Come interpretare correttamente il valore di P?

Luisa Zanolla, Maria Stella Graziani
Azienda Ospedaliera Universitaria Integrata di Verona

ABSTRACT

Understanding the P value. A number of recent contributions about the P value in statistics gave the spur to consider different aspects of its use in scientific papers. There is indeed a need of correct information on how and when to apply the P value to evaluate results of a scientific experiment and how to appropriately interpret the numerical value as well, considering that the P statistic is rather frequently reported in the medical literature. In this paper we first described the origin of the P value and its correct significance, using examples where the statistical significance of the P value does not necessarily mean a clinical significance. Then, we defined which kind of assumptions cannot be drawn from the a P value. The most common misuse is when a non-statistically significant P value is used to establish that two laboratory techniques are equivalent. Finally, some alternatives are given, the most common being the use of the confidence intervals. In spite of the interpretation mistakes commonly observed in scientific articles, the P value remains a useful statistic tool to be utilized and interpreted with better consciousness.

INTRODUZIONE

Recentemente è comparso su *Nature* un articolo di Regina Nuzzo, professore associato di statistica alla Gallaudet University di Washington e divulgatore scientifico "free lance", che mette in discussione l'utilizzo diffuso dei valori di P, l'indicatore comunemente usato nella definizione di significatività statistica (1). La posizione di Nuzzo è stata commentata da un editoriale della stessa rivista (2) e da un altro editoriale e un articolo apparsi su una delle più importanti riviste di laboratorio (3, 4). Anche una rivista scientifica di ambito non medico ha dedicato di recente un intero forum all'argomento (5).

La critica all'uso dei valori di P come discriminanti tra il "Santo Graal" della significatività statistica e l'opposta condanna della non significatività rimane quindi attuale, anche se la discussione ha inizio più di vent'anni orsono (6). Nonostante ciò, la frase canonica "statisticamente significativo (P <0,05)" compare ancora con grande frequenza nella letteratura scientifica.

Ma da dove nasce il valore di P? Cos'è esattamente e cosa significa? Cosa invece non è? Viene correttamente usato e interpretato? Esistono alternative al suo utilizzo? Questo contributo prova a rispondere a questi quesiti.

COS'È IL VALORE DI P

Il valore P fu proposto da sir Ronald Fisher negli anni '20 per misurare la forza di un risultato scientifico. Fisher lo propose come un indicatore della discrepanza tra i dati e l'ipotesi nulla (7).

Quando si formula un'ipotesi scientifica che asserisce, ad esempio, che il valore di colesterolemia è più basso nei soggetti trattati con una statina che nei soggetti trattati con placebo (per cui la differenza tra i due valori di colesterolemia non è nulla), la corrispondente ipotesi nulla è che il valore di colesterolemia sia uguale nei trattati con statina e nei trattati con placebo (cioè che la differenza sia nulla). Sui dati sperimentali viene calcolata una statistica test, che viene rapportata all'appropriata distribuzione statistica, e corrisponde a un valore di probabilità che è appunto denominato P (8). Il valore P è definito come la probabilità, sotto l'ipotesi nulla (cioè assumendo che l'ipotesi nulla sia vera), di ottenere un risultato uguale o più estremo di quello osservato.

Supponiamo di confrontare due gruppi di 40 pazienti ciascuno, assegnati per randomizzazione a trattamento con statina o placebo; alla fine di un anno di trattamento, la colesterolemia media del primo gruppo è 194,4 (DS 15,6) mg/dL, mentre quella del gruppo di controllo è 202,1 (DS 14,8) mg/dL (Tabella 1, esempio 1). Per il confronto di due medie calcolate su due campioni indipendenti, la statistica test appropriata è la t di

Corrispondenza a: Maria Stella Graziani, Vicolo S. Giovanni in Foro 5, 37121 Verona. E-mail mariastella@graziani.eu

Ricevuto: 19.04.2015 Revisionato: 24.05.2015 Accettato: 26.05.2015

Tabella 1.
Esempi di dati ottenuti all'interno di un ipotetico studio teso a confrontare i valori di colesterolemia in due gruppi di soggetti, trattati con statine o con placebo, dopo un anno di trattamento

		Gruppo 1 (statina)	Gruppo 2 (placebo)
Esempio 1	Media, mg/dL	194,4	202,1
	DS, mg/dL	15,6	14,8
	Numerosità	40	40
	Statistica test (t)	2,265	
	P	0,026	
	Differenza tra le medie, mg/dL	7,7	
Esempio 2	Media, mg/dL	194,4	202,1
	DS, mg/dL	15,6	14,8
	Numerosità	120	120
	Statistica test (t)	3,923	
	P	0,0001	
	Differenza tra le medie, mg/dL	7,7	
Esempio 3	Media, mg/dL	197,07	201,14
	DS, mg/dL	14,7	14,7
	Numerosità	100	100
	Statistica test (t)	1,963	
	P	0,051	
	Differenza tra le medie, mg/dL	4,07	
	Intervallo di confidenza, mg/dL	-0,02 ÷ 8,16	
Esempio 4	Media, mg/dL	197,30	201,14
	DS, mg/dL	14,70	14,65
	Numerosità	100	100
	Statistica test (t)	1,983	
	P	0,049	
	Differenza tra le medie, mg/dL	4,11	
	Intervallo di confidenza, mg/dL	0,023 ÷ 8,20	
Esempio 5	Media, mg/dL	197,14	201,14
	DS, mg/dL	14,65	14,60
	Numerosità	300	300
	Statistica test (t)	3,349	
	P	0,0009	
	Differenza tra le medie, mg/dL	4,0	
Esempio 6	Media, mg/dL	196,81	205,38
	DS, mg/dL	13,28	11,62
	Numerosità	50	50
	Statistica test (t)	3,434	
	P	0,0009	
	Differenza tra le medie, mg/dL	8,57	
Esempio 7	Media, mg/dL	197,07	201,14
	DS, mg/dL	14,7	14,7
	Numerosità	200	200
	Statistica test (t)	2,783	
	P	0,006	
	Differenza tra le medie, mg/dL	4,07	
	Intervallo di confidenza, mg/dL	1,195 ÷ 6,945	

Student. Dal calcolo emerge un valore di t di 2,265. Confrontare questo valore con la distribuzione t di Student è semplice con gli strumenti di calcolo attuali; ciò può essere fatto anche senza ricorrere a un programma statistico, con un foglio di calcolo. Il valore di P che risulta è 0,026 (<0,05). Qualsiasi pubblicazione scientifica riporterebbe questo risultato come "statisticamente significativo". Ma qual è il significato di quel valore di P? Sotto l'ipotesi nulla, cioè se non esistono differenze tra soggetti trattati e non trattati, la probabilità di un risultato come quello ottenuto o più estremo è pari a 26 su 1000, cioè piuttosto improbabile. Il risultato della verifica statistica ci induce quindi a concludere che è poco probabile che la differenza riscontrata tra i due gruppi sia dovuta al caso. Questo significa che si tratta anche di un risultato clinicamente rilevante? Non è la dimensione di P = 0,026 che risponde a questa domanda. Per questo, occorre fissare l'attenzione sulla differenza tra le due medie: in altre parole, una differenza di 7,7 mg/dL è clinicamente rilevante? La risposta va chiesta all'esperto dello specifico settore clinico, non alla statistica. Molti di noi probabilmente riterrebbero che una differenza di 7,7 mg/dL di colesterolemia sia clinicamente poco rilevante.

Occorre poi tenere presente che il valore di P è fortemente condizionato dalla numerosità del campione. Immaginiamo di aver ottenuto da uno studio i valori di colesterolemia precedentemente riportati, solo che la numerosità di ciascun gruppo era tripla, pari a 120 pazienti per campione. Il valore della statistica test t che si ottiene diviene 3,923, che corrisponde a una P = 0,0001 (Tabella 1, esempio 2). La probabilità del risultato, se l'ipotesi nulla è vera, diviene quindi notevolmente più bassa, ma la dimensione dell'effetto, che rappresenta l'informazione rilevante per il clinico, e cioè una differenza di 7,7 mg/dL, è sempre la stessa. Il valore di P molto più piccolo di quello precedentemente ottenuto nell'esempio 1 significa che la differenza fra i valori di colesterolemia è più significativa?

Una domanda ulteriore poi è perché consideriamo un valore di 0,05 quale soglia della "significatività statistica"? Fu Fisher che suggerì come convenzione il 5% come livello standard, con 1% come alternativa più stringente. Più piccolo è il valore di *P*, minore la probabilità di ottenere un valore più estremo di quello osservato se l'ipotesi nulla è vera. Ne deriva che più piccolo è il valore di P, più forte l'evidenza contro l'ipotesi nulla, che viene dichiarata non plausibile, cioè improbabile che si sia realizzata per effetto del caso, e viene quindi rifiutata. In caso di un P elevato (>0,05), l'ipotesi nulla non viene rifiutata, ma l'informazione che si ricava è che non vi sono dati sperimentali sufficienti per una conclusione.

Un decennio più tardi, Jerzy Neyman, un matematico polacco, ed Egon Pearson, il figlio del più famoso statistico Karl Pearson svilupparono il metodo che da loro prende il nome per scegliere tra due ipotesi. Lo scopo era quello di sostituire l'interpretazione in parte soggettiva della significatività statistica con un processo decisionale applicato ai risultati dell'esperimento, che

consentisse di sostenere che l'esperimento conferma o smentisce l'ipotesi sperimentale (9). Nel confronto tra l'ipotesi nulla, come precedentemente definita, e l'ipotesi alternativa, e precisamente che la differenza tra la colesterolemia nei trattati con statine e nei trattati con placebo non sia nulla, il risultato del test d'ipotesi (secondo Neyman e Pearson) è una decisione, non un'inferenza: rifiutare un'ipotesi e accettare l'altra, solo sulla base dei dati, e concludere che non si può asserire che esista differenza tra la colesterolemia del gruppo dei trattati con statine e quella dei trattati con placebo (mancato rifiuto dell'ipotesi nulla) o asserire che la differenza esiste (rifiuto dell'ipotesi nulla). Questa decisione espone il ricercatore al rischio di due tipi di errore (7). Comportarsi come se vi fosse una differenza tra le due terapie, quando di fatto non vi sono differenze, è un risultato falsamente positivo, o errore di tipo I o errore α (rifiutare l'ipotesi nulla quando essa sia vera). Concludere invece che le due terapie non differiscono, quando di fatto sono diverse, è un risultato falsamente negativo, o errore di tipo II o errore β (non rifiutare l'ipotesi nulla quando l'ipotesi alternativa è vera). Più usato è il complemento a 1 dell'errore di tipo II (1-β), definito potenza.

Il test di ipotesi non era inteso fornire una misura dell'evidenza dei risultati, un numero che riconducesse dai dati all'ipotesi sottostante. Il risultato è semplicemente l'accettare o il rifiutare l'ipotesi nulla, a un livello di probabilità prefissato, senza alcun tentativo di produrre un valore di P per stimare la forza dell'evidenza contro l'ipotesi nulla in un singolo studio. Nelle parole degli Autori (9): *"nessun test basato su una teoria probabilistica è in grado da solo di fornire una fondata evidenza sulla verità o falsità di una data ipotesi"* e *"senza sperare di poter conoscere se ogni singola ipotesi sia vera o falsa, possiamo cercare regole per indirizzare il nostro comportamento relativamente a esse e potremo essere confidenti sul fatto che, seguendole, nel lungo termine, non sbaglieremo molto spesso"*.

Con queste affermazioni Neyman e Pearson dichiarano che non si è in grado di valutare lo stato di realtà basandosi su un singolo esperimento. Una esemplificazione efficace di questa logica è fornita da Goodman (10), che lo paragona a un processo in un sistema giudiziario che non centra la sua attenzione su quale accusato sia giudicato colpevole o innocente (*"se ogni singola ipotesi sia vera o falsa"*), ma cerca invece di controllare il numero totale di verdetti errati (*"nel lungo termine, non sbaglieremo molto spesso"*); questa è ovviamente per il ricercatore una conclusione altamente insoddisfacente.

E' a questo punto che nello sviluppo della pratica statistica si è inserito l'uso del valore di P, come misura dell'evidenza derivante dal singolo esperimento, che intenderebbe non contraddire la logica di lungo termine del test d'ipotesi di Neyman e Pearson. Si è quindi generato una sorta di metodo combinato ibrido in cui:

1. prima dell'esperimento, si fissa il livello di errore di tipo I (praticamente sempre il 5%);
2. sempre prima dell'esperimento, si fissa la potenza (che è il complemento a 1 dell'errore di tipo II), praticamente sempre fissata all'80%;
3. si calcola quindi il valore *P* e si rifiuta l'ipotesi nulla se tale valore è inferiore al valore prefissato di errore di tipo I (punto 1).

Questo metodo combinato viene spesso presentato nei testi di statistica senza alcuna menzione della sua origine controversa. Importante sottolineare che il test di significatività può portare a rifiutare l'ipotesi nulla, ma non può mai dimostrarla o confermarla. Ciò implica la scorrettezza metodologica dei frequenti studi che, basandosi sull'assenza di una significatività statistica (cioè sul mancato rifiuto dell'ipotesi nulla), asseriscono l'equivalenza tra due (o più) gruppi studiati. Un esempio è rappresentato da uno studio recente che confronta 3 tecniche per l'elettroforesi delle proteine (11). Gli autori in particolare confrontano la specificità dei 3 metodi e, poiché non ottengono una significatività statistica (P >0,05), concludono che le caratteristiche analitiche sono equivalenti. In realtà, il mancato rifiuto dell'ipotesi nulla che deriva da questo risultato permette di affermare che i dati a disposizione non consentono di rifiutare l'ipotesi nulla (di equivalenza tra i 3 metodi), il che è diverso dal concludere che l'ipotesi nulla è vera e quindi che non ci sono differenze tra i metodi o, come riportato, che i 3 metodi possono essere considerati equivalenti. Si tratta di un uso radicalmente scorretto nell'interpretazione del valore di P, purtroppo piuttosto diffuso.

COSA NON È IL VALORE DI P?

Il quesito che lo sperimentatore solitamente si pone è del tipo "Dati questi risultati sperimentali, qual è la probabilità che l'ipotesi nulla sia vera?". Un valore di *P* pari a 0,026, come nell'esempio 1, a rigore ci dice che, se l'ipotesi nulla è vera, una differenza come quella che abbiamo osservato (o più estrema) si avrà con una probabilità del 2,6%. Il problema è che ci sembra abbastanza naturale invertire i termini della questione e concludere erroneamente che la probabilità dell'ipotesi nulla è 2,6%, dati i valori sperimentali osservati.

Fisher stesso aveva sottolineato come dal punto di vista della logica induttiva si cerchi di passare dal particolare (tipicamente un insieme di osservazioni sperimentali) a una regola generale (tipicamente una teoria applicabile all'esperienza futura) (12). Ma il processo inferenziale della classica teoria delle probabilità è deduttivo per sua natura, essendo costituito da affermazioni sul comportamento di un campione, estratto da una popolazione di cui sono note le proprietà.

Il valore P è definito come la probabilità di ottenere un risultato uguale o più estremo di quello osservato; non può quindi costituire una misura diretta della probabilità che l'ipotesi nulla sia falsa. Risulta quindi erronea la definizione, sostenuta da molti ricercatori, che un valore P di 0,05 significhi che l'ipotesi nulla ha una probabilità solo del 5% o inferiore al 5%.

Perché questa soglia di 0,05? Nel 1914, Karl Pearson aveva pubblicato le "Tables for statisticians and

biometricians" (13): per ciascuna distribuzione, Pearson riportava il valore di P per un'ampia serie di valori della variabile casuale. Quando invece nel 1925 Fisher pubblicò il libro "Statistical methods for research workers" (7), incluse tavole che presentavano i valori della variabile casuale solo per valori selezionati di P (0,05, 0,01, 0,001). Lo stesso approccio fu usato da questo Autore per la pubblicazione delle "Statistical tables for biological, agricultural, and medical research", insieme a Frank Yates nel 1938 (14). L'impatto fu notevole e negli anni '60 si diffuse come pratica standard l'indicare con un asterisco valori di P <0,05 e con due asterischi i valori di P <0,01. Occasionalmente, 3 asterischi erano usati per indicare P <0,001. Ancora oggi la maggior parte dei libri di statistica riproduce le tavole di Fisher.

L'artificiale dicotomia (o policotomia) che in questo modo veniva a crearsi risultò molto popolare, soprattutto per gli enti regolatori e i revisori delle riviste scientifiche alla ricerca di un criterio oggettivo per definire la significatività.

Fu sempre Fisher a proporre il termine "significativo" per valori piccoli di P: "*l'Autore preferisce fissare un limite inferiore di significatività al 5%... Un fatto scientifico dovrebbe essere considerato scientificamente dimostrato solo se un esperimento appropriatamente disegnato raramente manca di raggiungere tale livello di significatività".* Ma dal punto di vista scientifico è difficile pensare che i risultati di uno studio siano da interpretare in modo diametralmente opposto se il valore di P ottenuto è 0,055 invece di 0,045. Valori simili di P portano a conclusioni simili, non diametralmente opposte. Nel confronto tra due gruppi di 100 pazienti randomizzati a trattamento con statina o placebo, il colesterolo plasmatico dopo un anno di trattamento sia 197,07 (DS 14,7) mg/dL nel primo gruppo e 201,14 (DS 14,7) nel gruppo di controllo (Tabella 1, esempio 3); la statistica test è pari a 1,963, che corrisponde a un valore di P = 0,051. Il valore è più alto della soglia di 0,05, per cui non possiamo rifiutare l'ipotesi nulla e non possiamo asserire che vi sia differenza tra i due gruppi. La dimensione dell'effetto, cioè la differenza nella colesterolemia, è 4,07 mg/dL. Decidiamo di ripetere l'esperimento e nel gruppo trattato con statina otteniamo una colesterolemia di 197,30 (DS 14,7) mg/dL, mentre nel gruppo di controllo 201,14 (DS 14,65); la dimensione dell'effetto è 4,11 mg/dL, molto simile alla precedente (Tabella 1, esempio 4). Questa volta, però, la statistica test è pari a 1,983, cui corrisponde un valore di P di 0,049, che ci consente di rifiutare l'ipotesi nulla. Siamo realmente convinti che i risultati dei due studi portino a conclusioni opposte?

L'esempio focalizza l'attenzione sul fatto che un valore di P non dovrebbe mai essere riportato senza inserire anche la dimensione dell'effetto osservato. Tra le critiche riferite all'uso del valore di P, la più rilevante è probabilmente quella che esso non tiene conto della dimensione dell'effetto osservato: un effetto di piccola dimensione in uno studio con un'ampia dimensione campionaria può risultare nel medesimo valore di P di un effetto importante in uno studio di piccole dimensioni. Consideriamo due studi che confrontano pazienti trattati con statina con un gruppo di controllo trattato con placebo; entrambi gli studi riportano risultati statisticamente significativi e per entrambi la P è pari a 0,0009. Nel primo studio, i soggetti trattati con statina hanno una colesterolemia di 197,14 (DS 14,65), mentre il gruppo in placebo ha 201,14 (DS 14,60) mg/dL; la statistica test risulta 3,349 (Tabella 1, esempio 5). Nel secondo studio, il gruppo in statina ha colesterolemia 196,81 (DS 13,28), mentre quello in placebo 205,38 (DS 11,62) mg/dL; la statistica test risulta 3,434 (Tabella 1, esempio 6). La differenza nella statistica test, a parità di valore di P, è legata al diverso numero di gradi di libertà, che dipende sostanzialmente dalla numerosità delle unità studiate. Il primo studio, infatti, arruolava 600 soggetti, mentre il secondo ne arruolava solo 100. Se andiamo, tuttavia, a esaminare la dimensione dell'effetto, cioè la differenza nei livelli di colesterolo, questa è 4,0 mg/dL nel primo studio, mentre è 8,57 mg/dL, quindi più che doppia, nel secondo studio. L'enfasi sulla significatività statistica può portare il lettore non attento a ritenere sovrapponibili i due studi. E' da queste considerazioni che nasce l'uso degli intervalli di confidenza, che tengono in maggiore considerazione la dimensione dell'effetto.

In conclusione, il valore P è, e verosimilmente continuerà a essere, una presenza costante nella letteratura scientifica; la differenza dovrebbe risiedere nell'opinione degli autori e nell'interpretazione critica che il lettore dà dei risultati presentati.

ALTERNATIVE ALL'UTILIZZO DEL VALORE DI P: GLI INTERVALLI DI CONFIDENZA

Abbiamo visto come il valore di P non fornisca informazioni sull'entità della differenza nell'effetto tra due (o più) gruppi, anche se spesso il lettore è portato a pensare che quanto più piccolo è il valore di P, tanto maggiore è l'importanza del risultato. Questa errata interpretazione è così frequente che le linee guida per la valutazione statistica dei manoscritti presentati a questa rivista ritengono importante sottolineare esplicitamente la distinzione (15). Per esempio, una differenza di scarsa rilevanza clinica può essere interpretata come "significativa" perché è stata portata oltre il valore soglia di significatività (statistica) da una elevata dimensione campionaria.

È per far fronte a questo problema interpretativo che numerose riviste mediche sollecitano gli autori a presentare i risultati in termini di intervallo di confidenza, in alternativa, o meglio in aggiunta, ai valori di P. L'aspetto positivo dell'uso dell'intervallo di confidenza è quello di presentare la dimensione dell'effetto e l'intervallo di valori possibili. Torna al lettore la responsabilità di interpretare se un risultato, pur statisticamente significativo, risulti di fatto clinicamente rilevante.

Tuttavia, gli elementi che entrano nel calcolo dell'intervallo di confidenza sono gli stessi che servono al

calcolo dello stimatore statistico, che fornisce la stima della P. Se l'intervallo di confidenza, infatti, include il valore di non effetto, che sarà 0 per una differenza o 1 per un rapporto, ciò implica la presenza di una significatività statistica. Se riprendiamo gli esempi 3 e 4 in Tabella 1, nel secondo caso, nel quale si raggiungeva la significatività statistica valutando il valore di P, l'intervallo di confidenza ha come limite inferiore 0,023, per cui lo zero (che per una differenza rappresenta il valore di non effetto) non è incluso nell'intervallo dei valori possibili. Nel primo caso, invece, nel quale non si era raggiunto un valore di P corrispondente alla significatività statistica, il limite inferiore dell'intervallo di confidenza è pari a -0,02: il fatto che l'intervallo di confidenza includa lo zero conferma l'assenza di significatività statistica, ma la scarsa distanza del limite stesso dallo zero sta a suggerire che non ci si trova molto distante dal raggiungerla. In queste situazioni occorre tenere sempre presente il ruolo della potenza statistica, che all'atto pratico risente prevalentemente della dimensione campionaria. Se riprendiamo, infatti, gli stessi dati dell'esempio 3, ma li duplichiamo in modo da ottenere una dimensione campionaria di 200 casi (Tabella 1, esempio 7), non solo il valore della statistica test aumenterà a 2,783, corrispondente a una P pari a 0,006, ma l'intervallo di confidenza risulterà molto meno ampio, distante dal valore zero, anche se ovviamente, trattandosi di una media, rimarrà sempre simmetrico intorno al valore della differenza tra le medie, che è sempre 4,07 mg/dL.

ALTERNATIVE ALL'UTILIZZO DEL VALORE DI P: LA STATISTICA BAYESIANA

Un'altra alternativa proposta all'utilizzo del valore di P è rappresentata dalla statistica bayesiana (10). In estrema sintesi, i risultati derivanti dal singolo esperimento vengono pesati per l'"odds" *a priori* che l'ipotesi sia vera, formulata prima che i dati siano disponibili. Benché l'uso della statistica bayesiana si vada diffondendo nei modelli statistici più complessi, esso non ha mai guadagnato una popolarità incondizionata perché si è ritenuto che l'assegnazione *a priori* di una probabilità possa offrire un margine eccessivo di soggettività al giudizio, proprio in un ambito in cui si dovrebbe inseguire la massima oggettività.

CONCLUSIONE

Per il valore di P vale probabilmente il brocardo del diritto antico *"abusus non tollit usum"*: il fatto che sia impropriamente utilizzato e, soprattutto, interpretato non ne annulla l'utilità. La facilità con cui la significatività di un test di ipotesi correttamente interpretato può essere utilizzata come strumento decisionale è verosimilmente il motivo per cui il suo uso rimane diffuso nella maggior parte degli ambiti scientifici. Tuttavia, in un'epoca in cui sofisticati "software" statistici sono alla portata di qualunque ricercatore, una comprensione più approfondita del significato e dell'interpretazione del valore di P sono diventati troppo importanti per lasciarli confinati soltanto all'uso da parte degli statistici.

CONFLITTO DI INTERESSI

Nessuno.

BIBLIOGRAFIA

1. Nuzzo R. Scientific method: statistical errors. Nature 2014;506:150-2.
2. Anonimous. Number crunch. Nature 2014;506:131-2.
3. Boyd JC, Annesley TM. To P or not to P: that is the question. Clin Chem 2014;60:909-10.
4. Annesley TM, Boyd JC. The P value: probable does not mean practical. Clin Chem 2014;60:1021-3.
5. Ellison AM, Gotelli NJ, Inouye BD, et al. P values, hypothesis testing, and model selection: it's déjà vu all over again. Ecology 2014;95:609-10.
6. Harris EK. On P values and confidence intervals (why can't we P with more confidence?). Clin Chem 1993;39:927-8.
7. Fisher R. Statistical methods for research workers, 13th ed. New York: Hafner, 1958.
8. Zanolla L, Graziani MS. Glossario per il lettore di un articolo scientifico. Parte II: la statistica inferenziale. Biochim Clin 2014;38:314-25.
9. Neyman J, Pearson E. On the problem of the most efficient test of statistical hypotheses. Philosophical Transactions of the Royal Society. Series A 1933;231:289-337.
10. Goodman SN. Towards evidence-based medical statistics.1. The P value fallacy. Ann Intern Med 1999;130:995-1004.
11. Poisson J, Fedoriw Y, Henderson MP, et al. Performance evaluation of the Helena V8 capillary electrophoresis system. Clin Biochem 2012;45:697-9.
12. Fisher RA. The logic of inductive inference. J Roy Stat Soc 1935;98:69-78.
13. Pearson K. Tables for statisticians & biometricians. Cambridge University Press, 1914. https://archive.org/details/tablesforstatist00pearrich.
14. Fisher RA, Yates F. Statistical tables for biological, agricultural and medical research. 6th ed. 1938. https://digital.library.adelaide.edu.au/dspace/handle/2440/10701.
15. Zanolla L, Graziani MS. Glossario per il lettore di un articolo scientifico. Parte I: La statistica descrittiva. Biochim Clin 2014;38:129-35.