Guida al corretto uso della P e degli Intervalli di Confidenza nella lettura e presentazione dei risultati di uno studio clinico

P. Ravani[1], S. Andrulli[2]

[1]Divisione di Nefrologia e Dialisi, Azienda Ospedaliera, Cremona
[2]Divisione di Nefrologia e Dialisi, Ospedale A. Manzoni, Lecco

P value and confidence intervals to report and interpret the result of a clinical study

The main purpose of statistics in the analysis of clinical and epidemiological studies is to summarize data and information, as well as assess variability, trying to distinguish between chance findings and results that may be replicated upon repetition. Statistical analyses only convey the effect of chance element in data (random error). Statistics cannot control non-sampling errors concerning study design, conduct and methods adopted. At the end of the study, a result is defined statistically significant *if the observed difference in the outcome variable is too large to be attributed to chance. A small P value provides evidence against the null hypothesis (of no effect), since data have been observed that would be unlikely if the null hypothesis was true. However, confidence intervals estimate separate the two data dimensions (strength of the relation between exposure and disease, and precision with which the relation is measured), and add to the hypothesis testing useful information for finding interpretation and further research.(G Ital Nefrol 2006; 23: 490-501)*

KEY WORDS: Measures of central tendency and dispersion, P value and confidence intervals, Type I and II error rates, Random and systematic errors, Statistical test

PAROLE CHIAVE: Indici di tendenza centrale, Indici di dispersione, Stima puntuale, Stima intervallare, Errore di tipo I e II, Errore casuale e sistematico, Test di ipotesi

Importanza della variabilità biologica nei fenomeni di interesse medico

La variabilità biologica è una caratteristica propria di tutte le forme di vita e ad essa si attribuisce un ruolo chiave in senso evolutivo. Di questa variabilità è necessario tener conto sia nella descrizione delle caratteristiche di un campione (*statistica descrittiva*) che stimando le caratteristiche (i *parametri*) di una popolazione attraverso lo studio di un suo sottoinsieme (il *campione*). Come clinici siamo interessati alla *variabilità spiegabile* dei fenomeni biologici (imputabile a fattori o esposizioni) e a distinguerla dalla *variabilità casuale* (dovuta al caso).

Purtroppo, la variabilità in generale è spesso concepita come un disturbo da ignorare o nascondere al punto che anche in lavori pubblicati su riviste prestigiose, essa viene minimizzata utilizzando erroneamente indici numericamente più piccoli come l'*errore standard* anziché la *deviazione standard*. Le caratteristiche di un campione sono, infatti, riassunte mediante due indici statistici: *indici di tendenza centrale* e *indici di variabilità*. Ad esempio: l'informazione relativa al peso o alla pressione arteriosa di un campione di soggetti può essere riassunta con la media (indice di tendenza centrale) e la *deviazione standard* (indice di variabilità). A seconda della distribuzione dei valori della variabile considerata a volte è appropriato utilizzare altri indici statistici come la mediana e il *range* (valori estremi). Entrambi gli indici sono importanti per descrivere le caratteristiche del campione. Similmente esistono diversi *tests* statistici da applicare ai dati osservati a seconda delle caratteristiche della variabile considerata. Ad esempio se la variabile è "la sopravvivenza" allora le analisi di sopravvivenza, saranno appropriate; se la variabile è una variabile continua come i valori pressori (con distribuzione Normale) un t *test* o un'ANOVA saranno appropriati e così via. In ogni caso anche i *test* statistici tengono conto non solo dei valori più

© Società Italiana di Nefrologia

frequenti ma anche della dispersione dei dati. Pertanto, alla fine della lettura di questo articolo, risulterà chiaro che la variabilità dei dati è importante non solo per la *statistica descrittiva* ma anche per la *statistica inferenziale* nei due approcci complementari ed integrati di:

a) un *test di ipotesi nulla* che genera una P e

b) una *stima puntuale ed intervallare* che produce un intervallo di confidenza.

Errore casuale ed errore sistematico

Nella lettura critica di un lavoro scientifico ci poniamo delle domande relative al ruolo del caso nella generazione dei risultati. Fondamentalmente le domande sono: a) il risultato dello studio è statisticamente significativo?; b) se non lo è, lo studio è stato sufficientemente dimensionato? Infatti, la pianificazione di una ricerca medica è basata sull'idea di osservare un campione rappresentativo della popolazione (*campionamento*), elaborare delle stime su misure campionarie (per esempio medie e deviazioni *standard*) per poi generalizzare (fare *inferenza*) sulla popolazione da cui il campione è stato estratto. La variabilità nel campione dipenderà dalla variabilità originaria della popolazione campionata, dal ruolo del caso operante nel processo di campionamento e da fattori che influenzano la variabile di interesse.

La statistica ci permette di misurare questa variabilità nei dati campionari nell'intento di distinguere risultati casuali da quelli che si ripresenterebbero ripetendo lo studio. Tutto questo è necessario perché i fenomeni biologici non sono adeguatamente descritti dalle leggi della fisica e della chimica. Infatti essi sono chiamati fenomeni *stocastici* o *aleatori* (in contrasto ai fenomeni *deterministici*), in quanto producono risultati imprevedibili anche se ripetuti nelle stesse identiche condizioni. Tuttavia bisogna tener presente che gli errori che si possono commettere nella stima dei parametri veri non sono solo influenzati dal caso (*errore casuale*), ma anche da *errori sistematici* (*bias*), come ad esempio la presenza di confondenti, un campionamento non valido, un disegno peculiare dello studio, una non corretta taratura degli apparecchi di misura, ecc. Se immaginiamo che uno studio possa essere infinitamente aumentato di dimensione (aumentando la numerosità), l'errore casuale nella stima del parametro si ridurrebbe a zero, a mano a mano che il campione, crescendo, tende a coincidere con la popolazione. Al contrario, l'errore sistematico non si ridurrebbe, nemmeno in uno studio infinitamente grande (Fig. 1). Dobbiamo tenere ben presente che le misure di precisione che impareremo ad interpretare ci informano sul ruolo del caso (errore casuale) e non su altre distorsioni (errori sistematici) di cui soffrono molti lavori scientifici. Per mettere in evidenza questi ultimi, occorre una lettura attenta della parte metodologica, dove vengono descritte la modalità di selezione dei pazienti e le strategie di raccolta dei dati.

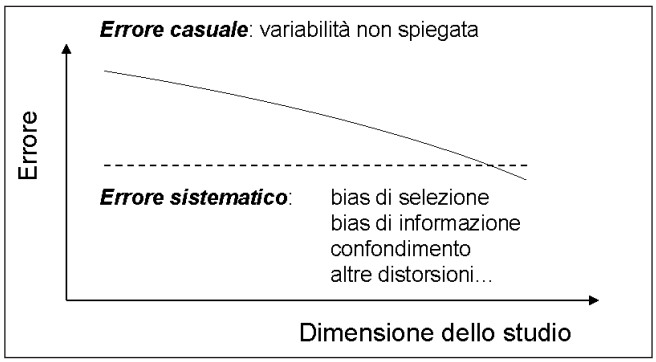

Fig. 1 - *L'errore casuale si riduce aumentando la numerosità campionaria mentre l'errore sistematico non cambia.*

Misure di precisione della stima

I *valori di P* e gli *intervalli di confidenza* sono misure di precisione della *stima* di un parametro. Consideriamo, per esempio, gli effetti di un'esposizione (un trattamento o un fattore di rischio) calcolata come "*risk ratio*" (RR), ovvero il rapporto tra il rischio di eventi tra pazienti trattati e controlli. Se tra 100 trattati si verificano 10 eventi, mentre tra i 100 controlli si verificano 20 eventi, il RR è 0.5 (0.1/0.2). Questa *stima puntuale* di RR *stima* il valore vero di RR di tutta la popolazione. Vicino a questa stima di effetto sono riportati un valore di P (0.04) e un intervallo di confidenza (al 95% da 0.25 a 0.99).

La comprensione degli indici di imprecisione della stima è molto importante per una corretta lettura degli studi clinici. Purtroppo i concetti statistici sottostanti sono complessi e possono scoraggiare il clinico se presentati in veste troppo tecnica. Per tale ragione eviteremo per quanto possibile l'utilizzo del gergo tecnico utilizzato dagli statistici, privilegiando un approccio pratico (1-5).

Il valore di P

Facciamo un esperimento semplice come il lancio di una monetina. L'esempio ci aiuterà a comprendere il significato di diversa probabilità (*distribuzione di probabilità*) dei risultati di un esperimento. Immaginiamo di lanciare una monetina non truccata, ossia con P di testa e croce identiche (50% ad ogni lancio), per 10 volte. Questa procedura di 10 lanci è il nostro *trial* (prova). Ora ripetiamo il *trial* per 10000 (ossia un numero elevatissimo di) volte e rappresentiamo la proporzione (percentuale) di ripetizioni che ha prodotto ciascun possibile numero di teste (esistono 11 possibili soluzioni): *trials* con 0 teste su 10 lanci (0/10); 1/10; 2/10; …; 10/10. La Figura 2 rappresenta la distribuzione di probabilità di ciascuna soluzione con un grafico a barre: ogni barra rappresenta la probabilità (P) di ciascuna soluzione (quante volte la soluzione si è verificata su 10000 prove, frequenza relativa). Va notato che la somma

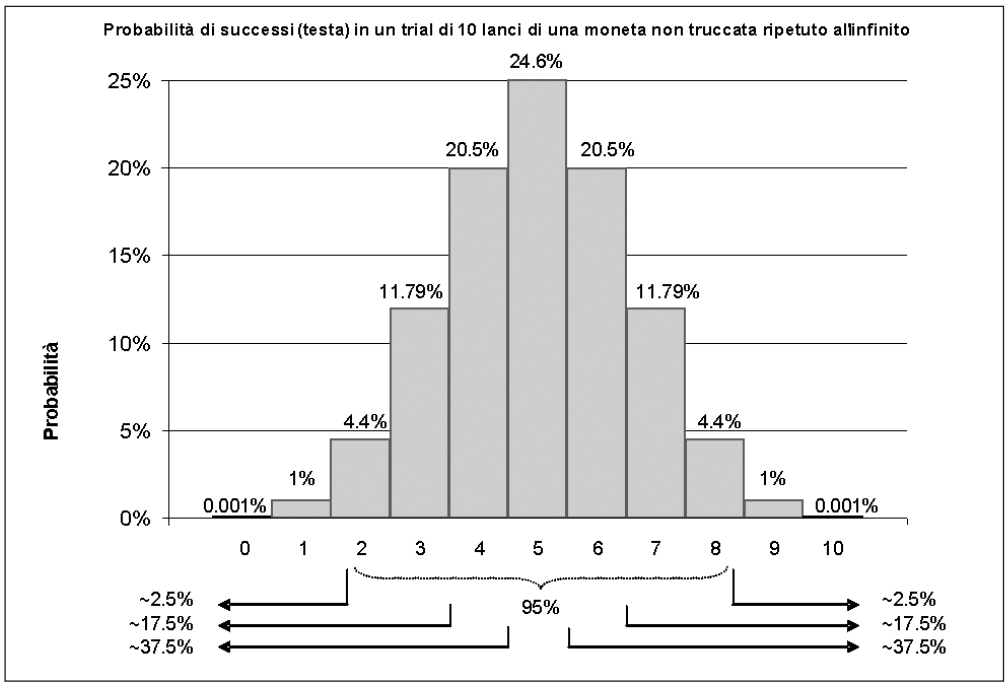

Fig. 2 - *Probabilità degli 11 possibili risultati della prova di 10 lanci di una monetina non truccata. La probabilità di ciascuna soluzione è la sua frequenza relativa in una serie infinita di ripetizioni della stessa prova.*

delle P di tutti i possibili risultati è pari al 100%. Adesso proviamo a porci qualche domanda: qual è la P di 6 o più teste per prova? La P (detta cumulata) di più risultati è ottenibile sommando le P di ciascun risultato: 20.5 + 11.8 + 4.4 + 1 + 0.001 = 37.7%. Ora che abbiamo descritto cosa succede in un numero molto elevato di *trials* (10000), ossia di lanci di una monetina non truccata, vediamo come possiamo interpretare il risultato di *una singola prova* (sempre di 10 lanci). La nostra ipotesi di lavoro potrebbe essere che esista un trucco e, ottenendo un risultato diverso da 5 teste e 5 croci vorremmo sapere se il dubbio dell'esistenza di un trucco sia ragionevole o meno. Immaginiamo di ottenere 7 teste e 3 croci. La domanda corretta da porsi è: qual è la probabilità di ottenere questo risultato o un risultato più estremo (8, 9, 10 teste, oppure 0, 1, 2, 3 teste) per il puro effetto del caso (in assenza di trucco)? In altre parole, qual è la P di 7 o >7 teste (oppure 3 o <3 teste) se la monetina non è truccata? La P è ~ 35%, ossia P = 0.35.

Questa probabilità elevata non mi consente di concludere che la moneta sia truccata (o meglio, che non sia non truccata) anche se in un *trial* di 10 lanci ho osservato 7 teste su 10 lanci, anziché 5. L'esperimento della monetina è simile ad uno studio clinico. L'assunto di partenza è che l'effetto del trattamento sperimentale sia nullo. Così come abbiamo generato la P di tutti i possibili risultati sotto l'assunto di mancanza di trucco (distribuzione di probabilità dei possibili successi in una serie infinita di prove), allo stesso modo in un *trial* clinico lo statistico genera una distribuzione teorica dei possibili risultati assumendo che i trattamenti in oggetto non siano diversi (ipotesi nulla). I valori di P del *test* statistico rappresentano le P (porzioni della distribuzione) corrispondenti a risultati estremi come (o anche più estremi di) quelli osservati nello specifico *trial*. Man mano che ci si sposta all'estremo delle code, ad esempio verso il risultato di 1 o 0 teste, la P si riduce. Ad un certo punto concorderemo nel ritenere che "la P di man-

TABELLA I - INTERVALLI DI CONFIDENZA (IC) DI STUDI CON IDENTICO RISULTATO MA DIMENSIONE PROGRESSIVAMENTE MAGGIORE. ALL'AUMENTARE DELLA NUMEROSITA CAMPIONARIA LA PRECISIONE DELLA STIMA MIGLIORA E L'IC SI STRINGE

				IC al 95%		
Soggetti	**R Trattati**	**R Controlli**	**RR**	**da**	**a**	**P**
8	1/4	2/4	0.5	0.07	3.65	0.4652
40	5/20	10/20	0.5	0.20	1.20	0.1025
80	10/40	20/40	0.5	0.26	0.92	0.0209
200	25/100	50/100	0.5	0.33	0.73	0.0003
2000	250/1000	500/1000	0.5	0.44	0.56	<0.0001

canza di effetto" (mancanza di trucco) sia così piccola che, dati i risultati osservati, l'ipotesi nulla non è più sostenibile (l'*ipotesi nulla* di monetina non truccata, ossia di effetto nullo, è rigettata).

Questo punto è stato convenzionalmente (e arbitrariamente) identificato nella P = 0.05. La regione complementare di "non rifiuto" è ovviamente pari al 95%.

Un esempio con una variabile continua

Facciamo ancora un esempio con l'aiuto della Figura 3.

Invece di confrontare gruppi, per semplicità poniamoci un quesito relativo ad *una singola osservazione*.

Supponiamo di dubitare che un paziente con determinati valori pressori (variabile continua) appartenga ad una certa popolazione (ipotesi dello studio) in cui la pressione arteriosa diastolica presenti valori di 80 (media) ± 10 (deviazione *standard-standard* deviation, SD) mmHg (curva superiore). La media descrive il valore centrale della distribuzione dei valori di PA nella popolazione in questione. In questa popolazione, ovviamente, non tutti i soggetti hanno PA = 80. La SD ci informa sulla dispersione dei valori attorno al valore centrale. Media e SD sono detti parametri della popolazione. L'area sotto la curva racchiude il 100% delle osservazioni. Sempre in Figura 3 un'altra curva (inferiore) descrive la funzione di densità di P in funzione del risultato di un *test* statistico (Z) dell'ipotesi nulla: la singola osservazione casuale appartiene alla popolazione (differenza nulla).

La statistica Z è lo strumento utilizzato dallo statistico per rispondere al quesito clinico che ci siamo posti. Il risultato (Z) è detta "deviata gaussiana standardizzata" perché ha una distribuzione gaussiana con media zero e SD unitaria ed è pari al rapporto tra la differenza del valore casuale osservato e la media (a numeratore) e la SD (a denominatore). Una caratteristica importante di Z è che, essendo adimensionale, può rappresentare lo strumento comune utile per costruire *test* di ipotesi nulla per *tutte* le variabili biologiche distribuite in modo normale.

Supponiamo ora che il soggetto in questione presenti valori di pressione diastolica compresi tra 70 e 90 mmHg. Il quesito clinico è che la sua PA sia *diversa* da quella dei soggetti della popolazione di riferimento.

Abbiamo imparato che la curva inferiore rappresenta la distribuzione dei valori di Z sotto l'ipotesi nulla che i valori di un'osservazione casuale non siano estranei alla popolazione. Dalla Figura 3 si vede che i soggetti con diastolica compresa tra 70 e 90 mmHg hanno un valore di Z compreso tra 0 e ± 1. A questo intervallo di valori di Z corrisponde una P di appartenere alla popolazione in questione di 0.683, per cui non possiamo rifiutare l'ipotesi nulla perchè non c'è sufficiente evidenza contro l'ipotesi nulla. Valori di pressione diastolica uguale o superiore a 100 rendono meno probabile l'ipotesi nulla perché (100-80) diviso 10 (la SD) generano un valore di Z ≥ 2 con una P dell'area

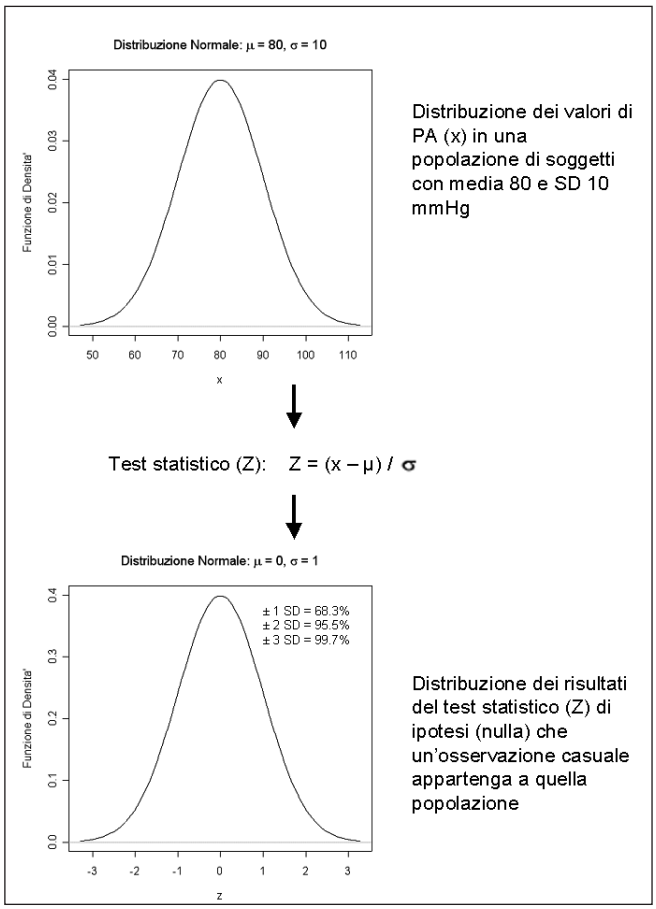

Fig. 3 - *Distribuzione normale standardizzata e probabilità (area sotto la curva) associata a multipli della deviazione* standard. *In ordinate c'è la densità di P perché in questo caso esistono valori infiniti di risultati del* test *statistico (variabile continua) e la P è calcolata per intervalli di Z (o di x) in quanto la P di un valore di Z (o di x) misurato con precisione infinita non è determinabile.*

sotto la curva ≤ 0.05 (vedi Fig. 3 curva inferiore). Poiché le code della distribuzione normale si estendono all'infinito, valori estremi di Z sia positivi che negativi benché associati a bassi valori di P, sono ancora compatibili con l'ipotesi nulla che quindi non si potrà mai definire vera o falsa. Tuttavia, possiamo *saggiare (misurare) la compatibilità dei dati con essa*. Se il risultato del *test* ha un valore elevato (basso P) noi abbiamo *due possibilità*: o si è verificato un errore dovuto al caso, oppure l'ipotesi nulla non è corretta. Di solito il limite viene posto ad un valore di differenza (numeratore) pari al doppio della misura di variabilità del denominatore (da cui il valore convenzionale di 0.05).

Struttura del *test* statistico (parametrico)

La stessa procedura può essere applicata per saggiare non singoli valori ma *medie campionarie* ponendo al denominatore l'*errore standard* al posto della SD. In generale

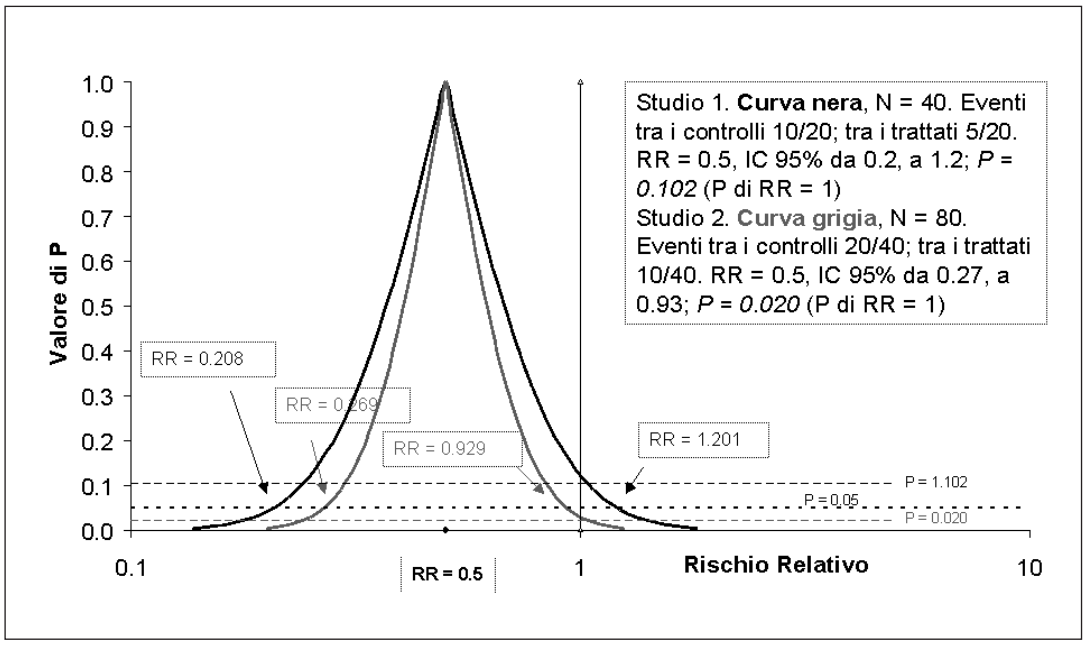

Fig. 4 - *Funzione "valore di P vs. Intervalli di Confidenza" di 2 studi di dimensione diversa ma con la stessa stima puntuale.*

un *test* statistico è il rapporto tra la stima della misura di effetto (ad esempio un rapporto o una differenza di rischi) e la stima della sua variabilità (errore *standard* dell'effetto) generate dai dati osservati. I risultati del *test* statistico hanno una *funzione di densità di probabilità* nota sotto l'ipotesi nulla (di mancanza di effetto).

Valori estremamente elevati (o bassi) del *test* corrispondono a valori di P estremamente piccoli sotto l'ipotesi nulla.

Riassumendo, sulla base del *test* statistico, l'ipotesi nulla è rigettata se la P è sufficientemente piccola, in genere, per convenzione, quando P è ≤ 0.05. Valori vicini a 0.05 offrono un'evidenza moderata mentre valori < 0.01 offrono evidenza considerevole contro l'ipotesi nulla.

Test P ad una o a due code?

I *test* a due code considerano che i valori osservati relativi all'effetto di un trattamento possano essere sia maggiori (trattamento migliore del controllo) che inferiori (trattamento peggiore del controllo) a quelli stabiliti dalla ipotesi nulla. Questi tipi di *test* sono i più utilizzati in quanto spesso siamo incerti sulla reale direzione dell'effetto di un trattamento, ma soprattutto perché sono più conservativi. Al contrario i *tests* ad una coda considerano una direzione soltanto e la P corrispondente sarebbe la metà (ad esempio se P = 0.06 a due code, P di un *test* ad una coda è 0.03). Essi sono applicabili solo se siamo certi che il trattamento non può associarsi ad aumento ma solo a riduzione del rischio di eventi. I *tests* a una coda, in pratica, non sono mai utilizzati, e quando questo succede vanno considerati con estrema prudenza (vedi l'esempio precedente).

Test di verifica

1) Un parametro è:
a. Una misura di tendenza centrale
b. Una caratteristica della popolazione stimata in un campione da essa estratto
c. Una misura di variabilità
d. Il risultato dell'evoluzione
e. A e b

2) Il valore di P è:
a. La probabilità che il risultato ottenuto sia scaturito solo per effetto del caso
b. Una misura di precisione della stima di un parametro
c. Generato dal *test* statistico sotto l'ipotesi nulla
d. L'errore che si commette nel rifiutare l'ipotesi nulla quando essa è vera
e. Tutte le precedenti

3) Un *test* statistico a due code:
a. Saggia i risultati ipotizzati di un *trial* di equivalenza
b. E più conservativo di un *test* ad una coda
c. E meno conservativo di uno ad una coda
d. Genera la P di ottenere per effetto del caso un risultato estremo come o anche più estremo di quello osservato nelle due direzioni possibili, ad esempio aumento e riduzione del rischio
e. B e d.

La risposta corretta alle domande sarà disponibile sul sito internet www.sin-italy.org/gin e in questo numero del giornale cartaceo dopo il Notiziario SIN

Gli intervalli di confidenza

Gli intervalli di confidenza offrono *informazioni complementari* a quelle del valore di P, entrambi utili per stimare, sulla base dei risultati osservati, il *valore vero* (per esempio il vero valore di RR nella popolazione generale).

Supponiamo che il RR di complicanze cardiovascolari associate ad un nuovo anti-ipertensivo (verso un farmaco *standard*) sia stato stimato a 0.8 (riduzione relativa del rischio del 20%), con un intervallo di confidenza al 95% tra 0.7 e 0.9 (riduzione relativa del rischio tra il 10 e il 30%) e un valore di P = 0.03. Cosa significa? Una P = 0.03 significa che, se il nuovo farmaco avesse la stessa efficacia di quello *standard*, noi dovremmo aspettarci *per il puro effetto del caso* una riduzione o un aumento di rischio relativo di almeno il 20% in soli 3 *trials* su 100 (in una serie ripetitiva di 100 identici *trials*). Tuttavia il valore di P non ci offre nessuna informazione sul *range di valori entro cui il valore vero verosimilmente giace*. Gli *intervalli di confidenza* ci danno proprio questa informazione. Vediamo come.

Supponiamo di studiare 8 soggetti (Tab. I, prima riga): 4 controlli (in trattamento convenzionale) e 4 trattati con un nuovo farmaco, e di osservare 2 eventi sfavorevoli tra i controlli e 1 tra i trattati (rischio di eventi 0.5 nei controlli e 0.25 nei trattati). Come misura di effetto usiamo ancora il rapporto tra rischi. Se il rischio non cambia per effetto del trattamento il rapporto è 1. Il rapporto dei rischi (RR) nel nostro *trial* è pari a 0.5 (1/4 diviso 2/4). Ci chiediamo: sulla base dei dati in nostro possesso possiamo sostenere che il trattamento dimezza il rischio di eventi rispetto al trattamento convenzionale? I nostri dati non sono sufficienti a dimostrarlo.

Infatti, la P di un risultato uguale o superiore a quello ottenuto è uguale a 0.46. Calcolando gli intervalli di confidenza (IC) al 95% vediamo che con 8 soggetti l'IC va da 0.07 a 3.65. Ci rendiamo conto quindi che quando pochi soggetti sono arruolati in uno studio, il ruolo del caso influenza pesantemente i risultati. Proviamo allora a calcolare gli IC in *trials* con risultati identici ma numerosità sempre maggiore (righe successive della Tab. I). Se il trattamento effettivamente comportasse un beneficio (ad esempio, RR di 0.5) arruolando 40 soggetti, oppure 80 o addirittura 2000, l'intervallo dei valori probabili si restringerebbe attorno alla stima del valore vero, per diminuzione del "peso" del caso (errore casuale). Al contrario, se non ci fosse alcun effetto (stima puntuale vicina ad 1), allora l'intervallo rimarrebbe impreciso (comprendendo il valore 1, corrispondente a differenza nulla dei rischi). In modo induttivo abbiamo costruito degli intervalli di valori verosimili attorno ad un valore plausibilmente vero. Abbiamo anche imparato che i limiti di confidenza tendono ad avvicinarsi al valore vero all'aumentare della numerosità campionaria.

Definizioni formali

Siamo pronti ad affrontare le *definizioni formali*. Il *test* di ipotesi (il *test* statistico che calcola il valore di P) e la stima intervallare (intervalli di confidenza) sono due modi per misurare la precisione della stima di un effetto osservato in un *trial* randomizzato o in uno studio con disegno più debole (non sperimentale). Il *test* di ipotesi determina la probabilità di un effetto nullo associato ad un'esposizione (probabilità dei dati sotto l'ipotesi nulla). Il *valore di P* viene presentato insieme alla stima più probabile dell'effetto (la stima puntuale, ossia il valore più probabile dell'effetto sulla base dei dati in nostro possesso). La *stima intervallare* permette di identificare un *range* di valori entro cui il vero effetto plausibilmente giace, dati i valori osservati. La probabilità di comprendere il valore dipende dal grado di fiducia con cui scommettiamo sull'intervallo, in genere (sempre per convenzione) pari al 95%. Va notato che 0.05 (1/20) è il complemento a 1 di 0.95 (19/20). Un modo approssimativo e rapido (ma non corretto) per esprimere il concetto è che l'intervallo di confidenza al 95% contiene il valore vero con una probabilità del 95%. Più correttamente, se l'effetto osservato fosse vero e noi ripetessimo lo studio molte volte, nel 95% delle ripetizioni il risultato risulterebbe compreso nell'intervallo calcolato.

Maggiore è la precisione della stima e minore l'ampiezza dell'IC. Per una stima più precisa (IC più ristretto) è necessario un maggior numero di pazienti (e quindi di eventi) nello studio.

Interpretazione degli intervalli di confidenza

Vediamo ora perché gli IC offrono informazioni interessanti per il clinico. Per agevolare la comprensione del testo rappresentiamo graficamente il concetto di precisione mettendo in relazione la P con gli IC (Fig. 4). La funzione P-IC descrive la P di diversi limiti di confidenza di RR sulla base dei dati osservati nel campione: in corrispondenza di P = 1 (ordinate) c'è solo la stima puntuale, poi al diminuire di P compaiono gli estremi di IC crescenti, da IC 0% a IC 100%, della stima del RR (ascisse). Questa relazione è molto utile perchè contrariamente a quanto alcuni erroneamente pensano non tutti i valori contenuti nell'IC sono ugualmente probabili. Ad esempio se un fattore di rischio è associato ad un RR di 2 per una malattia con IC da 1.3 a 2.7, i valori vicini agli estremi di IC (limiti di confidenza) sono meno probabili del valore 2. Il valore più verosimile è, infatti, la stima puntuale.

Valori progressivamente distanti sono meno probabili *sulla base dei nostri dati*, ma vanno riportati perché potrebbero essere veri nella popolazione da cui il nostro

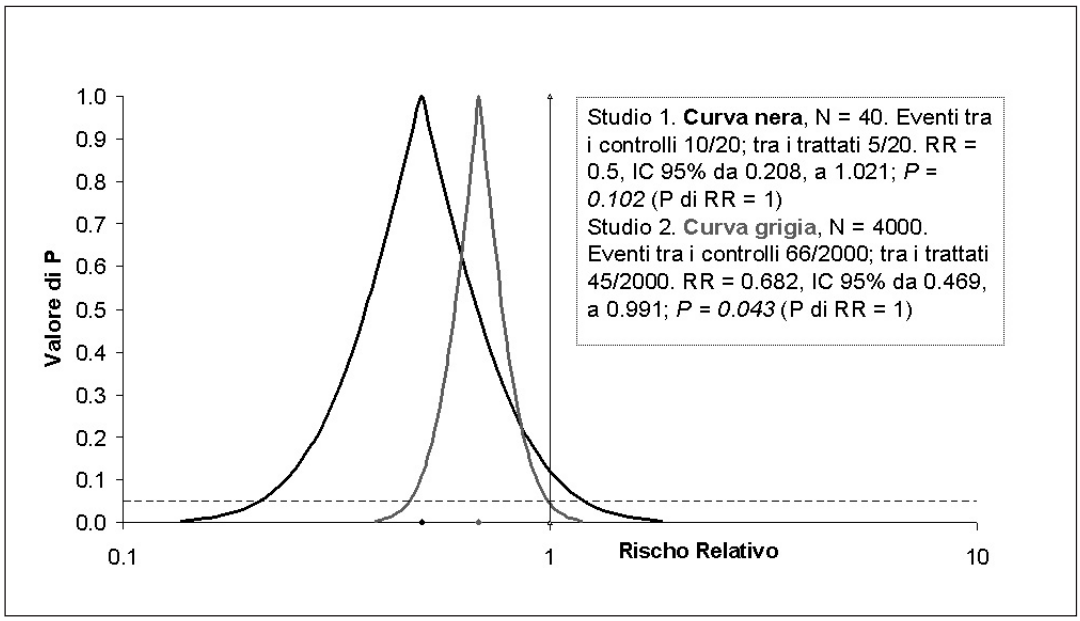

Fig. 5 - *Funzione "valore di P vs. Intervalli di Confidenza" di 2 studi di diversa dimensione e con diversa stima puntuale.*

campione proviene. Consideriamo allora due studi di diversa dimensione, uno con N = 40 e uno con N = 80.

Immaginiamo che un nuovo trattamento riduca il rischio di eventi del 50%, ossia il valore vero di RR (il parametro nella popolazione generale) sia 0.5 (o vicino ad esso). Nel grafico in Figura 4 sono riportate le funzioni di P-IC dei due studi. In corrispondenza del valore 1 delle ascisse una linea verticale indica l'effetto nullo. In corrispondenza del valore 0.05 delle ordinate una linea orizzontale indica il valore di P = 0.05.

Rappresentiamo adesso le funzioni di probabilità dei valori di RR nei due studi.

Per ciascun valore delle ordinate (P) avremo 2 valori di ascisse (RR) per ciascuno studio, corrispondenti al limite inferiore e superiore degli intervalli di confidenza (IC) al 10% (P = 0.9), al 20% (P = 0.8), al 30% (P = 0.7), ecc.

Infatti, abbiamo visto che il grado di fiducia degli IC è complemento ad 1 del valore di P. Noi siamo particolarmente interessati all'intervallo corrispondente alla P = 0.05, il rischio di errore (alfa, o di I tipo) che abbiamo deciso di correre rigettando l'ipotesi nulla qualora fosse vera: l'IC al 95%. Va notato che gli IC al 90% (corrispondenti ad una P di 0.1) sono più stretti e quindi più precisi, ma relativamente ad una maggiore P di errore di I tipo. Infatti, rispetto agli IC al 95%, più facilmente possono non comprendere l'effetto nullo e sono quindi meno conservativi.

Significato dei limiti di confidenza

Nell'esempio della Figura 4, i due studi producono la stessa stima puntuale dell'effetto del nuovo trattamento nella prevenzione di eventi sfavorevoli (RR 0.5). Infatti la stima puntuale è alla stessa distanza dal valore nullo (RR = 1). Il valore della stima puntuale, (RR 0.5) corrisponde al valore più probabile (P = 1, senza IC), ossia il più compatibile coi dati osservati. Quanto più i valori di RR si discostano (nelle due direzioni possibili) dal valore della stima puntuale tanto meno probabili diventano, come in ogni distribuzione. Nello studio più ampio (N = 80, colore grigio), gli eventi sono 20 (su 40) tra i controlli e 10 (su 40) tra i trattati. Nello studio più piccolo (curva nera) gli eventi sono 10 (su 20) tra i controlli e 5 (su 20) tra i trattati. Come discusso in precedenza, la stima è più precisa (l'incertezza minore) se N aumenta. Ma qual è il significato dei *limiti* degli IC? I limiti dell'IC per esempio al 95% (ma i concetti valgono per qualsiasi grado di fiducia) sono i valori che delimitano l'IC (estremi inferiore e superiore dell'IC) ed hanno una P pari al livello di errore di tipo I che abbiamo deciso di correre, in genere 0.05, quindi un RR pari al loro valore (di entrambi) ha una P = 0.05. Poiché rigettiamo l'ipotesi nulla solo se la P del *test* di ipotesi è < 0.05, allora (nell'esempio della Fig. 4) il valore del limite superiore dell'IC deve essere < 1 se la nostra stima puntuale è *significativamente diversa da 1*.

Vediamo per ciascuno dei due studi della Figura 4 il significato dei limiti dell'IC e il valore di RR che ha una P pari a quella del *test* dell'ipotesi nulla.

Supponiamo di avere fatto solo lo studio più piccolo. In questo studio la P della stima puntuale è 0.102 (*test* di ipotesi nulla). Cosa significa in termini di IC? I limiti dell'IC al 95% (corrispondenti a P = 0.05) includono il valore di RR = 1, ossia l'estremo superiore dell'IC al 95% (con P = 0.05), è maggiore di 1 (RR = 1.2). La P del *test* di ipotesi è la P (0.102) dell'IC che ha il limite superiore = 1. Che indicazioni ricaviamo da questo studio? Se

consideriamo solo il *test* di ipotesi nonostante il valore della stima puntuale (RR = 0.5), compatibile con un discreto beneficio, concludiamo che lo studio non rigetta l'ipotesi nulla e lo studio è negativo. Se consideriamo l'IC invece, vediamo non solo che l'intervallo è ampio, ossia la stima è imprecisa, ma anche che il valore inferiore (che ha la stessa P di quello superiore) suggerisce un effetto potenzialmente anche maggiore (un RR minore di 0.5). Infatti entrambi i valori di 0.2 e 1.2 hanno una P di 0.05. Lo studio degli IC ci permette di distinguere tra importanza della stima puntuale (stima dell'effetto) e precisione dello studio (significatività statistica) in entrambe le direzioni.

Gli IC, in questo caso, ci suggeriscono la necessità di ripetere lo studio con un maggior numero di pazienti.

Supponiamo di fare anche lo studio di dimensioni maggiori. Il *test* di ipotesi produce una P = 0.020. Gli IC al 95% hanno un valore superiore di 0.929 (i limiti con P = 0.05 sono infatti 0.269 e 0.929) nettamente inferiore ad 1. La P del *test* di ipotesi è la P che il limite superiore sia paria ad uno, una P inferiore ad 0.05 (esattamente 0.02). Possiamo dire che così come la P del *test* di ipotesi deve essere inferiore a 0.05 per poter rigettare l'ipotesi nulla (quando si è deciso di correre un rischio di errore di I tipo inferiore a 0.05), allo stesso modo il limite superiore dell'IC al 95% deve essere inferiore a 1. Un valore di P = 0.05 del *test* di ipotesi implica che un limite degli IC al 95% è pari a 1.

Grazie alla stima intervallare siamo autorizzati a ritenere che il valore vero stimato dal nostro studio sia compreso in un *range* di valori che va da 0.27 a 0.93.

Utilità di P e di IC

Gli IC permettono di distinguere l'informazione del risultato del *test* statistico di ipotesi nulla in due dimensioni: la forza della relazione e l'imprecisione della stima.

Sia il valore di P che gli IC sono utili per interpretare i risultati di uno studio, ed entrambi dovrebbero essere riportati in quanto rispondono a due diverse domande relative all'incertezza della stima. Il *test* di ipotesi calcola la P dell'ipotesi nulla mentre gli IC suggeriscono le P delle ipotesi alternative. Per esempio consideriamo un risk ratio (RR) come misura di effetto (associazione tra esposizione ed occorrenza di eventi).

La domanda a cui risponde il valore di P è: "*Quale è la probabilità, se il RR è 1 (ipotesi nulla: assenza di associazione), che uno studio come il nostro dia un risultato lontano da 1 come quello ottenuto, o anche più lontano?*" Per cui il valore di P è la probabilità condizionata dalla (sotto la) ipotesi nulla, di osservare un associazione forte come o anche maggiore di quella ottenuta. La domanda cui rispondono gli IC è: "*Quale è il range di valori di RR che includerebbe il valore vero nella predefinita proporzione di volte (di solito 0.95), se l'esperimento fosse ripetuto molte volte (idealmente all'infinito)?*" In altre parole, poiché noi stimiamo i parametri della popolazione studiando dei campioni, possiamo aspettarci che la stima campionaria non sia esattamente uguale al valore vero nella popolazione. Tuttavia, maggiore è la dimensione del campione e più vicina sarà la stima del parametro al valore del parametro stesso, e gli IC aiutano a quantificare l'incertezza di stima del parametro. Un IC al 95% della stima del parametro implica che se noi prendessimo 100 campioni di una pre-fissata numerosità e calcolassimo la stima del parametro con gli IC ogni volta, ci aspetteremmo che l'IC al 95% includerebbero il valore vero.

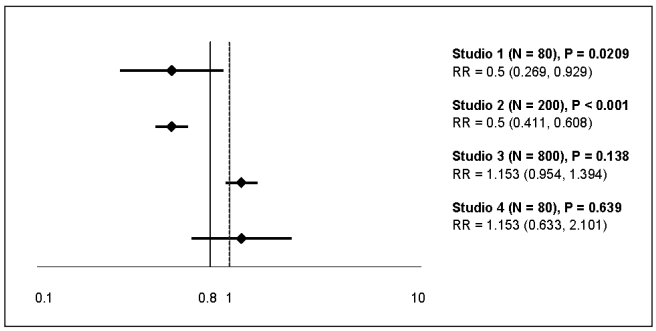

Fig. 6 - *Intervalli di confidenza al 95% di 4 studi di diversa dimensione e risultato (vedi testo).*

Entità della stima, studi positivi e definitivi

Supponiamo questa volta che il secondo studio (curva grigia, Fig. 5) non sia stato fatto nelle stesse condizioni del primo (curva nera, Fig. 2). Il finanziamento ha permesso di arruolare un numero 100 volte maggiore di pazienti. Tuttavia, in questo studio abbiamo un RR di minore entità anche se la precisione della stima è sufficiente (la P è inferiore a 0.05). Siamo soddisfatti? Sinceramente le condizioni dello studio precedente ci avevano resi più ottimisti, a meno che il nostro problema fosse solo il valore della P. Qui la P è inferiore a 0.05 (0.043), ma la curva è nettamente spostata verso l'effetto nullo. Lo studio è *positivo* in termini di significatività statistica ma con questo esempio intuiamo che anche l'impatto della misura di effetto che stiamo studiando va presa in considerazione. Dobbiamo capire meglio cosa significhi in realtà, dal punto di vista clinico, la modificazione del rischio dovuta al trattamento. Infatti uno studio positivo (in termini di significatività statistica) ma con effetto inferiore a quello riportato da uno studio negativo, potrebbe essere meno interessante dal punto di vista clinico.

Quando si dimensiona uno studio in realtà il riferimen-

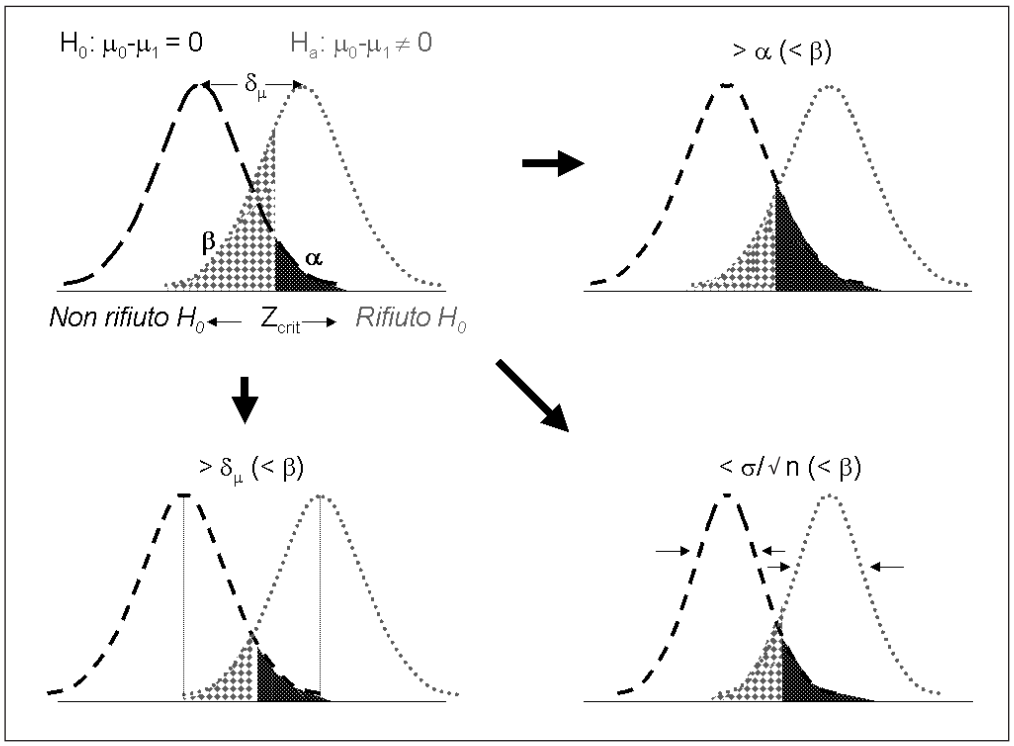

Fig. 7 - *Distribuzioni di campionamento sotto H_0 e H_a vs. i risultati di un* test *statistico ad una coda. La probabilità dell'errore di tipo II (β) è funzione dell'errore di tipo I (α), dell'entità dell'effetto (δ), e del rapporto tra variabilità e numerosità del campione (σ/√n).*

to per il calcolo della potenza è il cosiddetto "minimo effetto clinicamente rilevante" e non il RR = 1. Per esempio potrebbe essere 0.9 oppure 0.8. In base a questo effetto minimo rilevante dovremmo decidere se uno studio è anche *definitivo*, e non soltanto significativo. Per esempio se il valore di riferimento fosse 0.8 (effetto minimo clinicamente rilevante), in nessuno degli esempi precedenti avremmo ottenuto conferma degli effetti favorevoli del trattamento. Supponiamo siano condotti alcuni *trials* con ipotesi nulla di equivalenza di effetti (l'ipotesi degli studi è che un intervento sia protettivo, quindi sono *trial* di superiorità)[1]. *Trials* positivi (statisticamente significativi) con limite superiore minore di 0.8 sarebbero anche definitivi (Fig. 6). Ad esempio, un *trial* con RR = 0.50 (IC 95% da 0.27 a 0.93) dimostrerebbe un effetto significativo ma non sufficiente per essere considerato anche definitivo (studio 1), mentre un *trial* di dimensioni maggiori e con limite superiore dell'IC di 0.6 (studio 2) sarebbe anche definitivo. Se i *trials* dimostrassero un aumento anziché una riduzione del rischio associati ad un trattamento (può succedere!) bisognerebbe considerare il limite inferiore dell'IC: se nell'IC non fosse compreso il valore di 0.8 (il limite inferiore fosse > 0.8) il *trial* sarebbe definitivo (studio 3) altrimenti no (studio 4). In mancanza di risposte definitive nuovi *trials* opportunamente dimensionati sarebbero necessari.

[1] Da non confondere con i *trials* di equivalenza (o di non inferiorità), in cui l'ipotesi dello studio è che un nuovo trattamento non sia diverso (o peggiore) di un altro.

Test di verifica

1) Gli intervalli di confidenza (IC):
a. Sono informazioni complementari a quelle del valore di P
b. Sono misure di precisione della stima di un parametro
c. Suggeriscono le P delle ipotesi alternative
d. Sono il *range* di valori che includerebbe il valore vero del parametro in una pre-definita proporzione di volte se l'esperimento fosse ripetuto infinite volte
e. Tutte le precedenti

2) I valori di P e gli IC:
a. Non hanno alcuna relazione tra loro
b. Sono in rapporto tra loro e la P del *test* statistico è la P dell'effetto nullo (RR = 1)
c. Permettono di quantificare l'importanza clinica dell'effetto di un trattamento
d. Insieme aiutano a distinguere la forza dell'associazione e la precisione della stima
e. B e d

3) I limiti degli IC al 95%:
a. Sono un'alternativa al *test* di ipotesi nulla
b. Hanno entrambi una P pari a 0.05
c. Aiutano a stabilire se un risultato è positivo o negativo e se è definitivo o no
d. B e c
e. Tutte le precedenti.

La risposta corretta alle domande sarà disponibile sul sito internet www.sin-italy.org/gin e in questo numero del giornale cartaceo dopo il Notiziario SIN

Errore casuale di 1° (α) e 2° (β) tipo

Abbiamo imparato che *rigettando l'ipotesi nulla quando in realtà è vera*, si compie un errore detto di 1° tipo (errore α). Lo sperimentatore generalmente dichiara a priori il livello di errore che accetta di compiere, definendo il livello di *significatività* del *test* statistico (corrispondente alla P). Tuttavia, un risultato non-significativo ($P > 0.05$) non implica che l'ipotesi nulla (assenza di effetto) sia vera, piuttosto che i dati dello studio non forniscono l'evidenza sufficiente per confutarla. Rimane la possibilità che l'ipotesi nulla sia falsa nonostante lo studio non riesca a dimostrarlo (errore di 2° tipo, β). Poiché esiste non solo la possibilità di rigettare la ipotesi nulla quando essa è vera ma anche di *non rigettarla quando è falsa* (errore di 2° tipo β), ogni studio viene pianificato tenendo conto di questi due possibili errori. Come precedentemente anticipato, nella pianificazione viene innanzitutto definita una differenza minima di effetto (che si traduce in differenza tra stime di parametri) che rappresenta l'ipotesi di lavoro. Se lo studio non permette di rigettare l'ipotesi nulla ($P > \alpha$) quando essa è falsa, può essere commesso l'errore di tipo 2 (β). Anche il livello dell'errore di tipo 2 viene definito dal ricercatore nella pianificazione della ricerca. Di solito viene definito come complemento a 1 della potenza dello studio (dove la potenza è 1-β). Il potere dello studio e il complemento ad 1 dell'errore di tipo 1 (1-α) hanno lo stesso significato, rispettivamente, della sensibilità e della specificità di un *test* diagnostico per una diagnosi di malattia[2]. Infatti, essi rappresentano, rispettivamente, la probabilità di un risultato positivo condizionata dalla presenza di un effetto e la probabilità di un risultato negativo condizionata da assenza di effetto. Il concetto di potenza è rilevante nella pianificazione della ricerca (*potenza a priori*). Nell'interpretazione dei risultati di uno studio (*potenza a posteriori*), soprattutto quando lo studio si conclude negativamente (ossia con $P > 0.05$), può essere più intuitivo considerare gli IC.

Da cosa dipende l'errore di II tipo? La P di β (il rischio di errore di II tipo) è funzione di una serie di fattori di cui si tiene conto per calcolare la dimensione di uno studio: il valore di α, l'entità dell'effetto (δ), la numerosità del campione (n) e la variabilità dei dati (varianza, σ^2). Consideriamo le seguenti due ipotesi: la media (il valore medio di una variabile quantitativa qualsiasi come la pressione arteriosa) del campione 1 non è maggiore della media del campione 2 (ipotesi nulla); la media di 1 è maggiore di 2 (ipotesi alternativa). La Figura 7 mostra due distribuzioni di campionamento sotto l'ipotesi nulla (H_0) e alternativa (H_a) in funzione dei valori di un *test* statistico (Z) ad una coda (supponiamo per semplicità che il trattamento non possa aumentare ma solo diminuire la pressione). La posizione delle due curve dipende dalla distanza delle stime puntuali nei campioni, ovvero dall'effetto (δ), mentre la forma delle due curve dipende dalla varianza (σ^2). Come vediamo, passando dal riquadro in alto a sinistra a ciascuno degli altri, il rischio di β si modifica con le variazioni di ciascuno di questi fattori (modificati uno per volta): β è minore se α aumenta (ossia riducendo il valore critico Z), se la differenza media (δ) aumenta, se l'errore *standard* ($\sigma/\sqrt{n}$) si riduce, per esempio aumentando la numerosità del campione. Come si vede, data una certa variabilità nella popolazione, il solo modo di ridurre sia α che β è di aumentare la numerosità del campione.In pratica il livello di α e β sono scelti in base a criteri di costo-beneficio. Di solito i livelli di P α e β sono fissati ad un minimo di 0.05 e 0.2, rispettivamente, con valori maggiori più accettabili per β, per la stessa ragione per cui "assolvere un colpevole" (errore di 2° tipo β) sarebbe più accettabile che "condannare un innocente" (errore di 1° tipo α). In verità si tollera un errore β più grande solo negli studi di differenza (o superiorità), mentre si auspica un errore β più piccolo e pari all'errore α (ossia di 0.05) negli studi di equivalenza dove l'interesse del ricercatore è quello di *dimostrare e non confutare* l'assenza di differenza.

[2] La sensibilità è la proporzione dei positivi al *test* tra i malati. La specificità è il numero dei negativi al *test* tra i sani (non malati).

Test di verifica

1) L'errore casuale di I tipo:
a. E influenzato da quello sistematico
b. Non si può modificare
c. E il risultato del *test* di ipotesi alternativa
d. E l'errore che accettiamo di commettere rifiutando l'ipotesi nulla quando è vera
e. Tutte le precedenti

2) L'errore casuale di II tipo:
a. Non è in relazione con quello di primo tipo
b. E l'errore di accettare l'ipotesi nulla quando è falsa
c. E il complemento ad 1 della potenza dello studio
d. E influenzato dalle distorsioni nel disegno dello studio
e. B e c

3) La potenza dello studio:
a. E il complemento a 1 dell'errore di II tipo
b. Aumenta con l'aumentare della numerosità di N
c. Si riduce se riduciamo il rischio di errore di I tipo
d. Aumenta se l'effetto è maggiore
e. Tutte le precedenti.

La risposta corretta alle domande sarà disponibile sul sito internet www.sin-italy.org/gin e in questo numero del giornale cartaceo dopo il Notiziario SIN

Test di ipotesi vs stima degli intervalli di confidenza: un esempio in nefrologia

Come esempio della relazione tra valore di P ed IC, consideriamo i risultati del *National Cooperative Dialysis*

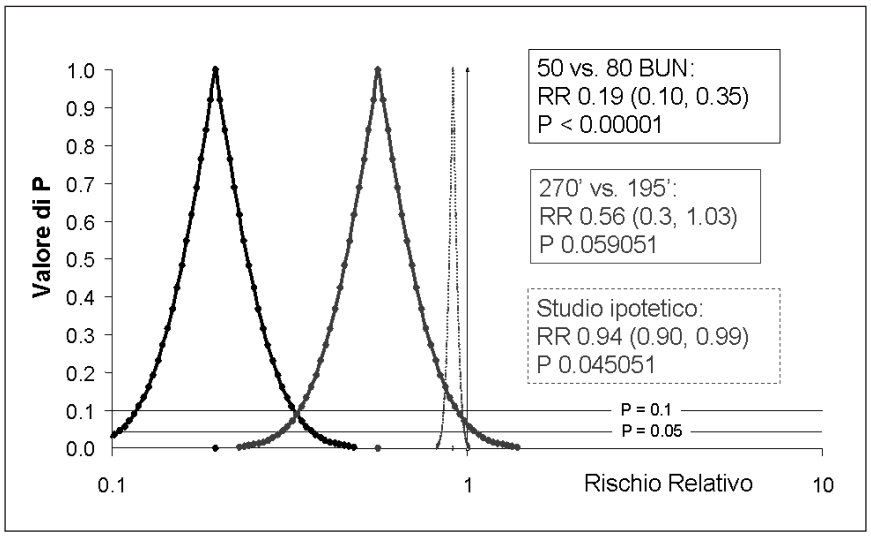

Fig. 8 - *Funzione "valore di P vs. Intervalli di Confidenza" del rapporto di rischio (RR, 0.95 IC) di ospedalizzazione associato con l'esposizione a diverso blood urea nitrogen (BUN, linea nera) e a diversa durata del trattamento (195' vs 270', linea grigia continua). I risultati di un ipotetico studio (linea grigia tratteggiata) a numerosità molto maggiore potrebbero essere statisticamente significativi nonostante un effetto clinicamente modesto.*

Study pubblicati da Lowrie et al. (6). Questo studio randomizzato della durata di sei mesi aveva lo scopo di saggiare due effetti o fattori (i livelli di BUN nel sangue e la durata della dialisi) sulla morbilità. Una delle conclusioni dello studio era che il tempo alla prima ospedalizzazione era "significativamente minore" (RR 0.19; P < 0.0001) nel gruppo assegnato a minore BUN. La seconda ipotesi nulla di uguaglianza di effetto di un tempo di dialisi maggiore non fu rigettata dato il livello non significativo della P (RR 0.56; P = 0.059). Sulla base di questi due *test* di ipotesi, gli Autori conclusero che soltanto minori livelli di BUN e non la durata del trattamento erano associati a minore morbilità.

Proviamo a tornare alle domande a cui rispondono i valori di P e gli IC rispetto alla seconda ipotesi dello studio. Innanzitutto: "Quale è la probabilità, se RR è 1 (H_0: nessun effetto della durata di dialisi sulla morbilità), che uno studio (come il NCDS) possa generare un RR di 0.56 o inferiore?" Risposta: la P di H_0 è 0.059. Inoltre: "Quale è il *range* di valori di RR che includerebbe il vero valore il 95% delle volte se la ricerca fosse ripetuta molte volte?" Risposta: l'intervallo di valori di RR che comprende il vero valore nella popolazione obiettivo con una probabilità del 95% se la ricerca fosse ripetuta molte volte, ha un limite inferiore di 0.3 e un limite superiore di 1.03. Detto in modo più semplice, esiste una probabilità del 95% che il valore vero sia compreso tra 0.3 e 1.03. La Figura 8 mostra la relazione tra P e IC dei RR. In questo grafico, la probabilità di ciascuno dei possibili valori di RR nei dati osservati viene rappresentata in funzione di RR. Per RR = 1 la curva produce il valore di mancanza di effetto (H_0).

Dove la curva raggiunge il suo massimo (P = 1), viene rappresentato il valore della stima puntuale, il valore di RR più compatibile coi dati. Allontanandosi dalla stima puntuale in entrambe le direzioni, i valori di P si riducono ad indicare minor compatibilità con i dati osservati di ciascuna ipotesi di RR. Maggiore è la distanza della stima puntuale dalla linea verticale corrispondente a RR = 1, più forte l'effetto del fattore "durata della dialisi"; più stretta è la curva e più precisa è la stima. Il primo risultato (curva nera) si riferisce al RR di (prima) ospedalizzazione in base a diversi valori di BUN: il rischio è ridotto dell'81% tra i pazienti assegnati a trattamento più efficiente.

L'associazione tra il fattore e l'*outcome* è molto forte (RR = 0.19) con significatività statistica evidente, nonostante gli IC siano ampi (*Standard Error* è relativamente elevato, denunciando un'alta variabilità nei dati, in parte spiegabile con le dimensioni del campione). Il secondo risultato (curva grigia, linea continua) si riferisce al RR di ospedalizzazione associato ad una maggior durata del trattamento (270' vs 195'): la riduzione del rischio è pari al 44%, un effetto minore del precedente, anche se clinicamente importante, ma di significatività "*borderline*" (P = 0.059).

Affidandosi alla significatività statistica soltanto, gli Autori hanno interpretato la mancanza di significatività statistica come evidenza di assenza di effetto, anche se questa interpretazione era contraddetta dalla stima puntuale (RR quasi dimezzato). Al contrario, considerando gli IC si vede che l'effetto è forte, anche se impreciso (RR da 0.3 a 1.03). L'imprecisione della stima dell'associazione tra il fattore "durata della dialisi" e la morbilità potrebbe essere imputabile a mancanza di potenza. In questo caso basare l'inferenza sul *test* di ipotesi soltanto è stato fuorviante. Il messaggio può essere anche più chiaro se consideriamo un'altra curva (curva grigia tratteggiata) che descrive un altro set di dati da un ipotetico studio più numeroso, con i seguenti risultati: RR = 0.94 (95%CI 0.90, 0.99), P = 0.045. Questa volta la funzione "valori di P-CI" è più stretta, ossia la stima è più precisa ma i dati suggeriscono un effetto meno importante. Nonostante ciò il risultato è statisticamente significativo (P = 0.045) e l'H_0 è rigettata. Anche in questo caso affidarsi alla significatività statistica soltanto condurrebbe ad interpretazione discutibile dei risultati. Pertanto l'effetto dell'aumento della durata della dialisi sulla riduzione del RR di morbilità nel NCDS è stato forte, benché non-significativo, e il risultato dello studio non

sembra conclusivo, vista l'imprecisione della stima. Al contrario nello studio ipotetico la stima dell'effetto è significativa, ma di scarso interesse, anche se precisa. E bene tener presente che piccole differenze di nessun interesse clinico possono essere significative se la numerosità del campione è sufficientemente elevata.

Sfortunatamente, la mancanza di significatività statistica è spesso interpretata come mancanza di effetto e la presenza di significatività come prova di effetto. Un *test* di ipotesi nulla che si conclude con un P valuta il ruolo del caso nel generare i risultati del campione, uno strumento utile nel processo decisionale. Tuttavia, come suggerisce l'esempio riportato di revisione critica dei dati del NCDS, la stima combinata e simultanea degli intervalli di confidenza può ridurre il rischio di giungere a conclusioni affrettate mettendo in evidenza il ruolo di possibili ipotesi alternative e può suggerire quindi ulteriori quesiti clinico-epidemiologici.

Test di verifica

1) Gli intervalli di confidenza (IC):
a. Quantificano la rilevanza del risultato
b. Sono misure di precisione della stima di un parametro
c. Suggeriscono le P dell'errore di II tipo
d. Sono un *range* di valori equamente probabili
e. Tutte le precedenti
2) Gli IC indicano che la stima è meno precisa se:
a. Il *range* di valori incluso è più stretto
b. Il *range* di valori incluso è più ampio
c. Il *range* include l'effetto nullo
d. Il *range* non include l'effetto nullo
e. B e d
3) Gli IC sono:
a. Superflui se il valore di P è riportato
b. Importanti anche se il valore di P non è riportato
c. Sempre utili da studiare insieme al valore di P
d. Indispensabili indipendentemente dal valore di P
e. Tutte le precedenti.

La risposta corretta alle domande sarà disponibile sul sito internet www.sin-italy.org/gin e in questo numero del giornale cartaceo dopo il Notiziario SIN

Conclusione

I valori di P e gli Intervalli di Confidenza (IC) sono strumenti di misura dell'imprecisione della stima campionaria di un parametro o caratteristica della popolazione. L'utilizzo contemporaneo dei valori di P e degli IC aiuta ad interpretare i risultati di uno studio clinico in relazione al ruolo del caso nella generazione dei dati. Infatti, mentre il valore di P è la probabilità dell'ipotesi nulla (nessun effetto), gli IC descrivono la probabilità delle ipotesi alternative (presenza di effetto), ossia la probabilità che il valore vero sia contenuto nei limiti stabiliti.

Riassunto

Le analisi statistiche vengono utilizzate per rappresentare in modo sintetico tutta l'informazione presente nei dati, con misure di tendenza centrale e con indici di variabilità. La statistica non può controllare distorsioni derivanti da errori di disegno e condotta dello studio. Se le differenze osservate in termini di misure di malattia tra esposti e non esposti (o trattati vs controlli) possono essere il risultato del caso con una probabilità molto bassa (in genere inferiore a 0.05) i risultati sono definiti *statisticamente significativi*. Un piccolo valore di P depone a sfavore dell'ipotesi nulla, perché i dati sarebbero improbabili se l'ipotesi nulla fosse vera. Tuttavia, il problema del *test* statistico è che il valore di P dipende dalla numerosità campionaria. Con un set di dati sufficientemente grande sarà quasi sempre possibile provare un effetto statisticamente significativo anche se piccolo. La *stima degli intervalli di confidenza* separa le due dimensioni dell'associazione testata: la forza dell'associazione e la precisione della sua stima. Pertanto, gli intervalli di confidenza aggiungono al *test* di ipotesi informazioni utili in termini di interpretazione dei risultati, inferenza causale e generazione di ulteriori ipotesi di lavoro.

Indirizzo degli Autori:
Dr. Pietro Ravani
Divisione di Nefrologia e Dialisi
Azienda Istituti Ospitalieri di Cremona
Largo Priori, 1
26100 Cremona
e-mail: p.ravani@ospedale.cremona.it

Bibliografia

1. Donaldson C, Mugford M, Vale L. eds: From effectiveness to efficiency: an introduction to evidence-based health economics. In: Evidence-based Health Economics. London: BMJ Books, 2002, pp. 1-9.
2. Sackett D, Haynes BR, Guyatt GH, Tugwell P. eds: Deciding for the best therapy. Clinical Epidemiology, a basic science for clinical medicine. Lippincott Williams & Wilkins, 1991, pp. 187-248.
3. Rothman KJ. ed: Random error and the role of statistics. In: Epidemiology: an introduction. New York: Oxford University Press, 2002, pp. 113-29.
4. Marubini E, Bossi A, Cortinovis I, Duca PG. Introduzione alla statistica medica. Nuova Italia scientifica, 1991.
5. Altman DG, Machin D, Bryant TN, Gardner MJ. eds: Confidence intervals rather than P values. In Statistics with confidence, 2nd edition. London: BMJ Books, pp. 15-27.
6. Lowrie EG, Laird NM, Parker TF, Sargent JA. Effect of the hemodialysis prescription of patient morbidity: report from the National Cooperative Dialysis Study. N Engl J Med 1981; 305(20): 1176-81.