Ann. Ist. Super. Sanità, vol. 33, n. 2 (1997), pp. 285-291

Alcune considerazioni sull'organizzazione dei dati in statistica applicata alla biologia

Carlo PETRINI

Laboratorio di Fisica, Istituto Superiore di Sanità, Roma

Riassunto. - Nella statistica medico-sanitaria l'acquisizione dei dati può essere o meno associata alla variabile tempo, e cioè al riferimento cronologico dell'istante in cui un dato è stato acquisito. La variabile tempo può essere un dato necessario, perché il problema lo richiede per sua natura. Può essere un dato utile, soprattutto nello studio dei sistemi multivariabili, per accrescere la selettività tra le variabili in gioco. Ci si riferisce soprattutto ai problemi che cercano le relazioni esistenti tra esposizioni ad agenti presunti come patogeni e danni biologici. Per identificare un modello matematico di un sistema in presenza di variazioni nel tempo, fatta distinzione tra metodi deterministici e statistici, si propone per i metodi statistici una interpretazione del loro significato, validità ed applicabilità. Il modello è confrontato con metodi quali la regressione lineare e gli studi di coorte, per illustrarne la contiguità e gli aspetti comuni. Sono prese in considerazione la numerosità dei dati, la frequenza e la durata nel tempo del campionamento. Sono anche presi in considerazione requisiti e prestazioni di un software che debba elaborare i dati. E' ricordata l'importanza dei protocolli di acquisizione, trasmissione e gestione dei dati.

Parole chiave: biostatistica, campionamento, mutua correlazione, sistema multivariabile, epidemiologia.

Summary (*Considerations on data organization in statistics applied to biology*). - In health statistics data collection may be associated or not to the variable time, that is the moment of data acquisition. The variable may be necessary, if the characteristics of the statistics problem require it. It may be a useful information, particularly in the study of multivariable systems, in order to increase the selectivity of the mutual relations between the variables involved. The main subject is the relation between presumed pathogenous agents and biological damage. In order to identify a mathemathical model of a system when variations in the time are present, drawn the distinction between deterministic an statistics methods, an intepretation of significance, validity, and applicability of statististical methods is proposed. For the purpose of showing affinities, the model is compared with other models, that is linear regression and cohort studies. The quantity of data, the frequency and the length in time of the sampling are considered. Characteristics and qualifications of a software for data processing are discussed. The importance of sampling, transmission and organization protocols of the data is also shown.

Key words: biostatistics, sampling, cross-correlation, multivariable system, epidemiology.

Introduzione

In epidemiologia cresce l'interesse per la conoscenza sugli agenti potenzialmente patogeni, con dosi di esposizioni individuali basse, patologie con insorgenze basse ed anche con rischi relativi (RR) attesi non elevati. La ricerca biologica, che sola può accertare e descrivere i rapporti di causa-effetto, è accompagnata, e spesso preceduta, da quella statistica. Il descrivere rischi bassi, l'escludere, quando occorra, la presenza di rischio, richiede precisione e, nel calcolo delle probabilità, elevato numero di osservazioni.

La possibilità di acquisire dati da eventi transitori e da situazioni in evoluzione desta l'interesse di misurare anche il tempo intercorrente tra l'esposizione ad agenti patogeni ed il manifestarsi di danni. Nel presente lavoro si cercherà, identificati con un modello generale sistemi che contengono le variabili tempo, di interpretare come casi particolari del modello sistemi che non le contengono.

La prima evidenziazione dell'esistenza di un legame tra due variabili epidemiologiche che tenga conto di un eventuale ritardo nel legame stesso è offerta dalla funzione di mutua correlazione. Questa funzione da sola non ne offre però né una misura né una valutazione. La descrizione e la misura del legame stesso sono offerte invece (entro le condizioni di applicabilità) dalla funzione di trasferimento, o dalla relativa rappresentazione grafica (diagrammi di Bode) (v. cap. "La terminologia e le espressioni matematiche").

Nel presente lavoro sono comparate tre situazioni: quella oggetto degli studi di coorte, la regressione lineare, e quella studiata attraverso il calcolo della funzione di trasferimento.

Degli studi di coorte si sottolineano attraverso un diagramma i requisiti di numerosità dei campioni.

Se il modello per essere identificato richiede una campionatura nel tempo, devono essere soddisfatti requisiti e condizioni essenziali, quali l'immutabilità nel tempo e la linearità del sistema osservato, la campionatura dei dati frequente e protratta nel tempo quanto necessario.

Questi aspetti, ed altri a questi strettamente legati (la natura deterministica o statistica del problema, la molteplicità delle variabili in gioco e la correlazione tra loro, l'organizzazione dei dati per il calcolo automatico, la disponibilità di software per le elaborazioni e per le relative valutazioni), hanno assunto una nuova connotazione con la comparsa di normative e di protocolli di raccolta dei dati ambientali e biologici.

Il modello della funzione di trasferimento (o il suo equivalente espresso attraverso equazioni integro-differenziali) non ha alternative nel rappresentare i fenomeni che esso descrive. I suoi limiti di applicabilità sono limiti fisici non eludibili. Il biologo ed il medico che studino un problema di tale natura si incontrano necessariamente con questa formulazione del problema stesso.

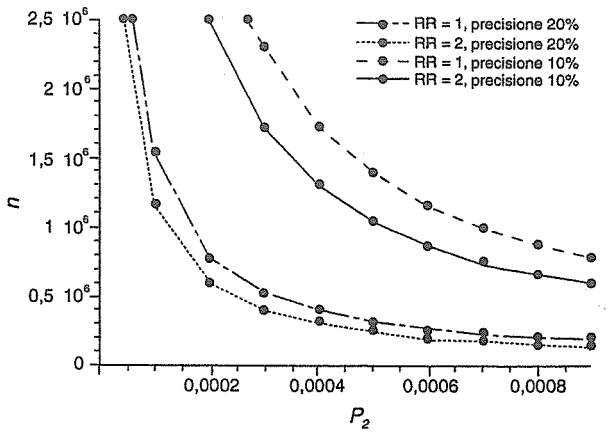

Fig. 1. - Numerosità di un campione in uno studio di coorte. P_2 = incidenza nella popolazione non esposta; n = numerosità del campione. Per tutte le curve il grado di confidenza è 95%.

La numerosità dei campioni negli studi di coorte

Come è noto, negli studi di coorte si definisce un RR come rapporto tra i casi percentuali di insorgenza in una popolazione esposta rispetto a una non esposta. L'assenza di rischio si ha con RR = 1. La Fig. 1 (v. anche "Diagramma di numerosità e rischio relativo" [1]), su due coordinate e con due parametri, descrive le relazioni esistenti tra quattro grandezze: percentuale di insorgenza in popolazione non esposta, numerosità del campione, RR, e precisione relativa. La figura consente, noti tre qualunque dei quattro parametri, di determinare il quarto.

Un richiamo sulla definizione di mutua correlazione nei problemi di regressione lineare

Questo paragrafo, nelle definizioni che vi ricorrono, è preparatorio al successivo paragrafo che introduce i dati cronologici nei problemi di biostatistica.

Il modello di regressione lineare semplice, partendo dall'osservazione di un diagramma di dispersione di dati relativi a due variabili cerca se vi sia e quale sia la retta migliore per descrivere graficamente le relazioni tra le due variabili. Il metodo generalmente impiegato è il metodo dei minimi quadrati e la retta è detta perciò retta dei minimi quadrati [2].

Il metodo, con l'espressione della geometria analitica dalla retta:

$$y = a + bx \qquad (1)$$

consente (ove vi ricorrano le circostanze statistiche) di determinare i valori migliori di a e di b.

Spesso nei problemi medico-sanitari, e sempre nei problemi di epidemiologia, le variabili x ed y assumono solo valori positivi o al più nulli (si pensi, per la variabile indipendente x, a dosi di esposizione ad agenti patogeni e, per la variabile y, a danni biologici, insorgenza di patologie, dati statistici sui decessi).

Per ragioni formali, e cioè per ricercare definizioni di grandezze fisiche applicabili a questo ed ai paragrafi che seguiranno, è necessario immaginare il diagramma di dispersione ed il segmento di retta che ne rappresenta il modello come traslati sui due assi in modo che il segmento passi per l'origine. Le grandezze x ed y e lo sciame dei loro punti devono essere misurati come scarto rispetto ai loro valori medi. In questo modo la nota espressione che consente di calcolare il valore di b diverrà

$$b = \frac{\Sigma x_i y_i}{\Sigma x_i^2} \qquad (2)$$

Il numeratore $\Sigma x_i y_i$ esprime, nel senso e con i limiti che si dirà, una correlazione tra le due variabili.

La circostanza che il numeratore assuma il valore zero può avere diversi significati: indicare che realmente non esiste un rischio, oppure che il numero di osservazioni non è sufficiente ad evidenziare un rischio esistente, o che l'eventuale relazione tra le variabili, sebbene forte, è di tipo non lineare.

La circostanza in cui il numeratore sia diverso da zero, e quindi maggiore di zero, non è in sé sufficiente a misurare l'intensità del legame tra x ed y, perché il valore del numeratore dipende dal numero di coppie (x; y) ed ovviamente dalle unità di misura di x ed y.

Sono stati perciò definiti in letteratura il "coefficiente di determinazione campionario" ed il "coefficiente di correlazione lineare", che consentono un confronto quantitativo tra situazioni epidemiologiche diverse e quindi una misura in quantità e qualità dell'influenza di x su y nella regressione lineare.

La mutua correlazione come funzione del tempo di ritardo

Un'importante generalizzazione del termine $\Sigma x_i y_i$ è l'estensione anche a coppie di valori per cui y sia in ritardo di un tempo τ rispetto ad x. Si dovrà anzitutto supporre che x(t) ed y(t) siano variabili continue nel tempo. Si può allora definire una funzione [3].

$$\varphi(\tau) = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} y(t+\tau)\, x(t)\, dt \qquad (3)$$

Per tutti i valori del tempo di ritardo per cui la funzione di mutua correlazione è maggiore di zero si deve concludere (con le cautele analoghe a quelle di cui sopra) che esiste una influenza di x(t) su una y(t+τ) e cioè con un ritardo τ.

La variabile tempo nell'epidemiologia medico sanitaria ed in biologia

Per identificare il modello matematico di un sistema di cui sia ingresso una variabile x(t) ed uscita y(t) è possibile usare una funzione di trasferimento. Con questo termine si vuole sottolineare che il modello deve consentire, partendo da qualunque storia x(t) ipotizzabile per la x, di ottenere l'evoluzione y(t) trasferita su y.

Se il sistema è di natura epidemiologica, si può pensare al caso in cui x rappresenta la dose di esposizione ad un possibile agente patogeno ed y il conseguente danno biologico.

Se sono noti i meccanismi fisici o biologici che legano y ad x il modello matematico potrà essere un algoritmo che attraverso operazioni algebriche, integrali o differenziali, interpreta il sistema come funzione del tempo; è possibile ed utile una sua trasformazione nell'algoritmo funzione di trasferimento come sotto definito.

Più spesso ricorre la circostanza opposta in cui si dispone di dati statistici e si desidera scoprire l'esistenza di una funzione di trasferimento, ci si ripromette di riconoscerla, sperando di giungere anche ad una conoscenza dei meccanismi fisici o biologici (e quindi ad un vero rapporto di causa-effetto) tra x ed y.

In questo secondo caso (dati ottenuti da rilevazioni statistiche) la funzione di trasferimento è identificabile per mezzo di due elaborazioni successive dei dati.

La prima è il calcolo della funzione di mutua correlazione definita da (3). In un sistema disturbato, come tipicamente sono quelli epidemiologici, le due variabili, ingresso ed uscita, sono affette da fluttuazioni casuali non correlate, e solo in parte da fluttuazioni correlate. La funzione di mutua correlazione per ogni valore del ritardo τ evidenzia (col meccanismo di somma dei prodotti) la sola componente correlata: in assenza di correlazione grandezze casuali a valore medio nullo danno origine ad una somma di prodotti nulla. La funzione di mutua correlazione deve essere misurata rapportandola ad un denominatore analogo a quello di formula (2). In questo caso all'autocorrelazione di x.

La seconda elaborazione dei dati è un cambio di variabile: da una rappresentazione del modello come funzione dei tempi di ritardo τ ad una rappresentazione del modello stesso come funzione di un parametro ω, come in seguito definito ed illustrato.

Il nuovo modello è una funzione di trasferimento nel senso sopra detto e cioè per qualunque evoluzione della variabile in ingresso esso consente di calcolare l'evoluzione in uscita del sistema.

La funzione di trasferimento definita mediante i diagrammi di Bode

Anche per le scienze naturali vale l'osservazione della fisica per cui, imprimendo ad una grandezza di un sistema sollecitazioni periodiche con frequenza f_1, si possono conseguire, in qualunque parte o parametro del sistema, solo effetti periodici, con la stessa frequenza f_1. Ciò vale per i sistemi lineari rappresentati sia come funzione del tempo t, sia come funzione del ritardo τ (come è appunto la funzione di mutua correlazione).

Osservando il sistema sottoposto a variazioni sinusoidali di frequenza f_1 (e quindi di pulsazione $\omega_1 = 2\pi f_1$) si può descrivere la relazione tra y ed x con due valori: il rapporto tra le ampiezze delle due sinusoidi $G(\omega)$, ed il loro ritardo di fase $\delta(\omega)$.

Estendendo l'indagine a tutto lo spettro di frequenze interessate si può ottenere una descrizione completa del legame tra y ed x [3-4].

Trattandosi di due variabili, ampiezza e fase, entrambe funzioni di ω, una rappresentazione efficace è quella grafica, proposta da Bode, che si serve di due diagrammi distinti.

La definizione di $G(\omega)$ come rapporto tra due ampiezze è interpretabile come una generalizzazione della b definita da (1), o meglio b è il valore particolare che assume $G(\omega)$ per $\delta = 0$, e quindi $\tau = 0$, o ancora molto spesso $\omega = 0$.

Le Fig. 2 e 3 vogliono essere esemplificative e quindi generali, ma possono rappresentare per la biologia anche un caso significativo e molto frequente, e cioè il caso in cui l'ampiezza si attenui per ω via via crescenti (variazioni via via più rapide) fino ad annullarsi. I ritardi crescono (Fig. 2 e 3) contestualmente all'attenuarsi di $G(\omega)$.

Origini statistiche e non statistiche dei dati per il tracciamento dei diagrammi di Bode

Come già detto, le Fig. 2 e 3 non sono necessariamente rappresentative di un modello statistico. Ricordando che la rappresentazione di Bode è nata per applicazioni nello studio delle comunicazioni e delle regolazioni, essa può nascere in contesti molto diversi:

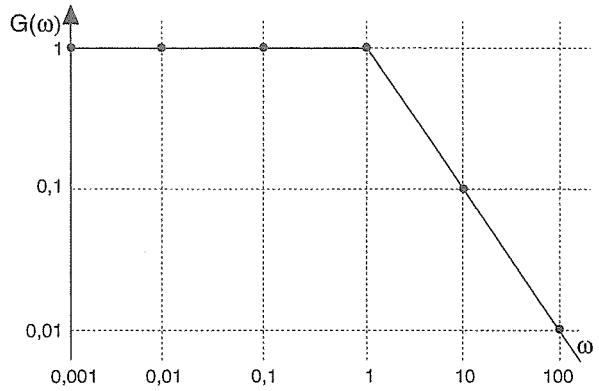

Fig. 2. - Esemplificazione di un diagramma di Bode per le ampiezze (approssimazione con una spezzata). Le ascisse rappresentano le pulsazioni (normalizzate rispetto al piede della spezzata). Le ordinate rappresentano il rapporto tra le ampiezze (normalizzato per il valore $\omega = 0$).

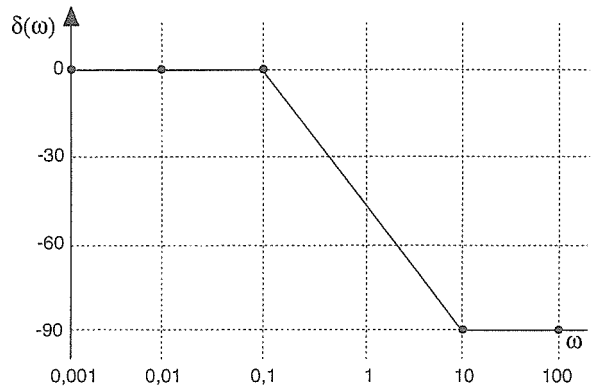

Fig. 3. - Esemplificazione di un diagramma di Bode per i ritardi di fase (approssimazione con una spezzata). Le ascisse rappresentano le pulsazioni. Le ordinate rappresentano i ritardi di fase (angolo misurato in gradi).

- il sistema (di comunicazione o di automazione) è progettato per avere la risposta $G(\omega)$ voluta;
- il sistema è identificato con misure sperimentali che possono usare metodi statistici o deterministici (questa via è ipotizzabile per sistemi fisici ed anche medico-sanitari, ma non epidemiologici). Possono essere assimilati alla identificazione per via sperimentale anche problemi di epidemiologia in occasione di osservazione di eventi eccezionali che producano perturbazioni assimilabili ad ingressi impressi sperimentalmente (ed eventualmente anche deterministici);
- il sistema è identificato con una acquisizione, campionata nel tempo, di dati aventi proprietà statistiche idonee;
- il sistema è già noto attraverso relazioni di causa-effetto, o attraverso la conoscenza dei meccanismi fisici, chimici e biologici che lo reggono. Anche questa origine può essere utilizzata per rappresentare il sistema attraverso la $G(\omega)$.

Esempio di un evento accidentale con rilevanti conseguenze sanitarie. Diagrammi cronologici e identificazione della funzione di trasferimento

Si supponga (l'esempio è volutamente teorico) che da un impianto industriale avvenga una accidentale immissione nell'atmosfera di un isotopo radioattivo. L'immissione, intensa e di breve durata, può essere rappresentata in un diagramma come un evento ad impulso (Fig. 4).

Si supponga che la radioattività dell'isotopo abbia un decadimento trascurabile nell'arco di tempo che interessa lo studio epidemiologico. La dose x(t) assorbita dalla popolazione esposta è perciò con sufficiente approssimazione un evento a gradino (Fig. 5). Nell'esempio teorico si supponga che prima dell'evento accidentale quell'isotopo sia stato sempre assente nell'atmosfera. Si supponga anche di poter individuare la popolazione esposta.

A questa esposizione con evoluzione a gradino si supponga ora che consegua un'accresciuta insorgenza di una ben identificata patologia. Si supponga che i mezzi di rilevamento, a fronte dell'evidenza del fenomeno, e quindi di una esauriente disponibilità di dati consentano di identificare con chiarezza l'evoluzione nel tempo del danno biologico y(t) misurato appunto come insorgenza. L'evoluzione sia quella di Fig. 6.

Un evento come quello preso ad esempio, e per cui si possa aver garanzia della chiarezza e precisione dei dati di Fig. 5 e Fig. 6 è a tutti gli effetti identificato. Le Fig. 5 e 6 rappresentano un caso particolare, ma sufficiente a dedurre i parametri epidemiologici.

Questi parametri sono:

- l'aver stabilito che la risposta a gradino è esponenziale,

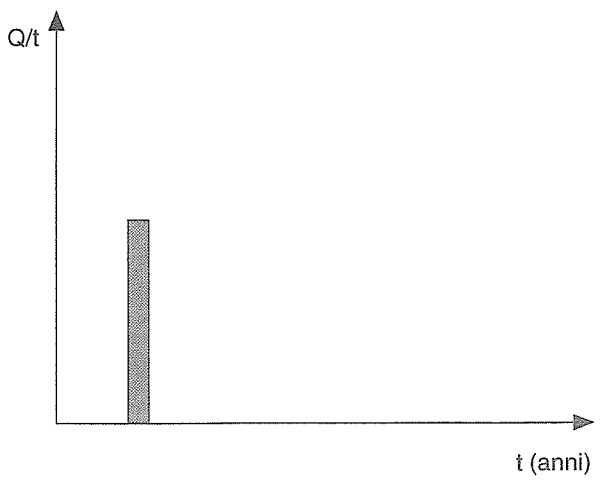

Fig. 4. - Evento ad impulso. L'area dell'impulso è la quantità di isotopo radioattivo complessivamente rilasciata nell'atmosfera.

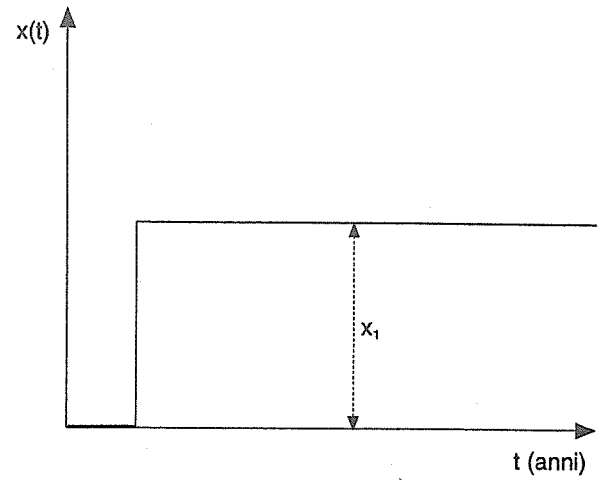

Fig. 5. - Evento a gradino. Parametro che misura l'esposizione all'agente patogeno per la popolazione esposta.

- che la sua costante di tempo è T_1,
- che a tempi infiniti il rapporto tra danno ed esposizione vale $(y_1 - y_2)/x_1$.

I parametri consentono di tracciare i due diagrammi di Bode e quindi definiscono la funzione di trasferimento.

La Fig. 7 è, salvo le scale di misura, uguale alla Fig. 2.

Analogamente, per i ritardi di fase $\delta(\omega)$ vale la figura 3, dove l'asse delle ascisse prende le stesse scale di misura di Fig. 7.

Identificazione della funzione di trasferimento in assenza di eventi accidentali rilevanti. Condizioni e limiti di esistenza ed identificazione

Nei paragrafi precedenti si è dato rilievo alla possibilità di conoscere per vie non statistiche i parametri identificativi del modello, ma si è costantemente fatto riferimento al caso più generale, in cui i dati statistici disponibili pur in assenza di eventi rilevanti consentono essi stessi l'identificazione.

Questo caso più generale riconduce all'analogia con il caso della regressione lineare in cui da sciami di osservazioni si identificava una retta. Per la funzione di trasferimento la presenza di un parametro in più (il tempo) accresce il numero dei dati, ma non necessariamente il numero delle osservazioni. Il tempo infatti aggiunge un elemento in più di discriminazione pur in presenza di condizioni di disturbo.

Vengono di seguito elencate quattro condizioni necessarie perché la funzione di trasferimento esista e sia identificabile:

- le relazioni devono essere lineari;
- la funzione $G(\omega)$ non deve evolvere durante il tempo della sua osservazione (immutabilità nel tempo del modello e cioè immutabilità delle proprietà statistiche del sistema osservato);
- il tempo di osservazione T_0 deve essere significativamente più lungo del tempo massimo t_M per cui esiste una correlazione tra y(t) ed x(t). Si è valutato che l'errore nel calcolare la $G(\omega)$ sia proporzionale a $\sqrt{\tau_M / T_0}$.
- il campionamento delle variabili y(t) ed x(t) deve essere più breve del tempo limite indicato dal teorema di Shannon [5]. Il teorema di Shannon dice che si devono operare campionamenti (per ognuna delle grandezze campionate) con una frequenza almeno doppia di quella della più alta frequenza interessante. La più alta frequenza interessante è valutabile sul diagramma di Bode esemplificato in Fig. 3 in corrispondenza di un significativo calo di ampiezza.

Ove siano verificate queste condizioni, l'identificazione è possibile attraverso le due elaborazioni di cui agli appositi paragrafi (le correlazioni e la trasformazione dal campo delle τ a quello delle ω) e con l'uso dei software e la gestione dei dati di cui nel seguito.

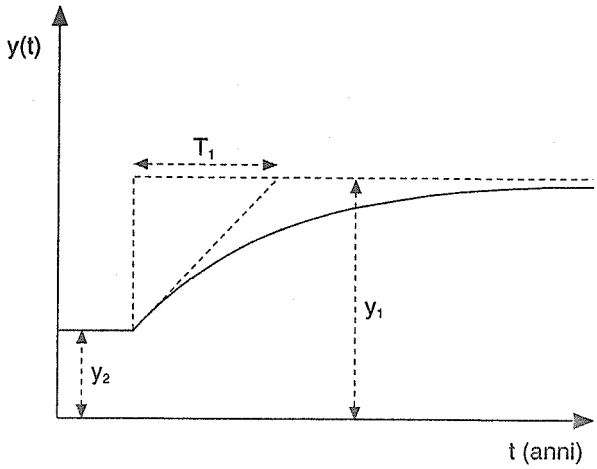

Fig. 6. - Evoluzione del danno biologico: incremento caratterizzato da una curva esponenziale con costante di tempo T_1, ed incremento y_1-y_2 raggiunto asintoticamente.

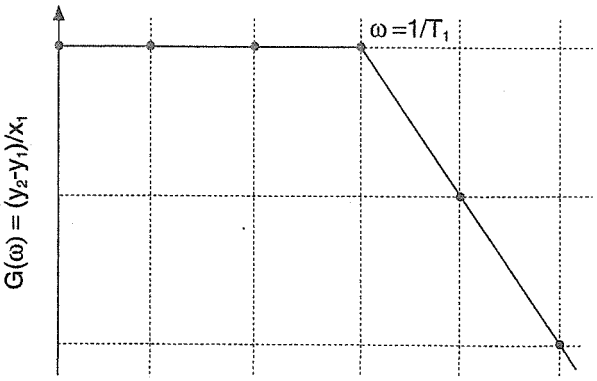

Fig. 7. - Rapporto tra danno biologico ed esposizione in funzione del parametro ω (diagramma di Bode per le ampiezze).

Sistemi multivariabili

La funzione G(ω) lega due variabili, nei casi di cui sopra, chiamate y ed x. E' possibile ricercare legami bilaterali anche nei sistemi multivariabili. Per sistema multivariabile (*multivariable* in inglese) si intende un sistema il cui modello è una matrice di funzioni di trasferimento. In quanto segue, le enunciazioni sono proposte tramite esempi, per semplicità e chiarezza.

In fisiologia umana si vogliano studiare le interazioni tra assunzione alimentare di calcio, magnesio e potassio sulle calcemia, magnesiemia, potassiemia. E' possibile identificare il sistema (per via sperimentale deterministica, oppure con un'indagine statistica) attraverso una matrice di funzioni G(ω): calcemia, in rapporto all'assunzione di calcio; magnesiemia ad assunzione di magnesio, ecc., ma anche calcemia ed assunzione di magnesio; calcemia ed assunzioni di potassio, e così via. Con la conoscenza di questa matrice è possibile distinguere e descrivere le ampiezze ed i tempi delle risposte. E' possibile che il sistema calcio-magnesio-potassio sia meglio descritto prendendo in considerazione anche altre grandezze, ad esempio il tasso nel sangue del paratormone, ed il conoscitore dei meccanismi di regolazione saprà costruire un "diagramma a blocchi" dove ogni blocco è individuato da una G(ω), e l'insieme individua il sistema.

Se tutte le G(ω) della matrice sono costanti per ogni valore di ω (oppure se sono costanti per tutti i valori di ω che interessano nella pratica) si può dire che si è identificata una matrice di valori costanti con le convenzioni dei punti precedenti. L'identificazione di una matrice di costanti può essere ottenuta direttamente con i metodi della regressione lineare [2]. E' degno di nota il maggior valore conoscitivo delle funzioni G(ω), anche per le potenzialità di discriminare tra loro più concause attraverso i loro diversi meccanismi temporali.

Requisiti e obiettivi di un programma di calcolo automatico. Disponibilità di software e di protocolli

Se l'identificazione è oggetto di un programma di calcolo automatico se ne possono elencare i requisiti e le prestazioni essenziali:

- i dati provengono da un campionamento limitato nel tempo, ed occorrerebbe invece disporre di un tempo infinito. E' quindi necessario costruire un artificio idoneo a superare questa limitazione senza alterare i contenuti statistici;
- il campionamento che rispetti il teorema di Shannon (v. sopra) permette di interpretare i dati discreti provenienti dal campionamento come rappresentativi di funzioni continue. Il software deve elaborarli come tali;
- il software deve verificare le proprietà statistiche dei dati e quindi insieme identificare e valutare il modello.

Quanto sopra vale per lo specifico algoritmo che elabora dati campionati nel tempo, correlati con una funzione di mutua correlazione che è funzione del tempo di traslazione τ, e che sono espressi in funzione non più del tempo, ma della pulsazione ω.

Occorre ricordare che il calcolo delle funzioni di auto e mutua correlazione e, insieme a questo, anche il calcolo della G(ω) si trovano spesso implementati su quelle strumentazioni industriali che utilizzano allo scopo microprocessori anche per acquisizione ed elaborazione di dati in tempo reale. A questi algoritmi ed ai relativi software sono da ritenersi riconducibili studi di fisiologia o patologia (si veda l'esempio di cui sopra del metabolismo del calcio, del magnesio e del potassio) che siano riconducibili a banche dati disponibili su mezzi di calcolo individuali e contenute nelle loro memorie.

Se invece i dati da elaborare provengono da campagne di rilevazione vaste per estensione geografica, per numerosità dei dati, e per estensione nel tempo, si deve supporre che essi costituiscano un patrimonio contenuto in database accessibili attraverso vie di comunicazione. L'utente sarà collegato ad un sistema, e così ne utilizzerà sia i dati, sia i programmi di calcolo automatico.

Se la ricerca è condotta utilizzando una rete locale di mezzi informatici del tipo *local area network* (LAN), l'utente si troverà ad avere accesso ad un contesto in cui sono unificati i protocolli della rete. Se la comunicazione dei dati avviene attraverso distanze dette "geografiche" (e cioè tra enti di ricerca lontani ed indipendenti) si deve ipotizzare che le comunicazioni avvengano attraverso un sistema classificato come *wide area network* (WAN), dove in generale esiste un ente che gestisce i canali di comunicazione, tramite protocolli di comunicazione normalizzati anch'essi da enti normativi. L'utente non ha in generale accesso diretto alla WAN, ma attraverso una LAN o altro sistema a sua volta connesso con la rete geografica. I protocolli di sistemi di trattamento dei dati (accesso, ricerca e trasmissione) sono specifici delle reti e dei loro gestori.

Un software che si prefigga l'identificazione di una funzione di trasferimento utilizzerà presumibilmente programmi già disponibili per il calcolo delle funzioni di mutua correlazione e di automazione per l'analisi in frequenza.

L'organizzazione dei dati, il controllo sulle loro proprietà statistiche, le interpolazioni, e tutte quelle interrogazioni che l'utente voglia porre difficilmente sono invece contenuti nei software già utilizzabili.

L'unificazione dei protocolli di raccolta dei dati epidemiologici, demografici ed ambientali è cura degli organismi sanitari nazionali, comunitari ed internazionali. Con protocolli normalizzati le banche dati saranno universali, e cioè integrabili tra loro e gestibili da chi vi abbia accesso.

La terminologia e le espressioni matematiche

In un contesto interdisciplinare con gerghi e letteratura specifici, si è cercato di usare il lessico minimo comune. Si è creduto possibile, dietro le convenzioni dei termini, cogliere i fatti fisici e biologici senza perdere in rigore. Alcune enunciazioni sono nuove così come sono state formulate, e come sono state raccolte e collegate fra loro. L'organizzazione dei dati in statistica applicata alla biologia è stata qui sviluppata ed esemplificata quasi esclusivamente con l'epidemiologia: l'epidemiologia è sempre preceduta, accompagnata e verificata nelle sue conclusioni dallo studio biologico-sanitario.

Si è cercato di evitare il ricorso alle formule matematiche, con l'eccezione della definizione di mutua correlazione: questa formula introduce il tempo di ritardo nel concetto di mutua correlazione.

Per la definizione della funzione di trasferimento si è ricorso alla rappresentazione grafica (di Bode), e si è evitato di citare la trasformata di Fourier. Si è però fatto cenno alla possibilità di scomporre un evento che si evolve nel tempo in una sovrapposizione di componenti elementari nel campo delle frequenze (analisi in frequenza). Delle correlazioni e dell'analisi in frequenza sono disponibili software descritti nei manuali d'uso.

Diagramma di numerosità e rischio relativo

La Fig. 1 è una interpolazione ed una estesa estrapolazione di tabulati contenuti in un manuale di epidemiologia edito dall'Organizzazione Mondiale della Sanità nel 1991 [1]. Di questo manuale si è anche adottáta la terminologia. La figura presuppone che l'approccio statistico, utilizzante come in questo caso tutti i dati disponibili, sia esaustivo, e cioè che non ci si possa attendere un approccio migliore con gli stessi dati.

Alcune considerazioni comparative sulle grandezze G(ω), b, RR

Nel definire la grandezza b e la funzione G(ω), si è notato come la prima diviene un valore particolare della seconda se si attribuisce alle due grandezze x ed y lo stesso significato.

Avendo preso come esempio e come riferimento il più volte citato problema di epidemiologia, si era invece finora lasciata aperta la scelta delle grandezze da comparare, e delle loro unità di misura.

Nello "studio di coorte" sono definite tutte le grandezze in gioco. Si ripete qui per immediatezza:

P_1 è l'incidenza della patologia nella popolazione esposta;

P_2 è l'incidenza nella popolazione non esposta da cui:

P_1-P_2 la maggior incidenza della popolazione esposta e quindi il danno (o il maggior danno) subito con l'esposizione;

$(P_1-P_2)/P_2$ è la stessa grandezza, misurata però avendo come riferimento (unità di misura) P_2. Questo rapporto rappresenta la grandezza y.

Nello studio di coorte la dose di esposizione è una grandezza che può assumere due soli valori: vale "zero" per la popolazione non esposta ed "uno" per quella esposta.

In questo caso x vale perciò 1 e si può scrivere:

$$b = \frac{(P_1-P_2)/P_2}{1} = RR-1 \qquad (4)$$

Si noti che il valore RR = 1 è l'assenza di rischio, per cui vale anche b = 0.

Ricevuto il 23 aprile 1996.
Accettato il 2 aprile 1997.

BIBLIOGRAFIA

1. LWANGA, S.K. & LEMESHOW, S. 1991. *Determination de la taille d'un échantillon dans les études sanometriques. Manual pratique*. Organisation Mondiale de la Santé, Genève.

2. WAYNE, W.D. 1995. *Biostatistics: a foundation for analysis in the health sciences*. John Wiley & Sons, Inc., New York.

3. DE COULON, F. 1990. *Théorie et traitement des signaux*. Presses Polytechniques Romandes, Lausanne.

4. PETRINI, C. 1992. Modelli matematici di sistemi in biologia ed epidemiologia. *Biologi italiani* 22(3): 14-15.

5. PIERCE, J.R. 1963. *La teoria dell'informazione*. Mondadori, Verona.