Uso della statistica nei contributi scientifici pubblicati su *Biochimica Clinica*: un ulteriore contributo

Bruno Mario Cesana
Sezione di Statistica Medica e di Biometria, Dipartimento di Scienze Biomediche e Biotecnologie, Facoltà di Medicina e Chirurgia, Università degli Studi di Brescia

Caro Editore,

ho letto con molto interesse l'articolo di Vidali e Bellomo recentemente pubblicato su *Biochimica Clinica* (1). Innanzitutto, vorrei congratularmi con i due Autori per aver "spezzato una lancia" a favore del corretto uso della statistica nell'ambito della Medicina di Laboratorio. Tuttavia, mi sembra che il messaggio fornito, sia pure in un contesto generalmente apprezzabile, risulti a volte incompleto e a volte confuso o addirittura scorretto, al punto da limitarne in alcuni casi la validità o, addirittura, risultare controproducente. Al fine di chiarire gli aspetti più rilevanti, desidero puntualizzare quanto segue.

Gli Autori riportano, citando Campbell (2), che il grafico a barre "dovrebbe essere utilizzato unicamente per rappresentare la distribuzione di variabili discrete o categoriche… ", ma l'uso fondamentale e primario di tale grafico è per rappresentare e studiare la distribuzione di variabili "continue" (istogramma "a canne d'organo giustapposte", come è anche definito in questo contesto). Mi sembra quindi verosimile che la citazione sia riportata in modo almeno parziale, in quanto è scontato che Campbell conosca e correttamente riporti i metodi di rappresentazione dei dati. Inoltre, il grafico a destra nella Figura 1 è indicato come "box-plot", ma più appropriatamente dovrebbe essere indicato come istogramma "a scatola e baffi" ("box and whiskers plot"), dove poi i "baffi" non terminano "al massimo" in corrispondenza di 1,5 volte l'intervallo interquartile (IQR) (sommato al terzo quartile – Q_3 – o sottratto al primo quartile – Q_1), come gli Autori un po' genericamente riportano, ma (usualmente) in corrispondenza dei cosiddetti "valori adiacenti": i valori osservati immediatamente maggiore e minore di 1,5 volte l'IQR. Quindi, i lettori non dovrebbero rimanere sorpresi se possono imbattersi in eccezioni della suddetta regola. Inoltre, un cenno sull'interpretazione di tale grafico sarebbe stato auspicabile: in caso di una distribuzione gaussiana ci si aspetta che circa il 99,3% dei dati sia entro ±1,5 volte l'IQR; altrimenti, la rilevanza dei "baffi" in questo grafico risulta poco comprensibile, a meno che ci si accontenti dell'aneddotica spiegazione che John Wilder Tukey (l'ideatore di tale rappresentazione) diede a chi gli chiedeva il perché del moltiplicare per 1,5 l'IQR: "perché una volta è poco e due volte sono troppe".

Ancora, in questo paragrafo si legge una non accettabile interpretazione della variabilità campionaria nei termini della distribuzione di riferimento. Il fatto che in un campione si sia ottenuta una misura della variabilità tale che sottraendo due volte la DS dalla media si ottenga come risultato un valore negativo, non permettere di concludere che "è abbastanza ovvio che la distribuzione delle osservazioni è asimmetrica e differente da quella gaussiana". La distribuzione delle osservazioni campionarie può presentarsi più o meno simmetrica (è praticamente impossibile che sia assolutamente simmetrica), ma si deve considerare quanto sia "verosimile" che un tale campione sia stato estratto da una popolazione gaussiana (simmetrica, per definizione), per cui dobbiamo attivare le apposite procedure inferenziali, adeguatamente applicabili al campione a cui gli Autori fanno riferimento. Infatti, riportare semplicemente che la "distribuzione delle osservazioni è differente da quella gaussiana" costituisce un'affermazione che non ha alcun significato statistico: la distribuzione delle osservazioni (ovviamente campionarie) è solo l'espressione di uno degli infiniti campioni di quella numerosità che possono essere (casualmente) estratti dalla popolazione di riferimento e quindi la distribuzione campionaria non è "gaussiana". Tra l'altro, è anche opportuno sottolineare come in simili ragionamenti debba essere inserita la dimensione campionaria per cui un campione con poche unità sperimentali non ci permetterà di trarre conclusioni plausibili sulla popolazione da cui è stato ottenuto, anche in presenza di valori negativi come risultato dal sottrarre due DS dalla media aritmetica. Un importante aspetto "didattico" è quello di sottolineare i limiti del ragionamento probabilistico alla base della statistica con, quindi, la dovuta cautela che deve essere presente nelle conclusioni derivate.

Difficile da commentare è la conclusione a pagina 603 alla fine del paragrafo sul fatto che l'intervallo di confiden-

Corrispondenza a: Bruno Mario Cesana, Sezione di Statistica Medica e di Biometria, Dipartimento di Scienze Biomediche e Biotecnologie, Facoltà di Medicina e Chirurgia, Università degli Studi, Viale Europa 11, 20153 Brescia. Tel./Fax 0303717530, E-mail cesana@med.unibs.it

Ricevuto: 19.01.2011 Revisionato: 21.01.2011 Accettato: 24.01.2011

za "permette di valutare se quella differenza è, oltre che statisticamente, anche clinicamente significativa". A parte l'uso di termini non proprio appropriato (meglio "significatività statistica" e "rilevanza clinica"), la suddetta interpretazione dell'intervallo di confidenza non è accettabile; l'intervallo di confidenza permette di avere una misura della "precisione" della stima intervallare del corrispondente parametro della popolazione; tale precisione è data dall'ampiezza dell'intervallo, avendo prefissato un determinato livello di confidenza (usualmente al 95%). Che poi il risultato del test di significatività statistica corrisponda ad un effetto "clinicamente rilevante" o meno, il ricercatore deve desumerlo dalla differenza che è stata trovata "statisticamente significativa" nel contesto di uno studio che deve essere basato, per essere considerato scientifico ed etico (requisito che non si ripeterà mai a sufficienza), su un appropriato calcolo della dimensione campionaria per ottenere un'adeguata potenza del test statistico. Il giudizio sulla rilevanza clinica di un risultato sarebbe, tuttavia, meglio formulato considerando la "dimensione dell'effetto" ("effect size"), costituito dal rapporto tra la differenza delle medie delle due popolazioni (ad es., trattamento sperimentale e standard) e la variabilità del fenomeno indagato. Si ricorda che sull'"effect size" si impernia il calcolo della dimensione campionaria per i vari modelli della ricerca biomedica.

Nel paragrafo "Correlazione e regressione", gli Autori discutono non correttamente il coefficiente di determinazione (r^2) nel contesto dell'analisi della correlazione. Infatti, tale coefficiente scaturisce dal contesto dell'analisi della regressione come rapporto tra la varianza spiegata dalla retta di regressione e la varianza totale e quindi come un criterio della bontà del modello lineare che, nel caso della regressione semplice (una sola variabile indipendente), è assolutamente vero che corrisponde al coefficiente di correlazione elevato al quadrato. Tra l'altro, appare abbastanza illogico, dopo aver correttamente sottolineato che le due analisi (della correlazione e della regressione) "significano cose differenti" e che il coefficiente di correlazione (r) non cambia scambiando le variabili, concludere che "una volta ottenuto r, possiamo calcolare anche il coefficiente di determinazione (r^2), che esprime la (quota percentuale della) variabilità della variabile dipendente spiegata dalla variabile indipendente".

Anche se l'affermazione che "l'uso della correlazione non è giustificato in situazioni di un insieme di dati non lineare" è sostanzialmente condivisibile, vorrei, tuttavia, sottolineare l'opportunità di non ridurre la problematica della gestione degli "outlier" in una sintetica affermazione che in loro presenza "l'uso della correlazione non è giustificato". Parimenti emblematico è anche l'altro concetto espresso in cui "l'uso della correlazione non è giustificato" in presenza di "cluster" di punti separati gli uni dagli altri; a parte il fatto tautologico che per definizione un "cluster" è un insieme di punti separati dagli altri, non si capisce come dovrebbe essere effettuata l'analisi dell'associazione tra due variabili in queste circostanze. Forse, gli Autori volevano affermare che un'associazione lineare statisticamente significativa può essere trovata anche in presenza di due "cluster" di dati molto distanti tra loro, purché si abbiano valori piccoli per entrambe le variabili in uno e, viceversa, valori elevati per entrambe le variabili nell'altro; anzi, è anche possibile che un risultato statisticamente significativo si abbia in presenza di un "cluster" (più o meno numeroso di punti in corrispondenza di valori bassi o alti) e di un solo dato distante con valori elevati o bassi per entrambe le variabili. Chiaramente, la presenza di tanti "cluster", disposti in modo lineare in corrispondenza di determinati valori delle variabili, non sembra costituire una condizione che non "giustifichi" l'uso della correlazione.

Relativamente al confronto tra due metodi, gli Autori non citano l'approccio di Bland e Altman (3, 4). Deve poi essere chiarito che le stime ottenute col metodo dei minimi quadrati per il modello della regressione semplice lineare sono assolutamente corrette sia per il caso della "regressione classica" (la variabile indipendente x è priva di errore casuale e quindi non è una "variabile casuale") e sia per il caso in cui la variabile indipendente è una variabile casuale (indipendentemente dal fatto che il suo errore casuale sia piccolo rispetto a quello della variabile dipendente e non "rispetto all'intervallo dei valori esaminati", come gli Autori riportano). Si hanno problemi invece quando la variabile indipendente, oltre ad essere una variabile casuale, presenta un'importante ulteriore fonte di variabilità costituita dall'errore di misura (analogamente a quanto ci si aspetta che accada alla variabile dipendente). In questo caso, le usuali stime dei minimi quadrati sono distorte e, in considerazione del problema da trattare, si deve ricorrere ad altri modelli. Ad esempio, come correttamente affermato dagli Autori, si potrebbe minimizzare la "distanza geometrica" dei punti dalla retta e non la distanza verticale del punto dalla retta di regressione come per il caso del metodo dei minimi quadrati. Deve poi essere spiegato che la regressione di Deming, tanto celebrata in ambito laboratoristico, non è nient'altro che uno dei tanti metodi (e, quasi certamente, neppure il più raccomandabile) per stimare l'intercetta e la pendenza della retta di regressione del modello strutturale quando entrambe le variabili siano casuali e misurate con errore. Infatti, in tal caso si pone che la variabilità dell'errore di misura del metodo nuovo (variabile dipendente) sia uguale a quella del metodo standard di riferimento (variabile indipendente) o, in modo alternativo, che il loro rapporto sia pari a 1. Si capisce, pertanto, quanto affermato in precedenza sulla scarsa validità di questo approccio che è, invece, a volte utilizzato in modo acritico (senza essere coscienti del suddetto assunto di uguaglianza dell'errore di misura dei due metodi). A tale proposito si puntualizza che sono proponibili altri metodi, tra cui quello di utilizzare una stima dell'errore di misura del metodo di riferimento o, meglio, di utilizzare delle repliche in modo che siano identificati tutti i parametri del modello con errore di misura per entrambe le variabili casuali (5). Diverso è il discorso dell'eteroscedasticità (la variabilità del metodo in ordinata non è costante) per tutto l'intervallo di valori esaminato, per cui altri test devono essere applicati. Gli Autori non chiariscono che in tale caso è opportuno ricorrere al metodo della regressione "pesata" con i pesi costituiti dal reciproco della variabilità (6) oppure, stante la complessità di questa procedura, alla regressione non parametrica di Passing e Bablok (7).

Come ulteriore aspetto, si trova frequentemente citata nell'articolo la necessità di verificare che gli assunti su cui si basano i vari test statistici siano soddisfatti; a tal proposito si suggerisce di ricorrere ad approcci formali (test statistici a loro volta come, ad esempio, il test di Shapiro-Wilk per l'assunto di gaussianità). Questa raccomandazione non è condivisibile in quanto tale modo di procedere porta ad effettuare dei test statistici "condizionati" al risultato di un altro test con conseguenti ripercussioni sia sull'errore di I° tipo (probabilità di rifiutare – non accettare – una "vera ipotesi nulla") che sull'errore di II° tipo (probabilità di accettare – non rifiutare – una "falsa ipotesi nulla"); inoltre, spesso il comportamento di tali test non è "ottimale", per cui è senz'altro più consigliabile affidarsi al fatto che i metodi statistici sono generalmente "robusti" rispetto alla violazione degli assunti su cui si basano, specialmente in caso di numerosità campionarie abbastanza consistenti.

Un altro punto che non mi sento di condividere con gli Autori è quello di affidarsi genericamente a "testi di statistica, anche gratuiti, stampati o consultabili "on-line""; infatti, anche se è possibile trovare dei testi accettabili, il più delle volte ci si imbatte in produzioni assolutamente spontaneistiche, non adeguatamente revisionate e quindi non affidabili. Anche tutte le referenze citate in bibliografia (dal 39 al 52) fanno riferimento a "note statistiche" pubblicate sul British Medical Journal che sono delle valide e corrette puntualizzazioni, ma che certamente non possono costituire un testo di riferimento. Invece, assolutamente condivisibile e senz'altro da realizzare è la proposta di creare un gruppo di lavoro che fornisca indicazioni sui metodi statistici da usare e che stabilisca anche i criteri per cui un potenziale contributo deve essere opportunamente revisionato o rifiutato in base all'analisi statistica impiegata.

BIBLIOGRAFIA

1. Vidali M, Bellomo G. Uso della statistica nei contributi scientifici pubblicati su *Biochimica Clinica*: problemi e proposte. Biochim Clin 2010;34:600-6.
2. Campbell MJ. How to present numerical results. In: Reece D, ed. How to do it. London: BMJ Publishing Group, 1995.
3. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986;327:307-10.
4. Altman DG, Bland JM. Measurement in medicine: the analysis of method comparison studies. The Statistician 1983;32:302-17.
5. Carstensen B. Comparing clinical measurement methods. Chichester: John Wiley & Sons, 2010.
6. Draper NR, Smith H, Applied regression analysis. 3rd ed., Toronto: John Wiley & Sons, 1998.
7. Passing H, Bablok W. A new biometrical procedure for testing the equality of measurements from two different analytical methods, Part I. J Clin Chem Clin Biochem 1983;21:709-20.