L'arte e la scienza di imparare dai dati: la biostatistica, dalle mutazioni genetiche all'intelligenza artificiale.

La statistica: l'unica scienza che permette a esperti diversi, usando gli stessi numeri, di trarne diverse conclusioni.
Attribuita a Evan Esar

Serena Arima *Dipartimento di Storia, Società e Studi sull'Uomo - Università del Salento*

La biostatistica è comunemente definita come l'applicazione di metodologie statistiche a dati di natura biologica o medica. In realtà, tale definizione è piuttosto riduttiva in quanto essa si occupa di sviluppare nuove metodologie che tengano conto della peculiarità dei dati sperimentali e che rispondano, in modo chiaro, alle domande poste dagli scienziati. In particolare, i metodi di biostatistica si sono particolarmente sviluppati per l'analisi di dati genetici che, per dimensione e per struttura di dipendenza, non possono essere analizzati con metodologie *standard*. La biostatistica nasce quindi come materia interdisciplinare che formalizza con un linguaggio statistico-probabilistico problemi di natura applicata e cerca di dare delle risposte che riflettano sia una significatività statistica che biologica.

Introduzione

La natura umana ci porta a riconoscere nell'osservato strutture, gruppi o schemi che ci permettano di generalizzare e formalizzare i fenomeni concreti. Tipico esempio di tale processo di apprendimento è il concetto stesso di probabilità: infatti, la probabilità è un concetto primario, una intuizione innata nell'essere umano. Nella vita di tutti i giorni accade spesso di dover prendere

delle decisioni in condizioni di incertezza e, seppur a digiuno di nozioni specifiche, ciascuno di noi pondera le proprie scelte minimizzando una funzione di rischio che rappresenta null'altro che una quantificazione della casualità di ciascun evento. Mediante tale funzione, gli individui orientano le proprie scelte.
Tale intuizione è, per sua definizione, soggettiva e pertanto non univoca. Per essere generalizzata e automatizzata, se ne definisce una formalizzazione specifica che passa attraverso assiomi, definizioni e teoremi che impongono una sorta di oggettività il cui obiettivo è convertire le osservazioni in conoscenza. La biostatistica può quindi essere definita come lo strumento di quantificazione della conoscenza che deriva da esperimenti biologici e/o medici. Essa attua una formalizzazione in termini matematico-probabilistici del problema sperimentale riportando all'utente finale una valutazione del rischio, in condizioni di incertezza, delle proprie decisioni.
Mai come in questo momento storico la valutazione del rischio e dell'incertezza legata al dato campionario svolgono un ruolo cruciale. Un esempio più che attuale è quello relativo ai *test* diagnostici. Un *test* diagnostico per una particolare condizione è detto positivo se stabilisce che la condizione è presente e negativo se sancisce che la condizione è assente. Un modo per stabilire l'accuratezza dei *test* diagnostici è calcolare la probabilità associata a due tipi di possibili errori:

- **falso positivo**: il *test* afferma che la condizione è presente, ma essa è effettivamente assente;

- **falso negativo**: il *test* sancisce che la condizione è assente, ma essa è effettivamente presente.

Il tri-*test* del sangue fornisce una stima, per una donna in gravidanza, del rischio che il bambino nasca affetto dalla sindrome genetica di Down. Questa sindrome ha una prevalenza pari a $1/800$ ma gli scienziati suppongono che tale prevalenza sia più elevata se le donne hanno più di 35 anni. Uno studio prospettico [1] condotto su 5282 donne con più di 35 anni confronta l'esito del *test* con l'effettivo stato di salute del nascituro, Tabella 1. La sensitività del *test* diagnostico, ossia la probabilità che l'individuo positivo al *test*

Sindrome di Down	Pos	Neg	**Tot**
Si	48	6	**54**
No	1307	3921	**5228**
	1355	**3927**	**5282**

Tabella 1: *Esiti dei test diagnostici riguardanti l'emergenza della sindrome di Down eseguiti su 5282 donne con più di 35 anni [1].*

sia realmente affetto dalla patologia in esame è definita nel seguente modo

$$P(Si|Pos) = \frac{P(Si \cap Pos)}{P(Pos)} = \frac{48/5282}{1355/5282} = 0.035$$

In sintesi, delle donne con *test* positivo, men del 4% ha effettivamente un feto con la sindrome di Down. Perchè una donna, allora, dovrebbe sottoporsi a questo *test* visto che la maggior parte dei risultati positivi sono falsi positivi? In base alla tabella vista prima, si ricava che la probabilità che una donna abbia un figlio con la sindrome di down, $P(Si) = 54/5282 = 0.0102$, sia circa pari all'1%. Dalla stessa tabella si evince che la specificità, $P(No|Neg) = 0.0015$, poco più di 1 su 1000.

Il teorema di Bayes

La Statistica si evolve e si modifica ai fini di riflettere e spiegare il pensiero dello scienziato. Tipico esempio è il teorema di Bayes, punto cardine della Statistica moderna, che formalizza il ragionamento alla base delle Scienze sperimentali quali la Biologia e la Medicina. Tale teorema, che ha dato origine ad un intero filone di pensiero statistico, la statistica Bayesiana, formalizza in termini statistico-probabilistici la conoscenza a-prioristica dello scienziato sullo studio in esame. Infatti, è inimmaginabile che lo scienziato che sta per condurre un esperimento sia completamente cieco rispetto ai potenziali risultati che potrebbe ottenere. Nell'impostazione Bayesiana, la probabilità di un evento, risultato di un esperimento aleatorio, è la congiunzione tra l'idea pre-sperimentale dello scienziato (detta a-priori) e l'esito, tramutato in dato, dell'esperimento stesso (detto verosimiglianza). È pertanto possibile definire, secondo questo approccio, la probabilità di un evento come l'aggiornamento delle ipotesi a-prioristiche con il dato sperimentale. Tale

probabilità è detta probabilità a-posteriori. Si noti, tuttavia, che l'esperimento deve essere ovviamente condotto in modo assolutamente equo rispetto alle ipotesi di base.

Si consideri il seguente esempio: supponiamo che un paziente manifesti un dolore al petto e prima di recarsi dallo specialista consulti, telefonicamente, il proprio medico di base. Quest'ultimo, riferiti i sintomi e l'anamnesi del paziente che conosce bene, immagina una serie di possibili cause di tale dolore (infarto, infreddatura, rottura di una costola ...) attribuendo a ciascuna di essere un peso, ossia una probabilità. Dopo aver visitato il paziente e aver collezionato una serie di informazioni oggettive, il medico giunge ad una diagnosi a-posteriori ottenuta ponderando le ipotesi di base e la visita strumentale del paziente.

Si noti che il teorema di Bayes è una mera applicazione della legge delle probabilità condizionate e, pertanto, la sua validità è indipendente dal tipo di approccio alla probabilità che si intende seguire. È dalla sua interpretazione, in termini di a-priori, verosimiglianza e a-posteriori che prende le mosse l'impostazione Bayesiana della probabilità.

In ambito inferenziale, le differenze tra impostazione classica e Bayesiana della probabilità emergono in modo chiaro. In particolare, supponiamo di aver a disposizione delle osservazioni campionarie $x_1, \ldots, x_n$ e vogliamo fare inferenza sui parametri della popolazione da cui tale campione è stato estratto. Se assumiamo una forma parametrica per il fenomeno in esame (ad esempio un modello teorico come quello Normale o Binomiale), $f(x;\theta)$, utilizziamo il dato campionario per stimare il parametro (o i parametri) θ della popolazione. La differenza sostanziale tra le due interpretazioni risiede nella caratterizzazione di θ: in ambito frequentista, θ è una quantità incognita ma fissa. Tale quantità può essere stimata con diversi metodi, fra cui il più famoso è il metodo della massima verosimiglianza che consiste nel massimizzare la funzione di verosimiglianza $L(\theta : x) = \prod_{i=1}^{n} f(x_i;\theta)$ associata ai dati (supponendo un campione indipendente e identicamente distribuito). Nell'impostazione Bayesiana, il parametro θ è incognito ma non fisso: in particolare, θ è trattato come una variabile aleatoria con una sua distribuzione di probabilità $\pi(\theta)$ che riassume la conoscenza a-priori sul parametro incognito della popolazione. Le procedure inferenziali sono basate sulla legge a-posteriori di θ che si ottiene come

$$\pi(\theta|x) = \frac{L(\theta;x)\pi(\theta)}{\int \pi(\theta)L(\theta;x)d\theta}$$

dove l'integrale può essere sostituito da una sommatoria se si ha a che fare con leggi di probabilità discrete. La quantità al denominatore svolge il ruolo di costante di normalizzazione (ossia fa sì che $\pi(\theta|x)$ sia una legge di probabilità) ma gioca un ruolo chiave nella statistica Bayesiana nel confronto tra modelli (fattore di Bayes). Tutte le procedure inferenziali si riducono quindi all'analisi della distribuzione a posteriori $\pi(\theta|x)$: gli intervalli di credibilità (controparte Bayesiana degli intervalli di confidenza) sono intervalli della legge di probabilità $\pi(\theta)$ che contengono una proporzione di osservazioni pari al livello di confidenza prefissato e i *test* di ipotesi si conducono valutando la plausibilità delle ipotesi che si vuole confrontare rispetto alla distribuzione a posteriori. In entrambi i casi non ci si rifà al principio del campionamento ripetuto, alla base delle procedure inferenziali classiche.

La presunta soggettività del metodo insieme alla complessità computazionale che si evince dalla formula precedente hanno reso la statistica Bayesiana meno comune rispetto agli altri approcci. Oggi tutte le metodologie sono ugualmente diffuse a livello di ricerca mentre in ambito applicato e sopratutto di statistica ufficiale i metodi classici sono più privilegiati (si veda, tra gli altri, [3] per una *review*).

Recentemente, in ambito biologico e medico, molti sono i lavori di Biostatistica che utilizzano la statistica Bayesiana anche perchè, come sarà chiaro nella sezione successiva, essa risulta particolarmente utile in situazioni in cui non si può avere un numero elevato di repliche sperimentali.

Modelli complessi

Dati genetici

Teorema di Bayes e studio di sensitività e specificità sono strumenti base essenziali per l'analisi del dati. Tuttavia, i dati provenienti da studi

sperimentali, sono complessi e difficilmente analizzabili in modo esaustivo utilizzando tecniche preconfezionate. Ecco perché la Biostatistica può essere definita come l'arte e la scienza che ci permette di apprendere dai dati. Le metodologie devono comprendere nella loro formalizzazione la complessità del dato, riconoscendo che alcun modello teorico possa riprodurre in modo fedele l'incredibile aleatorietà della scienza. Una celebre frase attribuita a George Box, uno dei padri della statistica, che in un lavoro del 1976 pubblicato su Journal of the American Statistical Association [2], riassume il significato della statistica moderna:

"*All models are wrong*".

Tale aforisma è stato esteso in "*All models are wrong but some are useful*": in esso si riconosce la fallibilità dei modelli statistici, come strumenti utilizzati per formalizzare fenomeni talmente complessi da non poter essere ingabbiati in una qualsiasi formalizzazione matematica. Il modello è sempre da considerare come uno strumento di studio di una parte, a volte piccola, di un fenomeno talmente complesso che anche comprenderne una minuscola porzione è un passo avanti per la scienza.

Questo è ciò che accade quando la statistica si applica a fenomeni complessi come l'analisi delle mutazioni genetiche. La determinazione di quali geni o proteine si modificano in parti del DNA è alla base della diagnosi e, auspicabilmente in un prossimo futuro, della terapia di malattie genetiche che, nella maggior parte dei casi, rendono la vita dell'individuo particolarmente difficile.

Grazie all'avvento di nuove tecnologie, quali `microarray genomici` e/o `proteomici`, negli ultimi anni è stata prodotta una grande mole di dati la cui analisi risulta complessa da diversi punti di vista. In primis, i dati, che sono una quantificazione del segnale genico (si veda Figura 1), vengono collezionati in matrici con un numero di righe (tipicamente i geni) estremamente elevato. Inoltre, visti i costi piuttosto elevati di replicabilità dell'esperimento e, in alcuni casi, il numero esiguo di pazienti reclutabili, il numero di repliche per ciascuna espressione è piuttosto esiguo. Si è quindi tipicamente in una situazione in cui per uno stesso gene si hanno poche repliche in diverse condizioni sperimentali (*course of dimensionality*). In tale contesto, qualsiasi *test* statistico è poco realistico perché la variabilità dovuta al campionamento, rumore sperimentale, domina sostanzialmente il segnale. È inoltre da considerare il problema della molteplicità e della struttura di correlazione naturalmente indotta dalla natura del dato genetico.

Introduco la notazione più utilizzata: sia y_{gcr} l'espressione genica del gene $g-$mo, r-ma replica (pazienti) e della $c-$ma condizione sperimentale ($g = 1, ..., n$; $r = 1, .., r_s$, $c = 1, ..., C$). Tipicamente, il numero di geni n è dell'ordine delle centinaia di migliaia mentre le repliche disponibili per ogni gene sono meno di una decina ($max(r_s) \leq 10$). Circa le condizioni sperimentali, nel contesto più semplice si ha $s = 1, 2$ condizioni, ossia confrontano le sequenze geniche di soggetti malati e quelle di soggetti sani. L'obiettivo è quello di valutare se esistono geni per i quali l'espressione relativa ai soggetti malati risulti alterata, significativamente più elevata (*over-expression*) o significativamente più bassa (*under-expression*), rispetto ai soggetti sani. Pur volendo ascrivere questo problema al semplice confronto tra campioni, l'esiguo numero di repliche, l'elevato numero di osservazioni e la complessa struttura di correlazione indotta tra osservazioni rende necessario l'uso di metodi alternativi. Se si lavora in un contesto di Statistica parametrica, si assume che il dato sperimentale sia ottenuto come un campione casuale da un modello teorico descritto da una legge di probabilità $f(x; \theta)$, dove θ è il vettore di parametri. Nel contesto più comunemente utilizzato, la legge di probabilità di riferimento è la legge Gaussiana i cui parametri $\theta = (\mu, \sigma^2)$ ne definiscono la media e la varianza. I dati osservati sono quindi considerati un campione casuale da tale legge di probabilità, e vengono utilizzati per stimare i parametri incogniti. In particolare, per valutare l'espressione differenziale, si assume

$$y_{gcr} \sim N(\alpha_g + 0.5(-1)^c \delta_g + \beta_{gcr}, \sigma^2_{gc})$$

dove α_g è l'effetto medio dovuto al gene g, δ_g è l'effetto differenziale tra due condizioni sperimentali; β_{gcr} cattura l'effetto *array*, ossia la parte del segnale osservato dovuto a rumore sperimentale, quali difetti del vetrino, quantifica-

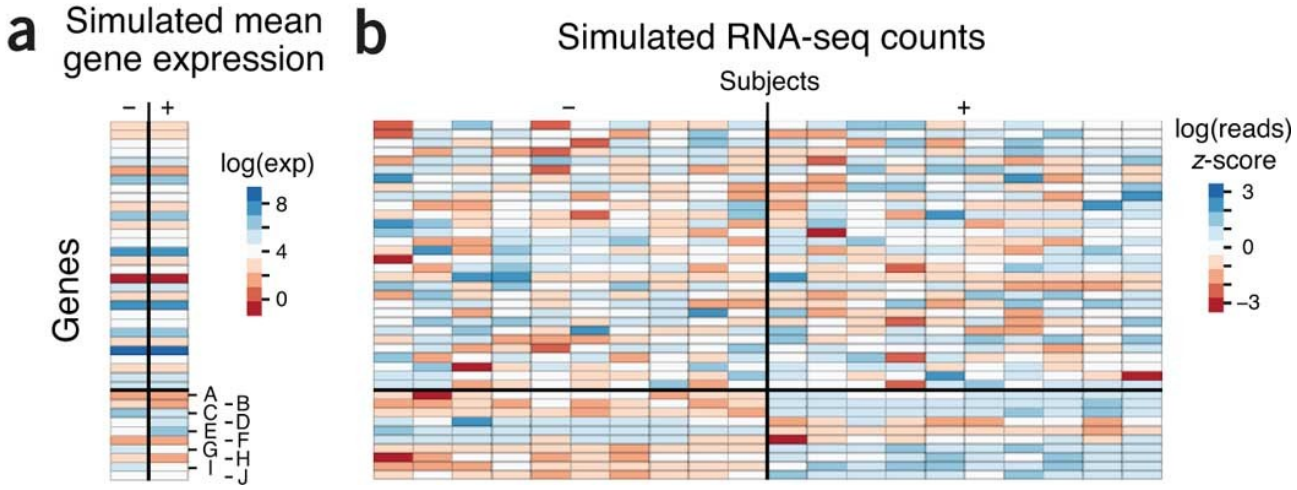

Figura 1: *Un esempio di microarray genetico*

zione errata del segnale di *background*, effetto dell'operatore. L'effetto *array* β_{gcr} si suppone agisca in modo **eteroschedastico**, ossia cresca in base al segnale corrispondente: è modellato, quindi, come $\beta_{gcr} = f_c(\alpha_g)$ dove f è una funzione tipicamente non parametrica, tipo *spline*, del livello di espressione stimato. Si noti che $\mu_{gc} = \alpha_g + 0.5(-1)^c\delta_g + \beta_{gcr}$ rappresenta la media della distribuzione Normale dalla quale supponiamo che i dati osservati siano stati selezionati. Per capire quindi se il gene g-mo è espresso in modo differenziale nelle due condizioni, bisogna confrontare μ_{g1} e μ_{g2}. Come chiaramente sottolineato in [4], la sovra/sottoespressione di un gene non è solo una questione di numeri, ossia dettata esclusivamente da fattori di natura statistica, ma deve tenere conto di una significatività prettamente biologica. Infatti, non è sufficiente che la differenza di espressione nelle due condizioni sia significativamente diversa per dichiarare un gene anomalo ma è anche necessario che tale gene superi un livello di espressione in ciascuna condizione che risulti significativo da un punto di vista biologico. Pertanto, un gene g è dichiarato differenzialmente espresso se

$$|\mu_{g1} - \mu_{g2}| Z\delta_{cut} \cap \alpha_g > \alpha_{cut}$$

dove δ_{cut} e α_{cut} sono soglie definite da esperti nel settore. In altre parole, α_{cut} rappresenta il livello medio minimo di espressione tale che un gene g possa essere considerato espresso, mentre δ_{cut} è la soglia tale che la differenza di espressione tra due condizioni sia biologicamente rilevante. Tipicamente δ_{cut} è posto pari a $log(2)$ in quanto da un punto di vista biologico un gene è differenzialmente espresso se la sua espressione in una condizione è il doppio che nell'altra. La soglia α_{cut} dipende dal tipo di *array* che si utilizza per la sperimentazione.

Questo tipo di modellistica, complicata nel corso degli anni, è un tipico esempio di come la Biostatistica coniuga la formalizzazione statistico-probabilistica con conoscenze e informazioni di natura biologica senza le quali le conclusioni, seppur corrette da un punto di vista metodologico, sarebbero vane nello specifico contesto di applicazione.

L'analisi di dati `proteomici` presenta una ulteriore complessità: il dato, che nei modelli precedenti viene linearizzato, quando si parla di proteine deve tenere in considerazione la correlazione tra i diversi siti e le relative distanze, nonché la conformazione tridimensionale della proteina come in Figura 2. (si veda, tra gli altri, [5]).

Filogenetica

La biostatistica trova larga applicazione in filogenetica, la scienza che si occupa di studiare il processo evolutivo degli organismi vegetali e animali dalla loro comparsa sulla Terra a oggi [6]. I metodi statistici permettono di costruire gli alberi filogenetici, diagrammi che mostrano le relazioni fondamentali di discendenza comune di gruppi tassonomici. Tali alberi vengono costruiti studiando le differenze relative tra sequenze geniche di diversi organismi e contando, mediante algoritmi complessi, tutte le possibili ricombina-

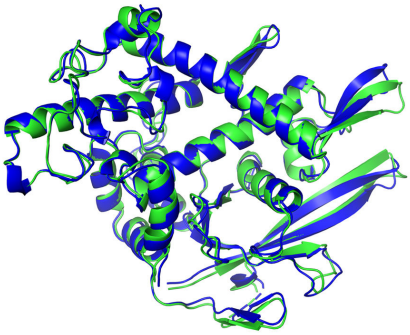

Figura 2: *Un esempio come appare la struttura 3D di una proteina. Le distanze linearizzate non rispecchiano chiaramente le distanze nella conformazione tridimensionale.*

zioni e relative distanze. In un albero filogenetico (si veda Figura 3), ciascun nodo rappresenta l'antenato comune più recente dei soggetti che si trovano ai nodi successivi e la lunghezza delle ramificazioni è proporzionale al tempo intercorso in termini evolutivi tra di essi.

A partire dal lavoro di Cavalli-Sforza [7] sono state proposte diverse metodologie di analisi che richiedono la definizione di processi stocastici (quali processi di `Branching`, processi `Markoviani` o `semimarkoviani`) che governano il meccanismo di mutazione delle sequenze geniche nel tempo. Questa tipologia di analisi è resa particolarmente complessa dalla grande mole di dati che coinvolge la cui analisi prevede l'utilizzo di algoritmi computazionali complessi anche da un punto di vista informatico. Algoritmi di percorso minimo o di massima parsimonia, come ad esempio l'algoritmo di Fitch [8] o di Sankoff [9], sono ancora attualmente oggetto di studio in diverse discipline per ottimizzarne la complessità matematica che computazionale.

Statistica e intelligenza artificiale

Il **Machine Learning** (ML), anche detto apprendimento automatico, è una delle applicazioni dell'intelligenza artificiale. Il ML è un insieme di algoritmi matematico/statistici che permettono ad un computer di apprendere dai dati e prendere decisioni senza ricevere istruzioni dirette. Il sistema informatico emula la logica del ragionamento umano e, mediante l'inserimento di nuovi dati, prende decisioni ottimizzando funzioni di rischio. La distinzione tra Statistica e

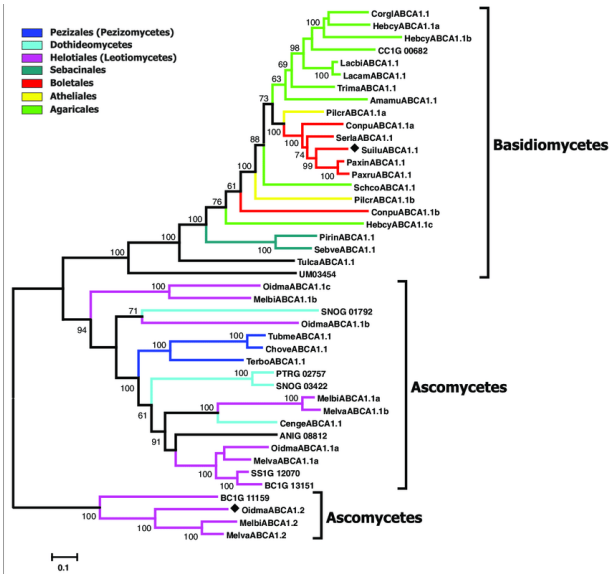

Figura 3: *Un esempio di albero filogenetico.*

ML è piuttosto oscura, ammesso che questa distinzione realmente esista. Nella letteratura, tale differenza è ascritta principalmente all'obiettivo delle due discipline. I metodi di ML sono basati su modelli ottimizzati dal punto di vista predittivo. I modelli statistici sono disegnati principalmente a fini esplicativi, ossia per valutare da un punto di vista inferenziale la significatività della relazione tra variabili. Sebbene tale definizione sia piuttosto condivisa, essa non è nè univoca nè esclusiva. Un recente articolo su Nature [10] evidenzia similarità e differenze tra Statistica e ML (così come *data science*) evidenziando che in realtà le differenze sono semplicemente di natura pratica e non concettuale. Il ML non esisterebbe senza la statistica, ma il ML è particolarmente utile in un'epoca come la nostra in cui la quantità di dati a nostra disposizione supera di gran lunga il concetto di campionamento tipico dei modelli statistici, nelle sue molteplici declinazioni.

Il ML ha trovato una grande applicazione in ambito medico, in particolare nelle Neuroscienze per lo studio degli stimoli cerebrali e le reazioni dei neuroni (si vedano recenti studi sull'Alzheimer fra cui [11]). Stesso dicasi per `Data Science` basato su metodi computazionali per l'analisi dei dati: tali metodi includono analisi esplorative che aiutano il *data scientist* a entrare nel dettaglio dei dati e del loro meccanismo generatore, a pulire e pre-processare i dati, intuire le prime relazioni tra variabili. Questo tipo di operazioni nell'epoca dei *big-data* richiede l'utilizzo di stru-

menti informatici specifici e più potenti dal punto di vista dell'esecuzione. Capire a quali aree della scienza attribuire ciascuna di queste definizioni è particolarmente complesso forse perché la scienza è per sua natura interdisciplinare.

Conclusioni

Molti ricercatori ritengono che la Biostatistica sia meramente un insieme di procedure da fare per soddisfare gli editori o *referee* di una rivisita o le autorità regolatorie o le agenzie di *funding*. Ma la necessità della Biostatistica è chiara quando si vuole considerare la relazione tra l'intuizione del ricercatore e un metodo formale.

Concludendo, l'interdisciplinarità e il ruolo della Biostatistica possono essere riassunti dall'affermazione di Laplace che ha osservato che la teoria della probabilità è [12]

> "*. . . at bottom only common sense reduced to calculus; it makes us appreciate with exactitude that which exact minds feel by a sort of instinct without being able ofttimes to give a reason for it. It leaves no arbitrariness in the choice of opinions and sides to be taken; and by its use can always be determined the most advantageous choice*"

In questo lavoro si è voluto evidenziare come le metodologie statistiche nascono da problemi di natura concreta. Esse si sono sviluppate e continuano ad adattarsi ed evolversi grazie ai problemi che sorgono dai dati reali. Le evoluzioni in ambito biologico e medico sono il volano per le evoluzioni metodologiche di analisi dei dati: il biologo/medico spiega il problema in esame eviscerandone le problematiche inerenti, lo statistico lo formalizza in un linguaggio statistico-probabilistico e, insieme , discutono i risultati valutandone la significatività a tutto tondo.

[1] J. Haddow et al.: *Reducing the Need for Amniocentesis in Women 35 Years of Age or Older with Serum Markers for Screening*, New England Journal of Medicine, 330 (1994) 1114.

[2] G. E. P. Box: *Science and Statistics*, Journal of American Statistical Association, 71 (1976) 791.

[3] C. P. Robert: *The Bayesian choice*, Springer, Berlino (2006).

[4] A. Lewin, S. Richardson, C. Marshall, A. Glazier,T. Aitman: *Bayesian modeling of differential gene expression*, Biometrics, 62 (2006) 1.

[5] P. Baldi, G.W. Hatfield: *DNA Microarrays and Gene Expression*, Cambridge University Press, Cambridge (UK) (2011).

[6] Rita Levi Montalcini: *La Galassia Mente*, Baldini & Castoldi, Milano (2001).

[7] L. L. Cavalli-Sforza, A.W.F. Edwards: *Phylogenetic analysis. Models and estimation procedures*, Am. J. Hum. Genet., 19 (1967) 233.

[8] W. M. Fitch: *The molecular evolution of Cytochrome c in Eukaryotes*, Journal of Molecula Evolution, 8 (1976) 13.

[9] D.Sankoff, R.J. Cedergren, G. Lapalme: *Frequency of insertion-deletion, transversion and transition in the evolution of 5S Ribosomal RNA*, Journal of Molecula Evolution, 7 (1976) 133.

[10] D. Bzdok, N. Altman, M. Krzywinski: *Statistics versus machine learning*, Nature, 15 (2018) 233-234.

[11] T. Jo, K. Nho e A.J. Sayking: *Deep Learning in Alzheimer's Disease: Diagnostic Classification and Prognostic Prediction Using Neuroimaging Data*, Frontier Aging Neuroscience, 20 (2019) . https://doi.org/10.3389/fnagi.2019.00220

[12] P.S. Laplace: *A Philosophical Essay on Probabilities*, Dover, New York (1951).

Serena Arima: è professore associato di Statistica per la ricerca sperimentale e tecnologica presso l'Università del Salento. Si occupa di statistica bayesiana e di metodi computazionali per la stima di modelli complessi.