevato nei maschi in maniera statisticamente significativa. Un rischio relativo di 0.6, con un intervallo di con-

fidenza al 95% di 0.3-0.9, indicherebbe che il rischio stimato è più basso nei maschi in maniera statisticamente significativa.
Per analogia, se le misure di interesse si riferiscono a differenze (supponiamo tra due valori medi), il valore che rappresenta l'indifferenza pari a zero, e due numeri la cui differenza è 0 sono uguali. In tal caso, una differenza di valori (per esempio, di medie) sarà considerata statisticamente significativa se il suo intervallo di confidenza non comprenderà lo 0.
Come calcolare, in pratica, gli intervalli di confidenza? Al momento, non avendo ancora affrontato in modo approfondito e analitico la distribuzione gaussiana (normale), cioè quella curva dalla caratteristica forma a campana a cui si è già accennato poco sopra, possiamo accontentarci di utilizzare un paio di formule molto semplici, che consentono la stima approssimata di un intervallo di confidenza di una proporzione e di una media.
Per una proporzione P, la formula che utilizziamo per calcolare con ragionevole approssimazione l'intervallo di confidenza al 95% per un campione di soggetti è:

$$P \pm 1.96 \sqrt{\frac{P(1-P)}{n}}$$

Questa equazione è approssimata, ma funziona sufficientemente bene, a patto che, nel campione da cui vogliamo calcolare l'intervallo di confidenza, almeno 5 soggetti siano in ciascuna delle due opzioni che stiamo considerando e che il numero totale dei soggetti (la numerosità campionaria) sia sufficientemente grande.
Facciamo un esempio: uno sperimentatore studia la percentuale di successi ottenuti curando con un farmaco X un campione di 120 pazienti affetti dalla malattia Y e osserva che 90 pazienti guariscono e 30 no, deducendo che la proporzione delle guarigioni (dei successi) nel suo campione è 90/120 = 0.75 = 75%. Volendo capire quale sia la percentuale di guarigione in tutta la popolazione dei pazienti malati, possiamo usare la formula approssimata fornita sopra, perché 90 pazienti su 120 sono guariti (e 90 è sicuramente maggiore di 5) e i rimanenti 30 pazienti su 120 non sono guariti (e anche 30 è maggiore di 5) e, quindi, l'equazione può essere adoperata con sufficiente sicurezza, visto che possiamo anche assumere che il campione complessivo di 120 pazienti sia sufficientemente grande. In questo caso, quindi, se vogliamo calcolare l'intervallo di confidenza al 95% con cui ci attendiamo di stimare la proporzione P di tutti i pazienti che guariscono grazie al farmaco X, sapendo che la proporzione nel nostro campione è stata $P = 0.75$ e che il numero di soggetti del campione era = 120, scriveremo, passaggio per passaggio e arrotondando alla seconda cifra decimale:

$$P \pm 1.96 \sqrt{\frac{P(1-P)}{n}} = 0.75 \pm 1.96 \sqrt{\frac{0.75 \times (1-0.75)}{120}} = 0.75 \pm 1.96 \sqrt{\frac{0.75 \times 0.25}{120}}$$

$$= 0.75 \pm 1.96 \sqrt{\frac{0.1875}{120}} = 0.75 \pm 1.96 \times \sqrt{0.0015625}$$

$$\approx 0.75 \pm 1.96 \times 0.04 = 0.75 \pm 0.08.$$

Quindi, l'intervallo di confidenza al 95% della proporzione di guarigioni con il farmaco X andrà da 0.75 – 0.08 = 0.67 a 0.75 + 0.08 = 0.83, cioè ci attendiamo, con il 95% di confidenza, che la percentuale di guariti in tutta la popolazione dei malati stia tra il 67% e l'83%. In altre parole, siamo sicuri al 95% che l'intervallo che va dal 67% all'83% contenga la vera proporzione dei guariti nella popolazione di tutti i malati.
L'intervallo di confidenza può essere calcolato anche con precisione maggiore: per esempio, se vogliamo un intervallo di confidenza del 99%, basta sostituire quel fattore 1.96 dell'equazione con 2.58 (capiremo nei prossimi articoli il significato di questi numeri), per ottenere:

$$0.75 \pm 2.58 \sqrt{\frac{0.75 \times (1-0.75)}{120}} = 0.75 \pm 2.58 \sqrt{\frac{0.75 \times 0.25}{120}}$$

$$= 0.75 \pm 2.58 \sqrt{\frac{0.1875}{120}} = 0.75 \pm 2.58 \times \sqrt{0.0015625}$$

$$\approx 0.75 \pm 2.58 \times 0.04 = 0.75 \pm 0.10.$$

Pertanto, l'intervallo di confidenza al 99% della proporzione dei guariti va da 0.65 a 0.85, cioè ci attendiamo, con il 99% di confidenza, che la percentuale di guariti in tutta la popolazione dei malati stia tra il 65% e l'85%: l'intervallo di confidenza al 99% è, quindi, più dilatato di quello al 95%, come è ovvio attendersi e come già ricordato precedentemente nel testo.
Anche per una media, è possibile calcolare in modo davvero semplice l'intervallo di confidenza al 95%, a patto di conoscere la deviazione standard s delle misure ottenute e di avere un numero sufficientemente grande di misure. In questo caso, la formula che si utilizza per calcolare con una ragionevole approssimazione l'intervallo di confidenza al 95% è:

$$m \pm T \frac{s}{\sqrt{n}},$$

dove compare il termine, che capiremo meglio un po' più avanti, ma di cui al momento non ci dobbiamo preoccupare. Basti sapere che, se il numero di soggetti che compongono il campione è almeno pari a 20, si può assumere che sia $T = 2$ senza sbagliarsi troppo.
Se un medico scolastico misura la PAS in un campione di 100 studenti che frequentano una determinata classe di un liceo, ottenendo una media = 121.5 mmHg e una deviazione standard $s = 9.2$ mmHg, può utilizzare l'equazione per calcolare con ragionevole approssimazione l'intervallo di confidenza al 95% per il valore della PAS nella totalità degli studenti di quella stessa età, assumendo che il campione da lui studiato sia rappresentativo dalla popolazione di riferimento; usando l'approssimazione $T = 2$, otteniamo:

$$m \pm T \frac{s}{\sqrt{n}} = 121.5 \pm 2 \times \frac{9.2}{\sqrt{100}} = 121.5 \pm 2 \times \frac{9.2}{10}$$

$$= 121.5 \pm 2 \times 0.92 \approx 121.5 \pm 1.8$$

Di conseguenza, l'intervallo di confidenza al 95% per la media della PAS degli studenti andrà da 121.5 − 1.8 = 119.7 a 121.5 + 1.8 = 123.3, cioè ci attendiamo, con il 95% di confidenza, che la PAS media di tutti gli studenti di quella determinata età stia tra 119.7 mmHg e 123.3 mmHg, quindi siamo sicuri al 95% che l'intervallo che va da 119.7 mmHg e 123.3 mmHg contenga la vera media della PAS degli studenti.
Ma che cosa è di preciso quel termine T che abbiamo approssimato a 2? È un valore che deriva dalla distribuzione t di *Student* e che dipende sia dal valore che vogliamo dare all'intervallo di confidenza sia dalla numerosità del campione che stiamo studiando. Per esempio, rimanendo su una confidenza del 95%, con un campione di 10 soggetti, abbiamo $T = 2.262$, mentre, se i soggetti sono 20, abbiamo $T = 2.093$, se sono 25, $T = 2.064$ e, se sono 30, $T = 2.045$; infine, quando i soggetti sono davvero tanti, al punto da averne un numero infinito, allora risulta $T = 1.96$, e questo valore dovrebbe ricordarci qualcosa che abbiamo già visto da poco.
Affronteremo, nei prossimi numeri della rivista, la distribuzione normale (la curva gaussiana a campana) e, allora, tutto inizierà a chiarirsi.

Riassunto

Viene presentata una introduzione molto semplice al concetto di intervallo di confidenza per una proporzione e per una media. L'intervallo di confidenza è utile quando si analizzano dati ottenuti da un campione estratto da una popolazione bersaglio, per capire entro quale intervallo di valori della variabile che stiamo studiando è lecito attendere con un prestabilito margine di sicurezza, sia posizionato il valore vero della proporzione e della media nella popolazione.

Parole chiave: Intervallo di confidenza, Media, Proporzione, Campione, Popolazione

Dichiarazione di conflitto di interessi: Gli Autori dichiarano di non avere conflitto di interessi.

Contributi economici agli Autori: Gli Autori dichiarano di non aver ricevuto sponsorizzazioni economiche per la preparazione dell'articolo.

Indirizzo degli Autori:
Dr. Michele Nichelatti
Servizio di Biostatistica
Dipartimento di Ematologia e Oncologia
Ospedale Niguarda Ca' Granda
Piazza Ospedale Maggiore 3
20162 Milano
michele.nichelatti@ospedaleniguarda.it