- Copyright - Il Pensiero Scientifico Editore downloaded by IP 3.17.74.227 Mon, 29 Apr 2024, 16:45:57



# Test statistici ed intervalli di confidenza


## Riassunto

Dopo aver presentato i concetti alla base degli intervalli di confidenza, sono state sottolineate le differenze con i test statistici, mostrando che non solo sono strumenti concettualmente distinti, ma anche che non possono ritenersi, in generale, equivalenti.

***Parole chiave.*** *Intervalli di confidenza, test statistici, inferenza sulla frequenza.*

## Summary

### *Confidence intervals and statistical test*

After showing the basic concepts of the confidence intervals, the differences between the confidence intervals and statistical tests were highlighted, also showing that their results are not ever equivalent.

***Key words.*** *Confidence intervals, statistical test, inference on the proportions.*


*Durante le prove orali di un concorso per l'ammissione ad un dottorato di ricerca, un professore ordinario della mia materia chiese al candidato se conoscesse un metodo alternativo al test per il confronto tra due frequenze per provare l'ipotesi nulla di uguale efficacia dei trattamenti, nel caso di risposta binaria. Dopo una lunga esitazione del candidato, sentii fornire la risposta dallo stesso docente: costruire un intervallo di confidenza per la differenza tra due frequenze e vedere se comprende lo zero; in caso affermativo si accetta l'ipotesi nulla, in caso negativo la si respinge. Sobbalzai.*

*Successivamente, purtroppo, ritrovai la stessa affermazione in una dispensa di Epidemiologia. Dato che negli ultimi 3 numeri di CASCO la rubrica è stata incentrata su vari aspetti dei test statistici, questo lontano ricordo mi ha indotto ad affrontare l'argomento della differenza tra test statistici ed intervalli di confidenza in questo numero della rubrica, nella convinzione – basata sull'esperienza – che non sempre i medici, nella loro preparazione nelle discipline a contorno della metodologia della ricerca clinica, sono seguiti da docenti del tutto affidabili.*

*Come al solito, pur cercando di rendere la lettura di questa nota quanto più possibile indipendente da ogni specifica pre-conoscenza, per riuscire a restare negli spazi stabiliti ci sarà bisogno di fare riferimenti ai contenuti di "Statistica per concetti" esposti nei numeri 5 e 6 di CASCO.*

## Gli intervalli di confidenza

**Esempio.** In uno studio di fase 2, condotto su 100 pazienti, la percentuale di successi terapeutici ottenuta con il nuovo trattamento è del 60%. Questa è la migliore stima che si può fornire della percentuale di successi che potremmo osservare nella popolazione target (cioè la migliore stima del parametro il cui valore è sconosciuto). Essa prende il nome di "stima puntuale".

La stima puntuale fornisce la massima informazione sul valore del parametro in quanto con il numero dedotto dalle osservazioni si stima un numero sconosciuto. Il suo problema, però, è che la probabilità che essa coincida con il parametro è, in generale[1], pari a zero (v. CASCO 5: Stime e loro variabilità).

1. È sempre uguale a zero quando lo stimatore è rappresentato da una variabile casuale continua, può essere non nulla quando lo stimatore è una variabile casuale discreta.

**Esempio (prosec.).** La migliore stima della percentuale di successi nella popolazione target, P, cioè nella popolazione di tutti i pazienti presenti e futuri affetti da quella certa malattia, simili a quelli osservati (per criteri di eleggibilità ed esclusione) e che ricevano il nuovo trattamento è pari al 60% (massima informazione sul parametro). Però, la probabilità che la stima coincida con il parametro è pari a zero. Si può allora considerare un intervallo entro cui cada il valore di P, ad esempio, tra il 58% e il 62%. In tal caso la probabilità che P vi sia compreso è maggiore di zero, ma, così facendo, si è rinunciato a fornire il massimo dell'informazione. Naturalmente possiamo considerare un intervallo più ampio, diciamo tra il 55% e il 65%: l'informazione diminuisce (perché il parametro può assumere un qualunque valore di tale intervallo: rispetto al caso precedente, ne sappiamo di meno), ma la probabilità che P sia interno a tale intervallo è ben maggiore. Il caso estremo è rappresentato dall'intervallo compreso tra 0 e 100%: la probabilità che P vi sia compreso è uguale a 1 (cioè è certo che P vi cada), ma il contenuto di informazione è pari a zero (è stato inutile eseguire lo studio, in quanto tale informazione era nota anche prima di condurlo).

L'intervallo di confidenza è un compromesso tra le due situazioni estreme (massima informazione, ma probabilità nulla che sia vera; informazione nulla ma probabilità 1 che sia vera).

Più precisamente, **l'intervallo di confidenza ha il significato di intervallo, costruito intorno alla stima puntuale, in cui, con alta probabilità, cade il valore del parametro.** Convenzionalmente, tale "alta probabilità" viene fissata al 95% ed è detta "coefficiente dell'intervallo", sebbene non siano rarissimi i casi in cui si fissino altri valori, ad esempio 99% o 90%.

L'intervallo di confidenza viene costruito a partire: a) dalla stima puntuale, b) dal corrispondente stimatore, c) dall'errore standard dello stimatore (v. CASCO 5).

**Esempio (prosec.).** I dati sono: n = 100, f = 0,6. Sapendo che nel caso dei "grandi" campioni (come nel nostro caso)

- Copyright - Il Pensiero Scientifico Editore downloaded by IP 3.17.74.227 Mon, 29 Apr 2024, 16:45:57



la frequenza relativa si distribuisce normalmente, stimandone l'errore standard, si può calcolare che, con il 95% di probabilità, la percentuale di successi nella popolazione target è compresa tra il 50,4% e il 69,6%.

Se l'intervallo di confidenza fosse ritenuto poco informativo (perché troppo ampio), sarebbe necessario aumentare il numero dei pazienti osservati. Se ne avessimo arruolati 400, anziché 100 (sempre con f = 0,6), l'intervallo di confidenza al 95% sarebbe stato compreso tra il 57,6% e il 62,4%: ben più informativo del precedente!

Gli intervalli di confidenza, come l'errore standard e i test statistici, sono strumenti per valutare l'imprecisione della stima, ossia la sua variabilità dovuta unicamente al caso, cioè al campionamento. Tali strumenti, quindi, non possono valutare l'inaccuratezza della stima, ossia la variabilità indotta da distorsioni sistematiche. Ad esempio, se per motivi, noti o sconosciuti, uno studio clinico fosse basato su pazienti in mediamente migliori condizioni rispetto a quelli che si presentano all'osservazione nella pratica clinica, un intervallo di confidenza, così come un test statistico, non potrebbe comunque mai essere riferito alla totalità dei pazienti, ma solo ai pazienti in condizioni analoghe a quelle degli arruolati nello studio (questo è il principale problema delle stime ottenute con gli studi di fase II che possono dar luogo a risultati entusiasmanti, spesso, però, destinati ad essere vanificati nella successiva fase III).

In sede di pianificazione di uno studio osservazionale o sperimentale (come, ad es., uno studio di fase II), la dimensione dello studio viene fissata o in relazione alla potenza del test statistico che sarà successivamente usato per il confronto con un dato esterno, o all'imprecisione massima tollerabile, cioè all'ampiezza massima dell'intervallo che si reputa ancora informativa sul valore del parametro. Nell'esempio, per n = 100, se l'intervallo f ±10% fosse considerato poco informativo, si dovrebbe aumentare la dimensione dello studio fino a giungere ad un intervallo che si reputa informativo (ad es., f ±3%).

## Intervalli di confidenza e test statistici

A ben guardare, la tecnica di costruzione di un test statistico è simile a quella che è alla base degli intervalli di confidenza (v. CASCO 6), e ciò può aver indotto l'equivoco di cui si è parlato in premessa, ma non sono strumenti equivalenti. Le differenze sostanziali sono due, una di natura concettuale, la seconda più tecnica, ma decisamente più importante.

Premettiamo che due sono le aree di interesse dell'inferenza statistica: la stima dei parametri e la prova delle ipotesi. Gli intervalli di confidenza appartengono alla prima area, mentre i test statistici, che appartengono alla seconda, si propongono un diverso obiettivo: provare l'ipotesi che il parametro assuma un determinato valore. Quindi, sono due strumenti, concettualmente differenti, che rispondono ad esigenze diverse.

La seconda differenza è che **entrambi gli strumenti poggiano sul valore dell'errore standard, ma, mentre gli intervalli di confidenza, in generale, richiedono una sua stima, i test statistici si basano sull'errore standard calcolato sotto l'ipotesi nulla**.

Ci sono casi in cui i due metodi producono lo stesso valore per l'errore standard (quando, ad esempio, nel test il valore ipotizzato per il parametro non ha influenza sul valore stimato dell'errore standard, come nell'inferenza sulla media) – e, in questo caso, il collega citato in premessa avrebbe potuto aver ragione – ma non è così nei problemi di inferenza relativi alle frequenze. Infatti, nel caso della frequenza relativa – come anche nel caso del confronto tra due frequenze – l'errore standard stimato non coincide con quello calcolato sotto l'ipotesi nulla e quindi l'intervallo di confidenza ed il test statistico potrebbero dar luogo a risultati contrastanti.

Ad esempio, nel nostro lavoro sugli antiemetici (IGAR, Ondansetron + dexamethasone vs metoclopramide + dexamethasone + diphenhydramine in prevention of cisplatin-induced emesis. Lancet 1992; 340: 96-9), l'intervallo di confidenza (al 95%) della differenza dell'incidenza della nausea acuta tra i due gruppi (stima puntuale della differenza = 0,116 = 11,6%) era compreso tra 0,8% e 22,4% (lasciando così intravvedere una significatività della differenza, in quanto tale intervallo non include lo zero). Il livello di significatività del test statistico era, invece, risultato pari a 0,051, quindi superiore al 5%, per cui non si può concludere che i due trattamenti abbiano una diversa efficacia nella prevenzione della nausea acuta.

In tali casi a chi dare retta? Senza dubbio al test statistico in quanto lo studio è specificamente programmato per basarsi su di esso: gli intervalli di confidenza hanno solo il ruolo di aggiungere un'informazione – pur importante, ma non decisiva – sulla dimensione dell'effetto.

***Enzo Ballatori***