Carlo Ferreri

Approcci per il problema dell'instabilità nell'analisi della dispersione non-poissoniana

Quaderni di Dipartimento

Serie Ricerche 2008, n.2

ISSN 1973-9346

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Dipartimento di Scienze Statistiche "Paolo Fortunati"

Indice

1 – Premessa 3

2 – Sull'esigenza di un inquadramento euristico della problematica della sovradispersione " 13

2.1 – Un esempio d'uso incauto dei modelli di sovradispersione " 13

2.2 – Effetti d'instabilità e struttura dell'evidenza statistica di processi di nascita: casi illustrativi " 15

3 – Due processi di nascita per inferenze sulla non-poissonianità. " 18

3.1 – Il processo iperbinomiale e procedure inferenziali sui relativi *process-trend* e *process-index function* " 18

- Il processo binomiale negativo " 18
- L'extended Pólya process " 20
- Il processo iperbinomiale (HBP) " 21
- Letture stratificatorie della distribuzione iperbinomiale negativa (NHBD) " 23
- Estimazione di ML della distribuzione iperbinomiale (HBD) " 27

3.2 – Una estensione della distribuzione di Pólya-Aeppli (PAD) come distribuzione di un birth process a dispersione non-poissoniana " 29

- Caratteristiche e ruolo di una riparametrizzazione della PAD nell'analisi della dispersione sovrapoissoniana " 29
- Interpretazione stratificatoria della PAD " 31
- Estimazione di ML della PAD riparametrizzata " 32
- Estensione della PAD in distribuzione di processo di nascita per l'analisi della dispersione non-poissoniana " 34

4 – Passi di un'indagine, parametrica, su certa incidentalità degli operai di una industria metalmeccanica " 39

4.1 – Distribuzioni statistiche dei dati di conto e risultati dell'impiego

dei *p.m.* HBD e HPAD. " 39

4.2 – Attenuazione degli effetti d'instabilità tramite estimazione di ML basata su distribuzioni statistiche di perequazione per medie mobili " 42

4.3 – Trend interpolatorio della funzione di dispersione non-poissoniana di entrambi i processi di nascita HBP e HPAP. " 45

5 – Confronto tra i valori d'interpolazione $\tilde{\alpha}_{Spl}(t)$ basati sull'HBP e i corrispondenti risultati di estimazioni di diversa concezione " 49

5.1 – Estimazione d'impostazione jackknife della process-index function $\alpha(t)$ ai tempi t_i " 49

- Delineazione di alcune procedure pseudo-jackknife " 49
- Stime pseudo-jackknife della funzione di dispersione $\alpha(t)$ " 51

5.2 – Delineazione di una procedura tipo bootstrap per l'estimazione della process-index function $\alpha(t)$, ai tempi t_i, " 58

- Presupposti e caratteristiche della procedura " 58
- Risultati della CB-procedura " 58

6 – Quadro conclusivo " 60

Appendice: " 64

Tav. I – Distribuzioni osservate e corrispondenti perequate (Campioni di n=108).

Tav. II – Stime di ML dei ϑ_i e degli α_i e corrispondenti valori di pseudo-spline.

Tav. III – Numerosità teoriche della HPAD corrispettive delle stime $\hat{\vartheta}_i$ e $\hat{\vartheta}_{Si}$.

Tav. IV – Numerosità teoriche della HBD corrispettive delle stime $\hat{\alpha}_i$ e $\hat{\alpha}_{Si}$.

Tav. V – Stime di ML, jackknife e bootstrap dei livelli $\alpha(t_i)$ di HBP.

Riferimenti bibliografia " 74

Approcci per il problema dell'instabilità nell'analisi della dispersione non-poissoniana

"Bada che se, studiando un fenomeno,
ne disancori gli eventi
dai tempi del loro accadimento,
ti metti fuori della realtà.
E se ti balza che è sacrosanto dire:
«la verità è la realtà che si accetta»,
vedi di non farla rimanere
l'aquilone dei tuoi sogni."

Irving Fisher

1 – *Premessa*

Nella stragrande maggioranza degli ambiti di ricerca capita sovente che l'obiettivo del ricercatore consista nel cogliere caratteristiche delle sequenze di un certo evento rilevabili pressoché contemporaneamente in "analoghi" set di unità d'osservazione, oppure registrabili lungo un dato intervallo temporale $(0,T]$ su ciascuna delle unità d'indagine considerate

Come si pone mente alle sequenze di entrambi i casi, recependole sulla scorta delle peculiarità dell'esperimento casuale che si progetta in rapporto alla natura del fenomeno e ai requisiti delle unità d'osservazione avendo di mira gli obiettivi, è risaputo che difficilmente al ricercatore non balza una situazione di riferimento, un *benchmark*, che non sia l'espressione di una regolarità ideale, di una situazione di perfezione. Più spesso egli, nel riconoscerla irraggiungibile, quasi senza volerlo o rendersene pienamente conto finisce anzi per porsela come una sorta di meta ideale.

Compiuto l'esperimento, nel passare a far tesoro dell'evidenza statistica rilevata per inferenze sugli aspetti d'interesse, solo in un momento di disattenzione egli potrebbe perciò ritenere reale la situazione di riferimento congetturata, non fosse altro perché ogni avvenimento di questo mondo è unico, così come lo è l'unità d'osservazione, qualunque ne sia la natura.

(*) Il presente quaderno è comprensivo di gran parte di un *report* predisposto per un seminario tenuto nell'anno 2000, presso il Dipartimento di Scienze Statistiche dell'Università di Bologna, nell'ambito di un incontro tematico organizzato dalla prof.ssa Alessandra Giovagnoli.

È pertanto più che naturale che la sua attenzione vada a polarizzarsi sulla discrepanza della configurazione empirica, delineata in termini dell'evidenza statistica, dal profilo che il riferimento concepito corrispondentemente implica.

E poiché questo è, in fondo, l'atteggiamento dell'umano per capire e commisurarsi con gli aspetti della realtà in cui si trova calato, era inevitabile che lo statistico si cimentasse in molteplici modi, per via parametrica, non parametrica, distribuzionale, adistribuzionale, ecc., per mettere a punto approcci e strumenti che gli consentissero, in chiave probabilistica, di prendere posizione nel riguardare tale discrepanza troppo elevata per stare al riferimento ai fini degli obiettivi di ricerca perseguiti.

▷ Un semplice esempio concreto della prima delle due circostanze indicate all'inizio si ha immaginando: 1) un parco di $k \geq 2$ macchine progettate, realizzate e messe a punto per consentire prodotti di un dato standard; 2) che, per saggiarne "l'omogeneità" e la "conformità" a una difettosità non eccedente un dato livello, dai prodotti di ciascuna j-esima macchina, per $j=1,2,..,k$, visti nell'ordine di fabbricazione si prelevi simultaneamente un segmento di n elementi, estraendoli con una certa cadenza; 3) che tutti i segmenti (campioni) vengano poi inviati al punto di controllo, dove ogni loro elemento sia classificato in non difettoso ($\bar{D}$) o difettoso (D).

Un tale esperimento casuale (o statistico che dir si voglia) ovviamente induce la considerazione dei k vettori casuali (uno per macchina)

$$\boldsymbol{X}_j = (\, X_{1j}, .., X_{ij}, .., X_{nj} \,)', \qquad \text{per} \quad j = 1, .., k, \tag{1.1}$$

ciascuno elemento dei quali è un indicatore bernoulliano, visto cioè del tipo

$$X_{ij} \sim (1 - p_{ij})^{1 - x_{ij}} \, p_{ij}^{x_{ij}}, \quad \text{con} \quad x_{ij} \in \{0,1\} \quad \text{e} \quad P(X_{ij} = 1) = p_{ij}, \tag{1.2}$$

$$\text{dove} \quad X_{hj} \perp X_{il} \quad \text{per} \quad j, l = 1, .., k \quad \text{con} \quad j \neq l \quad \text{ed} \quad h, i = 1, .., n,$$

se si dà per scontato che a ciascun elemento prodotto sia associato una variabile casuale (v.c.) a risposta dicotomica: 0 oppure 1 a seconda che esso non sia o invece sia dichiarato difettoso, e che il tipo di esperimento faccia accogliere l'assunzione d'indipendenza tra gli indicatori bernoulliani di vettori diversi.

Se si ritiene che ciascuna macchina operi in condizioni di sottocontrollo, in cui cioè gli accadimenti "difettoso" siano riguardabili essenzialmente accidentali almeno durante l'arco temporale contemplato dall'esperimento, quale situazione di riferimento balza ovviamente quella espressa dall'ipotesi dell'omogeneità (contro non-H_0)

$$H_0 : p_1 = \cdots = p_k = p, \tag{1.3}$$

$$\text{dove} \quad p_j = P(X_{ij} = 1)\text{ , per } i = 1,..,n, \text{ con } j = 1,..,k$$

$$\text{e} \quad X_{hj} \perp X_{il} \text{ per} \begin{cases} h \neq i \quad \text{con} \quad j = l \\ h,i = 1,2,\cdots,n \text{ con } j \neq l \end{cases},$$

che riguarda i campioni $\boldsymbol{X}_j$, corrispettivi delle macchine, non solo tra loro indipendenti: $\boldsymbol{X}_j \perp \boldsymbol{X}_l$ per $j \neq l$, ma anche campioni casuali dal medesimo modello di probabilità (*p.m.*) bernoulliano di parametro p.

Una chiara cognizione della rigidità di questo primo fondamentale riferimento può evincersi anche solo pensando che presume invariabile nel tempo la possibilità di aversi un difettoso. A parte questo, ancorché sia vista accoglibile la "omogeneità" non solo tra gli elementi di una macchina ma anche tra quelli di macchine diverse, del problema affrontato rimane comunque da saggiare la "conformità" alla difettosità prevista.

Come si sa, quale indicatore di discrepanza per tale caso si impone anzitutto il ben noto quoziente empirico di divergenza di Lexis-Bortkiewicz, espresso da

$$Q_f^2 = \frac{nk-1}{k-1}\frac{n\sum_{j=1}^{k}(f_j-f)^2}{nkfg}, \quad \text{ossia} \quad Q_f^2 = \frac{1}{k-1}\frac{n\sum_{j=1}^{k}(f_j-f)^2}{MS_T}, \quad \text{con} \quad MS_T = \frac{nkfg}{nk-1} \qquad (1.4)$$

per il quale vale la relazione asintotica $(k-1)Q_f^2 \sim_{\text{a}} \chi_{k-1}^2$, cui viene fatto ricorso con "grandi" campioni. Ovviamente, $f_j = \frac{1}{n}\Sigma_{i=1}^{n} X_{ij}$ designa la frequenza dei difettosi (D), cioè degli "1", riscontrabili nelle n unità rilevate per la macchina j-esima; $f = \frac{1}{nk}\sum_{i,j} X_{ij}$ quella degli "1" risultante nell'insieme di tutti i campioni, mentre $g = 1 - f$.

La seconda espressione di Q_f^2 è sovente preferita per il fatto che rimarca il riferimento alla situazione di omogeneità totale caratterizzata dal *p.m.* di Bernoulli $X_{ij} \sim \mathcal{B}(p)$, della cui varianza pq la "mean square" MS_T è stimatore corretto così come, sotto tale ipotesi, lo è il numeratore $\frac{1}{k-1}n\Sigma_{j=1}^{k}(f_j - f)^2$. Ovviamente, si ragiona in termini della prima delle (1.4) quando il riferimento è identificato nella distribuzione binomiale dato che, nella stessa ipotesi (1.3), si ha $S = \Sigma_{j=1}^{k}\Sigma_{i=1}^{n} X_{ij} \sim \mathcal{B}(nk, p)$, ossia che la somma S di tutti gli indicatori casuali è di distribuzione binomiale di parametri nk e p.

E poiché si è tanto più portati a respingere l'ipotesi di tale omogeneità quanto più

il quoziente empirico Q_f^2 risulta diverso da 1, per Q_f^2 "significativamente" *maggiore* di 1, in generale si suole parlare di sovradispersione o, per accentuare il riferimento, di *dispersione sovrabinomiale* od anche di *sovrabinomialità*, mentre per Q_f^2 significativamente *minore* di 1 si usa il termine sottodispersione e si parla quindi di *dispersione sottobinomiale* o di *sottobinomialità*, se non anche di dispersione sottobernoulliana.

In entrambe le situazioni più spesso si evince così l'esigenza di approfondire il discorso (Ferreri, 2002) con l'intento di arrivare ad una configurazione fenomenica che supporti un *p.m.* che, in termini della medesima evidenza statistica, dia modo di motivare plausibilmente la dispersine non-bernoulliana quanto meno per gli aspetti di rilevante interesse per gli obiettivi d'indagine e, nella fattispecie, nei riguardi della suddetta supposizione di indipendenza dal tempo della possibilità di difettoso.

▷ Un esempio concreto della seconda circostanza inizialmente citata può vedersi in una ricerca sulla natura degli incidenti cosiddetti "ripetibili" capitati, nel rapporto uomo-macchina durante un dato intervallo di tempo (0,*T*], agli *n* operatori delle macchine di un reparto di un'industria meccanica, in cui svolgono mansioni ritenute indifferenziabili almeno rispetto alla tipologia degli incidenti.

In questo caso, dato che tra gli obiettivi di ricerca di solito non manca quello di vedere se l'incidentalità può riguardarsi essenzialmente accidentale, nell'articolare l'esperimento casuale in primo luogo viene infatti da inquadrare quanto accaduto su ciascuna delle *n* unità d'osservazione (operatori) nell'ottica di un processo di nascita del tipo $\{ X(t \mid \xi_h), t > 0 \}$, per *h*=1,2,..,*n*, dove la v.c. discreta $X(t \mid \xi_h)$ designa il numero degli eventi (incidenti) capitabili, nell'arco di tempo (0,*t*], all'unità *h*-esima, vista caratterizzata da una propria tipica propensione ξ_h all'evento, generalmente dipendente sia dal numero degli incidenti successi anteriormente a *t*, che dal tempo *t*.

In tale ottica, l'analogia nelle mansioni e nella preparazione tecnica degli operatori e detto obiettivo della ricerca, fanno ben presto balzare la situazione che riguarda gli operatori essenzialmente di pari livello di propensione all'incidente, oltre che di incidentalità puramente accidentale. Situazione che segna l'altro fondamentale riferimento che contempla, per ogni singolo soggetto, il medesimo processo di nascita $\{ X(t), t > 0 \}$ di funzione d'intensità costante

$$p_x(t) = \lambda, \quad \text{con} \quad \lambda > 0, \tag{1.5}$$

e cioè caratterizzato da una propensione all'incidente indipendente dal tempo *t* e, ad

ogni istante di tempo, dal numero degli eventi anteriormente successi. Funzione che per la v.c. $X(t)$, comporta, come è noto, la distribuzione di Poisson

$$X(t) \sim P_x(t) = \frac{\lambda t e^{-\lambda t}}{x!}, \quad \text{con } x = 0,1,2,..., \tag{1.6}$$

il cui valore medio

$$\mu(t) = \mathsf{E}[X(t)] = \lambda t, \tag{1.6a}$$

comunemente detto *process-trend*, è nella fattispecie rappresentato da una semiretta uscente dall'origine e crescente nel 1° quadrante del piano cartesiano $Ot\mu$.

Essendo ben noto che la distribuzione di probabilità (1.6) è di media uguale alla varianza, da $\mu(t) = Var[X(t)] = \sigma^2(t) = \lambda t$, nella situazione di riferimento il cosiddetto *indicatore di dispersione*, quale rapporto della varianza alla media, risulta

$$I_D(t) = \frac{\sigma^2(t)}{\mu(t)} = 1, \quad \forall t > 0, \tag{1.7}$$

cioè uguale ad 1 per ogni valore di $t > 0$.

Anche se i dati raccolti sugli incidenti successi ai soggetti d'osservazione per lo più consistono delle rispettive sequenze dei tempi d'accadimento, tra i modi di avvalersene per compiere inferenze su caratteristiche dell'incidentalità, più spesso si privilegia la considerazione della distribuzione statistica delle *n* unità d'osservazione rispetto al numero, *x*, degli incidenti loro capitati durante l'intero arco temprale (0,*T*], limitando poi l'attenzione alla situazione di sovradispersione che essa comunemente manifesta.

Come ci è già capitato di rimarcare (Ferreri, 1983), non si vede però la ragione di rifarsi unicamente al profilo distributivo relativo al momento $t = T$, visto che in tal modo si ignora volutamente l'informazione circa lo svolgimento dei processi d'incidentalità dei soggetti, e quindi rilevante ai fini degli obiettivi dell'indagine.

Per tenerne conto almeno in parte e volendo, per coerenza, utilizzare in tali termini l'evidenza statistica rilevata, il riferimento poissoniano in questione anzitutto suggerisce:

1) di operare un'opportuna segmentazione dell'intero periodo d'osservazione (0,*T*] in successivi intervalli $(t_{i-1}, t_i]$, per $i = 1,2,..,r$, dove $t_0 = 0$ e $t_r = T$;

2) di riguardare, come campione casuale, il set $(X_{1i},..,X_{hi},..,X_{ni})$ dei numeri degli incidenti rilevabili per gli *n* operatori nel periodo $J_i = (0, t_i]$, con $i = 1,2,..,r$;

3) di considerare, a partire da un tempo t_1 che lo motivi, la sequenza delle distribuzioni statistiche di numerosità

$$\mathcal{D}_{n,i}(x) = \{\ (x, n_{xi});\ x = 0,1,..,c_i\ ,\ n = \sum_{x=0}^{c_i} n_{xi}\ \},\quad i = 1,2,..,r \quad \text{e} \quad t_r = T\,, \quad (1.8)$$

la *i*-esima delle quali, in quanto implicata dalla realizzazione del campione casuale corrispondente al tempo scelto t_i, è la distribuzione delle *n* unità d'osservazione rispetto al numero, *x*, di incidenti loro capitati nell'intervallo J_i, $c_i = max(x)$ essendo il massimo numero di incidenti in esse riscontrato.

Calcolando, per ciascuna *i*-esima distribuzione statistica (1.8), con *i*=1,…*r*, la media campionaria $M_i = \frac{1}{n}\Sigma_{x=1}^{c} x n_{xi}$, il corrispettivo 2° momento $m'_{2i} = \frac{1}{n}\Sigma_{x=1}^{c} x^2 n_{xi}$ e quindi la varianza campionaria $\tilde{\sigma}_i^2 = m'_{2i} - M_i^2$, per l'indicatore di dispersione campionario si perviene alla sequenza di valori

$$(\tilde{I}_{D,1},..,\tilde{I}_{D,i},..,\tilde{I}_{D,r})\,, \qquad \text{dove} \quad \tilde{I}_{D,i} = \frac{\tilde{\sigma}_i^2}{M_i}\,, \qquad (1.9)$$

che quasi inevitabilmente porta a riguardarne l'andamento rispetto alla semiretta orizzontale $I_D(t) = 1$, che la (1.7) indica per la situazione di riferimento configurata.

Anche solo tenendo conto che ogni situazione, ogni cosa, ogni singola forma di vita, ogni accadimento è da vedersi unico, ovviamente non ci si meraviglia che le distribuzioni $\mathcal{D}_{n,i}(x)$ implicate dai valori rilevati, le quali per semplicità saranno comunque dette osservate, per l'indicatore di dispersione campionario $\tilde{I}_D$ comportino valori generalmente diversi da uno.

Quando, anche grazie alla relazione asintotica $n\tilde{I}_D \sim_a \chi^2_{n-1}$, un valore di $\tilde{I}_{D,i}$, per *i*=1,…*r*, è ritenuto talmente maggiore di 1 da far riguardare la rispettiva distribuzione statistica di frequenza quale configurazione empirica di certa distribuzione di probabilità di un processo di nascita a dispersione maggiore di quella di Poisson (*PD*), nuovamente si parla di sovradispersione o, per chiarezza, di *dispersione sovrapoissoniana*. Si parla invece di sottodispersione o di *dispersione sottopoissoniana* quando avviene il contrario, vale a dire allorché il valore positivo dell'indicatore di dispersione $\tilde{I}_D$ è visto sostanzialmente inferiore ad 1.

Se sovente capita di leggere che le distribuzioni statistiche di dati di conto del tipo in questione comunemente appalesano sovradispersione, è perciò essenzialmente perché si dà per sottinteso di essersi riguardato unitario l'arco temporale (0,*T*] di rilevazione e, come si è detto, incentrata l'attenzione unicamente sulla distribuzione statistica relativa al tempo $t_r = T = 1$, la quale contempla così il totale degli eventi-

incidente registrati con l'indagine, portando a considerare la distribuzione di Poisson $P_x = \frac{\lambda e^{-\lambda}}{x!}$, con $x = 0,1,2,\ldots.$, cui si riduce la (1.6) per t=1.

Con una segmentazione abbastanza fine dell'intervallo d'osservazione si ha infatti modo di constatare che, al passare da ogni t_i al successivo o, equivalentemente, all'aumentare del numero, $S_i = \Sigma_{x=0}^{c_i} x n_{xi}$, degli eventi nel tempo, i valori della sequenza (1.9) – assunti dall'indicatore $\tilde{I}_D(t)$ – vanno segnalando, dapprima, una sottodispersione tendenzialmente sempre meno spiccata e, poi, una dispersione sempre più nettamente sovrapoissoniana, a partire da un tempo τ_P di cambiamento, che, in generale, attira l'attenzione anche per il fatto di mostrarsi come un sorta di punto di equilibrio instabile.

E non può essere che così poiché, per il tempo t, con $t = t_i$, in cui la rispettiva distribuzione statistica (1.8) presenta $c_t = 1$, consegue $\tilde{\sigma}^2 < M$ ed $\tilde{I}_D(t) < 1$; e, con $c_t = 2$, si ha $\tilde{\sigma}^2 < M$ sia per $f_0 = n_0 / n < 0.5$ che per $f_2 < f_0 - \sqrt{2 f_0 - 1}$ se $f_0 > 0.5$.

Il che induce a congetturare una schematizzazione fenomenica che supporti un modello statistico che, almeno per la durata $(0,T]$ dell'indagine, sia capace di una "lettura" plausibile del reale processo di nascita $\{ X(t), t > 0 \}$ e perciò di distribuzioni di probabilità ai tempi t_i, delle quali, al crescere di t_i, quelle statistiche di frequenza corrispettive delle (1.8), col loro tipico andamento di dispersione non-poissoniana, siano sempre più riguardabili configurazioni empiriche.

Tale andamento, che a partire da detto tempo τ_P è contemplante sovradispersione sempre più netta, va perciò tenuto ben distinto dalla situazione di sovradispersione considerata in tante delle molteplici ricerche in cui gli eventi registrati sono visti sostanzialmente ricondotti al tempo T=1. Ricerche che possono comunque dirsi sempre di "moda" visto che gli statistici, senza muovere obiezioni a detta *ottica atemporale*, per lo più tacitamente l'accolgono continuando a proporre o a sostenere approcci che consentono motivazioni della sovradispersione studiata, e non solo con riferimento alla distribuzione di probabilità di Bernoulli o di Poisson.

Questo, a noi, è successo nei primi anni di vita universitaria, allorquando si avverte particolarmente forte il timore di non riuscire a raccapezzarsi in modo da dire poi qualcosa non del tutto privo d'interesse aspirando a perseguire finalità effettivamente euristiche. Con tale intento e l'attenzione rivolta alle tendenze della letteratura statistica in proposito, ci è però capitato di renderci conto che la sovrapoissonianità anda-

va vista in rapporto al numero degli eventi accaduti, e che il problema cruciale stava non tanto nella costruzione o nella scelta del modello di probabilità (*p.m.*) quanto nell'opportunità di aderire a detta ottica mirando a certi obiettivi e, poi, nella possibilità di usare certi strumenti dell'inferenze a causa dei cosiddetti effetti d'instabilità.

Coerentemente con quanto da più parti si andava sempre più rimarcando, una volta messo a punto il modello di probabilità ritenuto adeguato alla circostanza, in situazioni del genere anche a noi non è infatti mancata l'occasione di constatare che una minima modificazione nelle distribuzioni statistiche utilizzate poteva comportare variazioni talmente rilevanti nelle stime dei parametri del modello usato da vanificare l'iter inferenziale progettato per gli obiettivi della ricerca.

Sulla scorta dei molteplici approcci che la letteratura statistica indicava (Olkin ed al., 1981; Carroll e Lombard, 1985) e soprattutto delle premesse fatte da Hall (1994) nello sviluppare una teoria asintotica delineante la "erratic performance" degli stimatori classici, si è così tentato, sia pure pressoché limitatamente alle circostanze recanti al riferimento poissoniano, di approfondire il discorso avanzando criteri per contrarre gli effetti dell'instabilità (Ferreri, 1996) radicandolo sempre più nel supporto di base alla scelta o alla costruzione del modello di probabilità (*p.m.*), naturalmente anche alla luce delle caratteristiche della forma assunta per l'evidenza statistica rilevata.

Col passaggio all'ottica del processo di nascita è naturalmente emersa l'esigenza, da un lato, di modificare modelli di probabilità usati a livello atemporale per renderli adeguati alla circostanza di dispersione non-poissoniana e, da un altro lato, di tentar di cogliere portata e implicazioni degli effetti d'instabilità sugli stimatori di ML dei modelli di probabilità ritenuti idonei per tale circostanza, quanto meno in rapporto sia a requisiti salienti delle distribuzioni statistiche costruite che ai lineamenti della configurazione congetturata a supporto del modello scelto.

La rassegna compiuta della vasta letteratura statistica disponibile sul primo punto ci ha così sollecitato a discernere il copioso filone delle ricerche che non tengono in qualche modo conto degli effetti d'instabilità. Filone concretatosi, come si sa, con studi essenzialmente teorici che, ancorché affinatissimi e di livello statistico-matematico e probabilistico davvero rilevante – se non pure inutilmente disarmante –, solo di rado mostrano però un utilizzo del modello statistico che non consista nell'ingabbiatura del fenomeno in una struttura squisitamente formale non solo di stile classico: là dove si passa al contesto bayesiano, il "sovraccarico formale" che ne consegue, per lo più sembra anzi affrontato per fruire del "marchio di qualità".

Da un altro lato, non si è potuto fare a meno di inquadrare il filone delle analisi di carattere descrittivo o poco più, che, ancorché sottili e dotte – e talora pure non scevre

da artificiosa ricercatezza –, non hanno contribuito granché al congetturarsi di ipotisi di lavoro su un piano sostanziale e quindi favorito notevoli avanzamenti nel processo di conoscenza del fenomeno d'interesse, nonostante a tal fine ricerche come quello di Greenwood e Yule (1920) avessero aperto da tempo una strada maestra.

Scopo di questo lavoro è pertanto di riferire sui passi compiuti, nel portare euristicamente avanti una ricerca tesa a cogliere la natura dell'incidentalità nel rapporto uomo-macchina degli operai di una certa azienda metalmeccanica, facendo tesoro di procedure per attenuare gli effetti d'instabilità sulle stime da calcolare per i modelli di processo di nascita ritenuti abbastanza adeguati alla circostanza.

Per evidenziare, a tal fine, l'opportunità dei distinguo fatti con la rassegna della letteratura statistica in proposito e per accentuare come la problematica della scelta o della costruzione del modello statistico non vada affrontata che sul piano più sostanziale possibile, nel § 2 non si ritiene fuori luogo, dapprima, richiamare uno studio di detto filone teorico che, a nostro avviso, emblematicamente finisce per non segnare nemmeno una mera esercitazione formale illustrativa di un caso di sovradispersione e, poi, indicare esempi che mostrino come il problema dell'instabilità vada generalmente inquadrato quanto meno sulla scorta degli aspetti salienti della forma con cui ci si limita ad utilizzare l'evidenza statistica disponibile.

Dopo un tale prolungamento della premessa, stando alla circostanza di dispersione non-poissoniana segnalata dalla sequenza dei valori dell'indicatore di dispersione $\tilde{I}_D(t)$ implicati, ai tempi t_i, i=1,..,r, dalle distribuzioni statistiche degli operai rispetto al numero degli incidenti subiti fino agli stessi tempi, nel § 3 ci si dedica alla delineazione di configurazioni fenomeniche che supportino, per le v.c. del modello di nascita, distribuzioni di probabilità che diano modo di arrivare, tramite un'accoglibile soluzioni del problema dell'instabilità, ad una plausibile quantificazione della discrepanza tra il profilo d'attenzione congetturato sulla base di tali distribuzioni statistiche e il corrispondente comportato dal processo di Poisson (PP) assunto a riferimento.

Nella prima parte del paragrafo, partendo da verosimili modificazioni delle assunzioni su cui è visto basato il PP viene indicato, in particolare, l'iter seguito per giungere, col tempo, al processo di nascita $\{ X(t), t > 0 \}$ denominato *processo iperbinomiale* (HBP), probabilizzato da una famiglia di distribuzioni espressa in termini della funzione *process-trend* $\mu(t) = \mathsf{E}[X(t)]$ e da una funzione, detta di *dispersione non-poissoniana* o *process-index function*, in quanto ne indica l'andamento nel tempo rispetto a quella della situazione di riferimento. Processo che è perciò visto capace di fornire una ragionevole "lettura" di quello fenomenico in molti ambiti di ricerca.

Nella seconda parte dello stesso paragrafo è invece preso in considerazione un

processo di nascita che abbiamo denominato *iperPólya-Aeppli process*, e siglato con HPAP, che si propone e s'impone come alternativa all'HBP essendo analogamente ottenibile: (*i*) muovendo dalla nota distribuzione di probabilità di Pólya-Aeppli, peraltro similmente interpretabile sul piano sostanziale come distribuzione mistura o composta di Poisson; (*ii*) tramite una riparametrizzazione che permette di giungere ad una distribuzione della v.c. $X(t)$ in termini del relativo *process-trend* e di una analoga funzione di dispersione non-poissoniana. Processo che, sotto ogni profilo, in pratica si mostra idoneo a fornire una "lettura" del processo fenomenico circa altrettanto, se non più condivisibile, di quella consentita dall'HBP.

Per la distribuzione di probabilità delle v.c. di entrambi i processi viene ovviamente descritto il modo di avvalersene a cominciare dalla estimazione di massima verosimiglianza (ML)

Nel § 4, viene anzitutto presentata l'evidenza statistica in termini di una sequenza di 31 distribuzioni come la (1.8), la *i*-esima delle quali, per *i*=1,2..,31, fornisce la distribuzione degli *n*=108 operatori a certe macchine rispetto al numero, *x*, di incidenti (ripetibili) a loro capitati nel periodo $(0,t_i]$, col tempo misurato in trimestri. Dopodichè, per entrambi processi si costruiscono i plot delle ML-stime che tali distribuzioni implicano per dette funzioni di processo agli stessi tempi, dei quali plot si segnalano poi gli aspetti differenziali.

Al fine di arrivare a stime meno influenzate da fattori d'instabilità, i cui effetti sono particolarmente rilevanti là dove l'indicatore di dispersione indica sottopoissonianità, in una 2^ parte del paragrafo, si procede a perequare le numerosità delle distribuzioni statistiche tramite medie mobili rispetto ai tempi t_i. E, una volta basata sulle distribuzioni statistiche perequate l'estimazione di ML della funzione di dispersione non-poissoniana (ai medesimi tempi t_i) di entrambi i processi di nascita HBP e HPAP, in una 3^ parte si determinano infine rispettive semplici spline-curve per cogliere l'evolversi della non-poissonianità quanto meno lungo il periodo dell'indagine.

Poiché la tendenza emersa fa pensare che gli effetti d'instabilità non si siano soddisfacentemente ridotti, nel § 5 si passa ad estimazioni di diversa concezione. In una 1^ parte, limitatamente all'HBP, onde evitare ripetizioni di discorso, si ricorre ad un approccio estimativo di tipo jackknife (Cfr.: Ferrante e Ferreri 1996) di struttura tale da comportare l'attenuazione degli effetti dei fattori d'instabilità ritenuti più salienti.

In una 2^ parte, relativamente a ciascuna delle *r*=31 distribuzioni statistiche considerate viene invece richiamato (Cfr.: Ferrante e Ferreri 1997) e praticato un approccio consistente in una procedura di estimazione di tipo bootstrap, che è siglata con CB (da

central bootstrap) in quanto imperniata su un criterio di dichiarazione di *outlier* imperniato su soglie di esclusione connesse tanto al numero *c* delle classi della distribuzione statistica quanto alla relativa media aritmetica.

Nel § 6 viene dato un quadro riassuntivo del lavoro in cui sono richiamate le conclusioni tratte dai confronti basati sui risultati dei diversi approcci seguiti. Conclusioni che fanno quanto meno notare, in primo luogo, che al crescere t_i o, volendo, del numero degli incidenti praticamente subiti dai soggetti, gli effetti d'instabilità vanno riducendosi fino a divenire trascurabili intanto che la sovrapoissonianità diviene sempre più accentuata e, poi, che l'estimazione degli ultimi due approcci si appalesa preferibile grazie ai profili che comportano per la *process-index function*.

A parte il tipo di estimazione, ne consegue comunque che, quando la circostanza fenomenica porta ad assumere a riferimento il processo di Poisson, decisioni prese tramite analisi della sovradispersione basate sull'ottica atemporale vanno riguardate perlomeno insufficientemente motivate, visto che tanto il tipo della dispersione non-poissoniana della distribuzione statistica delle unità d'osservazione rispetto al numero degli eventi su di esse registrati quanto, poi, l'entità della sovradispersione, si manifestano connessi al numero degli accadimenti del reale processo di nascita.

2 − ***Sull'esigenza di inquadramento euristico della problematica della sovradispersione.***

2.1 − *Un esempio d'uso incauto dei modelli di sovradispersione*.

A motivare l'opportunità di distinguere il filone delle ricerche squisitamente teorico-formali rispetto a quello delle indagini d'intento euristico, si pensa (Ferreri, 2000) sia sufficiente uno degli esempi dati da Gelfand & Dalal (1990). Costoro, dopo aver constatato che certe distribuzioni statistiche manifestavano sovradispersione rispetto alla famiglia uniparametrica di modelli di probabilità assunta a riferimento, stando all'usuale ottica atemporale hanno sviluppato una procedura test per saggiarla e basare poi sui dati l'individuazione del membro della famiglia a due parametri ritenuta capace della stessa sovradispersione.

Per illustrare la procedura, e magari sollecitare il lettore a non scoraggiarsi davanti a tanta poca semplicità, si sono serviti delle due distribuzioni statistiche del prospetto qui riportato.

La 1^ dà la distribuzione, rispetto al numero dei maschi, di 6115 traghettamenti (sibships) di 12 unità per volta, avvenuti in Sassonia nel periodo 1876-85, nel qual

caso come schema di riferimento è stata considerata la distribuzione binomiale.

La 2^ fornisce la distribuzione di 9461 guidatori belgi rispetto al numero degli incidenti di cui sono stati vittima in un anno, per la sovrapoissonianità della quale, la procedura di detti autori ha implicato le frequenze teoriche $p_x^{(B)}$.

1^) *Sibship data (Sokal and Rohlf, 1973) with fitted probabilities*

	n_x	f_x	$p_x^{(A)}$	$p_x^{(PA)}$	$p_x^{(NB)}$
0	3	0.0005	0.0004	0.0003	0.0003
1	24	0.0039	0.0038	0.0034	0.0029
2	104	0.0170	0.0177	0.0166	0.0153
3	286	0.0468	0.0520	0.0505	0.0492
4	670	0.1096	0.1088	0.1079	0.1086
5	1033	0.1689	0.1706	0.1707	0.1746
6	1343	0.2196	0.2057	0.2060	0.2102
7	1112	0.1818	0.1922	0.1923	0.1927
8	829	0.1356	0.1380	0.1388	0.1351
9	478	0.0782	0.0743	0.0760	0.0721
10	181	0.0296	0.0285	0.0301	0.0288
11	45	0.0074	0.0070	0.0074	0.0083
12	7	0.0011	0.0008	0.0000	0.0017
13					0.0002
	χ^2		(14.54)	13.72	17.98
	ν		(10)	8	9

2^) *Accident data (Seal, 1969) with fitted probabilities*

	n_x	f_x	$p_x^{(B)}$	$p_x^{(PA)}$	$p_x^{(NB)}$
0	7840	0.8287	0.8286	0.8292	0.8294
1	1317	0.1392	0.1325	0.1357	0.1362
2	239	0.0253	0.0302	0.0282	0.0271
3	42	0.0044	0.0068	0.0056	0.0057
4	14	0.0015	0.0015	0.0011	0.0012
5	4	0.0004	0.0003	0.0002	0.0003
6	4	0.0004	0.0001		0.0001
7	1	0.0001			
	χ^2		(25.92)	15.41	8.81
	ν		(3)	2	2

Per ciascuna distribuzione statistica, di fianco alle frequenze osservate, oltre alle corrispondenti teoriche $p_x^{(A)}$ e $p_x^{(B)}$, si sono riportate quelle desunte impiegando, quali modelli a due parametri, la *distribuzione binomiale negativa* (NBD) e la *distribuzione di Pólya-Aeppli* (PAD).

Ebbene, indipendentemente dalle indicazioni fornite dal χ_ν^2 o da qualche altro test di analoghe finalità, un semplice confronto tra la 2^ distribuzione statistica e le rispettive teoriche non può che portare a chiedersi *come si possa poggiare l'adeguatezza di*

una procedura come quella degli autori citati o altro tortuoso procedimento del genere (Lindsay, 1986) *sull'adattamento che si raggiunge sulla coda destra della distribuzione statistica quando questa coda – vista relativa alle ultime 5 classi – quota appena lo 0.69%, circa, del totale dei soggetti.*

In situazioni del genere, lo spostamento di una sola unità da una classe all'altra della coda destra può invero essere sufficiente a causare il ribaltamento della conclusione tratta, nonostante i fenomeni d'instabilità siano alquanto attutiti quando il numero degli eventi (incidenti) è elevato come quello dell'esempio in questione.

2.2 − *Effetti d'instabilità e struttura dell'evidenza statistica di processi di nascita: casi illustrativi.*

L'esempio di detti Autori induce così ad indirizzare l'attenzione sui "benedetti" effetti d'instabilità, anche se raramente i ricercatori mostrano di ritenere che valga la pena tenerne conto nell'investigazione in termini di distribuzioni statistiche di dati di conto frutto di manifestazioni di fenomeni che si svolgono secondo un processo di nascita, benché tali effetti possano essere responsabili di interpretazioni inattendibili, se non aberranti, della situazione investigata.

Un'idea di questa possibilità può evincersi anche dal confronto delle due seguenti distribuzioni statistiche:

$$\mathcal{D}_A = \left\{ \begin{array}{cccccccc} \text{x} & : & 0 & 1 & 2 & 3 & 4 & 5 \\ n_{A,x} & : & 37 & 32 & 31 & 4 & 3 & 1 \end{array} \right. , \qquad \mathcal{D}_B = \left\{ \begin{array}{cccccccc} \text{x} & : & 0 & 1 & 2 & 3 & 4 & 5 \\ n_{B,x} & : & 36 & 33 & 31 & 4 & 3 & 1 \end{array} \right.$$

($n_A = n_B = 108$, $M_A = 1.1389$ $M_B = 1.1481$), la seconda delle quali, tratta dalla Tav. I − e quindi una delle r=31 distribuzioni, qui considerate, di 108 operai di un'azienda metalmeccanica rispetto al numero degli incidenti loro capitati entro i rispettivi tempi t_i −, è vista ottenuta dalla prima spostando un'unità dalla 1^ alla 2^ classe.

L'indicatore di dispersione campionario $\tilde{I}_D$ (rapporto della varianza campionaria $\tilde{\sigma}^2$ alla corrispondente media M), nell'assumere rispettivamente i valori $\tilde{I}_{D,A}$ = 1.0155 e $\tilde{I}_{D,B}$ =0,9970, fa invero prendere atto di come lo spostamento di una sola unità da una classe alla contigua, tra quelle più numerose, possa comportare il passaggio da dispersione sovrapoissoniana a quella sottopoissoniana.

E, di fronte a due distribuzioni statistiche del genere, riguardabili relative a momenti successivi di un reale processo di nascita $\{ \mathcal{X}(t), t > 0 \}$, non basta certo pensare

che la sovrapoissonianità o la sottopoissonianità è più spesso tale che, per prescinderne, non è nemmeno il caso di avvalersi di test di significatività. Quando un tale processo è l'oggetto d'indagine, la presa di posizione non può infatti prescindere né dall'andamento dell'indicatore di dispersione campionario $\tilde{I}_D(t)$, né da un'analisi delle plausibili fonti d'instabilità. Ad attestarlo, nella fattispecie, vi è del resto che l'indicatore campionario $\tilde{I}_D$ presenta un andamento discorde col crescere del numero degli eventi contemplati dalle distribuzioni statistiche $\mathcal{D}_A(x)$ e $\mathcal{D}_B(x)$.

Un esempio illustrativo sotto questo aspetto è offerto dalle due seguenti distribuzioni statistiche

$$\mathcal{D}_C = \left\{ \begin{array}{lccccc} \text{x} & : 0 & 1 & 2 & 3 & 4 \\ n_{C,x}: & 40 & 37 & 25 & 4 & 2 \end{array} \right. , \qquad \mathcal{D}_D = \left\{ \begin{array}{lcccccc} \text{x} & : 0 & 1 & 2 & 3 & 4 & 5 \\ n_{D,x}: & 40 & 37 & 25 & 3 & 2 & 1 \end{array} \right. .$$

($n_C = n_D = 108$, M_C=0.99074, M_D=1.00926), la seconda delle quali – pure della Tav. I – differisce dalla prima per comprendere, in più, la classe (c=5, n_c=1), ottenuta collocando un operatore con 3 incidenti nella classe di 5 incidenti. Anche in tal caso l'instabilità affiora con lo spostamento di una sola unità, ma questa volta a prolungamento della coda della distribuzione. E poiché ne consegue $\tilde{I}_{D,C}$=0.9251 contro $\tilde{I}_{D,D}$=1.0183, è dato rilevare che lo spostamento comporta, sì, ancora un cambiamento nel tipo di dispersione, a partire però da una situazione di ben netta dispersione sottopoissoniana. Il che fa dire che se sussiste, com'è verosimile, un effetto d'instabilità sui valori dell'indicatore $\tilde{I}_D$, è comunque tale da non inficiarne un andamento concorde con quello del numero degli accadimenti.

Non si ha invece un cambiamento nel tipo di dispersione (in termini di $\tilde{I}_D$) ottenendo in modo analogo dalla distribuzione $\mathcal{D}_E$, desunta moltiplicando per 3 le numerosità della $\mathcal{D}_C$, la distribuzione $\mathcal{D}_F$

$$\mathcal{D}_E = \left\{ \begin{array}{lccccc} \text{x} & : 0 & 1 & 2 & 3 & 4 \\ n_{E,x} & : 120 & 111 & 75 & 12 & 6 \end{array} \right. , \quad \mathcal{D}_F = \left\{ \begin{array}{lcccccc} \text{x} & : 0 & 1 & 2 & 3 & 4 & 5 \\ n_{F,x}: & 120 & 111 & 75 & 11 & 6 & 1 \end{array} \right. .$$

($n_E = n_F$ =324, M_E=0.99074, M_F=0.99691). Da $\tilde{I}_{D,C} = \tilde{I}_{D,E} = 0.9251$ si passa invero al rapporto $\tilde{I}_{D,F} = 0.9566$, che lascia intuire il ruolo che può giocare un incremento nel numero delle unità d'osservazione in una situazione pressoché di simiglianza.

Anche se molteplici sono i quesiti e le perplessità che questi esempi possono sollevare, davanti ad evidenza statistica in termini di distribuzioni statistiche considerate per un reale processo di nascita $\{ \mathcal{X}(t), t > 0 \}$, è comunque il caso di rammentare come, allorquando la media M di una di esse non è di molto superiore ad 1 e il relativo numero di accadimenti è non troppo elevato, la cautela sia da riguardare pressoché d'obbligo in quanto *il tipo di dispersione implicato dalla distribuzione statistica non può dirsi senz'altro sufficiente per decidere in favore di un modello che lo contempli*.

A segnalarlo, tra i "grossi" nomi, c'è anche l'Efron (1986): per la fenomenologia che stava investigando, anziché preoccuparsi di rivedere la procedura usata in modo che prevedesse non solo sovradispersione, in primo luogo enfatizzò infatti ragioni fisiche di vario genere capaci di motivare la possibilità di effettiva sottodispersione.

Nell'analisi di distribuzioni statistiche di dati di conto del genere, trarre conclusioni sull'adeguatezza di un modello a prescindere da effetti d'instabilità può, insomma, dirsi quanto meno inopportuno, specie in mancanza di accentuata dispersione sovrapoissoniana e particolarmente là dove si fa ricorso a modelli che contemplano regressori con l'intento di individuarne il ruolo rispetto alle caratteristiche della medesima sovradispersione, dato che questa può essere alquanto sensibile a fattori d'instabilità non trascurabili.

Non ci si può pertanto che stupire quando, ciò nonostante, in lavori di ricerca concreta portati avanti tramite simulazioni imperniate sulla distribuzione binomiale negativa

$$P_x = \binom{-1/\alpha}{x} (1+\alpha\mu)^{-1/\alpha} \left(\frac{-\alpha\mu}{1+\alpha\mu}\right)^x, \; x=0,1,\ldots, \quad ; \quad \mu > 0, \quad \alpha > 0, \qquad (2.1)$$

di valore medio $\mathsf{E}(X) = \mu$ e parametro di sovradispersione α, con $\alpha > 0$, si accettano *sic et sempliciter* le stime positive di α e non si avanzano riserve di sorta per escludere i casi che comportano stime negative.

Va detto però che c'è stato anche chi (ad esempio, Piegorsch, 1990) non ha fatto il "salto della quaglia", pur limitandosi ad operare un'integrazione quasi esclusivamente tramite l'impiego della distribuzione binomiale come forma approssimata di quella binomiale positiva.

In particolare, non può essere comunque scordato che, per rimarcare l'imprescindibilità del problema dell'instabilità nell'analisi dei dati di conto – e quindi anche per le finalità del discorso che qui ci si propone di svolgere –, è stato fatto presente che:

> "*l'instabilità* (del contesto in questione), *oltre che peculiare del magico gioco degli accadimenti che si scolpiscono sullo scenario del nostro orizzonte, va vista anche*

come un qualcosa che s' innesta nell'evidenza statistica con l'articolazione che se ne dà per il fenomeno, direi proprio come se la si annodasse ai risvolti della struttura che l'umano crea alla base dei dati di conto per coglierne la manifestabilità".

3 − ***Due processi di nascita per inferenze sulla non-poissonianità.***

3.1 − *Il processo iperbinomiale e procedure inferenziali sui relativi process-trend e process-index function*.

▭ Il processo binomiale negativo. Quando l'oggetto d'attenzione è un reale processo di nascita $\{ X(t), t > 0 \}$ visto a riferimento poissoniano, secondo le constatazioni fatte il primo problema che affiora non concerne tanto la messa a punto di criteri per saggiare la discrepanza dalla situazione di poissonianità, come invece può forse evincersi dalla letteratura statistica, quanto l'inferimento sull'andamento della dispersione non-poissoniana nel tempo durante il periodo $(0.T]$ di osservazione, ed eventualmente l'individuazione di archi temporali in cui la non-poissonianità può ritenersi essenzialmente trascurabile.

Poiché tale andamento generalmente non gioca a sostegno del processo di Poisson e, cioè, di accadimenti puramente accidentali, ci si trova così a pensare come modificare le assunzioni che ne stanno alla base, le quali consistono, come si è detto, nel riguardare il set $(X_1(t),..,X_h(t),..,X_n(t))$ dei numeri degli eventi (incidenti) capitabili nel periodo $(0,t]$ alle n unità d'osservazione quale campione casuale, cioè costituito da v.c. indipendenti e identicamente distribuite (*i.i.d.*), di modello di probabilità poissoniano (1.6).

La via grandemente privilegiata per circostanze simili a quella di incidenti ripetibili nel posto di lavoro, cui si è detto di limitare il discorso, muove dal non ritenersi irragionevole che la soggettività delle unità d'osservazione (operatori) nel rapporto col mezzo motivi la supposizione di una propensione all'evento incidente generalmente diversa da soggetto a soggetto, ancorquando ciascuno possa vedersi caratterizzato da un processo di tipo poissoniano.

In tal modo, si abbandona la suddetta assunzione di campione casuale, in quanto viene meno il presupposto dell'identica distribuzione (*i.d.*), e in primo luogo si va a ritenere verosimile quella che riguarda:

(*i*) la v.c. $X(t)$ (numero degli incidenti nel periodo $(0,t]$) associata a ogni operatore ancora poissoniana, ma secondo una rispettiva tipica propensione ξ all'evento

incidente, che la (1.6a) fa ovviamente identificare nel saggio medio di eventi per intervallo unitario di tempo;

(*ii*) la propensione individuale quale livello di una variabile casuale Ξ di distribuzione plausibilmente unimodale sul supporto $\Re^+$, per la quale viene privilegiata la distribuzione gamma $\Xi \sim \mathcal{G}(\alpha\lambda, 1/\alpha)$ avente $\alpha\lambda$ e $1/\alpha$ rispettivamente quali parametri di scala e di forma. Distribuzione che in termini di funzione di densità (*d.f.*) fa perciò scrivere

$$\Xi \sim f_{\mathcal{G}}(\xi\,;\,\alpha\lambda,\tfrac{1}{\alpha}) = \frac{1}{(\alpha\lambda)^{1/\alpha}\Gamma(\alpha)} e^{-\frac{1}{\alpha\lambda}\xi} \xi^{\frac{1}{\alpha}-1}, \quad \text{con} \quad \begin{matrix}\xi > 0 \\ \alpha,\lambda > 0\end{matrix}. \qquad (3.1)$$

Su tale presupposto, alla generica unità d'osservazione resta così associata la v.c. doppia mista (X_t, Ξ) – discreta in $X_t = \mathrm{X}(t)$ e continua in Ξ – distribuita secondo la congiunta *d.f.* (quantica rispetto ad X_t e di probabilità nei riguardi Ξ)

$$(X_t, \Xi) \sim \mathcal{P}_t(x \mid \xi)\ f_{\mathcal{G}}(\xi\,;\,\alpha\lambda,\tfrac{1}{\alpha}) = \frac{(\xi t)^x e^{-\xi t}}{x!} \frac{1}{(\alpha\lambda)^{1/\alpha}\Gamma(\alpha)} e^{-\frac{1}{\alpha\lambda}\xi} \xi^{\frac{1}{\alpha}-1}, \quad \begin{matrix} x = 0,1,\cdots \\ \xi > 0 \end{matrix},$$

dove $\mathcal{P}_t(x \mid \xi)$ sta a designare la probabilità che, al tempo *t*, la v.c. poissoniana $X_t = X(t)$ assuma il valore intero non negativo $X_t = x$ subordinatamente al livello ξ della v.c. Ξ.

Per la distribuzione marginale $\int_0^\infty \mathcal{P}_t(x \mid \xi) f_{\mathcal{G}}(\xi\,;\,\alpha\lambda,\tfrac{1}{\alpha}) d\xi$ della v.c. $\mathrm{X}(t)$ consegue, come si sa, la distribuzione binomiale negativa (NBD)

$$P_{NB,x}(t) = \frac{\Gamma(\frac{1}{\alpha}+x)}{x!\Gamma(\frac{1}{\alpha})} (1+\alpha\lambda t)^{-\frac{1}{\alpha}} \left(\frac{\alpha\lambda t}{1+\alpha\lambda t}\right)^x, \quad \text{con} \quad \begin{matrix}\alpha,\lambda > 0 \\ x = 0,1,\cdots\end{matrix}, \qquad (3.2)$$

che, in quanto riguardabile valore medio $\mathsf{E}_{\mathcal{G}}[\mathcal{P}_t(x \mid \Xi)]$ dalla distribuzione poissoniana condizionata, è detta pure *distribuzione composta o mistura* della Poisson avente per misturante la distribuzione gamma (3.1).

Poiché la funzione generatrice dei momenti (*m.g.f.*)

$$\varphi_X(\tau) = \left(1 + \alpha\lambda t - \alpha\lambda t e^{\tau}\right)^{-\frac{1}{\alpha}}, \quad \text{con } \alpha,\lambda > 0, \qquad (3.3)$$

è la corrispettiva della NBD (3.2), è facile verificare che

$$\mu(t) = \mathsf{E}[X(t)] = \lambda t, \qquad \sigma^2(t) = Var[X(t)] = \mu(t)(1+\alpha\lambda t), \qquad (3.4)$$

da cui l'indicatore di dispersione

$$I_D(t) = \frac{\sigma^2(t)}{\mu(t)} = 1 + \alpha\lambda t\,, \qquad (3.5)$$

che, per $\alpha \to 0$, tende al valore unitario implicato dalla distribuzione (1.6) di Poisson, alla quale ad un tempo si riduce la NBD (3.2). Ovviamente, per t=T=1, la (3.2) va a coincidere con la (2.1) in quanto

$$\binom{-1/\alpha}{x}(-1)^x = \binom{1/\alpha + x - 1}{x} = \frac{\Gamma(1/\alpha + x)}{x!\,\Gamma(1/\alpha)}$$

e quindi $\mu = \mu(1) = \mathsf{E}[X(1)] = \lambda$.

Stando a tale configurazione fenomenica, il set $(X_1(t),..,X_h(t),..,X_n(t))$ dei numeri degli eventi (incidenti) capitabili nel periodo $(0,t]$ alle n unità d'osservazione torna nuovamente ad imporsi come campione casuale, ma di modello di probabilità (*p.m.*) dato dalla NBD (3.2), per inferenze sui cui parametri è ormai disponibile una vasta letteratura statistica.

▭ L'extended Pólya process. A questo punto, non può dirsi però che, con la distribuzione binomiale negativa (3.2), si sia giunti alla soluzione del problema emerso nella premessa dato che, come mostra la (3.5), la NBD contempla α>0. E in quanto caratterizzata da un valore positivo del parametro α, detto di dispersione (Anraku and Yanagimoto, 1990) o *process-index*, può al più assolvere il ruolo di modello di probabilità per le circostanze in cui l'indicatore campionario $\tilde{I}_D(t)$ segnala dispersione sovrapoissoniana.

Inoltre, anche là dove essa è vista in qualche misura assumibile, in pratica solo raramente la circostanza è tale da ritenersi ragionevole stare al *process-trend* lineare $\mu(t) = \lambda t$, come comporta il processo poissoniano omogeneo.

Per superare questa limitazione, si è pensato (Ferreri, 1983) di modificare la parte (*ii*) dell'assunzione che supporta la NBD (3.2) in modo da aversi, come *process-trend*, non già una specificata espressione, ma una funzione $\mu(t)$, naturalmente dai seguenti requisiti

$$\mu(0) = 0\,, \quad \mu'(0) \geq 0\,; \qquad \mu(t),\ \mu'(t) \geq 0\,, \quad \text{per } t > 0\,, \qquad (3.6)$$

richiesti dal ruolo, sulla quale compiere inferenze in termini dell'evidenza statistica nella forma considerata.

A tal fine si è visto che in qualità di misturante, anziché la (3.1), bastava assumere la distribuzione gamma

$$\Xi(t) \sim f_{\mathcal{G}}(\xi;\ \alpha\frac{\mu(t)}{t}, \frac{1}{\alpha})\,, \quad \text{con } \alpha > 0\,, \qquad (3.7)$$

avente $\alpha\frac{\mu(t)}{t}$ quale parametro di scala e sempre $\frac{1}{\alpha}$ come parametro di forma.

Procedendo come si è fatto per arrivare alla (3.2), con quest'ultima distribuzione misturante si giunge infatti a

$$P_{EP,x}(t)=\frac{\Gamma(\frac{1}{\alpha}+x)}{x!\Gamma(\frac{1}{\alpha})}\left[1+\alpha\mu(t)\right]^{-\frac{1}{\alpha}}\left[\frac{\alpha\mu(t)}{1+\alpha\mu(t)}\right]^{x},\quad \text{con}\quad \begin{matrix}\alpha>0\\ x=0,1,\dots\end{matrix}, \qquad (3.8)$$

che si è convenuto di denominare *extended Pólya distribution* (EPD) in quanto estende la distribuzione del Pólya, implicata dall'omonimo processo imperniato sulla funzione di intensità che discende, per $\mu(t)=\lambda t$, da quella $p_x(t)=\frac{1+\alpha x}{1+\alpha\mu(t)}\mu'(t)$ su cui può essere vista basata la (3.8).

A questa ovviamente corrisponde la funzione generatrice di probabilità (*p.g.f.*) $\pi_X(z)=\left[1+\alpha\mu(t)-\alpha\mu(t)z\right]^{-\frac{1}{\alpha}}$ e quindi la *m.g.f.*

$$\varphi_X(\tau)=\left[1+\alpha\mu(t)-\alpha\mu(t)e^{\tau}\right]^{-\frac{1}{\alpha}},\quad \text{con } \alpha>0, \qquad (3.9)$$

che fa ottenere

$$\mathsf{E}[X(t)]=\mu(t),\quad \sigma^2(t)=\mu(t)[1+\alpha\mu(t)]\quad \text{e}\quad I_D(t)=\frac{\sigma^2(t)}{\mu(t)}=1+\alpha\mu(t). \qquad (3.10)$$

Poiché dall'espressione della varianza discende

$$\alpha=\frac{\sigma^2(t)-\mu(t)}{\mu^2(t)}, \qquad (3.10a)$$

calcolandola in termini della media M_i e della varianza $\tilde{\sigma}_i^2$ a partire da ciascuna delle distribuzioni statistiche (1.8) corrispondenti ai tempi t_i per i quali si ha $\sigma^2(t_i)>\mu(t_i)$, cioè sovradispersione – come richiesto dalla (3.8) –, si perviene ad un *set* ordinato di valori $\tilde{\alpha}_i$ utile per l'indicazione che può fornire sull'andamento dei rispettivi α_i rispetto a quello costante contemplato dalla EPD

▭ Il processo iperbinomiale (HBP). Se è vero che la EPD (3.8) non è utilizzabile per inferenze sull'andamento di non-poissonianità segnalato dall'indicatore di dispersione campionario $\tilde{I}_D(t)$ in quanto, con $\alpha>0$ e indipendente dal tempo, consente di approfondire il discorso sul *process-trend* unicamente nelle circostanze di sovrapoissonianità costante, è pur vero che induce a chiedersi: se si è fatto in modo di superare la rigidità in merito al *process-trend* rendendolo oggetto d'inferenza, perché non pro-

cedere analogamente nei riguardi del *parametro di dispersione* α dato che, in pratica, si appalesa alquanto inverosimile supporlo costante nel tempo?

Non ha, del resto, gran senso limitare l'attenzione all'ipotesi H_o di poissonianità visto che, nella realtà di un processo di nascita, generalmente segna soltanto un particolare punto dell'andamento della relativa funzione di dispersione non costante durante il periodo del proprio svolgimento. Naturalmente, il senso riaffiora prediligendo, come si usa fare, l'ottica atemporale, per la quale H_o va a segnare una sorta di stato di idealità o di situazione limite, in armonia con quanto suggerisce il parametro di forma della distribuzione gamma misturante (3.7) visto che tende a $+\infty$ per $\alpha \to 0^+$.

Prendendo le mosse da queste considerazioni si è passati ad introdurre (Ferreri, 1990, 1992) un processo di nascita $\{X(t), t > 0\}$, denominato *processo iperbinomiale* (HBP), la distribuzione della cui v.c. $X(t)$ sul piano formale può vedersi implicata dalla funzione generatrice di probabilità (*p.g.f.*)

$$\pi_X(z) = \left[1 + \alpha(t)\mu(t) - \alpha(t)\mu(t)z\right]^{-\frac{1}{\alpha(t)}} \tag{3.11}$$

desunta da quella della EPD assumendo, anziché α costante, una funzione $\alpha(t)$ dipendente dal tempo e dai requisiti imposti dal ruolo.

Dalla (3.11) è dato infatti pervenire alla distribuzione iperbinomiale (HBD)

$$P_x(t) = \binom{\frac{-1}{\alpha(t)}}{x} \left[1 + \alpha(t)\mu(t)\right]^{\frac{-1}{\alpha(t)}} \left[\frac{-\alpha(t)\mu(t)}{1 + \alpha(t)\mu(t)}\right]^x, \quad \text{per } x = 0,1,..,\kappa(t), \tag{3.12}$$

$$\text{dove evidentemente} \quad \kappa(t) = \begin{cases} \infty & \text{se } \alpha(t) > 0 \\ [-1/\alpha(t)] & \text{se } \alpha(t) < 0 \end{cases}, \tag{3.12a}$$

$[-1/\alpha(t)]$ indicando la parte intera di $-1/\alpha(t)$.

Come in altre occasioni, tenendo presente che la funzione rischio cumulativo (*c.h.f.*) è definita da $\Lambda(t) = -\log P_0(t)$, dove $P_0(t) = P[X(t) = 0] = R(t)$, anche qui ci si limita a supporre che le due funzioni $\mu(t)$ e $\alpha(t)$ siano differenziabili e tali che:

(*i*) $\mu(t), \mu'(t) > 0$, $\forall t > 0$, con $\lim_{t \downarrow 0} \mu(t) = 0$;

(*ii*) $\alpha(t) > -1/\mu(t)$, $\forall t > 0$, nonché implicante

$$\Lambda(t) = \frac{1}{\alpha(t)} \log\left[1 + \alpha(t)\mu(t)\right] \quad \text{con } \Lambda'(t) > 0, \ \forall t > 0, \text{ e } \lim_{t \downarrow 0} \Lambda(t) = 0, \tag{3.12b}$$

come sempre si ha se è valida la relazione

$$\frac{\mu'(t)}{\mu(t)} \geq -\frac{\alpha'(t)}{\alpha(t)} < 0 \quad \text{con} \quad \lim_{t\downarrow 0} \alpha(t)\mu(t) = 0 \,. \tag{3.12c}$$

La denominazione usata, e quindi la sigla HBP, può dirsi imposta dal fatto che:

1) quando $\alpha(t) > 0$, la distribuzione (3.12) si riduce alla distribuzione iperbinomiale negativa (NHBD)

$$P_x(t) = \frac{\Gamma(\frac{1}{\alpha(t)}+x)}{x!\Gamma(\frac{1}{\alpha(t)})}\left[1+\alpha(t)\mu(t)\right]^{-\frac{1}{\alpha(t)}}\left[\frac{\alpha(t)\mu(t)}{1+\alpha(t)\mu(t)}\right]^x, \quad \text{con} \quad \begin{matrix}\alpha(t)>0\\ x=0,1,\dots\end{matrix}, \tag{3.13}$$

dell'omonimo processo iperbinomiale negativo di *process-trend* $\mu(t)$ e *process-index function* $\alpha(t)$, la quale distribuzione comporta

$$\mathrm{E}[X(t)] = \mu(t) \;; \qquad \sigma^2(t) = Var[X(t)] = \mu(t)[1+\alpha(t)\mu(t)] \,; \tag{3.14}$$

2) quando $\alpha(t) < 0$, l'HBP s'identifica invece col processo iperbinomiale positivo (PHBP), la cui distribuzione (3.12) implica $S(t) = \Sigma_{x=0}^{\kappa(t)} P_x(t) \leq 1$.

Ovviamente, il segno uguale vale soltanto se $-1/\alpha(t)$ è un numero naturale $\kappa(t)$, nel qual caso, la (3.12) si riduce alla distribuzione di tipo binomiale (TBD)

$$P_x(t) = \binom{\kappa(t)}{x}\left(1-\frac{\mu(t)}{\kappa(t)}\right)^{\kappa(t)}\left(\frac{\mu(t)}{\kappa(t)-\mu(t)}\right)^x \quad , \quad x=0,1,\dots,\ \kappa\,(\mathrm{t}) \,, \tag{3.15}$$

considerata dal Binet (1986), per la quale sono valide le espressioni (3.14), la seconda delle quali nella fattispecie segnala una circostanza di dispersione sottopoissoniana. Ponendo $\mu(t) = \kappa(t)\,p(t)$, dalla $\mathrm{TBD}(\kappa(t), \mu(t))$ ovviamente discende l'usuale distribuzione binomiale $\mathrm{BD}(\kappa(t), p(t))$, che fa capire come in pratica le (3.14) possono essere assunte per $\kappa(t) < -1/\alpha(t)$ soltanto nella misura consentita dall'ordine di grandezza del termine $P_{\kappa(t)+1}(t) = 1 - S(t)$, che viene perciò comunemente scritto a completamento della distribuzione;

3) quando $\alpha(t) \to 0$ per $t \to \tau_P > 0$, la (3.12) tende alla distribuzione del processo di Poisson non-omogeneo avente $\mu(\tau_P)$ come process-trend a τ_P.

▭ Letture stratificatorie della distribuzione iperbinomiale negativa (NHBD). Anche se la distribuzione (3.12) è stata qui introdotta su un piano squisitamente formale, per $\alpha(t) > 0$ essa è desumibile sulla base di più approcci (Ferreri, 1990) e, in particolare,

in infiniti modi come distribuzione composta di Poisson avente per misturante una distribuzione gamma con parametro di forma $1/\alpha(t)$.

È facile verificare infatti che la (3.13) può essere espressa secondo lo schema di stratificazione

$$P_x(t) = \int_0^\infty \frac{1}{x!}\left(\frac{\alpha(t)\mu(t)}{\sigma_{\xi}(t)}\xi\right)^x e^{-\frac{\alpha(t)\mu(t)}{\sigma_{\xi}(t)}\xi} \cdot \frac{[\sigma_{\xi}(t)]^{-\frac{1}{\alpha(t)}}}{\Gamma(\frac{1}{\alpha(t)})} e^{-\frac{1}{\sigma_{\xi}(t)}\xi} \xi^{\frac{1}{\alpha(t)}-1} d\xi, \quad \begin{matrix} x=0,1,\cdots, \\ \xi>0 \end{matrix},$$

che viene usualmente segnalato tramite la relazione

$$\text{NHBD}\left(\mu(t),\alpha(t)\right) = \text{Poisson}\left(\frac{\alpha(t)\mu(t)}{\sigma_{\xi}(t)}\xi\right) \underset{\xi}{\Lambda} \text{Gamma}(\sigma_{\xi}(t), \tfrac{1}{\alpha(t)}), \qquad (3.16)$$

dove il parametro di scala della distribuzione gamma, anziché univocamente determinato, può essere identificato in una funzione $\sigma_{\xi}(t) = \sigma(\alpha(t),t)$, $t>0$, positiva e differenziabile e tale che

1) $\frac{\mu'(t)}{\mu(t)} \geq \frac{\sigma'_{\xi}(t)}{\sigma_{\xi}(t)} - \frac{\alpha'(t)}{\alpha(t)}$,

2) $\lim_{t\downarrow 0}\sigma_{\xi}(t) = 0$ con $\lim_{t\downarrow 0}\frac{\sigma'_{\xi}(t)}{\alpha'(t)} > 0$ se $\lim_{t\downarrow 0}\alpha(t) = 0$, allorché la distribuzione gamma misturante si riduce alla distribuzione degenere.

Per $\sigma_{\xi}(t) = \alpha(t)\frac{\mu(t)}{t}$, dalla (3.16) si desume

$$\text{NHBD}\ \left(\mu(t),\alpha(\text{t})\right) = \text{Poisson}\left(\xi t\right) \underset{\xi}{\Lambda} \text{Gamma}(\alpha(t)\tfrac{\mu(t)}{t}, \tfrac{1}{\alpha(t)}) \qquad (3.16a)$$

indicante, da un lato, che la distribuzione misturata è quella (per lo più considerata) di un processo di Poisson omogeneo e, da un altro lato, che in nessun caso la distribuzione gamma misturante può essere indipendente dal tempo *t* se il parametro di forma $1/\alpha(t)$ ne è invece dipendente, in armonia col fatto che, in uno schema di stratificazione, la misturante va concepita come una funzione di stato (Ferreri, 1984).

Per $\sigma_{\xi}(t) = \alpha(t)$, dalla (3.16) discende invece

$$\text{NHBD}\ \left(\mu(t),\ \alpha(\text{t})\right) = \text{Poisson}\left(\mu(t)\xi\right) \underset{\xi}{\Lambda} \text{Gamma}(\alpha(t),\ \tfrac{1}{\alpha(t)}), \qquad (3.16b)$$

da cui appare che, mentre il valore medio della poissoniana misturata è $\mu(t)\Xi$, la distribuzione gamma misturante va a dipendere soltanto da $\alpha(t)$, segnando così una

forma della gamma ampiamente usata nell'analisi statistica (Cfr., ad esempio, Ferreri, 1983; Regazzini, 1983; Grogger, 1990).

Per $\sigma_{\varphi}(t) = \mu(t)$, dalla (3.16) consegue infine

$$\text{NHBD}\left(\mu(t), \alpha(\text{t})\right) = \text{Poisson}\left(\alpha(t)\xi\right) \underset{\xi}{\Lambda} \text{Gamma}(\mu(t), \tfrac{1}{\alpha(t)}), \qquad (3.16\text{c})$$

che, se solleva perplessità per il valore medio della distribuzione di Poisson misturata, può attirare invece attenzione quando il valore medio $\mu(t)$ del modello di probabilità $\text{NHBD}\left(\mu(t), \alpha(\text{t})\right)$ è visto in termini di covariate dato che, in tal caso, ne risulta dipendente il parametro di scala della distribuzione gamma misturante.

Quest'ultima constatazione fa capire però che, nelle circostanze in cui si ritiene di formalizzare le supposizioni ritenute ragionevoli in termini di espressioni parametriche della distribuzione gamma, non è detto che la NHBD che rimane implicata sia della forma (3.13). Il che porta a distinguere le formalizzazioni della distribuzione gamma che rientrano nella (3.16) da quelle in cui ciò non avviene.

Tra le riparametrizzazioni della NHBD che, per l'impiego che trovano, vanno quantomeno tenute presenti:

A) quella basata sulla posizione $\alpha(t) = \frac{1}{\gamma(t)}$, che dalla (3.16) fa ottenere la relazione

$$X(t) \sim \text{NHBD}\left(\mu(t), \gamma(t)\right) = \text{Poisson}\left(\frac{\mu(t)}{\gamma(t)\sigma_{\varphi}(t)}\xi\right) \underset{\xi}{\Lambda} \text{Gamma}(\sigma_{\varphi}(t), \gamma(t)), \quad (3.17)$$

con $\text{E}[X(t)] = \mu(t)$ e $\sigma^2(t) = Var[X(t)] = \mu(t)[1 + \frac{1}{\gamma(t)}\mu(t)]$. Relazione che viene considerata soprattutto per $\sigma_{\varphi}(t) = \frac{\mu(t)}{t\gamma(t)}$, allorché si riduce a

$$\text{NHBD}\left(\mu(t), \gamma(\text{t})\right) = \text{Poisson}\left(\xi t\right) \underset{\xi}{\Lambda} \text{Gamma}(\tfrac{\mu(t)}{t\gamma(t)}, \gamma(t)).$$

Ovviamente, tale fattorizzazione non segna granché di vantaggioso dato che consiste soltanto nel considerare la sovradispersione della NHBD in termini di $\gamma(t)$;

B) la seguente

$$X(t) \sim \text{NHBD}\left(\mu(t), \beta(t)\right) = \text{Poisson}\left(\frac{\beta(t)}{\sigma_{\varphi}(t)}\xi\right) \underset{\xi}{\Lambda} \text{Gamma}\left(\sigma_{\varphi}(t), \frac{\mu(t)}{\beta(t)}\right), \quad (3.18)$$

con $\mathsf{E}[X(t)] = \mu(t)$ e $\sigma^2(t) = Var[X(t)] = \mu(t)[1+\beta(t)]$, implicata dal porre $\alpha(t) = \frac{\beta(t)}{\mu(t)}$ nella (3.16). Non può sorprendere il favore trovato dalla (3.18), spesso indicata come Form I della NHBD (Cfr.: Cameron and Trivedi 1986) dato che l'indicatore di sovradispersione $I_D(t) = \sigma^2(t)/\mu(t) = 1+\beta(t)$ dipende unicamente dall'omonima funzione-parametro $\beta(t)$.

Per $\sigma_{\mathcal{G}}(t) = \frac{\beta(t)}{t}$ la (3.18) si riduce a

$$\text{NHBD}\left(\mu(t), \beta(t)\right) = \text{Poisson}\left(\xi t\right) \underset{\xi}{\wedge} \text{Gamma}\left(\frac{\beta(t)}{t}, \frac{\mu(t)}{\beta(t)}\right),$$

cui viene spesso rivolta l'attenzione in pratica quando le supposizioni vengono incentrate sulla distribuzione gamma misturante;

C) lo schema stratificatorio (Cfr.: Husman, Hall and Griliches 1984)

$$X(t) \sim \text{NHBD}\left(\delta(t), \phi(t)\right) = \text{Poisson}\left(\frac{1}{\delta(t)\sigma_{\mathcal{G}}(t)}\xi\right) \underset{\xi}{\wedge} \text{Gamma}(\sigma_{\mathcal{G}}(t), \phi(t)), \quad (3.19)$$

con $\mathsf{E}[X(t)] = \frac{\phi(t)}{\delta(t)}$ e $\sigma^2(t) = Var[X(t)] = \mu(t)[1+\frac{1}{\phi(t)}\mu(t)]$, che consegue da quello (3.16) per $\mu(t) = \frac{\phi(t)}{\delta(t)}$ e $\alpha(t) = \frac{1}{\phi(t)}$, ovviamente con $\delta(t), \phi(t) > 0$ e $\frac{\phi'(t)}{\phi(t)} - \frac{\delta'(t)}{\delta(t)} > 0$ per $t > 0$.

Assumendo, in particolare, $\sigma_{\mathcal{G}}(t) = \frac{1}{t\delta(t)}$ la (3.19) porta alla relazione

$$\text{NHBD}\left(\delta(t), \phi(t)\right) = \text{Poisson}\left(\xi t\right) \underset{\xi}{\wedge} \text{Gamma}\left(\frac{1}{t\delta(t)}, \phi(t)\right).$$

Anche se le parametrizzazioni appena richiamate sono state suggerite soprattutto da considerazioni in merito alla gamma misturante, per lo più questa è stata vista come unica, nonché tacitamente ancorata alla distribuzione di un processo di Poisson omogeneo. Dato però che assai di frequente le situazioni oggetto di studio sono tali da far configurare un processo poissoniano non omogeneo, la (3.17), la (3.18) e la (3.19) danno modo di capire come l'esplicitazione della gamma misturante, e cioè della funzione-parametro di scala $\sigma_{\mathcal{G}}(t)$, vada concepita in relazione al tipo di non-omogeneità del processo poissoniano.

È poi quasi inutile aggiungere che dai requisiti formali indicati per la funzione $\sigma_{\varphi}(t)$ della relazione (3.16) attinente alla (3.12) con $\alpha(t) > 0$, conseguono facilmente quelli delle funzioni-parametro specificanti le parametrizzazioni indicate per la NHBD.

▭ Estimazione di ML della distribuzione iperbinomiale (HBD). Per l'estimazione di massima verosimiglianza (ML) della (3.12) in termini di ciascuna *i*-esima delle distribuzioni statistiche (1.8), torna utile rilevare che, calcolata la rispettiva media aritmetica $M = M_i = \frac{1}{n}\Sigma_{x=0}^{c} x n_{xi}$, con $c = c_i$, la funzione di log-verosimiglianza può essere scritta nella forma (Ferreri, 1996)

$$l(\mu,\alpha) = \sum_{x=0}^{c} n_x log\left\{(-\mathrm{O})^x \binom{-1/\alpha}{x}\right\} - \frac{n}{\alpha} log(1+\alpha\mu) + nM \cdot log \frac{\mathrm{O}\alpha\mu}{1+\alpha\mu} \qquad (3.20)$$

dove O= -1 per $\alpha = \alpha_i < 0$ oppure O= +1 per $\alpha = \alpha_i > 0$.

Ma, sia per $\alpha > 0$ che per $\alpha < 0$, è facile verificare che la prima derivata di $l(\mu,\alpha)$ rispetto a μ porta allo stimatore $\hat{\mu} = M$ e che, in termini di questo stimatore, l'equazione di ML di α è esprimibile nella forma

$$\phi(\alpha) = \phi_1(\alpha) - \phi_2(\alpha) = 0, \qquad (3.21)$$

dove

$$\phi_1(\alpha) = -\alpha \sum_{h=0}^{c-1} \frac{1-F_h}{1+\alpha h}, \quad \phi_1'(\alpha) = -\sum_{h=0}^{c-1} \frac{1-F_h}{(1+\alpha h)^2}, \quad \phi_1''(\alpha) = 2\sum_{h=0}^{c-1} \frac{(1-F_h)h}{(1+\alpha h)^3}; \qquad (3.21a)$$

$$\phi_2(\alpha) = -\log(1+\alpha M), \quad \phi_2'(\alpha) = -\frac{M}{1+\alpha M}, \quad \phi_2''(\alpha) = \left(\frac{M}{1+\alpha M}\right)^2 \qquad (3.21b)$$

con $F_h = F_{hi} = \Sigma_{x=0}^{h} f_{xi}$, essendo $f_{xi} = n_{xi}/n$.

Per risolvere iterativamente l'equazione (3.21) di ML, torna utile rammentare, da un lato, che con la (3.10a) il metodo dei momenti (MM) designa, quale valore iniziale, quello fornito dallo stimatore

$$\tilde{\alpha} = \frac{\tilde{\sigma}^2 - M}{M^2} \qquad (3.21c)$$

e, da un altro lato, che per tale equazione va considerato (Ferreri, 1997)

$$\alpha > max(-\frac{1}{c-1}, -\frac{1}{M}), \qquad \text{con} \quad c > 1, \qquad (3.22)$$

(ovviamente, per $c = 1$ l'equazione avrebbe la sola radice $\alpha_0 = 0$) poiché:

(i) quando $\tilde{\sigma}^2 > M$, essa ha una sola radice $\alpha_0 > 0$, cioè positiva;

(ii) quando, invece, $\tilde{\sigma}^2 < M$, il caso di $M < c-1$ va distinto da quello di $M > c-1$.

Infatti,

a) per $\tilde{\sigma}^2 < M$, con $M < c-1$, l'equazione (3.21) ha sempre un'unica radica negativa $\alpha_0 > -\frac{1}{c-1}$

b) per $\tilde{\sigma}^2 < M$, con $M > c-1$, la stessa equazione (3.21) o non ha soluzione reale o ne ha due negative nell'intervallo $(-\frac{1}{M}, 0)$, la più grande delle quali dà la stima $\hat{\alpha}$ di ML;

c) per $\tilde{\sigma}^2 < M$, con $M = c-1$, l'equazione (3.21) ha ovviamente una sola radice negativa.

È evidente che, per $\tilde{\sigma}^2 < M$, l'assunzione della radice α_0 della (3.21) come stima $\hat{\alpha} = \hat{\alpha}_t$ di ML va vista congiuntamente alla relazione $\mu_t = \mathsf{E}[X(t)]$, che è valida soltanto se il valore di $-1/\alpha(t)$ è intero, come si evince dal fatto che la (3.20) è considerabile come $logL(\mu, \alpha)$ nella misura che la (3.15) approssima una distribuzione completa. Del resto, la funzione $l(\mu, \kappa) = logL(\mu, \kappa)$, a cui ci si riduce, concerne la stessa (3.15), che, sotto il profilo dell'estimazione, si pone continua rispetto a μ , ma discreta con riguardo all'argomento κ , portando così all'impiego di criteri che contemplino κ intero.

In pratica, pertanto, con $\tilde{\sigma}^2 < M$, è opportuno scegliere anzitutto di rifarsi al modello (3.13), che contempla α come parametro, o a quello (3.15). Infatti, nel primo caso ci si trova nell'estimazione di ML e, con ciò, ad affrontare le situazioni dianzi indicate, mentre nel secondo è sufficiente (Ferreri, 1996) trattare, come vuole la (3.15), l'estimazione di ML in contesto discreto. In questo ambito, una volta ottenuto lo stimatore $\hat{\mu} = M$ di ML da $\partial l(\mu, \kappa)/\partial \mu = 0$, l'intero positivo κ che massimizza $l(\kappa) = l(\kappa, M)$ ovviamente segue dalla disuguaglianza $l(\kappa - 1) \le l(\kappa) > l(\kappa + 1)$ con $\kappa \ge c = max(x)$, la quale porta ad identificare la stima di κ nel più grande numero naturale per cui vale la relazione

$$\Delta l(\kappa) = n\left\{ \sum_{x=0}^{c} f_x log \frac{\kappa+1}{\kappa-x+1} + \Delta[(\kappa - M)\log(\kappa - M)] - \Delta(\kappa \log \kappa) \right\} < 0 , \qquad (3.23)$$

$$\text{con} \quad \kappa \ge c .$$

È, del resto, evidente che riguardare κ continuo equivale a considerare il modello (3.13) e poi ad assumere il valore intero tramite arrotondamento.

La matrice quadrata diagonale di elementi

$$I_{11}(\mu,\alpha)=\frac{n}{\mu(1+\alpha\mu)}$$
$$I_{22}(\mu,\alpha)=\frac{n}{\alpha^2}\left\{-\sum_{h=0}^{\kappa}\frac{1+2\alpha h}{(1+\alpha h)^2}(1-\Phi_h)+\frac{2}{\alpha}log(1+\alpha\mu)-\frac{\mu}{1+\alpha\mu}\right\}, \qquad (3.24)$$

dove $\Phi_h = P(X \le h) = \Sigma_{x=0}^{h} P_x$, dà infine la matrice d'informazione del Fisher ammesso che essa sia similarmente riguardata quando $\tilde{\sigma}^2 < M$ con $M < c-1$: soltanto su tale base l'estimazione di ML e il calcolo della matrice d'informazione non richiede la preliminare distinzione del caso di $\alpha>0$ da quello di $\alpha<0$.

3.2 − *Una estensione della distribuzione di Pólya-Aeppli (PAD) come distribuzione di un birth process a dispersione non poissoniana.*

▭ Caratteristiche e ruolo di una riparametrizzazione della PAD nell'analisi della dispersione sovrapoissoniana. L'intento che sollecita ad introdurre e ad avvalersi dell'HBP per investigare sul process-trend $\mu(t)$ e sull'andamento della dispersione non-poissoniana di un reale processo di nascita, porta quasi inevitabilmente a ripercorrere l'iter con riguardo alla distribuzione di Pólya-Aeppli, siglata con PAD, mirando ad ottenere un'alternativa all'HBP.

Dato che, in pratica, la PAD non ha goduto dell'attenzione rivolta alla distribuzione binomiale negativa, non si ritiene inutile ricordare anzitutto che è stata scritta in diversi modi in termini di funzione generatrice di probabilità (*p.g.f.*), interpretata variamente sia come distribuzione generalizzata che come mistura, ed impiegata in vari ambiti di ricerca.

Anche se la letteratura statistica su di essa è rilevante (Cfr., ad esempio, Evans, 1953; Kemp, 1978; Douglas, 1980, 1986; Johnson et al., 1992), può però dirsi che solo il lavoro dell'Evans è davvero fondamentale, non fosse altro che per aver fornito utili relazioni ricorrenti per calcolare le probabilità P_x, per $x = 0,1,\ldots$; e che comunque la PAD (a due parametri) non ha avuto il successo che avrebbe meritato soprattutto a causa della sua scarsa maneggevolezza rispetto ai modelli usualmente considerati alternativi (distribuzione binomiale negativa e distribuzione type A del Neyman).

Da un'analisi comparativa dei risultati acquisiti si è allora cercato (Ferreri, 2000)

di vedere se c'era una parametrizzazione capace di agevolarne la maneggevolezza, specie nell'affrontare l'estimazione di ML. E, dopo aver battuto diverse strade, sotto tale profilo si è giunti a riguardare alquanto opportuna, se non ottimale, l'esplicitazione della *p.g.f.* della PAD nella forma

$$\pi(z;\mu,\vartheta)=\exp\left(\mu\frac{z-1}{1+\vartheta-\vartheta z}\right)\ ,\quad \text{con}\quad \mu>0,\quad \vartheta>0\,, \tag{3.25}$$

suggerita, in fondo, dall'Evans (1953) dato che la usò in termini dei parametri $m=\mu$ e $a=\vartheta/2$, considerati però solo raramente in seguito.

Tale parametrizzazione, nell'implicare

$$\mathsf{E}(X)=\mu\ ,\quad \sigma^2=Var(X)=\mu(1+2\vartheta)\quad \text{e}\quad I_D=\frac{\sigma^2}{\mu}=1+2\vartheta>1\,, \tag{3.25a}$$

indica ϑ come *parametro di sovradispersione*.

Ovviamente, per $\vartheta=0$ la (3.25) si riduce alla *p.g.f.* della distribuzione di Poisson.

Una volta ottenute le probabilità

$$P_0=\exp\left\{-\tfrac{\mu}{1+\vartheta}\right\}\ ,\qquad P_1=P_0\frac{\mu}{(1+\vartheta)^2}\ , \tag{3.26}$$

con un semplice programma di calcolo è dato infatti calcolare senza difficoltà quelle P_x, per $x=2,3,..$, tramite l'espressione

$$P_x=\frac{P_1}{(1+\vartheta)^{x-1}}\sum_{j=0}^{x-1}C_{x,j}\left(\frac{\mu}{1+\vartheta}\right)^{x-1-j}\vartheta^j \tag{3.27}$$

in termini dei coefficienti numerici $C_{x,j}$, $j=0,...,x-1$, della seguente matrice triangolare infinita

$$\mathbf{C}=\begin{pmatrix} 1 \\ 1/2! & 1 \\ 1/3! & 1 & 1 \\ 1/4! & 1/2 & 3/2 & 1 \\ 1/5! & 1/6 & 1 & 2 & 1 \\ 1/6! & 1/24 & 5/12 & 5/3 & 5/2 & 1 \\ \dots\dots\dots\dots \end{pmatrix},$$

che ha: (*i*) $C_{x,x-1}=1$ per $x=1,2,...$; (*ii*) $C_{x,0}=1/x!$ come 1° elemento della *x*-

esima riga, contemplante l'insieme dei coefficienti per il calcolo della probabilità P_x. Coefficienti che sono desumibili tramite il concatenamento espresso da

$$C_{x,j} = \frac{2x-1-j}{x} C_{x-1,j-1} + \frac{1}{x} C_{x-1,j} \ , \quad \text{per } j = 1,2,..,x-2 \, . \qquad (3.28)$$

Facendo tesoro della relazione

$$(x+1)P_{x+1} = \frac{1}{1+\vartheta}\left(\frac{\mu}{1+\vartheta} + 2\vartheta x\right)P_x - \left(\frac{\vartheta}{1+\vartheta}\right)^2 (x-1)P_{x-1} \ , \quad x = 1,2,... \, , \qquad (3.29)$$

che consegue da una formula ricorrente di Evans (1953), è poi facile verificare che il calcolo degli elementi della matrice è facilitato:

da un lato, dal ricorso alla relazione

$$C_{x,j} = \frac{x-1}{j} C_{x-1,j-1} \ , \qquad \text{per } \ j = 1,...,x-2 \, , \qquad (3.30)$$

in quanto fornisce l'elemento $C_{x,j}$ in termini di quello $C_{x-1,j-1}$ della riga precedente;

da un altro lato, dalla formula ricorrente

$$C_{x,j} = \left\{\frac{x(x+1)}{j} - (2x+1-j)\right\} C_{x,j-1} \ , \quad \text{per } j = 1,...,x-1 \, ,$$

di grande aiuto in pratica perché consente di ottenere tutti i termini della *x-esima* riga della matrice partendo dal primo di essi: $C_{x,0} = 1/x!$ o, viceversa, dall'ultimo: $C_{x,x-1} = 1$.

▭ Interpretazione stratificatoria della PAD. Sotto il profilo interpretativo è dato verificare (Ferreri 2008) che, al pari della distribuzione binomiale negativa, anche quella di Pólya-Aeppli può essere vista come una distribuzione composta di Poisson quanto meno secondo la relazione

$$PAD(\mu,\vartheta) = \text{Poisson}(\xi) \underset{\xi}{\wedge} \text{GP}(\mu,\vartheta) \, , \qquad (3.31)$$

che, quale misurante, indica la distribuzione continua $\text{GP}(\mu,\vartheta)$ di funzione generatrice dei momenti $\varphi_{\Xi}(\tau;\mu,\vartheta) = exp\left(\frac{\mu\tau}{1-\vartheta\tau}\right)$, implicante $\mathsf{E}(\Xi) = \mu$ e $Var(\Xi) = 2\mu\vartheta$, che è corrispettiva della *d.f.*

$$f_{GP}(\xi)d\xi = \exp\left\{-\frac{(\xi^{\frac{1}{2}} - \mu^{\frac{1}{2}})^2}{\vartheta}\right\} \exp(-2\frac{\mu^{\frac{1}{2}}}{\vartheta}\xi^{\frac{1}{2}}) I_1(2\frac{\mu^{\frac{1}{2}}}{\vartheta}\xi^{\frac{1}{2}})\ d(2\frac{\mu^{\frac{1}{2}}}{\vartheta}\xi^{\frac{1}{2}}), \qquad (3.31a)$$

dove $I_1(x)$ con $x = 2\mu^{\frac{1}{2}}\xi^{\frac{1}{2}}/\vartheta$ sta a designare la funzione modificata di Bessel del primo tipo di ordine 1.

In parecchie situazioni, specie quando il modello PAD è impiegato in termini di covariate, viene fatto uso della trasformazione $\Xi = \mu Z$, che fa scrivere la relazione

$$\text{PAD}(\mu, \vartheta) \sim \text{Poisson}(\mu\zeta) \underset{\zeta}{\Lambda} \text{GP}(\frac{\mu}{\vartheta}) . \qquad (3.32)$$

Dalla *m.g.f.* $\varphi_Z(\tau) = exp(\frac{\tau}{1-\tau\vartheta/\mu})$ della misturante GP($\frac{\mu}{\vartheta}$), che comporta $\mathsf{E}(Z) = 1$ e $Var(Z) = 2\vartheta/\mu$, appare evidente che la corrispettiva *d.f.* $f_{GP}(\zeta)$ risulta dipendente unicamente da un solo parametro dato dal rapporto $\phi = \vartheta/\mu$, cosi come avviene se la medesima trasformazione è usata per la distribuzione binomiale negativa visto che porta alla (3.16b) considerata per $\alpha(t) = \alpha$ costante.

La (3.31a), benché poco familiare sotto il profilo formale, fa ovviamente intuire che non può che essere di particolare interesse pratico grazie alla vasta gamma di forme di cui è capace. Il relativo primo termine esponenziale richiama infatti alla mente la ben nota *folded normal distribution*, mentre la funzione positiva $e^{-x}I_1(x)$, va prima crescendo a partire da zero per poi decrescere asintoticamente sempre a 0, come del resto appare chiaramente dalla descrizione grafica data in Abramowith and Stegun (p. 375).

▭ Estimazione della PAD riparametrizzata. Per l'estimazione di massima verosimiglianza della PAD(μ, ϑ) in termini di ciascuna *i*-esima distribuzione statistica (1.8), vista come realizzazione di un campione casuale di cui la stessa PAD sia assunta come *p.m.*, torna utile rilevare che dalla (3.25) tenendo conto della relazione ricorrente (3.29) è dato ottenere

$$\begin{aligned} \mu\frac{\partial}{\partial\mu}\log P_x &= -\frac{\mu}{\vartheta} - x + \frac{1+\vartheta}{\vartheta}(x+1)\frac{P_{x+1}}{P_x} \\ \vartheta\frac{\partial}{\partial\vartheta}\log P_x &= \frac{\mu}{\vartheta} + 2x - \frac{1+2\vartheta}{\vartheta}(x+1)\frac{P_{x+1}}{P_x} \end{aligned} \quad , \text{ con } \begin{matrix} P_x = P_x(\mu,\vartheta) \\ x = 0,1,2,... \end{matrix} . \qquad (3.33)$$

Sulla base di queste ultime è infatti facile verificare che il sistema di ML può essere scritto nella forma (Ferreri 2000)

$$\begin{cases}(1+\vartheta)H(\mu,\vartheta)=\mu+M\vartheta\\(1+2\vartheta)H(\mu,\vartheta)=\mu+2M\vartheta\end{cases},\qquad \text{con}\quad H(\mu,\vartheta)=\sum_{x=0}^{c}f_x(x+1)\frac{P_{x+1}(\mu,\vartheta)}{P_x(\mu,\vartheta)},\quad (3.34)$$

dove $f_x=n_x/n$ ed $M=\Sigma_{x=0}^{c}xf_x$.

E dato che dal sistema si ottiene immediatamente $\hat{\mu}=M$, l'estimazione di ML della PAD si completa col risolvere l'equazione in ϑ

$$\psi(\vartheta)=H(\vartheta)-M=0,\quad \text{con } \psi'(\vartheta)=\frac{2}{\vartheta}H(\vartheta)-\frac{1+2\vartheta}{\vartheta^2}\sum_{x=0}^{c}f_x\tilde{\pi}_x(\tilde{\pi}_{x+1}-\tilde{\pi}_x),\ (3.35)$$

$$\text{dove}\quad \tilde{\pi}_x=(x+1)\frac{P_{x+1}(\vartheta)}{P_x(\vartheta)},\qquad (3.35a)$$

$$\text{essendo}\quad P_x(\vartheta)=P_x(M,\vartheta),\quad \text{e}\quad H(\vartheta)=H(M,\vartheta)=\sum_{x=0}^{c}f_x\tilde{\pi}_x.\qquad (3.35b)$$

Giova notare che l'uso della relazione ricorrente

$$\tilde{\pi}_x=\tilde{\pi}_0+\frac{\vartheta x}{1+\vartheta}\left(2-\frac{\vartheta}{1+\vartheta}\frac{x-1}{\tilde{\pi}_{x-1}}\right),\qquad \text{with}\quad \tilde{\pi}_0=\frac{M}{(1+\vartheta)^2}\quad \text{ed}\quad x=1,2,\ldots,\quad (3.36)$$

implicata dalla (3.29) sulla base delle (3.26), dà modo di ridurre notevolmente gli effetti d'arrotondamento rispetto a quelli che comporterebbe l'operare in termini delle probabilità.

Le seguenti espressioni

$$\begin{aligned}I_{11}&=n\left[\frac{1+4\vartheta+2\vartheta^2}{\mu\vartheta^2}-\frac{1}{\mu^2}\left(\frac{1+\vartheta}{\vartheta}\right)^2D\right],\qquad \text{with}\quad D=Var(X)-Var(\pi_X)\\I_{12}&=n\left[-\frac{(1+\vartheta)(1+4\vartheta)}{\vartheta^3}+\frac{(1+\vartheta)(1+2\vartheta)}{\mu\vartheta^3}D\right]\\I_{22}&=n\left[\frac{(1+2\vartheta)(1+4\vartheta)}{\vartheta^4}\mu-\left(\frac{1+2\vartheta}{\vartheta^2}\right)^2D\right],\end{aligned}\qquad (3.37)$$

possono essere infine ottenute per gli elementi della matrice $\boldsymbol{I}(\mu,\vartheta)$ d'informazione del Fisher dato che, con $\pi_x=(x+1)\frac{P_{x+1}(\mu,\vartheta)}{P_x(\mu,\vartheta)}$, può verificarsi che

$$\mathsf{E}[\sum_{x=0}^{c}f_x\pi_x(\pi_{x+1}-\pi_x)]=D-\mu=2\mu\vartheta-Var(\pi_X),$$

dove: $\mathsf{E}(X) = \mathsf{E}(\pi_X) = \mu$, $Var(X) = \mu(1+2\vartheta)$ e $\mathsf{E}(\pi_X^2) = \Sigma_{x=0}^{\infty}\pi_x(x+1)P_{x+1}$, che per la distribuzione di Poisson ($\vartheta = 0$) si riduce ad $\mathsf{E}(\pi_X^2) = \mu^2$ implicando $D - \mu = 0$.

Dalla procedura d'estimazione appena descritta si evincono chiaramente le ragioni che hanno guidato la scelta della parametrizzazione (3.25). Il fatto che il valore medio $\mathsf{E}(X) = \mu$ sia un parametro stimato dalla media campionaria *M* riduce, invero, drasticamente le difficoltà calcolatorie rispetto a quanto generalmente implica una parametrizzazione recante ad un sistema da risolvere, nel suo insieme, per iterazioni (Cfr, ad esempio, Douglas, 1986, alla voce: "Pólya-Aeppli distribution" nella Encyclopedia of Statistical Sciences,7 pp. 56-59, Eds. S.Kotz, N.L.Johnson and C.B.Read). È anche per favorirne il confronto che ci si è avvalsi soprattutto del simbolismo del Douglas, peraltro ormai di rito nella letteratura statistica in proposito.

Ovviamente, tramite gli stimatori $\tilde{\mu} = M$, $\tilde{\vartheta} = (\tilde{\sigma}^2 - M)/(2M)$, con $Var(\tilde{\mu}) = \frac{1}{n}\mu(1+2\vartheta)$, $Var(\tilde{\vartheta}) \approx \frac{1}{n}\{\frac{1}{2}(1+2\vartheta)^2 + \frac{1}{\mu}\vartheta(1+\vartheta)(1+2\vartheta)\}$ e $Cov(\tilde{\mu},\tilde{\vartheta}) \approx \frac{1}{n}\vartheta(1+\vartheta)$ (Cfr.: Evans, 1953), il metodo dei momenti fornisce il valore iniziale per risolvere iterativamente l'equazione (3.35).

Giova notare che tra lo stimatore (3.21c) di MM del parametro α della distribuzione binomiale negativa nella forma (2.1), oppure (3.8) con $\mu(t) = \mu$, e l'analogo stimatore di ϑ della PAD in questione sussiste la relazione

$$\tilde{\alpha} = 2\frac{\tilde{\vartheta}}{M}, \qquad (3.38)$$

che sostanzia il parallelismo cui, per certi versi, si mirava con la considerazione del modello di Pólya-Aeppli.

▭ Estensione della PAD in distribuzione di processo di nascita per l'analisi della dispersione non-poissoniana. La relazione (3.38), nel richiamare alla mente le considerazioni che dalla distribuzione binomiale negativa hanno portato alla distribuzione del processo iperbinomiale, a questo punto induce quasi inevitabilmente a ripercorrere l'iter. Del resto, se non vi sono state mai riserve all'estensione della NBD in modo da contemplare entrambi i tipi di dispersione non-poissoniana, non vi è motivo che ne insorgano di fronte ad un'analoga estensione della distribuzione di Pólya-Aeppli da utilizzare nella ricerca concreta in alternativa all'HBD, specie se si mostrerà meno sensibile ai fattori d'instabilita.

Alla facile obiezione che, sul piano fenomenico, sarebbe difficoltoso giustificare l'estensione implicante dispersione sottopoissoniana, ci pare che basterebbe ribattere

che lo stesso potrebbe dirsi per la corrispettiva distribuzione iperbinomiale positiva (NHBD), anche se questa può sollevare meno riserve in quanto comprende, come caso particolare, l'usuale distribuzione binomiale.

A parte atteggiamenti "d'abitudine", per entrambe le estensioni c'è invece che, mentre i modelli biparametrici a sovradispersione poissoniana generalmente godono di varie interpretazioni, e spesso di quella stratificatoria, le distribuzioni a dispersione sottopoissoniana di solito consentono assai meno.

Ma, se questo fa pensare che, in casi del genere, la "lettura" dei risultati richiede una cautela davvero particolare, fa pure dire che solo chi è poco avvezzo alla costruzione e all'uso di modelli può provare sensi di inopportunità davanti a certi strumenti di lavoro, non fosse altro perché è stato autorevolmente rilevato che

> «*Lo statistico, come ogni altro umano nel suo vivere di tutti i giorni, se arriva a costruirsi una vanga per dissodare un certo terreno, ci arriva attraverso il cosiddetto* "*tormento della ricerca*"*. Pertanto, se è sempre consapevole che ci può essere ben di meglio e di ben più appropriato e motivato, è pure cosciente che, usando con perizia lo strumento che si è costruito nel quadro di un preciso contesto problematico e alla luce del proprio supporto teorico, può aver modo di aggiungere una "pietruzza" nel selciato del suo cammino; e che senza strumenti rimarrebbe al palo.*»

Volendo, con questa consapevolezza, tentare di rendere il modello di Pólya-Aeppli alternativo all'HBD bisogna pertanto vedere anzitutto come ampliare lo spazio parametrico di ϑ, dato che è stato finora visto come parametro di sovradispersione in quanto contemplata dalla PAD per $\vartheta > 0$.

A tal fine, l'espressione (3.25a) dell'indicatore di dispersione fa ovviamente prendere atto che si ha $I_D = \sigma^2 / \mu < 1$ solo se in quella $\sigma^2 = \mu(1 + 2\vartheta)$ della varianza è $-\frac{1}{2} < \vartheta < 0$. E dato che l'ampliamento può essere costituito dai valori dell'intervallo $\mathrm{I} = (-\frac{1}{2}, 0)$ o di parte di esso rappresentata da un intorno sinistro dello zero, la relativa determinazione va a consistere nell'individuare i valori del parametro ϑ compresi nell'intervallo I, per quali si ha una distribuzione di probabilità tutt'al più incompleta o impropria, così come corrispondentemente avviene per l'HBD.

Poiché, con $\vartheta \in \mathrm{I}$, le espressioni (3.26) presentano sempre valori positivi P_0^*, P_1^*; mentre, per $x \geq 2$, la (3.27) può assumere anche valori negativi, non resta pertanto che incentrare l'attenzione su quest'ultima. Ma è facile verificare che

$$P_2^* > 0 \;\text{ se }\; \mu \geq \tfrac{1}{2} \qquad \text{e, se } \mu < \tfrac{1}{2}, \;\text{ per }\; \vartheta > \vartheta_2 = -\tfrac{1}{2} + \tfrac{1}{2}(1 - 2\mu)^{1/2};$$

$P_3^* > 0$ se $\mu \geq 1/L_3$ e, se $\mu < 1/L_3$, per $\vartheta > \vartheta_3 = -\frac{1}{2} + \frac{1}{2}(1 - L_3\mu)^{1/2}$,

dove $L_3 = 4/(3+\sqrt{3})$ e $\vartheta_3 > \vartheta_2$; e quindi che P_2^* e P_3^* sono entrambi positivi quando:

a) $\mu \geq 1/L_3 = 1.1830$; b) $\mu < 1/L_3$ se $\vartheta > \vartheta_3$.

E siccome, accolto un valore per μ, è dato arrivare ad un intero *m* per il quale ci sarà un limite L_m tale che, con $\mu < L_m$, si otterrà $P_2^*, P_3^*, ..., P_m^* > 0$ e $P_{m+1}^* < 0$, in generale può dirsi che dalla (3.27) conseguono consecutivamente più di due termini positivi se il valore di ϑ rientra in un intorno sinistro di $\vartheta = 0$ opportunamente piccolo, il cui raggio dipende dall'ordine di grandezza del valore medio μ.

Allorché μ e ϑ sono tali che la sequenza dei successivi termini positivi finisce con un P_κ^* da riguardarsi praticamente trascurabile, allora l'insieme di coppie

$$\{ (x, P_x^*);\ x = 0, ..., \kappa \ \text{ e } \ T_\kappa = \sum_{x=0}^{\kappa} P_x^* \cong 1 \}$$

praticamente fornisce pertanto una distribuzione di probabilità che estende la PAD in quanto a dispersione sottopoissoniana;

Quando, invece, il resto *negativo* della sequenza dei P_x^* è tale da aversi $T_\kappa > 1$ in modo non trascurabile, per lo più torna utile considerare, come probabilità, i valori normalizzati $P_x^\circ = P_x^* / T_\kappa$. Nelle analisi compiute questa circostanza non ci è però mai capitata. Si è allora fatto in modo di provocarla più volte al fine di averne piena contezza, anche perché, se l'incompletezza della distribuzione iperbinomiale positiva (PHBD) in pratica comunemente non ingenera problemi, tale circostanza potrebbe comportarne, specie nella estimazione.

Il fatto che di solito l'intorno sinistro di $\vartheta = 0$, per il quale si ha dispersione sottopoissoniana, sia alquanto piccolo, non può comunque sollevare perplessità, non fosse che per l'abitudine ad usare la HBD: si è visto infatti che la relativa estimazione di ML sulla base di una distribuzione statistica $\mathfrak{D}_n(x;c)$ di indicatore di dispersione campionario $\tilde{I}_D = \tilde{\sigma}^2 / M < 1$ presuppone $-1/\mu < \alpha < 0$ e che la stima del parametro α è compresa in un piccolo intorno sinistro di $\alpha = 0$ in quanto soddisfa la relazione $\hat{\alpha} > max(-1/(c-1), -1/M)$,

In perfetto parallelismo con quanto si è fatto estendendo il supporto della distribuzione binomiale negativa (NBD), si parlerà così di *distribuzione iperPólya-Aeppli*,

siglandola con HPAD, per indicare quella di Pólya-Aeppli a supporto esteso; e si userà la sigla UPAD (da under-Poissonian dispersion) solo per segnalare la PAD limitatamente alla situazione di dispersione sottopoissoniana, cioè per ϑ compreso in un intorno sinistro dello zero.

Ai fini pratici torna comunque utile rammentare che:

(*i*) la UPAD (definita per $\vartheta < 0$) ha soltanto una moda: ad $x = 0$ se $\mu < 1$, oppure ad $x > 0$ se $\mu > 1$;

(*ii*) le espressioni (3.25a) rispettivamente del valor medio e della varianza della distribuzione di Pólya-Aeppli possono essere assunte anche per il caso di $\tilde{I}_D < 1$ nella misura in cui la (3.25) può essere approssimativamente riguardata come funzione generatrice di probabilità della UPAD;

(*iii*) con analoga approssimazione, ci si può avvalere anche dei suddetti stimatori di massima verosimiglianza, i quali però risultano solitamente ben più instabili di quelli del metodo dei momenti, già descritto per PAD.

Sotto il profilo computazionale, l'estimazione di ML può pertanto essere compiuta abbastanza facilmente tanto per l'HBD quanto per l'HPAD, che ovviamente si pone in alternativa. In entrambi i casi questi stimatori di solito mostrano effetti d'instabilità rilevanti, per non dire talora cruciali, almeno fin quando il valore dell'indicatore di dispersione campionario $\tilde{I}_D = \tilde{\sigma}^2 / M$ è minore o prossimo ad 1. In situazioni del genere, una piccola variazione nei dati considerati (visti implicati da una realizzazione di un campione casuale) si traduce infatti in una elevata, se non davvero abnorme, variazione sia nella stima di κ dell'HBD, e quindi dei parametri della distribuzione binomiale (Cfr. Olkin et al., 1981), che in quella di ϑ dell'HPAD, benché sulla stima di quest'ultimo parametro gli effetti sovente sembrino un po' attutiti.

Va perciò tenuto quanto meno presente "l'inopportunità" di basare l'inadeguatezza di un modello sulla coda destra di una distribuzione statistica osservata, almeno finché questa non segni una quota consistente dei casi d'osservazione, supposto naturalmente che il loro numero sia abbastanza elevato, e non se ne sia ben soppesata la portata sotto il profilo dell'instabilità. Ben lo mostra, del resto, l'esempio di Gelfand e Dalal descritto nel § 2, il quale fa balzare che "*sparare alla mosca col cannone*" si corre il rischio di incorrere in un inconveniente ben maggiore di quello che può provocare la mosca, che è, anche se lo si dimentica troppo spesso, *un inquilino*, come noi, della realtà che troppo spesso diciamo "nostra".

Finora l'HPAD è stata considerata con riguardo ad una singola distribuzione stati-

stica di dati di conto, e non ad una sequenza di analoghe distribuzioni come quella (1.8), la *i*-esima delle quali, per $i=1,..,r$, è implicata dalla realizzazione di un *set* di v.c. ($X_{1i},..,X_{hi},..,X_{ni}$), la *h*-esima del quale sta a denotare il numero degli eventi, nella fattispecie incidenti, capitabili alla rispettiva unità d'osservazione nell'arco temporale $J_i=(0,t_i]$

L'intento che ha sollecitato ad introdurre l'HBP per investigare tanto sul process-trend $\mu(t)$ quanto sull'andamento della dispersione non-poissoniana, porta così ad ampliare analogamente il discorso in modo da vedere nella HPAD=HPAD(*t*) un famiglia di distribuzioni di probabilità (indicizzate da *t*) di un processo di nascita alternativo all'HBP.

E poiché questo processo è stato praticamente suggerito dal contesto in cui si è inquadrato l'andamento delle stime $\hat{\alpha}_i=\hat{\alpha}_{t_i}$ di ML della distribuzione iperbinomiale (HBD) usata per ciascun membro della sequenza $\{\mathcal{D}_t,\ t=t_1,..,\ t_r\}$ delle distribuzioni statistiche (1.8), lo stesso contesto non può che indurre a riguardare similmente la HPAD così da utilizzare le analoghe stime $\hat{\vartheta}_i=\hat{\vartheta}_{t_i}$ del parametro ϑ per cogliere, tramite l'attenuazione degli effetti d'instabilità, aspetti essenziali dell'andamento della *dispersione non-poissoniana.*

Alla considerazione dell'HPAD(*t*) quale distribuzione di probabilità della v.c. $X(t)$ di un processo di nascita $\{X(t),t>0\}$ si perviene pertanto se, invece di μ e ϑ, nella (3.25) e nella (3.27) vengono rispettivamente inserite due funzioni $\mu(t)$ e $\vartheta(t)$ del tempo, per le quali ragioni di semplicità suggeriscono ancora una volta di stare alla supposizione che siano differenziabili per $t>0$ e tali che:

(*i*) $\mu(t),\ \mu'(t)>0$ per $t>0$, con $\lim_{t\downarrow 0}\mu(t)=0$, dove $\mu(t)=E[X(t)]$ denota il *process-trend*;

(*ii*) per $\vartheta(t)<0$ con $t>0$, una condizione come $\vartheta(t)>\vartheta_L(\mu(t))$, con $-\frac{1}{2}<\vartheta_L(\mu(t))<0$, consenta di ottenere un'adeguata UPAD(*t*);

(*iii*) sia valida la relazione $\frac{\mu'(t)}{\mu(t)}>\frac{\vartheta'(t)}{1+\vartheta(t)}\ \ \forall t>0$, nel qual caso si ha

$$\Lambda(t)=\frac{\mu(t)}{1+\vartheta(t)}\ ,\ \text{con}\ \Lambda'(t)>0,\ \forall t>0,\ \text{e}\ \lim_{t\downarrow 0}\Lambda(t)=0\,.$$

Al pari della funzione $\alpha(t)$ dell'HBP, quando, all'aumentare di t a partire da *t*=0,

quella $\vartheta(t)$ va crescendo da valori negativi fino ad annullarsi ad un certo tempo $\tau_P > 0$, anche la HPAD(t) va riducendosi alla distribuzione di Poisson di parametro $\mu(\tau_P) = \lim_{t \to \tau_P} \mu(t)$. Naturalmente, il fatto che la $\vartheta(t)$ sia riguardabile funzione di dispersione non-poissoniana della distribuzione di probabilità del processo di nascita HPAP fa capire come sia opportuno denominarla pure *process-index function*.

A questo punto, non ci sono pertanto ragioni per non riconoscere nella HPAD($\mu(t), \vartheta(t)$) una valida alternativa alla HBD($\mu(t), \alpha(t)$), tanto più che, come si avrà modo di constatare, sovente l'impiego di taluni procedimenti volti ad attenuare gli effetti d'instabilità sulle stime di ML relative ai tempi t_i considerati sembrano favorire l'inferimento sulla funzione di dispersione $\vartheta(t)$ rispetto a quello sulla corrispondente funzione $\alpha(t)$ dell' HBD($\mu(t), \alpha(t)$).

4 – *Passi di un'indagine, parametrica, su certa incidentalità degli operai di una industria metalmeccanica.*

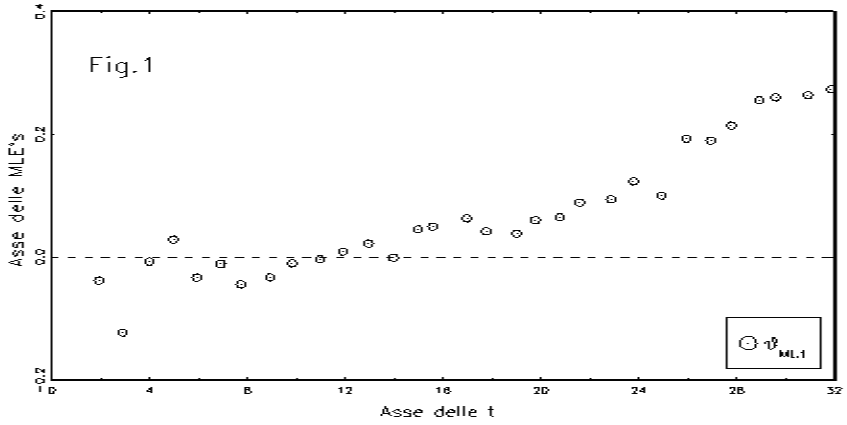
4.1 – *Distribuzioni statistiche dei dati di conto e risultati dell'impiego dei p.m. HBD e HPAD.*

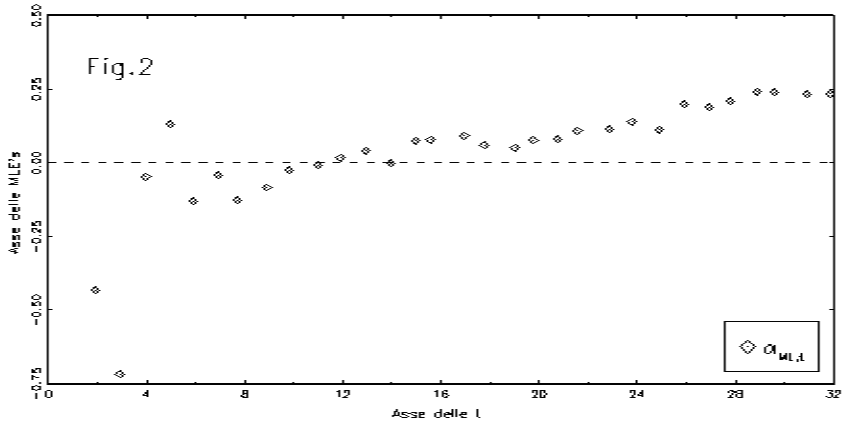
Come già è stato detto, questo quaderno nasce dall'intento di dare una risposta plausibile (in quanto interpellati) in merito alla natura dell'incidentalità riscontrata nel rapporto uomo-macchina praticato dagli operai di un'industria metalmeccanica relativamente al periodo d'osservazione (0,*T*] (otto anni), durante il quale erano stati registrati, per ciascuno, i tempi d'accadimento.

È del resto per ciò che, con la 2^ delle due fondamentali circostanze con cui si è iniziata la premessa, si è ritenuto di premettere la struttura di supporto al discorso che s'intendeva svolgere e di chiarire subito perché l'evidenza statistica disponibile sarebbe stata utilizzata in termini di distribuzioni statistiche, nonostante l'ottica temporale che ne segnava, con la base, l'orizzonte, potesse far preferire altra forma.

Una volta operata la segmentazione del periodo d'osservazione in trimestri, in primo luogo si è così fatto tesoro dei dati rilevati per costruire la sequenza di r=31 distribuzioni $\mathcal{D}_{n,i}(x) = \mathcal{D}_{n,t_i}(x)$, per $i = 1,..,31$ in quanto t_i=2, 3, ..., 32, degli $n = 108$ operai considerati (costituenti un campione osservazionale) rispetto al numero x di incidenti successi a ciascuno nei corrispettivi intervalli di tempo $J_i = (0, t_i]$. Distribu-

zioni che in seguito saranno pure dette osservate, benché costruite nel modo anzidetto con riguardo alla segmentazione adottata per il periodo d'osservazione, e che sono leggibili nella Tav. I dell'Appendice a fronte dell'ultimo tempo d'incidente dei corrispondenti intervalli J_i, nonostante il trimestre fosse l'unità di tempo pensata per discretizzare il periodo d'osservazione (1980-87).





A monte si è ritenuto inoltre che i criteri seguiti per assegnare gli operai alle macchine, potevano motivare che si riguardassero praticamente "omogenei" rispetto agli eventi incidente della tipologia considerata, il cui più grande numero a loro capitato in

J_i è indicato con c_i.

L'andamento nel tempo dei valori, $\tilde{I}_{D,i}$, dell'indicatore di dispersione campionario, leggibili nella seconda colonna della Tav. II dell'Appendice, nell'andare da livelli minori 1 a livelli sempre più nettamente maggiori di 1, ovviamente traduce la constatazione che, tramite una opportuna discretizzazione dell'intervallo di osservazione, la dispersione delle successive distribuzioni $\mathcal{D}_i(x)$ va dalla sottopoissoniana a quella sovrapoissoniana, sempre meno influenzata da fattori di instabilità.

È del resto per questo che, se, da un lato, si è segnalato che impieghi di criteri di significatività per saggiare la conformità di una distribuzione osservata al modello di riferimento vanno visti discutibili allorché compiuti a prescindere dallo svolgimento del processo fenomeno; da un altro lato, ci si è trovati a dover estendere i modelli di probabilità considerati in modo da renderli adeguati alla circostanza.

Stando alla distribuzione iperbinomiale (3.12) contemplante dispersione non-poissoniana, corrispettivamente ai tempi t_i della prima colonna della Tav. I si sono così ottenute (Ferrante e Ferreri, 1996) le stime $\hat{\alpha}_i = \hat{\alpha}(t_i)$ di ML, leggibili nella Tav. II e graficamente rappresentate, nella Fig. 2, dal $(t, \hat{\alpha}_t)$-plot della HBD.

Le distribuzioni statistiche $\mathcal{D}_i(x)$ sono state poi riutilizzate per determinare, tramite la (3.35), le stime di ML $\hat{\vartheta}_i = \hat{\vartheta}(t_i)$ della HPAD($\mu(t), \vartheta(t)$) estendente la distribuzione di Pólya-Aeppli. Come per l'HBD($\mu(t), \alpha(t)$), nessuna difficoltà è emersa neanche nei casi in cui l'indicatore di dispersione campionario $\tilde{I}_{D,t} = \tilde{\sigma}_t^2 / M_t$, per $t = t_i$ – da ora sarà sempre supposto in situazioni analoghe –, risultava di valore inferiore ad 1. A mostrare che le condizioni viste per ϑ, con $\vartheta < 0$, in pratica non possono dirsi restrittive, vi è che la differenza $\Delta = T_k - 1$ è stata pari a $\Delta_1 \cong 5.9 \cdot 10^{-4}$ e a $\Delta_2 \cong 2.8 \cdot 10^{-5}$ ordinatamente per le prime due distribuzioni D_1 e D_2, per le quali $c_1 = c_2 = 2$, e che per tutte le altre si è avuto $\Delta_i < 10^{-7}$ per $c_i > 2$.

Il $(t, \hat{\vartheta}_t)$-plot delle stime $\hat{\vartheta}_i$ di ML della HPAD, leggibili nella prima colonna della terza parte della Tav. II, è stato disegnato nella Fig. 1, posta sopra la Fig. 2 al fine di rimarcarne gli aspetti differenziali.

Confrontando i due plot, sulle prime potrà forse balzare che, nella fattispecie, i due modelli usati sono riguardabili pressoché parimenti idonei a descrivere l'andamento della dispersione non-poissoniana delle $r = 31$ distribuzioni statistiche della

sequenza $\{ \mathcal{D}_1(x),..,\mathcal{D}_i(x),..,\mathcal{D}_r(x) \}$ ancorata all'articolazione $\{ 0, t_1,..,t_i,..,t_r \}$ che si è ritenuto di considerare per il periodo d'osservazione. Quanto meno in termini della probabilità $\Pr = P[\chi^2_\nu > X^2_{\nu,obs}]$, in favore della HPAD vi è tuttavia che:

(*i*) all'aumentare del numero $c+1$ delle classi della distribuzione osservata a partire da 7 ($t_i \approx 20.76$) a 9 ($t_j \approx 31.85$), la probabilità percentuale $\%\Pr_{HPAD}$ passa da 21.9 a 89.8, intanto che quella $\%\Pr_{HBD}$ va da 21.6 soltanto a 85.0.

(*ii*) quando l'indicatore di dispersione campionario, dato dal rapporto della varianza campionaria alla corrispondente media, è minore o prossimo ad uno, come in effetti è fino a $c_t \leq 5$ con $n_c = 1$ ($t < 10$), il $(t, \hat{\vartheta}_t)$-plot rivela un andamento meno instabile, sia pure lievemente, di quello delle stime $\hat{\alpha}_t$.

Da una semplice occhiata alla Fig. 1 si evince comunque che, anche nella prima parte del plot delle stime $\hat{\vartheta}_i$, l'andamento è cosi irregolare da indurre a ricercare una qualche strada che consenta di addivenire ad una tendenza recepibile quale forma approssimata della funzione $\vartheta(t)$ di dispersione non-poissoniana e, quindi, ragionevolmente riguardabile come verisimile risultato di "attenuazione" degli effetti di instabilità.

4.2 – *Attenuazione degli effetti d'instabilità tramite estimazione di ML basata su distribuzioni statistiche di perequazione per medie mobili.*

Molteplici sono ovviamente le procedure congetturabili per attenuare gli effetti d'instabilità al punto da ottenere trend verosimili per $\mu = \mu(t)$, $\alpha = \alpha(t)$ e $\vartheta = \vartheta(t)$. Non ci può essere però strumento metodologico che, sulla falsariga degli indicatori di bontà dell'adattamento, possa indirizzare la scelta di una di esse. Sono infatti essenzialmente implicazione di un'idea a priori inverificabile che, se può trovare forza nella ragionevolezza delle tendenze, specie rispetto al ridursi degli effetti di instabilità, soprattutto si regge sul contenuto profondo della saggia e inconfutabile affermazione: "*la verità è la realtà che si accetta*", che ogni umano non può scordare senza incorrere nel rischio di ridursi a perseguire idoli, magari costruendosi schemi che talora non esita a proporre in ottica oggettiva o seguendo "l'aquilone" dei suoi sogni.

Mirando a formulare una plausibile procedura atta allo scopo non si può quindi che incominciare, da un lato, dal domandarsi se, nonostante le ampie possibilità tanto del processo iperbinomiale quanto di quello iperPólya-Aeppli, sia veramente il caso di

prescindere dai modelli competitori che la letteratura statistica pone a disposizione; e, da un altro, dal prendere atto che la constatazione che l'instabilità ha rilevanti radici nella struttura che si conferisce all'evidenza statistica con l'articolazione cui viene ancorato l'osservato per il fenomeno, suggerisce di invertire l'ordine nel da farsi, e cioè di affrontare il problema dell'instabilità a monte, vale a dire prima della specificazione del modello e della rispettiva estimazione.

Con l'occhio rivolto a quanto offre la letteratura statistica, pur stando – ma non per opportunità – ai due modelli di probabilità messi a punto, non ci si può così esimere dal proporsi quanto meno di:

(*i*) delineare innanzitutto una procedura di "rettifica" delle distribuzioni statistiche della sequenza $\{ \mathcal{D}_1(x),..,\mathcal{D}_i(x),..,\mathcal{D}_r(x) \}$, che dia modo di attenuare sostanzialmente l'instabilità riscontrata nelle stime $\hat{\alpha}_i$ e $\hat{\vartheta}_i$ di ML;

(*ii*) tentare di cogliere profili delle funzioni $\alpha = \alpha(t)$ e $\vartheta = \vartheta(t)$ direttamente dalle sequenze $\{ \hat{\alpha}_{R1},..,\hat{\alpha}_{Ri},..,\hat{\alpha}_{Rr} \}$ e $\{ \hat{\vartheta}_{R1},..,\hat{\vartheta}_{Ri},..,\hat{\vartheta}_{Rr} \}$ delle stime ordinatamente implicate, per i due modelli di probabilità usati, dalle distribuzioni statistiche rettificate (che danno ragione dell'indice R);

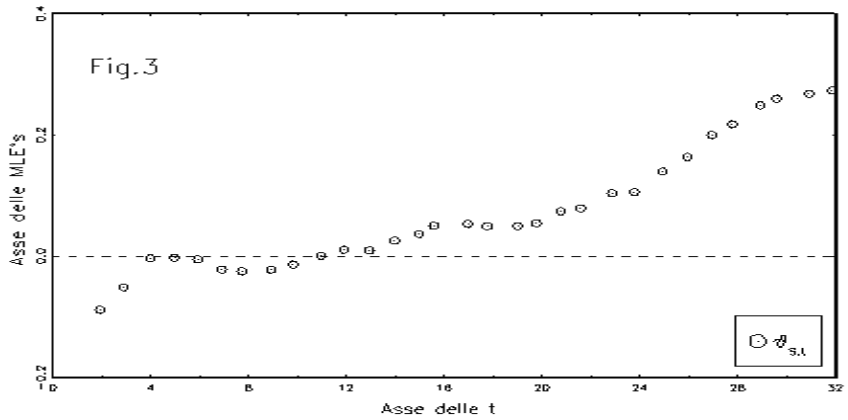
(*iii*) effettuare confronti anche con sequenze di stime desumibili con altri procedimenti di analoghe finalità, come sono taluni che si concretano in certi stimatori pseudo-jackknife o che consistono in procedure tipo bootstrap.

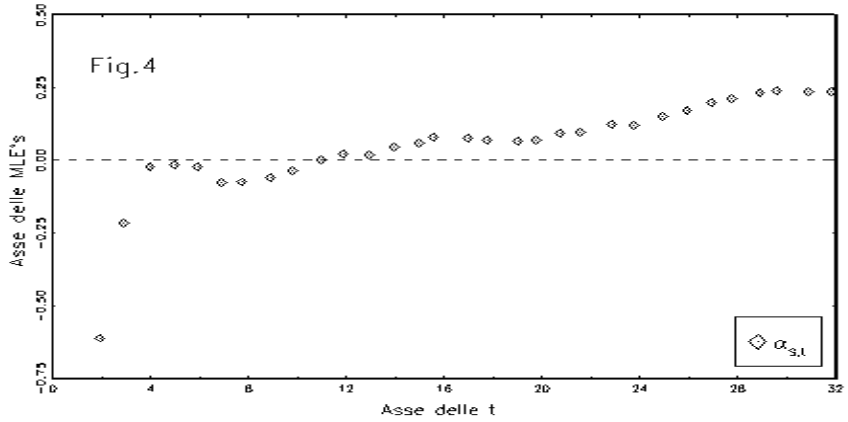
L'intento di utilizzare uno strumento che non facesse perdere notevolmente il senso del dato, ha anzitutto fatto balzare l'opportunità di affrontare il punto (*i*) ricorrendo ad una semplice procedura di perequazione, ancorché sovente ritenuta "quasi fuori moda".

In questo contesto, per la rettifica delle distribuzioni statistiche $\mathcal{D}_i(x)$ è parso soddisfacente operare una perequazione per medie mobili ponderate di tre termini su ciascuna delle sequenze di numerosità n_{xi}, i=1,..,31, facendo però in modo da aversi, per ogni *i*-esima distribuzione, $n_{xi} > 0$ per $x = 0,1,..,max(c_i)$ e $n_{xi} = 0$ per $x > c_i$. È evidente che, così procedendo, l'operazione di rettifica può dirsi conclusa soltanto una volta rinormalizzate ad *n* le numerosità calcolate in corrispondenza di ciascuna di quelle della $\mathcal{D}_i(x)$.

Alla luce dei pro e dei contro di diversi tipi di perequazione del genere, si è ritenuto di assumere, come numerosità di perequazione corrispettiva della n_{xi} osservata, il

numero, ad un solo decimale, fornito dalla retta adattata col metodo dei minimi quadrati al *set* delle tre numerosità consecutive $n_{x,i-1}$, $n_{x,i}$, $n_{x,i+1}$.





Per dare inizio alla procedura, quale numerosità di perequazione al tempo iniziale t_1 si è convenuto di adottare, quella fornita dalla prima retta adattata (relativa a t_2) se positiva oppure la rispettiva numerosità osservata. Come ultima distribuzione statistica di perequazione, cioè per la 31^, è stata invece assunta quella osservata dato che, al tempo $t = T$, gli effetti d'instabilità sono stati riguardati praticamente irrilevanti.

Le distribuzioni statistiche così calcolate per ogni valore di t_i – naturalmente anche compiendo il dimensionamento ad $n = 108$ di quelle di perequazione –sono state riportate nella Tav. I ordinatamente nelle righe sottostanti a quelle delle rispettive distribuzioni statistiche osservate in modo da renderne immediato il confronto.

A completamento dello stesso punto (*i*), tali distribuzioni di perequazione sono state utilizzate per l'estimazione sia della HPAD($\mu(t), \vartheta(t)$) che dell'HBD($\mu(t), \alpha(t)$). Le rispettive sequenze $\{ \hat{\vartheta}_{Si};\ i = 1,..31 \}$ e $\{ \hat{\alpha}_{Si};\ i = 1,..31 \}$ delle stime di ML così ottenute, e riportate nella Tav. II singolarizzandole con l'indice *S* (da smoothing), sono state infine usate per disegnare sia il $(t_i, \hat{\vartheta}_{Si})$-plot che il $(t_i, \hat{\alpha}_{Si})$-plot, che ordinatamente compaiono nella Fig. 3 e nella Fig. 4.

Nonostante le distribuzioni statistiche di perequazione per lo più fossero risultate talmente prossime alle corrispondenti osservate da far ritenere pressoché inutile la l'operazione di rettifica, i due plot colpiscono per la rilevante regolarità dei rispettivi andamenti, specie se visti rispetto a quelli delle stime $\hat{\vartheta}_i$ and $\hat{\alpha}_i$ implicate dalle distribuzioni osservate.

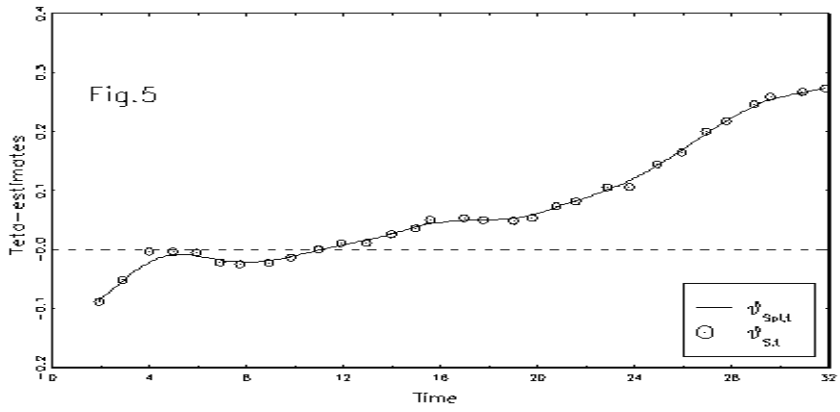
Il che ha portato a pensare che, *anche se i plot delle stime* $\hat{\vartheta}_i$ *e* $\hat{\alpha}_i$ *potevano forse indurre a ritenere che non fosse il caso di operare qualche tipo di spline, il ricorso ad una tecnica come quella usata non è quanto meno da ignorare mirando all'ottenimento di stime* $\hat{\vartheta}_{Si}$ *e* $\hat{\alpha}_{Si}$ *significative sotto il profilo dell'attenuazione dell'instabilità.*

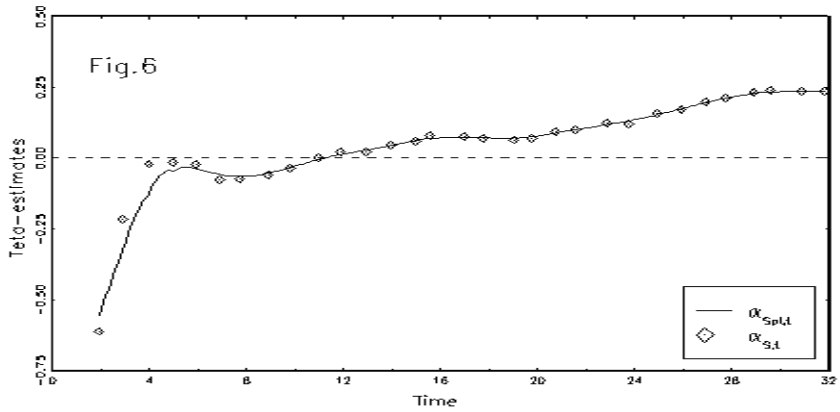
4.3 – *Trend interpolatorio della funzione di dispersione non-poissoniana di entrambi i processi di nascita HBP e HPAP*

Il proposito di non dar spazio a tecniche che, per molti versi, svuotano irrimediabilmente i dati della loro sostanzialità, ma di addivenire comunque ad una linea continua in qualche misura riguardabile forma approssimata della *process-index function* $\vartheta(t)$ del processo di nascita iperPólya-Aeppli ha poi suggerito di applicare anche alla sequenza delle stime ($\hat{\vartheta}_{S1},..,\hat{\vartheta}_{Si},..,\hat{\vartheta}_{Sr}$), con $r = 31$, il semplice procedimento di rettifica impiegato per le numerosità delle distribuzioni statistiche osservate.

Perequando, in modo analogo, *due volte* le stime $\hat{\vartheta}_{S,i}$ di ML si è ottenuta la sequenza ($\tilde{\vartheta}_{p,1},..,\tilde{\vartheta}_{p,i},..,\tilde{\vartheta}_{p,r}$), $r = 31$, dei valori $\tilde{\vartheta}_{p,i} = \tilde{\vartheta}_{Spl,i}$ riportati nell'ultima colonna della Tav. II, i quali hanno fatto prendere atto che pressoché gli stessi risultati si

sarebbero potuti ottenere perequando analogamente una sola volta le stesse stime $\hat{\vartheta}_{S,i}$ utilizzando cubiche.



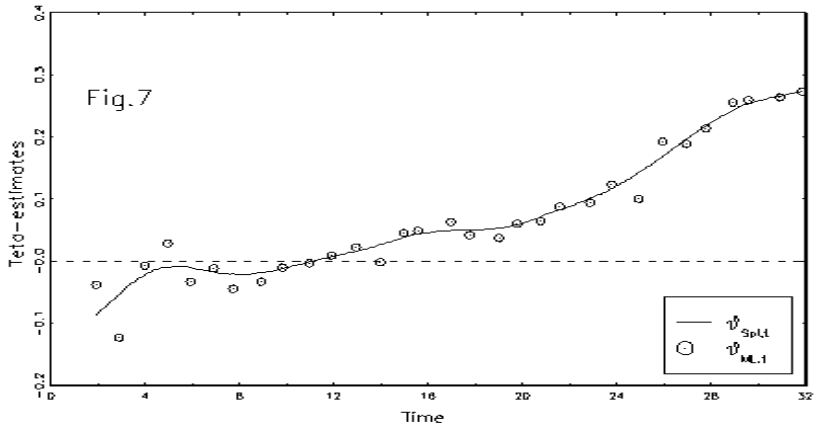


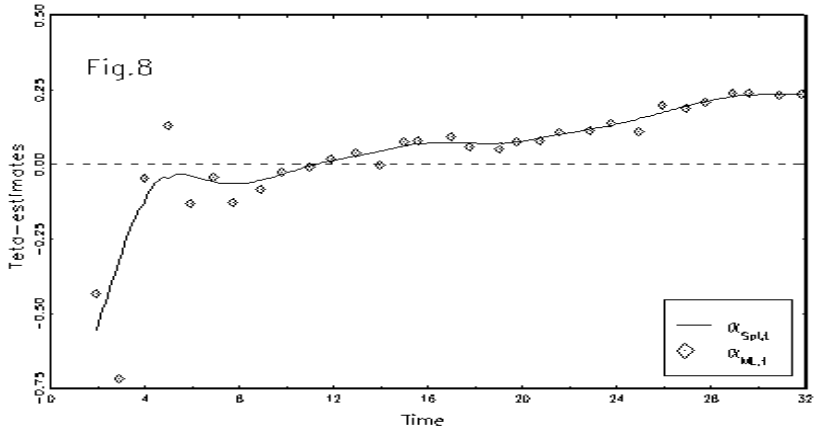
Tra ogni coppia di valori consecutivi di tale sequenza si è infine interpolato tramite la media ponderata

$$\tilde{\vartheta}_{Spl}(t) = \gamma_t \tilde{\vartheta}_{p,i-1}(t) + (1-\gamma_t)\tilde{\vartheta}_{p,i}(t), \quad \text{con } i = 2,..,31, \qquad (4.1)$$

dove, essendo scritta relativamente al generico sotto-intervallo $(t_{i-1}, t_i]$, ovviamente

si ha: $\gamma_t = (t_i - t)/(t_i - t_{i-1})$, con $t_{i-1} < t \le t_i$, e $\tilde{\vartheta}_{p,k}(t) = a_{0,k} + a_{1,k}t$ è la retta che dà $\tilde{\vartheta}_{p,k}$ al tempo t_k, con $k = i-1,\ i$.





La curva continua di punti $(t, \tilde{\vartheta}_{Spl}(t))$ è stata tracciata nella Fig. 5 tra quelli del $(t_i, \hat{\vartheta}_{S,i})$-plot, mentre nella Fig. 7 è stata ridisegnata rispetto al $(t_i, \hat{\vartheta}_i)$-plot delle stime di ML basate sulle distribuzioni statistiche $\mathcal{D}_{n,i}(x)$ non tanto per agevolare i confronti

quanto per meglio valutarne l'accoglibilità come linea estimativa di quella della funzione di dispersione non-poissoniana $\vartheta(t)$.

L'opportunità di cogliere la portata dell'approssimazione fornita dalla $\tilde{\vartheta}_{Spl}(t)$ si evince anche dalla Tav. III, dove per ogni valore t_i, a fonte della distribuzione teorica implicata dalla stima $\hat{\vartheta}_i$ di ML basata sulla distribuzione statistica $\mathcal{D}_{n,i}(x)$ osservata, si è riportata (nella riga sottostante) quella teorica implicata dalla corrispondente stima $\hat{\vartheta}_{S,i}$ di ML desunta in termini della rispettiva distribuzione statistica di perequazione scritta nella Tav. I. I confronti effettuabili per ciascun t_i non avrebbero infatti incoraggiato il ricorso al procedimento di tipo spline con cui si è giunti alla linea continua di $\tilde{\vartheta}_{Spl}(t)$.

La strada seguita per ottenere l'approssimazione $\tilde{\vartheta}_{Spl}(t)$ della *process-index function* $\vartheta(t)$ dell'HPAP è stata, pari pari, battuta nei riguardi della funzione di dispersione non-poissoniana dell'HBP. La sequenza $(\tilde{\alpha}_{p,1},..,\tilde{\alpha}_{p,i},..,\tilde{\alpha}_{p,r})$, con $r=31$, dei valori $\tilde{\alpha}_{p,i}=\tilde{\alpha}_{Spl,i}$ così desunti sono leggibili nella 3^ colonna della parte centrale della Tav. II.

Tali valori sono stati analogamente utilizzati per giungere a quelli d'interpolazione $\tilde{\alpha}_{Spl}(t)$ usati per tracciare la linea continua sia nella Fig. 6 che nella Fig. 8, al fine di vederne l'accoglibilità, in qualità di approssimazione della funzione $\alpha(t)$, tanto rispetto al $(t_i,\hat{\alpha}_{S,i})$-plot quanto al $(t_i,\hat{\alpha}_i)$-plot delle stime di ML implicate dalle distribuzioni statistiche $\mathcal{D}_{n,i}(x)$ osservate.

La constatazione fatta a proposito della Tav. III, trova ovviamente conferma nella Tav. IV, dove, per ogni valore t_i, nella riga sottostante a quella in cui compaiono le numerosità della distribuzione teorica imperniata sulla stima $\hat{\alpha}_i$ di ML (calcolata in termini della distribuzione statistica osservata), si sono riportate le numerosità teoriche ottenute in termini della rispettiva stima $\hat{\alpha}_{S,i}$ implicata dalla corrispondente distribuzione di perequazione della Tav. I.

Nella "lettura" dei risultati, non va comunque mai scordato, da un lato, che il "lisciamento" raggiunto con le linee di pseudo-spline $\tilde{\vartheta}_{Spl}(t)$ e $\tilde{\alpha}_{Spl}(t)$ tracciate nella Fig. 7 e nella Fig. 8, è in qualche misura legato alla scelta della segmentazione operata

dell'intero periodo d'osservazione e, da un altro lato, che aspetti del loro evolversi possono essere imputabili agli effetti d'instabilità di fattori connessi al modificarsi della struttura delle distribuzioni statistiche $\mathcal{D}_{n,i}(x)$ specie fin quando queste si appalesano sottopoissoniane.

5 – *Confronto tra i valori d'interpolazione $\tilde{\alpha}_{Spl}(t)$ basati sull'HBP e corrispondenti risultati di estimazioni di diversa concezione.*

5.1 – *Estimazione d'impostazione jackknife della process-index function $\alpha(t)$ ai tempi t_i.*

▭ Delineazione di alcune procedure pseudo-jackknife. L'iter seguito per giungere a detta approssimazione della funzione di dispersione non-poissoniana di entrambi i processi di nascita HPAP e HBP e le considerazioni svolte sulla base di confronti tra dette figure, a nostro avviso, giocano a sostegno dell'opportunità, in situazioni del genere, di incominciare col perequare similmente, o altrimenti, le distribuzioni statistiche $\mathcal{D}_i(x)$, anziché dal basare su di esse la determinazione delle sequenze delle stime di ML ($\hat{\vartheta}_1,..,\hat{\vartheta}_i,..,\hat{\vartheta}_r$) e ($\hat{\alpha}_1,..,\hat{\alpha}_i,..,\hat{\alpha}_r$).

E ciò non fosse altro perchè, essendo il profilo di una distribuzione statistica sottopoissoniana scarsamente consistente e quindi fortemente capace d'effetti d'instabilità, si riconoscono, in fondo, molto fragili le ragioni di una tale estimazione. Tant'è che gli andamenti dei plot di punti $(t_i,\hat{\vartheta}_i)$ e $(t_i,\hat{\alpha}_i)$, con i=1,..,r, si sono mostrati, come avviene di solito, talmente riflettenti effetti d'instabilità connessi alla struttura delle distribuzioni statistiche $\mathcal{D}_i(x)$ da non ritenersi bastevole il ricorso a tecniche per rettificare le sequenze delle stime di ML. È, del resto, per questo che si è passati a determinare linee di pseudo-spline, riguardandone l'andamento degno di una certa attenzione come approssimazione delle rispettive *process-index function* $\vartheta(t)$ e $\alpha(t)$.

Le riserve mosse in proposito hanno comunque indotto a pensare che, pur partendo dalle stime di ML basate sulle $\mathcal{D}_i(x)$, non era da escludersi di potersi arrivare alla messa a punto di qualche procedura implicante, per le stesse funzioni, un'approssimazione all'incirca della regolarità che viene da configurare, specie per la parte a dispersione non-poissoniana.

Il che, nel far balzare l'ambito dei *procedimenti jackknife*, ha ingenerato l'intento

di articolarne uno, radicato sì sugli stimatori di ML basati sulle $\mathcal{D}_i(x)$, ma capace di comportare stime quasi insensibili ai fattori d'instabilità (Cfr: Ferreri, 1996 e, soprattutto, Ferrante e Ferreri, 1996), cioè stime non già prive di tali effetti – perché sarebbe privo di senso pensarlo – ma da riconoscere loro una portata ridotta al punto da valutare basso il rischio di risultati aberranti.

Si è perciò ripresa l'analisi dei plot delle stime $\hat{\alpha}_t$ e $\hat{\vartheta}_t$ di ML delle rispettive *process-index function* $\alpha(t)$ e $\vartheta(t)$ ai tempi $t = t_i$, per *i*=1,..,31, al fine inquadrare quelle cause d'instabilità da cui non fosse dato prescindere.

Tra gli aspetti particolarmente salienti, di questi plot ha nuovamente colpito più che altro la presenza di un salto d'attenzione ogni volta che, nella sequenza delle distribuzioni osservate, una è più lunga della precedente per l'aggiunta di una classe del tipo ($c,\ n_c = 1$); salto da dirsi poi tanto più accentuato quanto più rilevante la sotto-poissonianità caratterizzante la parte iniziale degli stessi plot.

Su di questo si è pertanto ritenuto di imperniare il tentativo di delineare, con detto intendimento, un approccio di estimazione jackknife.

Per richiamarne la tecnica con riferimento alla circostanza in questione, ma in modo da contemplare entrambi i modelli di probabilità in precedenza utilizzati, per semplicità si designa con θ il livello della funzione di dispersione non-poissoniana ad un dato istante (t_i) del processo di nascita adottato e si indica con $\hat{\theta}$ la relativa stima basata sulla distribuzione statistica

$$\mathcal{D}_n(x) = \{\ (x, n_x); x = 0,1,..,c\ ,\ n = \sum_{x=0}^{c} n_x\ \} \qquad (5.1)$$

della sequenza (1.8) costruita in termini dell'evidenza statistica rilevata non oltre lo stesso tempo t_i del periodo d'indagine.

Come si sa, dandosi per scontato di assumere $\bar{\theta} = \hat{\theta}_n$, cioè di muovere dalla stima di ML di θ, la classica procedura jackknife fa, dapprima, considerare la distribuzione di frequenza

$$\mathcal{D}_n(\hat{\theta}) = \{\ \hat{\theta}_{n-1,x}, n_x\ \}\ ,\quad \text{con}\quad x = 0,\ 1,...,c, \qquad (5.2)$$

dove $\hat{\theta}_{n-1,x}$ denota la stima implicata da ciascuna distribuzione statistica delle n-1 unità di rilevazione che rimangono escludendone, via via, una di quelle caratterizzate da x eventi (incidenti); e poi vedere in

$$\hat{\theta}_{JK} = n\,\hat{\theta}_n - (n-1)\,\bar{\theta}_{n-1}\ ,\quad \text{dove}\quad \bar{\theta}_{n-1} = \frac{1}{n}\sum_{x=0}^{c} \hat{\theta}_{n-1,x} n_x\ , \qquad (5.3)$$

lo stimatore jackknife di θ.

Ovviamente, stando allo stimatore $\hat{\theta}_n$, la (5.3) non può essere usata se $c = 2$ con $n_2 = 1$ $(n_0, n_1 > 1)$ dato che, in tal caso, la procedura jackknife implicherebbe una distribuzione con $c = 1$, per la quale, come si è detto a proposito della (3.22), non è dato rifarsi all'estimazione di ML.

A parte questo, quando in pratica il numero c, con $c > 2$ come da ora sarà sempre supposto, è comunque piccolo, ad esempio $c = 3$, la (5.3) può modificare gli effetti d'instabilità delle stime di ML al punto da comportare, per $\hat{\theta}_{JK}$, un valore positivo ancorché siano negative tanto $\hat{\theta}_n$ quanto tutte le stime $\hat{\theta}_{n-1,x}$ della distribuzione (5.2), e cioè da renderli più accentuati di quelli delle stime di ML.

Questo capita nel caso che, per l'estimazione dell'HBD(t_i), ci si avvalga della distribuzione statistica della Tav. I corrispondente a $t_i = 6.91$: a fronte sia della stima $\hat{\alpha} = -0.0421$ di ML che di $\hat{\alpha}_{n-1,x} < 0$, per $x = 0,..,4$, lo stimatore jackknife (5.3) comporta infatti $\hat{\alpha}_{JK} = 0.01405$. Un confronto tra le stime $\hat{\alpha}_{n-1,x}$ (negative) mostrerebbe però che il valore di $\hat{\alpha}_{JK}$ è principalmente dovuto ad $\hat{\alpha}_{n-1,4} = -0.29161$.

E poiché, ripetendo il calcolo per parecchi casi analoghi, la circostanza è risultata imputabile essenzialmente al valore di $\hat{\theta}_{n-1,c}$, sulla scorta delle conclusioni tratte con l'analisi sulle cause d'instabilità si è convenuto di spostare l'attenzione sulla distribuzione $\mathcal{D}_n^{(c)}(\hat{\theta})$ che consegue escludendo l'ultima classe in quella $\mathcal{D}_n(\hat{\theta})$ definita dalla (5.2), nonché di assumere, quale stimatore jackknife,

$$\hat{\theta}_{JK}^{(c)} = n\,\hat{\psi}_n - (n-1)\,K_{\hat{\theta}}\,\overline{\theta}_{n-1}^{(c)}, \tag{5.4}$$

che balza col prendersi atto che:

(*i*) la media

$$\overline{\theta}_{n-1}^{(c)} = \frac{1}{n-n_c}\sum_{x=0}^{c-1}\hat{\theta}_{n-1,x}n_x$$

della distribuzione $\mathcal{D}_n^{(c)}(\hat{\theta})$ è suggerita, come si è detto, dal proposito di mitigare l'effetto dominante dell'ultima classe della coda destra della distribuzione statistica presa in considerazione;

(*ii*) la combinazione lineare

$$\hat{\psi}_n = v\overline{\theta}_{n-1} + (1-v)\hat{\theta}_n \tag{5.5}$$

è preferibile a $\hat{\theta}_n$ per dar modo di attutirne gli effetti d'instabilità tramite un opportuno peso υ, da specificare;

(*iii*) il coefficiente

$$K_{\hat{\theta}} = \frac{(n-1)(n-n_C)}{n(n-1)-n_C(n-\phi/\hat{\theta}_n)} \tag{5.6}$$

consegue dall'imporre che valga la relazione $\hat{\theta}_{JK}^{(c)} = \hat{\theta}_n$ allorquando, invece del parametro θ di dispersione stimato da $\hat{\theta}_n$, si considera il valore medio μ e, per questo, la media aritmetica campionaria $M = M_n$ con $\phi = c$.

Stando al parametro θ, sovente l'assunzione della media $\overline{\theta}_{n-1}$ per ϕ si appalesa comunque abbastanza giustificata dato che la (5.4) comporta $\hat{\theta}_{JK}^{(c)} \cong \hat{\theta}_n$ se $\overline{\theta}_{n-1} \cong \hat{\theta}_n$.

L'impiego della (5.4) ha, ad ogni modo, portato a constatare che, per υ=0, 0.5 e qualche altro valore, l'assunzione di $\overline{\theta}_{n-1}^{(c)}$ in luogo della media $\overline{\theta}_{n-1}$ della distribuzione (5.2) di frequente comporta un effetto maggiore di quello voluto.

Per aggirare l'inconveniente si è passati a considerare la media ponderata

$$\hat{\theta}_{JKw} = w\hat{\theta}_{JK}^{(c)} + (1-w)\hat{\theta}_{JK}\ , \tag{5.7}$$

tra lo stimatore jackknife $\hat{\theta}_{JK}^{(c)}$ e il corrispettivo classico $\hat{\theta}_{JK}$, da usare però con un peso w tale da conferire a $\hat{\theta}_{JK}^{(c)}$ un ruolo sempre meno rilevante al crescere di *c*.

Se la (5.7) poteva segnare un certo passo sotto l'aspetto che l'ha suggerita, di fatto ha complicato la vita in quanto richiede di specificare un particolare criterio per l'assegnazione del valore al peso w.

L'intento di darne una formulazione in termini delle caratteristiche della distribuzione statistica considerata ha però fatto rammentare che la stima $\tilde{\Lambda}_i = \tilde{\Lambda}(t_i)$, del metodo dei momenti, della funzione rischio cumulativo (3.12b) dell'HBP può essere espressa dal prodotto $\tilde{\Lambda}_i = M_i^2\delta_i$, dove δ_i sta ad indicare il valore assunto al tempo t_i dal rapporto

$$\tilde{\delta} = \frac{\log\tilde{\sigma}^2 - \log M}{\tilde{\sigma}^2 - M} > 0, \tag{5.7a}$$

che si è mostrato (Ferreri, 1992) di notevole aiuto nell'analisi dell'instabilità degli stimatori classici.

E, in termini di δ e di c, si è giunti a constatare che, in pratica, l'espressione

$$w = \frac{\tilde{\delta}}{\tilde{\delta} + \log c / \tilde{\delta}} \quad \text{per} \quad c \geq 2\,, \tag{5.8}$$

può dirsi alquanto accoglibile sia per il fatto di dipendere dalla differenza tra $\tilde{\sigma}^2$ ed M tramite $\tilde{\delta}$ attribuendo così a $\hat{\theta}_{JK}^{(c)}$ un ruolo che va decrescendo come la stima $\hat{\theta}$ di θ cresce a partire da un valore negativo, sia perché rende lo stimatore pseudo-jackknife (5.7) in grado di fornire, per la *process-index function* del processo di nascita, valori d'attenzione in corrispondenza dei livelli t_i del periodo (0,*T*] d'osservazione.

Le considerazioni fatte circa l'andamento di w rispetto a c e il fatto che, con c piccolo, i valori di $\overline{\theta}_{n-1}$ e $\hat{\theta}_n$ spesso non differiscono di molto, sono stati poi ritenuti per lo più bastevoli a motivare il ricorso alla (5.5) col peso

$$v = \frac{\tilde{\delta}}{2\tilde{\delta} + c - 2}\,, \qquad 2 \leq c < \infty\,, \tag{5.9}$$

visto che si riduce a 0.5 per c =2 e decrescente al crescere di c .

Ancorché molteplici siano le obiezioni che potrebbero essere mosse alla (5.7), anche perché necessita di qualche specificazione a partire dalla (5.5) e dal coefficiente (5.6) quando $\hat{\theta}_n$ non coincide con la media campionaria M , ci è parso meritevole del ruolo di opportuno estimatore quanto meno mirando, come si è detto, ad attenuare l'instabilità che caratterizza i trend delle stime di ML della funzione di dispersione non-poissoniana dei due modelli di probabilità adottati per il processo di nascita.

Per il caso di c =2 con $n_c = 1$, non potendosi usare neanche la (5.4) se la media $\overline{\theta}_{n-1}$ è assunta sia nella (5.5) che nella (5.6) invece di ϕ, si è visto che il più delle volte l'ostacolo è soddisfacentemente superabile ponendo $\overline{\theta}_{n-1}^{(c)}$ nelle (5.5), (5.6) e (5.7) rispettivamente al posto di $\overline{\theta}_{n-1}$, ϕ e $\hat{\theta}_{JK}$.

Constatato che sono parecchie le versioni semplificate della procedura conclusasi con la (5.7) che possono trovare consenso in pratica, per chiarezza, se non per evitare equivoci, si è convenuto (Ferrante e Ferreri, 1996):

a) di indicare con A0 la procedura che si concreta con la (5.7) in termini delle funzioni peso (5.8) e (5.9) per $c > 2$ e per $c = 2$ se $n_c > 1$ allorché insorge l'opportunità di porre $\overline{\theta}_{n-1}$ invece di $\hat{\theta}_{JK}$ nella (5.7); e di riservare invece la sigla A1 al caso particolare di *c*=2 con $n_c = 1$ per il quale generalmente si sta alle suddette posizioni;

b) di designare con B0 e B1 le versioni semplificate che discendono ordinatamente da A0 ed A1 ponendo $v = 0$ nella (5.5);

c) di usare C0 e C1 per indicare la procedura che, oltre a $v = 0$, contempla pure l'assunzione di $\overline{\theta}_{n-1}$ o di $\overline{\theta}_{n-1}^{(c)}$ al posto di $\hat{\theta}_{JK}$ nella (5.7).

In pratica, anche se di solito gli elementi suggeriti dalla circostanza esaminata non sono decisivi ai fini della scelta di una delle procedure indicate o per formularne di nuove, non va comunque scordato, da un lato, che il peso

$$v_1 = \frac{\tilde{\delta}}{2(\tilde{\delta} + c - 2)}, \qquad 2 \le c < \infty, \tag{5.10}$$

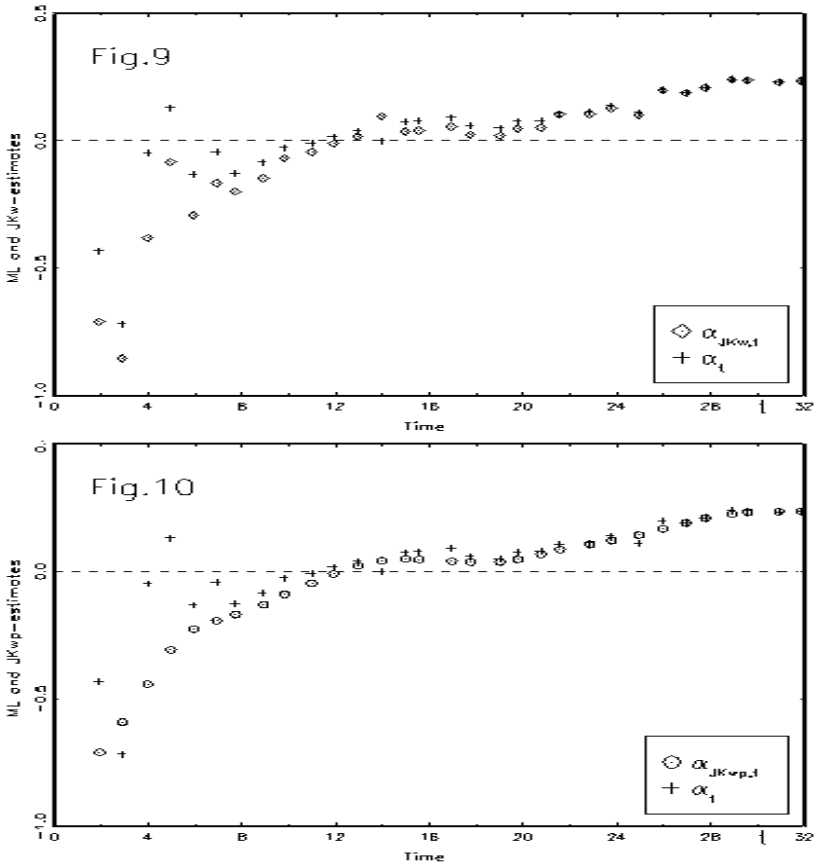
può essere preferibile a quello v definito dalla (5.9) e, da un altro lato, che nella (5.5), invece di $\overline{\theta}_{n-1}$, talora affiora l'opportunità di porre la media $u\overline{\theta}_{n-1} + (1-u)\overline{\theta}_{n-1}^{(c)}$, con $0 \le u \le 1$ o, in particolare, con u=0.5 qualora si ritenga che sia il caso di ottenere valori più prossimi alle stime $\hat{\theta}_n$ di ML.

▭ Stime pseudo-jackknife della *process-index function* $\alpha(t)$. Finora ci si è avvalsi delle distribuzioni $\mathcal{D}_{n,i}(x)$, per *i*=1,..,31, per compiere inferenze sulla funzione di dispersione non-poissoniana di entrambi i processi HBP e HPAP. A questo punto, onde evitare pressoché la ripetizione del discorso partendo dalle rispettive stime $\hat{\alpha}_i$ e $\hat{\vartheta}_i$ di ML, ci si limita a dar conto, dapprima, dell'impiego di qualcuna delle suddette procedure di tipo jackknife unicamente con riferimento al processo iperbinomiale e, poi, dell'ottenimento, in termini delle rispettive stime di $\alpha(t_i)$, che saranno dette *pseudo-jackknife*, di quelle da confrontare con le stime $\tilde{\alpha}_{Spl}(t_i)$ plottate nella Fig. 8.

Stando, anzitutto, alla procedura per così dire più generale, siglata con A0, sulla base di ciascuna delle distribuzioni statistiche $\mathcal{D}_{n,t}(x)$, per $t = t_i$ con $i = 1,..31$, tramite la (5.7) si è giunti alle stime $\hat{\alpha}_{JKw,t}$ scritte nella 5^ colonna della Tav. 5 a fronte di quelle $\hat{\alpha}_t$ di ML. Aderendo alla procedura A1, la media $\overline{\theta}_{n-1}^{(c)}$ è stata invece utilizzata nel modo suddetto per ottenere le stime pseudo-jackknife dalle prime due distribuzioni statistiche caratterizzate da *c*=2 con $n_c = 1$.

Nella Fig. 9, al $(t, \hat{\alpha}_t)$-plot delle stime di ML è stato sovrapposto il $(t, \hat{\alpha}_{JKw,t})$-plot delle stime pseudo-jackknife così ottenute al fine di evidenziare come la differenza

$\hat{\alpha}_t - \hat{\alpha}_{JKw,t}$ vada diventando sempre più trascurabile al crescere di t grazie al corrispondente aumentare di $c(t)$ a seguito del crescente numero degli accadimenti.



L'andamento di tale differenza, anche se potrebbe far dire che da t=13 – allorché le stime di α_t presentano un trend di valori positivi – ci si potrebbe limitare alla considerazione delle sole stime di ML, ovviamente sollecita a notare che le corrispondenti stime pseudo-jackknife mostrano una tendenza quanto meno di pari accoglibilità, specie se vista a prolungamento di quella per t<13.

Benché, in questa prima parte dei plot della Fig. 9, le "irregolarità" delle stime

pseudo-jackknife siano in qualche misura meno accentuate di quelle delle stime di ML, tra gli effetti d'instabilità riflessi da quelle $\hat{\alpha}_{JKw,t}$ almeno fino a $t<5$ appaiono comunque sempre rilevanti quelli imputabili all'aggiungersi, nelle distribuzioni statistiche $\mathcal{D}_{n,t}(x)$, di una nuova classe, soprattutto se questa è di numerosità $n_c=1$.

Anche alla luce della Fig. 9 si è così avvertita l'opportunità di operare sulle stime $\hat{\alpha}_{JKw,t}$ un procedimento di perequazione per medie mobili sulla falsariga di quanto si è fatto a proposito di quelle $\hat{\alpha}_i$ di ML del processo iperbinomiale.

Il procedimento che, alla luce di precedenti esperienze (Ferreri, 1992), si è ritenuto di impiegare è espresso dalla relazione

$$Y_i = a_{-1,i}X_{i-1} + a_0 X_i + a_{1,i}X_{i+1}\,, \quad \text{con} \quad a_{-1,i} + a_0 + a_{1,i} = 1 \tag{5.11}$$

$$\text{dove:}\quad a_{-1,i} = a_0 - (3a_0 - 1)\frac{t_i - t_{i-1}}{t_{i+1} - t_{i-1}}\,, \qquad a_{1,i} = a_0 - (3a_0 - 1)\frac{t_{i+1} - t_i}{t_{i+1} - t_{i-1}}\,. \tag{5.12}$$

Per dati X_t equidistanti, com'è usuale nelle serie storiche, la (5.11) comporta la relazione $0 < a_{-1} = a_1 < a_0$ (indipendente da *i*) dato che si suppone a_0>1/3. Naturalmente, con a_0=1/2, la (5.11) fornisce i valori desumibili tramite interpolazione lineare.

Nella Fig. 10 compare il plot dei valori $\hat{\alpha}_{JKwp,t}$ ottenuti perequando due volte le stime $\hat{\alpha}_{JKw,t}$ tramite la (5.11) con a_0=0.4, scelto con l'intento di attenuare gli effetti d'instabilità in modo però da coglierne abbastanza attendibilmente il trend. E poiché la (5.11) non fornisce un valore per il primo e per l'ultimo termine della sequenza delle stime su cui essa viene operata, ogni volta si sono assunti i rispettivi valori iniziali onde evitare di ridurre la lunghezza del plot.

Non può sorprendere che un confronto tra i plot della Fig. 9 con i corrispondenti della Fig. 10 possa portare, da un lato, a riguardare lo stimatore pseudo-jackknife impiegato come uno strumento per rendere "robusti" certi stimatori classici e, da un altro lato, a continuare a far uso della perequazione per medie mobili quando s'intende cogliere tendenze con scarsa perdita di "contenuto" dell'evidenza statistica nella forma utilizzata.

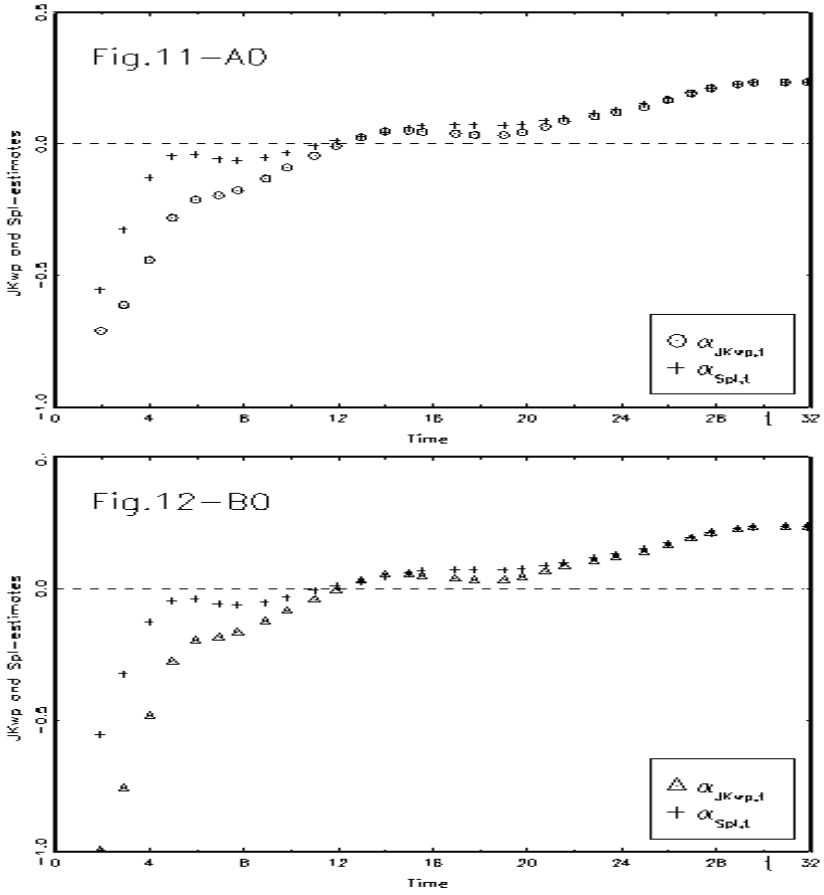
Tutto questo si evince anche dalle stime

1.407, 1.633, 2.263, 3.560, 4.725, 5.122, 5.667, 7.586, 11.12, 21.97, 121.7,

che i valori implicati dalla perequazione, e leggibili nella Tav. V, comportano per $-1/\alpha(t_i)$ quando $\alpha(t_i)$<0: le rispettive stime di $\kappa+1$ della (3.12) risultano infatti

$\tilde{\kappa}+1$=2, 2, 3, 4, 5, 6, 6, 8, 12, 22, 122

contro i corrispondenti valori di $c = max(x)$ = 2, 2, 3, 3, 3, 4, 4, 4, 5, 5, 5 delle distribuzioni statistiche $\mathcal{D}_{n,t}(x)$ per le quali n_c =1. Ciò che è peraltro indicativo della diversità di effetti che i fattori d'instabilità hanno sui *p.m.* (3.12) e (3.13).



Non può sfuggire che, se il confronto tra i plot corrispondenti delle Figg. 9 e 10 aiuta ad avere contezza dello strumento usato, induce pure a vedere come le stime pseudo-jackknife $\tilde{\alpha}_{JKwp,t}$ si pongono rispetto alle rispettive stime $\tilde{\alpha}_{Spl,t}$ di pseudo-

spline della *process-index function* dell'HBP; stime scritte nella Tav. V insieme a quelle $\hat{\alpha}_{JKwp,t}$ di entrambe le procedure A0 e B0 onde rimarcarne le differenze.

Nella Fig. 11 sono stati perciò disegnati tanto il $(t_i, \tilde{\alpha}_{JKwp,t_i})$-plot implicato dalla procedura A0 quanto quello di punti $(t_i, \tilde{\alpha}_{Spl,t_i})$, i cui andamenti fanno segnalare tratti apprezzabilmente differenti quando le stime $\tilde{\alpha}_{JKwp,t_i}$ ed $\tilde{\alpha}_{Spl,t_i}$ sono negative, ed una tendenza a coincidere al crescere di *t*, cioè all'aumentare del numero degli accadimenti, dal momento che tali stime divengono positive.

Per cogliere in quale misura ciò poteva dirsi imputabile alla tipicità della procedura jackknife usata, si è passati ad impiegare pure quella B0, che ha portato ai valori corrispondentemente leggibili nella 7^ e nella 8^ colonna della Tav. V, utilizzati per la costruzione dei plot della Fig. 12.

Col prendersi atto che quest'ultima figura sostanzialmente avvalora le conclusioni tratte alla luce della Fig. 11, è venuto però da chiedersi non solo a quale dei due profili, tra quello delle stime $\tilde{\alpha}_{Spl,t_i}$ e quello pseudo-jackknife di base A0, era da accordare la preferenza, ma anche se era il caso di porsi la domanda, e non tanto perché la forma evolutiva del secondo può, a nostro avviso, indurre a privilegiarlo.

5.2 − *Delineazione di una procedura tipo-bootstrap per l'estimazione della process-index function* $\alpha(t)$ *ai tempi* t_i.

▭ Presupposti e caratteristiche della procedura. Poiché il quesito emerso fa indirizzare l'attenzione su un approccio alternativo, ci si è rifatti alla cosiddetta tecnica bootstrap con l'intento di modificarla in modo da renderla adeguata alla circostanza, visto che gli effetti d'instabilità sono da ritenersi connessi ad aspetti che vanno tipicamente manifestando nel tempo le distribuzioni osservate al crescere del numero degli accadimenti (incidenti) del reale processo di nascita.

Secondo la nota tecnica bootstrap (Cfr.: Efron, 1979 ed Efron and Tibshirani, 1986), ciascuna distribuzione statistica di numerosità $\mathcal{D}_{n,i}(x)$, per $i = 1,..,31$, è stata così campionata (Cfr. Ferrante e Ferreri, 1997) estraendo casualmente senza sostituzione da essa *n*=108 valori e, poi, parimenti ricampionata non però un numero fisso di volte, bensì fino ad aversi *m*=200 distribuzioni statistiche con $max(x)$ maggiore di 1 onde ottenere, da ciascuna, la stima di ML del parametro $\alpha_t = \alpha(t)$, con $t = t_i$, della distribuzione di probabilità (3.12) dell'HBP.

In termini delle stime $(M_{ij}, \hat{\alpha}_{ij})$, $j = 1,..,m$, implicate dalle *m*=200 distribuzioni di ricampionamento da $\mathfrak{D}_{n,i}(x)$, per ciascun i, con $i = 1,..,r = 31$, si è quindi proceduto a costruire, prendendo 0.10 come ampiezza di classe costante a partire da –1, la distribuzione di frequenza per classi delle bootstrap-stime $\hat{\alpha}_{ij}$ al fine d'aver modo di cogliere il modificarsi, nel tempo (cioè al crescere del numero degli eventi incidente), degli effetti d'instabilità verosimilmente caratterizzanti le distribuzioni per classi $\mathfrak{D}_{n,i}(\hat{\alpha}_{ij})$ costruite con le stime $\hat{\alpha}_{ij}$ corrispettive delle $\mathfrak{D}_{n,i}(x)$.

E si è visto che, con le stime $\hat{\alpha}_{ij}$ di ML per lo più negative, come avviene per ciascuno dei primi valori di *i*, l'entità della dispersione non poissoniana attribuibile a fattori d'instabilità poteva dirsi così rilevante da implicare una media $\bar{\alpha}_i$ tale da rifletterne effetti perfino della portata riconoscibile alle relative stime di ML.

Per superare l'inconveniente si è allora deciso di fissare, per ogni *i*, un relativo intervallo (L_i, U_i) e di riguardare come *outlier* da scartare ciascuna stima $\hat{\alpha}_{ij}$ non compresa in esso.

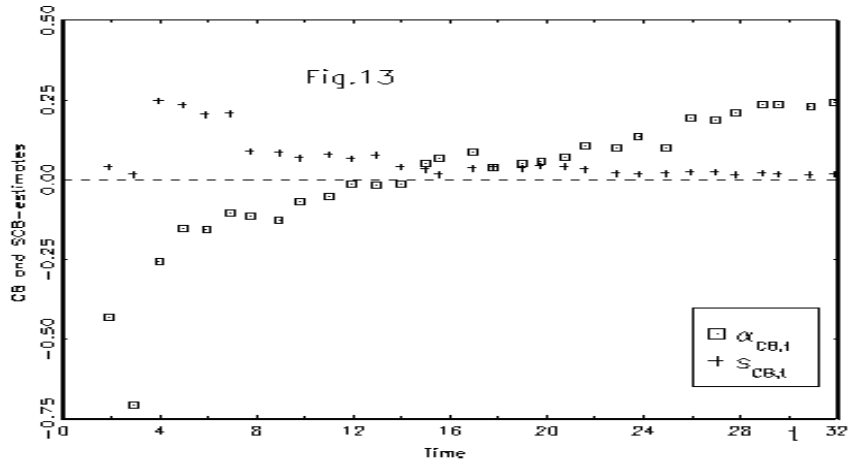
Per la scelta, non certo semplice, degli estremi L_i ed U_i si è partiti dalla constatazione che gli effetti d'instabilità vanno attenuandosi all'aumentare sia di $c_i = max(x)$ della distribuzione statistica $\mathfrak{D}_{n,i}(x)$ che dell'indicatore di dispersione campionario $\tilde{I}_D = \tilde{\sigma}^2 / M$ (Cfr.: Olkin et al., 1981) a cominciare da un valore minore di uno.

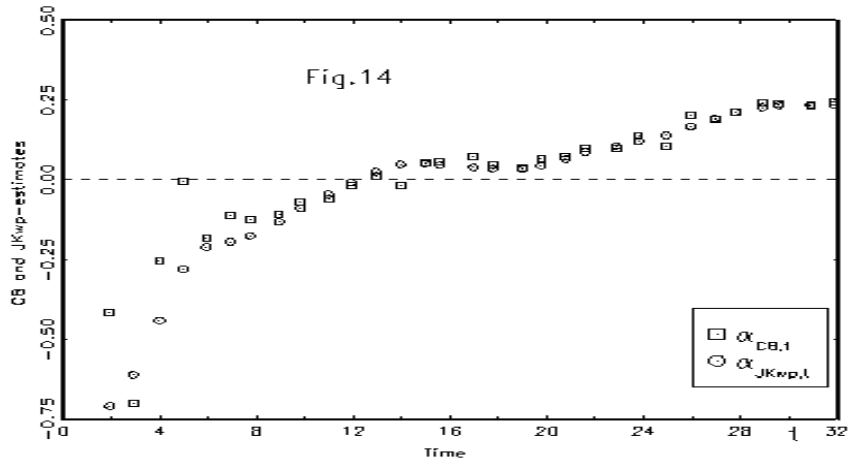
Quali estremi dell'intervallo di non rifiuto delle stime $\hat{\alpha}_{ij}$, si è quindi convenuto di assumere i quantili $Q_{p,i}$ e $Q_{2p+0.5,i}$ con $p = 0.25C_i$, dove per C_i si è adottata la media geometrica

$$C_i = \left(\frac{c_i - 1}{c_i + 1} \frac{e^{M_i} - 1}{e^{M_i} + 1} \right)^{0.5} \qquad \text{dei due indici} \qquad \frac{c_i - 1}{c_i + 1} \quad \text{e} \quad \frac{e^{M_i} - 1}{e^{M_i} + 1}.$$

E ciò perché questi, essendo, per ogni *i*, entrambi compresi tra 0 ed 1 e crescenti ordinatamente rispetto sia a c_i e ad M_i, danno corrispettivamente modo di ottenere un intervallo di "accettazione" di limite inferiore L_i tendente alla mediana $Q_{0.50,i}$ delle *m* stime $\hat{\alpha}_{ij}$ e di limite superiore U_i congiuntamente crescente verso l'ultima statistica d'ordine, non molto maggiore della stessa mediana. Mediana che, essendo peraltro assai meno instabile della media aritmetica delle stime $\hat{\alpha}_{ij}$ – e perciò da pre-

ferire a quest'ultima –, è stata denominata *central bootstrap estimate* e quindi siglata con CBE, nonché indicata con $\bar{\alpha}_{CB,i}$, onde distinguerla da quella $\bar{\alpha}_{B,i}$ relativa al caso in cui non viene scartata alcuna bootstrap-stima $\hat{\alpha}_{ij}$.





Poiché la CB-procedura non consente di valutare l'accuratezza delle stima $\bar{\alpha}_{CB,i}$, per ciascun i si è provveduto a replicarla k=20 volte ed a considerare poi ordinatamente la media aritmetica e la standard deviation

$$\bar{\alpha}_{CB,i} = \tfrac{1}{k}\sum_{s=1}^{k} \hat{\alpha}_{CBs,i} \ , \qquad s_{CB,i} = [\tfrac{1}{k-1}\sum_{s=1}^{k} (\hat{\alpha}_{CBs,i} - \bar{\alpha}_{CB,i})^2]^{1/2} \ ,$$

nonché le corrispondenti $\bar{\alpha}_{B,i}$ e $s_{B,i}$ implicate da tutte le stime $\hat{\alpha}_{Bs,i}$.

▭ Risultati della CB-procedura. I valori ottenuti dalle 31 distribuzioni statistiche osservate tramite la CB-procedura sono leggibili nelle ultime due colonne della Tav. V. Quelli corrispondentemente forniti dagli stimatori $\bar{\alpha}_{B,i}$ e $s_{B,i}$ non vengono qui trascritti in quanto, com' è possibile constatare alla luce dei plot disegnati in Ferrante e Ferreri (1997), sono fortemente influenzati da effetti d'instabilità.

Sulla base dei risultati riportati, nella Fig. 13 sono stati sovrapposti i due $(t_i, \bar{\alpha}_{CB,i})$- e $(t_i, s_{CB,i})$-plot al fine di rimarcare che: 1) i primi due valori della standard deviation sono piccoli perché le rispettive distribuzioni statistiche osservate presentano c_i=2 con n_c=1 e le corrispondenti distribuzioni bootstrap risultano di coda destra veramente simili; 2) i valori più elevati di $s_{CB,i}$ appaiono per t_i da 3 a 7 a causa dei notevoli effetti d'instabilità che si hanno per c_i=3 con n_c=1 e per il fatto che, passando a c_i=4, si ha in particolare $n_3 = n_4 = 1$; 3) al crescere di t_i, la standard deviation $s_{CB,i}$ presenta un trend decrescente in armonia col corrispondente comportamento degli effetti d'instabilità, come lascia cogliere l'alquanto regolare andamento crescente del diagramma punteggiato delle BC-stime $\bar{\alpha}_{CB,i}$.

É anzi proprio la dinamica di queste stime che ha suggerito di sovrapporre, nella Fig. 14, il $(t_i \bar{\alpha}_{CB,i})$-plot a quello di punti $(t_i, \tilde{\alpha}_{JKwp,i})$. Questi mostrano infatti andamenti quasi coincidenti a partire $t_i \cong 6$ benché non si sia ritenuto di applicare un procedimento di perequazione sulle stime $\bar{\alpha}_{CB,i}$, dato che sono il risultato di una depurazione da outlier.

Stando alle distribuzioni statistiche $\mathcal{D}_{n,i}(x)$ considerate come osservate, può così dirsi che il confronto tra i due plot della Fig. 14 fornisce una risposta al quesito che ha suggerito l'estimazione bootstrap. Rispetto allo spline operato su stime di ML, per vari aspetti la procedura bootstrap e quella pseudo-jackknife utilizzata si prospettano infatti entrambe preferibili per inferenze sulla funzione di dispersione non-poissoniana, la quale non può che essere vista basilare quando s'incentra l'attenzione su distribuzioni statistiche pure con l'intento di saggiarne la poissonianità.

6 – *Quadro conclusivo.*

Sebbene sia alquanto vasta la letteratura statistica sull'analisi delle distribuzioni statistiche di dati di conto tramite l'impiego di modelli di probabilità, sono davvero pochi i lavori che riguardano i relativi eventi ai tempi del periodo in cui si sono registrati sulle *n* unità di rilevazione.

Solitamente, infatti, una distribuzione statistica del genere è costruita in termini dell'evidenza statistica che si è acquisita durante un intervallo di tempo $(0,T]$ come se gli accadimenti fenomenici d'interesse si fossero identicamente verificati allo stesso istante T. E poiché le distribuzioni statistiche così ottenute presentano per lo più sovradispersione rispetto al modello di riferimento concepito, si è andata sviluppando una metodologia d'analisi concernente, per gran parte, test per saggiare la "conformità" alla situazione di riferimento e modelli di probabilità per una "lettura" dell'alternativa configurata.

Insomma, tutto questo è avvenuto senza tenersi granché conto delle implicazione di quella che, per semplicità, può dirsi *ottica atemporale*, aderire alla quale equivale a scegliere di scordarsi che gli eventi del fenomeno d'attenzione – ancorché riguardabili identici – vengono colti, registrati ed inquadrati rispetto al tempo di accadimento secondo un processo di nascita seguito lungo l'intero periodo di osservazione $(0,T]$.

Definito inequivocabilmente a monte l'evento incidente, stando a questa realtà sono stati rilevati i tempi degli eventi-incidente capitati, nel rapporto uomo-macchina, ad n=108 operai (unità statistiche di rilevazione) di un'azienda metalmeccanica e, come situazione di riferimento, si è considerato il processo di nascita di Poisson. E dato che gli studi in tema di sovradispersione si avvalgono essenzialmente, per non dire sempre, della distribuzione statistica implicata da detta ottica, si è pensato di superarla passando ad articolare l'intervallo d'osservazione in r successivi intervalli $(t_{i-1}, t_i]$, per $i = 1,2,..,r$ con $t_0 = 0$ e $t_r = T$, ed a costruire le distribuzioni statistiche $\mathcal{D}_{n,i}(x)$, sempre per i=1,..,r, delle n unità di rilevazione rispetto al numero degli eventi-incidente loro successi nell'intervallo $(0, t_i]$.

Ma, come si è calcolato per ciascuna distribuzione statistica $\mathcal{D}_{n,i}(x)$ l'indicatore di dispersione campionario $\tilde{I}_{D,i} = \tilde{\sigma}_i^2 / M_i$, si è avuto modo di constatare che la sequenza dei relativi valori, al crescere di t_i, andava, più o meno regolarmente, manifestando un trend evolutivamente crescente da una situazione di sottopoissonianità ad una di sovrapoissonianità.

E poiché si è appurato che si era in presenza di una circostanza da dirsi tipica non del caso oggetto di studio, ma della sequenza delle distribuzioni statistiche in tal modo ottenibili per il processo di nascita almeno per una segmentazione abbastanza fine del periodo d'osservazione, ci si è trovati a dovere inevitabilmente tralasciare l'ottica atemporale per inquadrare l'andamento di $\tilde{I}_{D,i}$ nel contesto della non-poissonianità nel tempo d'osservazione del fenomeno e, quindi, la situazione di riferimento poissoniana essenzialmente in un punto di tale andamento.

Questo contesto ha naturalmente reso più complesso il problema della costruzione del modello statistico poiché doveva essere impostato per un processo di nascita. Le congetture con cui si è giunti a configurare lo svolgimento del fenomeno ha però indotto a riguardare ragionevoli per la circostanza tanto il cosiddetto processo iperbinomiale (HBP) quanto l'iperPólya-Aeppli process (HPAP), che, come si è visto, comportano una *process-index function* capace di contemplare l'andamento che l'evidenza statistica fa cogliere per la dispersione non poissoniana.

Ma se può essere stata facile la messa a punto dei due modelli di processo di nascita accoglibili per la circostanza, non lo è stata certo altrettanto l'estimazione delle relative *funzioni di dispersione non-poissoniana* in termini delle distribuzioni statistiche $\mathcal{D}_{n,i}(x)$ della sequenza ancorata ai tempi t_i considerati. A farla da padrone è stata invero l'instabilità mostrata dallo stimatore di ML: nella parte iniziale dei plot costruiti con le stime di ML corrispettivamente ottenute per i due modelli – le quali con un valore negativo denotano dispersione sottopoissoniana –, gli effetti d'instabilità si sono infatti mostrati così rilevanti da indurre a portare avanti il discorso con approcci che potessero consentirne un'attenuazione tale da permettere l'individuazione di un trend riguardabile ragionevolmente attendibile per la *process-index function* del processo di nascita scelto.

E questo sulla scorta della constatazione che, da un lato, al crescere di t_i (i=1,..,r) a partire dall'inizio di ciascun plot, e cioè all'aumentare del numero degli eventi nelle singole unità statistiche, gli effetti d'instabilità andavano riducendosi fino a divenire praticamente trascurabili passando alle stime positive; e che, da un altro lato, la portata di tali effetti era comunque da vedersi ancorata alla struttura con si utilizzava l'evidenza statistica. Da un'analisi delle cause d'instabilità sulle stime di ML è emerso infatti che l'effetto conseguente da una distribuzione statistica che, rispetto alla precedente della sequenza, presenta una ulteriore classe, è, specie se del tipo $(c, n_c = 1)$, particolarmente rilevante per piccoli valori di c, in presenza di sottopoissonianità e fin quando la rispettiva media aritmetica è, al più, di poco maggiore di 1.

Le caratteriste riscontrate nell'andamento delle stime di ML della funzione di dispersione non-poissoniana di entrambi i modelli di processo di nascita formulati ha così portato a considerare, come *primo approccio*, una procedura di attenuazione degli effetti d'instabilità, specie fino al punto di poissonianità o in prossimità di esso, consistente di procedimenti di perequazione per medie mobili. Tramite il loro impiego ed opportune interpolazioni si è giunti ad ottenere una sorta di spline-curva di andamento riguardabile di prima approssimazione della funzione di dispersione non-poissoniana oggetto d'attenzione.

Poiché, come appare dalle Figg. 5 e 6, tale curva mostra tratti e flessioni da vedersi, a nostro avviso, attribuibili a fattori d'instabilità, non fosse altro perché la *process-index function* è, in fondo, concepibile monotonicamente crescente, si è passati alla messa a punto di un *secondo approccio* consistente in un procedimento di estimazione pseudo-jackknife articolato in modo da attenuare particolarmente gli effetti dei suddetti fattori d'instabilità. Per evitare ripetizioni di discorso, l'approccio è stato impiegato solo con riferimento allo stimatore di ML del processo iperbinomiale. Coi risultati conseguiti è stato costruito il plot della Fig. 9. Nella Fig. 10, tale plot è stato sovrapposto a quello dei punti relativi ai tempi t_i di detta spline-curva al fine di rimarcarne le differenze.

Il confronto suggerito da quest'ultima figura ha però portato a chiedersi su quale dei due andamenti era il caso di propendere e di impostare quindi un discorso che fosse quanto meno di chiarimento.

Ci si è così impegnati nel delineare, come *terzo approccio*, una procedura di tipo bootstrap imperniata su un criterio di dichiarazione di *outlier* che, per ciascuna distribuzione statistica $\mathcal{D}_{n,i}(x)$, comporta l'esclusione di ogni bootstrap-stima di ML della *process-index function* al corrispondente tempo t_i sulla base di soglie connesse al congiunto numero c_i+1 delle classi di $\mathcal{D}_{n,i}(x)$ e alla corrispettiva media aritmetica. Con le stime fornite da tale procedura, qui siglata con CB (da central booastrap), si è costruito il CB-plot della Fig. 14 sovrapponendolo a quello delle stime pseudo-jackknife del precedente approccio.

Come emerge dalla Fig. 14, il plot delle stime bootstrap ovviamente risponde meglio all'idea di un andamento pressoché monotonicamente crescente. Stando a quanto detto sotto tale profilo, esso si guadagna perciò la preferenza specie rispetto a quello delle suddette spline-stime per il fatto che queste ultime risulterebbero maggiormente sensibili ai fattori preminenti d'instabilità. È però evidente che se, nella fattispecie, gli ultimi due approcci possono ritenersi preferibili per inferire sull'andamento della *pro-*

cess-index function, è soprattutto, per non dire soltanto, per la parte di sottopoissonianità allorché il processo iperbinomiale va ad identificarsi in quello iperbinomiale positivo.

Qualunque sia l'approccio seguito, non è comunque azzardato rimarcare come l'*ottica atemporale* di inquadramento degli eventi di un processo di nascita, e quindi la considerazione della sola distribuzione statistica relativa al tempo finale *T* del periodo d'osservazione, sia a rigore improponibile per saggiare ipotesi di conformità ad uno schema di riferimento; e come sia quindi scarsamente euristica la vasta letteratura statistica sulla cosiddetta sovradispersione.

Per vedere se la conclusione poteva essere vista in qualche misura subordinata al modo con cui è qui utilizzata l'evidenza statistica, l'analisi è stata portata avanti (Ferreri, 1992), oltre che in termini di distribuzioni statistiche, anche tramite alcune delle strade indicate dall'ampia metodologia sulle funzioni di sopravvivenza. Così, tra l'altro, si è utilizzato lo stimatore di Kaplan-Meier per ottenere le stime di $\widehat{\Lambda}_h$ ai tempi t_h del primo evento di ciascuna unità d'osservazione, nonché basato sulle distribuzioni statistiche $\mathcal{D}_{n,s}(x)$, relative a tutti i tempi t_s d'incidente, la determinazione delle corrispondenti stime $\tilde{\Lambda}_s = \tilde{\Lambda}(t_s)$ della funzione rischio cumulativo (3.12b) del processo di nascita iperbinomiale, facendo tesoro della relazione $\tilde{\Lambda}_s = \tilde{\mu}_s^2 \cdot \tilde{\delta}_s$ in termini del rapporto δ espresso dalla (5.7a).

Com'era prevedibile, si è giunti però alla conferma della conclusione tratta in quanto praticamente ininfluenti si sono potuti dire tanto il tipo di segmentazione del periodo (0,*T*] di osservazione (purché abbastanza fine) quanto la considerazione dell'HBP in termini della coppia di funzioni $(\mu(t), \Lambda(t))$ – che sotto il profilo degli effetti d'instabilità è parsa quasi svantaggiosa – anziché di quella $(\mu(t), \alpha(t))$ nella *process-index function*.

APPENDICE

Tavole richiamate:

Tav. I – Distribuzioni osservate e corrispondenti perequate (Campioni di n=108)

t_i	n_{0i}	n_{1i}	n_{2i}	n_{3i}	n_{4i}	n_{5i}	n_{6i}	n_{7i}	n_{8i}	ν	% Pr_{HPAD}	% Pr_{HBD}
1.93	90	17	1									
	89.7	17.6	0.7									
2.90	83	24	1									
	83.3	22.8	1.6	0.3								
3.99	76	28	3	1								
	77.1	26.4	3.8	0.7								
4.97	73	27	7	1								
	71.6	29.4	6.0	1.0								
5.93	66	33	8	1								
	65.7	33.0	8.0	1.0	0.3							
6.91	58	39	9	1	1							
	57.5	37.6	11.2	1.0	0.7							
7.73	50	40	16	1	1							
	52.8	39.1	13.8	1.3	1.0							
8.92	49	38	18	2	1							
	47.8	39.3	17.8	2.1	0.6	0.4						
9.82	45	40	19	3	-	1				1	36.5	37.8
	45.9	39.0	18.9	3.2	0.4	0.6				″	42.8	43.2
10.97	43	39	20	5	-	1				1	56.0	54.8
	42.5	38.6	21.5	3.7	0.7	1.0				″	30.5	29.7
11.90	40	37	25	3	2	1				1	11.5	11.7
	40.4	37.1	23.9	4.0	1.6	1.0				″	20.2	19.7
12.94	38	35	27	4	3	1				1	9.1	9.1
	38.0	35.0	27.7	3.7	2.6	1.0				″	6.2	6.3
13.97	36	33	31	4	3	1				1	2.0	2.0
	36.0	34.0	29.3	4.3	2.7	1.7				″	4.4	4.5
14.97	34	34	30	5	2	3				1	4.4	4.3
	34.5	32.9	30.7	5.2	2.2	2.5				″	3.1	3.0
15.57	34	32	31	6	2	3				1	3.0	3.0
	33.6	32.7	30.4	6.3	1.8	3.2				2	10.8	10.8
16.96	32	31	30	10	1	4				2	23.6	23.3
	32.1	30.0	31.9	9.0	1.2	3.8				″	10.2	9.9

(Segue Tav. I)

t_i	n_{0i}	n_{1i}	n_{2i}	n_{3i}	n_{4i}	n_{5i}	n_{6i}	n_{7i}	n_{8i}	ν	% Pr_{HPAD}	% Pr_{HBD}
17.96	31	28	34	10	1	4				2	4.4	4.4
	31.1	29.2	32.2	10.5	1.0	4.0				″	10.8	10.6
19.00	30	28	33	12	1	4				2	8.3	8.4
	29.9	27.6	32.8	12.3	1.0	4.4				″	8.4	8.2
19.77	29	27	32	14	1	5				2	10.4	10.3
	28.8	27.7	32.1	13.5	1.0	4.9				″	12.5	12.7
20.76	27	28	31	15	1	6				2	21.9	21.6
	27.6	27.3	31.7	14.3	1.4	5.3	0.4			″	16.0	15.6
21.58	27	27	32	14	2	5	1			2	16.2	15.9
	26.5	27.9	31.1	14.7	1.9	5.1	0.8			″	24.5	24.7
22.86	25	29	30	15	3	4	2			2	43.2	42.6
	25.5	27.3	30.6	15.1	3.5	3.9	2.1			″	30.2	29.9
23.77	25	26	30	16	5	3	3			3	37.9	37.2
	24.4	27.4	30.3	15.3	4.6	3.4	2.6			2	39.2	39.3
24.93	23	27	31	15	6	3	3			3	51.7	51.7
	23.6	26.7	28.2	16.4	6.4	3.4	3.0	0.3		″	69.3	69.1
25.94	23	27	24	18	8	4	3	1		3	89.3	86.7
	22.6	27.0	26.0	17.0	8.0	3.7	3.0	0.7		″	92.0	90.2
26.93	22	27	23	18	10	4	3	1		3	91.1	87.1
	22.3	26.6	23.3	18.4	8.3	5.1	3.0	1.0		″	87.3	83.7
27.77	22	26	23	19	7	7	3	1		3	62.7	58.6
	22.0	26.1	23.3	16.9	9.8	5.6	3.3	1.0		″	95.7	91.8
28.92	22	25	24	13	13	6	4	1		4	73.5	68.6
	21.6	25.7	22.5	15.3	12.0	6.2	3.3	1.4		″	92.9	88.2
29.59	21	26	21	15	14	6	3	2		4	74.2	68.3
	21.2	25.4	22.3	14.1	13.8	6.2	3.2	1.8		″	74.4	69.4
30.91	20	25	21	15	15	7	2	3		4	61.1	54.3
	19.9	24.9	21.4	15.4	13.9	7.1	2.7	2.3	0.4	″	80.2	73.9
31.85	19	24	22	16	13	8	3	2	1	4	89.8	85.0

Tav. II – Stime di ML dei ϑ_i e degli α_i e corrispondenti valori di pseudo-spline

t_i	$\tilde{I}_{D,i}$	M_i	M_{Si}	$\hat{\alpha}_i$	$\hat{\alpha}_{Si}$	$\tilde{\alpha}_{Spl,i}$	$\hat{\vartheta}_i$	$\hat{\vartheta}_{Si}$	$\tilde{\vartheta}_{Spl,i}$
1.93	0.929	0.1759	0.1759	-0.4322	-0.6115	-0.5539	-0.0384	-0.0887	-0.0874
2.90	0.836	0.2407	0.2491	-0.7179	-0.2153	-0.3242	-0.1233	-0.0516	-0.0540
3.99	0.982	0.3426	0.3343	-0.0473	-0.0216	-0.1264	-0.0077	-0.0036	-0.0233
4.97	1.047	0.4074	0.4111	0.1313	-0.0145	-0.0456	0.0282	-0.0031	-0.0109
5.93	0.942	0.4815	0.4926	-0.1313	-0.0224	-0.0386	-0.0334	-0.0054	-0.0107
6.91	0.970	0.5926	0.6093	-0.0421	-0.0758	-0.0570	-0.0115	-0.0221	-0.0175
7.73	0.901	0.7315	0.6907	-0.1268	-0.0751	-0.0622	-0.0444	-0.0250	-0.0204
8.92	0.937	0.7778	0.7926	-0.0844	-0.0599	-0.0516	-0.0332	-0.0225	-0.0183
9.82	0.974	0.8519	0.8426	-0.0261	-0.0349	-0.0324	-0.0103	-0.0141	-0.0119
10.97	0.992	0.9167	0.9306	-0.0077	0.0008	-0.0061	-0.0034	0.0003	-0.0015
11.90	1.018	1.0093	1.0028	0.0172	0.0214	0.0124	0.0085	0.0105	0.0070
12.94	1.043	1.0926	1.0824	0.0398	0.0179	0.0275	0.0219	0.0097	0.0155
13.97	0.997	1.1481	1.1556	-0.0027	0.0453	0.0449	-0.0016	0.0261	0.0265
14.97	1.096	1.2222	1.2148	0.0754	0.0596	0.0615	0.0453	0.0361	0.0379
15.57	1.098	1.2500	1.2556	0.0785	0.0810	0.0690	0.0492	0.0506	0.0436
16.96	1.119	1.3426	1.3389	0.0920	0.0779	0.0727	0.0627	0.0531	0.0493
17.96	1.078	1.3889	1.3806	0.0589	0.0691	0.0721	0.0421	0.0490	0.0509
19.00	1.068	1.4259	1.4454	0.0507	0.0657	0.0717	0.0375	0.0494	0.0538
19.77	1.105	1.5000	1.4898	0.0767	0.0698	0.0757	0.0602	0.0540	0.0588
20.76	1.116	1.5648	1.5565	0.0799	0.0916	0.0883	0.0647	0.0736	0.0709
21.58	1.163	1.5926	1.5935	0.1082	0.0961	0.0990	0.0882	0.0784	0.0813
22.86	1.183	1.6481	1.6657	0.1137	0.1227	0.1175	0.0941	0.1039	0.1004
23.77	1.226	1.7315	1.7120	0.1391	0.1221	0.1306	0.1232	0.1059	0.1152
24.93	1.190	1.7685	1.8056	0.1116	0.1523	0.1527	0.0998	0.1399	0.1414
25.94	1.345	1.9074	1.8833	0.1992	0.1709	0.1733	0.1926	0.1633	0.1670
26.93	1.329	1.9630	1.9639	0.1887	0.1997	0.1953	0.1891	0.1997	0.1953
27.77	1.369	2.0093	2.0130	0.2091	0.2120	0.2111	0.2141	0.2175	0.2170
28.92	1.423	2.0833	2.0852	0.2409	0.2323	0.2284	0.2556	0.2464	0.2430
29.59	1.433	2.1389	2.1324	0.2391	0.2392	0.2336	0.2598	0.2594	0.2535
30.91	1.436	2.2222	2.2269	0.2327	0.2367	0.2360	0.2636	0.2673	0.2663
31.85	1.470	2.2963	2.2963	0.2354	0.2354	0.2354	0.2730	0.2730	0.2730

Tav. III – Numerosità teoriche della HPAD corrispettive delle stime $\hat{\vartheta}_i$ e $\hat{\vartheta}_{Si}$

$t = t_i$	n_{0i}	n_{1i}	n_{2i}	n_{3i}	n_{4i}	n_{5i}	n_{6i}	n_{7i}	n_{8i}	ν	% Pr_{HPAD}
1.93	89.9	17.2	0.9								
	89.0	18.8	0.2								
2.90	81.9	25.7	0.4								
	83.1	23.0	1.9								
3.99	76.5	26.6	4.4	0.5							
	77.2	26.0	4.3	0.5							
4.97	72.7	28.0	6.2	1.0	0.1						
	71.5	29.6	6.0	0.8	0.1						
5.93	65.6	33.8	7.6	0.9	0.1						
	65.8	32.8	8.0	1.3	0.1						
6.91	59.3	35.9	10.5	2.0	0.3						
	57.9	36.9	10.9	2.0	0.3						
7.73	50.2	40.2	14.3	2.9	0.4						
	53.2	38.6	13.1	2.7	0.4						
8.92	48.3	40.2	15.4	3.5	0.5	0.1					
	48.0	39.8	15.6	3.8	0.7	0.1					
9.82	45.7	39.7	16.9	4.7	0.9	0.1				1	36.5
	46.0	39.8	16.7	4.5	0.9	0.1				″	42.8
10.97	43.1	39.7	18.2	5.5	1.3	0.2				1	56.0
	42.6	39.6	18.4	5.7	1.3	0.3	0.1			″	30.5
11.90	39.7	39.4	19.9	6.8	1.7	0.4	0.1			1	11.5
	40.0	39.3	19.7	6.7	1.8	0.4	0.1			″	20.2
12.94	37.1	38.8	21.1	8.0	2.3	0.6	0.1			1	9.1
	37.0	39.2	21.2	7.8	2.2	0.5	0.1			″	6.2
13.97	34.2	39.4	22.6	8.6	2.5	0.6	0.1			1	2.0
	35.0	38.4	22.1	8.8	2.8	0.7	0.2			″	4.4
14.97	33.5	37.5	22.6	9.7	3.3	1.0	0.3	0.1		1	4.4
	33.5	37.8	22.7	9.6	3.2	0.9	0.2	0.1		″	3.1
15.57	32.8	37.2	22.9	10.1	3.5	1.1	0.3	0.1		1	3.0
	32.7	37.2	22.9	10.1	3.6	1.1	0.3	0.1		2	10.8
16.96	30.6	36.3	23.7	11.2	4.3	1.4	0.4	0.1		2	23.6
	30.3	36.6	23.9	11.2	4.2	1.3	0.4	0.1		″	10.2
17.96	28.5	36.4	24.8	11.9	4.5	1.4	0.4	0.1		2	4.4
	29.0	36.3	24.5	11.7	4.5	1.5	0.4	0.1		″	10.8
19.00	27.3	36.2	25.3	12.4	4.8	1.5	0.4	0.1		2	8.3
	27.2	35.8	25.1	12.6	5.0	1.7	0.5	0.1		″	8.4
19.77	26.2	35.0	25.3	13.2	5.5	1.9	0.6	0.2	0.1	2	10.4
	26.2	35.2	25.4	13.1	5.4	1.9	0.6	0.2		″	12.5

(Segue Tav. III)

t_i	n_{0i}	n_{1i}	n_{2i}	n_{3i}	n_{4i}	n_{5i}	n_{6i}	n_{7i}	n_{8i}	ν	% Pr_{HPAD}
20.76	24.8	34.3	25.8	13.9	6.0	2.2	0.7	0.2	0.1	2	21.9
	25.3	34.2	25.5	13.8	6.0	2.2	0.7	0.2	0.1	″	16.0
21.58	25.0	33.6	25.3	14.0	6.3	2.5	0.9	0.3	0.1	2	16.2
	24.6	33.8	25.6	14.1	6.3	2.4	0.8	0.3	0.1	″	24.5
22.86	23.9	33.0	25.5	14.6	6.8	2.7	1.0	0.3	0.2	2	43.2
	23.9	32.6	25.4	14.6	6.9	2.9	1.1	0.4	0.2	″	30.2
23.77	23.1	31.7	25.3	15.1	7.5	3.3	1.3	0.5	0.2	3	37.9
	23.0	32.1	25.6	15.1	7.3	3.1	1.2	0.4	0.2	2	39.2
24.93	21.6	31.6	26.0	15.7	7.8	3.3	1.3	0.5	0.2	3	51.7
	22.2	30.8	25.2	15.6	8.1	3.7	1.5	0.6	0.3	″	69.3
25.94	21.8	29.3	24.3	15.9	8.8	4.4	2.0	0.9	0.6	3	89.3
	21.4	29.8	24.9	16.0	8.7	4.2	1.8	0.7	0.5	″	92.0
26.93	20.7	28.8	24.5	16.3	9.3	4.7	2.2	0.9	0.6	3	91.1
	21.0	28.7	24.3	16.2	9.2	4.7	2.2	1.0	0.7	″	87.3
27.77	20.6	28.1	24.1	16.4	9.5	5.0	2.4	1.1	0.8	3	62.7
	20.7	28.1	24.1	16.3	9.5	5.0	2.4	1.1	0.8	″	95.7
28.92	20.5	27.2	23.5	16.3	9.9	5.4	2.8	1.3	1.1	4	73.5
	20.3	27.2	23.7	16.5	9.9	5.4	2.7	1.3	1.0	″	92.9
29.59	19.8	26.6	23.5	16.6	10.2	5.7	3.0	1.4	1.2	4	74.2
	19.9	26.7	23.4	16.6	10.2	5.7	2.9	1.4	1.2	″	74.4
30.91	18.6	25.9	23.4	17.0	10.7	6.1	3.3	1.6	1.4	4	61.1
	18.6	25.8	23.4	17.0	10.7	6.2	3.3	1.6	1.4	″	80.2
31.85	17.8	25.2	23.3	17.2	11.1	6.5	3.5	1.8	1.6	4	89.8

Tav. IV – *Numerosità teoriche della HBD corrispettive delle stime* $\hat{\alpha}_i$ *e* $\hat{\alpha}_{Si}$

$t = t_i$	n_{0i}	n_{1i}	n_{2i}	n_{3i}	n_{4i}	n_{5i}	n_{6i}	n_{7i}	n_{8i}	ν	% $\Pr_{HBD}$
1.93	89.9	17.1	0.9	0.1							
	89.7	17.7	0.6								
2.90	82.9	24.1	1.0								
	83.6	22.0	2.3	0.1							
3.99	76.5	26.6	4.4	0.5							
	77.2	26.0	4.3	0.5							
4.97	72.6	28.1	6.2	1.0	0.1						
	71.5	29.6	6.0	0.8	0.1						
5.93	65.7	33.7	7.5	1.0	0.1						
	65.8	32.8	8.0	1.3	0.1						
6.91	59.3	36.0	10.5	1.9	0.3						
	57.9	37.0	10.9	2.0	0.2						
7.73	50.1	40.4	14.2	2.9	0.4						
	53.1	38.7	13.1	2.7	0.4						
8.92	48.3	40.2	15.3	3.5	0.6	0.1					
	47.9	39.9	15.6	3.8	0.7	0.1					
9.82	45.6	39.8	16.9	4.6	0.9	0.1				1	37.8
	45.9	39.9	16.7	4.5	0.9	0.1				″	43.2
10.97	43.1	39.7	18.2	5.5	1.3	0.2				1	54.8
	42.6	39.6	18.4	5.7	1.4	0.3				″	29.7
11.90	39.7	39.4	19.9	6.8	1.7	0.4	0.1			1	11.7
	40.0	39.3	19.7	6.7	1.8	0.4	0.1			″	19.7
12.94	37.1	38.8	21.1	8.0	2.3	0.6	0.1			1	9.1
	37.0	39.2	21.2	7.8	2.2	0.5	0.1			″	6.3
13.97	34.2	39.4	22.6	8.6	2.5	0.6	0.1			1	2.0
	35.0	38.5	22.1	8.8	2.7	0.7	0.2			″	4.5
14.97	33.5	37.5	22.6	9.7	3.3	1.0	0.3	0.1		1	4.3
	33.4	37.9	22.7	9.6	3.2	0.9	0.3			″	3.0
15.57	32.8	37.3	22.9	10.0	3.5	1.1	0.3	0.1		1	3.0
	32.7	37.2	22.9	10.1	3.6	1.1	0.3	0.1		2	10.8
16.96	30.5	36.4	23.7	11.2	4.3	1.4	0.4	0.1		2	23.3
	30.2	36.7	23.9	11.2	4.2	1.3	0.4	0.1		″	9.9
17.96	28.4	36.5	24.8	11.9	4.5	1.4	0.4	0.1		2	4.4
	28.9	36.4	24.5	11.7	4.5	1.5	0.4	0.1		″	10.6
19.00	27.3	36.3	25.3	12.4	4.7	1.5	0.4	0.1		2	8.4
	27.2	35.8	25.2	12.5	5.0	1.7	0.5	0.1		″	8.2
19.77	26.1	35.1	25.4	13.2	5.4	1.9	0.6	0.2	0.1	2	10.3
	26.2	35.3	25.5	13.1	5.3	1.8	0.6	0.2		″	12.7

(Segue Tav. IV)

t_i	n_{0i}	n_{1i}	n_{2i}	n_{3i}	n_{4i}	n_{5i}	n_{6i}	n_{7i}	n_{8i}	ν	$\%\,\mathrm{Pr}_{HBD}$
20.76	24.7	34.4	25.8	13.9	6.0	2.2	0.7	0.2	0.1	2	21.6
	25.2	34.3	25.6	13.7	6.0	2.2	0.7	0.2	0.1	″	15.6
21.58	24.8	33.7	25.4	14.0	6.3	2.5	0.9	0.3	0.1	2	15.9
	24.5	33.9	25.6	14.1	6.3	2.4	0.8	0.3	0.1	″	24.7
22.86	23.8	33.1	25.6	14.5	6.8	2.7	1.0	0.3	0.2	2	42.6
	23.7	32.8	25.5	14.6	6.9	2.8	1.1	0.4	0.2	″	29.9
23.77	22.9	31.9	25.4	15.1	7.5	3.2	1.3	0.5	0.2	3	37.2
	22.8	32.3	25.6	15.1	7.3	3.1	1.2	0.4	0.2	2	39.3
24.93	21.5	31.8	26.1	15.7	7.7	3.3	1.3	0.4	0.2	3	51.7
	21.9	31.0	25.3	15.6	8.0	3.7	1.5	0.6	0.4	″	69.1
25.94	21.4	29.6	24.6	15.8	8.7	4.3	2.0	0.9	0.7	3	86.7
	21.1	30.1	25.1	16.0	8.6	4.1	1.8	0.7	0.5	″	90.2
26.93	20.3	29.1	24.8	16.3	9.1	4.6	2.1	0.9	0.8	3	87.1
	20.6	29.1	24.6	16.2	9.1	4.6	2.2	1.0	0.6	″	83.7
27.77	20.2	28.6	24.4	16.3	9.4	4.9	2.4	1.1	0.7	3	58.6
	20.2	28.5	24.4	16.3	9.4	4.9	2.4	1.1	0.8	″	91.8
28.92	20.0	27.7	23.8	16.3	9.8	5.3	2.7	1.3	1.1	4	68.6
	19.7	27.7	24.0	16.4	9.8	5.3	2.7	1.3	1.1	″	88.2
29.59	19.2	27.2	23.8	16.6	10.1	5.6	2.9	1.4	1.2	4	68.3
	19.3	27.2	23.8	16.6	10.1	5.6	2.9	1.4	1.1	″	69.4
30.91	18.0	26.4	23.8	17.0	10.6	6.0	3.2	1.6	1.4	4	54.3
	18.1	26.3	23.7	17.0	10.6	6.0	3.2	1.6	1.5	″	73.9
31.85	17.2	25.7	23.6	17.3	11.0	6.4	3.4	1.8	1.6	4	85.0

Tav. V – Stime di ML, jackknife e bootstrap dei livelli $\alpha(t_i)$ di HBP

$t = t_i$	M_t	$\hat{\alpha}_t$	$\tilde{\alpha}_{Spl,t}$	A0 e A1		B0 e B1		$\overline{\alpha}_{CB,t}$	$s_{CB,t}$
				$\tilde{\alpha}_{JKw,t}$	$\tilde{\alpha}_{JKwp,t}$	$\tilde{\alpha}_{JKw,t}$	$\tilde{\alpha}_{JKwp,t}$		
1.93	0.1759	-0.4322	-0.5539	-0.7109	-0.7109	-0.9952	-0.9952	-0.4323	0.0402
2.90	0.2407	-0.7179	-0.3242	-0.8537	-0.5924	-0.9922	-0.7659	-0.7080	0.0174
3.99	0.3426	-0.0473	-0.1264	-0.3822	-0.4430	-0.3385	-0.3633	-0.2576	0.2481
4.97	0.4074	0.1313	-0.0456	-0.0832	-0.3081	-0.0756	-0.2635	-0.1529	0.2352
5.93	0.4815	-0.1313	-0.0386	-0.2927	-0.2275	-0.2781	-0.1783	-0.1558	0.2052
6.91	0.5926	-0.0421	-0.0570	-0.1672	-0.1942	-0.1551	-0.1533	-0.1069	0.2085
7.73	0.7315	-0.1268	-0.0622	-0.1992	-0.1699	-0.1887	-0.1527	-0.1160	0.0894
8.92	0.7778	-0.0844	-0.0516	-0.1470	-0.1312	-0.1422	-0.1371	-0.1274	0.0845
9.82	0.8519	-0.0261	-0.0324	-0.0683	-0.0898	-0.0613	-0.0918	-0.0693	0.0694
10.97	0.9167	-0.0077	-0.0061	-0.0446	-0.0466	-0.0408	-0.0345	-0.0523	0.0797
11.90	1.0093	0.0172	0.0124	-0.0119	-0.0096	-0.0109	0.0005	-0.0156	0.0664
12.94	1.0926	0.0398	0.0275	0.0181	0.0222	0.0186	0.0377	-0.0162	0.0784
13.97	1.1481	0.0000	0.0449	0.0964	0.0420	0.1140	0.0594	-0.0125	0.0402
14.97	1.2222	0.0754	0.0615	0.0354	0.0484	0.0357	0.0593	0.0512	0.0331
15.57	1.2500	0.0785	0.0690	0.0410	0.0458	0.0412	0.0557	0.0655	0.0173
16.96	1.3426	0.0920	0.0727	0.0556	0.0396	0.0557	0.0386	0.0845	0.0368
17.96	1.3889	0.0589	0.0721	0.0243	0.0359	0.0244	0.0293	0.0377	0.0409
19.00	1.4259	0.0507	0.0717	0.0193	0.0375	0.0194	0.0302	0.0496	0.0364
19.77	1.5000	0.0767	0.0757	0.0470	0.0471	0.0471	0.0439	0.0566	0.0446
20.76	1.5648	0.0799	0.0883	0.0521	0.0661	0.0522	0.0625	0.0715	0.0423
21.58	1.5926	0.1082	0.0990	0.1037	0.0855	0.1037	0.0822	0.1049	0.0332
22.86	1.6481	0.1137	0.1175	0.1055	0.1053	0.1055	0.1059	0.0984	0.0200
23.77	1.7315	0.1391	0.1306	0.1283	0.1215	0.1284	0.1201	0.1346	0.0183
24.93	1.7685	0.1116	0.1527	0.1020	0.1427	0.1020	0.1410	0.1006	0.0203
25.94	1.9074	0.1992	0.1733	0.1982	0.1668	0.1983	0.1691	0.1942	0.0243
26.93	1.9630	0.1887	0.1953	0.1878	0.1912	0.1878	0.1947	0.1871	0.0249
27.77	2.0093	0.2091	0.2111	0.2082	0.2100	0.2082	0.2142	0.2103	0.0166
28.92	2.0833	0.2409	0.2284	0.2399	0.2249	0.2400	0.2300	0.2368	0.0210
29.59	2.1389	0.2391	0.2336	0.2370	0.2316	0.2370	0.2370	0.2350	0.0187
30.91	2.2222	0.2327	0.2360	0.2297	0.2347	0.2297	0.2297	0.2277	0.0163
31.85	2.2963	0.2354	0.2354	0.2356	0.2356	0.2356	0.2356	0.2411	0.0187

Riferimenti bibliografici

Abramowitz, M. and Stegun, I.A. eds (1972). Handbook of mathematical functions. General Publ. Co. Toronto.

Anraku K. and Yanagimoto, T. (1990). Estimation for the negative binomial distribution based on the conditional likelihood. *Communications in Statistics-Simulatyion and Computation*, *19*, 771-806.

Binet, F. E. (1986). Fitting the negative binomial distribution. *Biometrics*, *42*, 989-992.

Cameron, A.C. and Trivedi, P. K. (1996). Count data models for financial data. In *Handbook of Statistics*, Vol. 14,. G. S. Maddala & C. R. Rao, eds., Amsterdam: Elsevier, 363-391.

Carroll, R. J. and Lombard, F. (1985). A note on *n* estimators for the binomial distribution. *Journal of the American Statistical Association*, *80*, 423-426.

Douglas, J. B. (1980). *Analysis with Standard Contagious Distribution*, Fairland, MD: International Co-operatve Publishing Hause.

Douglas, J. B. (1986). "Pólya-Aeppli distribution," in *Encyclopedia of Statistical Sciences,* Vol.7, eds.: S. Kotz, N.L.Johnson and C.B.Read, New York :Wiley, 56-59.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, *7*, 1-26.

Efron, B. and Tibshirani, R. J. (1986). Bootstrap methods for standard errors, confidence intervals and other measures of statistical accuracy. *Statistical Science*, *1*, 54-77.

Evans, D. A. (1953). "Experimental evidence concerning contagious distributions in ecology," *Biomrtika*, *40*, 186-211.

Ferrante, M. R. and Ferreri, C. (1996). Pseudo-jackknife estimators for a hyperbinomial process. *Statistica*, *56*, 2, 175-187.

Ferrante, M. R. and Ferreri, C. (1997). A bootstrap-type procedure for an underdispersion analysis. *Statistica*, *57*, 2, 199-210.

Ferreri, C. (1983). On the extended Pólya process and some its interpretations. *Metron*, *41*, 1/2, 11-27.

Ferreri, C. (1990). Nonhomogeneous negative binomial process with time-dependent index. *Statistica*, *50*, 3, 316-326.

Ferreri, C. (1992). On the role of a hyperbinomial process. *Metron*, *50*, 1/2, 3-18.

Ferreri, C. (1996). On a hyperbinomial process. *Communications in Statistics-Theory and Methods*, *25*, 1, 83-103.

Ferreri, C. (1997). "On the ML-estimator of the two-parameter both positive and negative binomial distribution," *Statistics & Probability Letters*, 33, 129-134.

Ferreri, C. (2000). A reconsideration of the Pólya-Aeppli model. *Statistica*, *60*, 1, 15-24.

Ferreri, C. (2002). Relazione tra il test Q del Cochran e il quoziente Q_f^2 di Lexis-Bortkiewicz. In "Studi in onore di Angelo Zanella", 293-305.

Gelfand, A. E. & Dalal, S. R. (1990). A note on overdispersed exponential families. *Biometrika*, *77*, 55-64.

Greenwood, M. and Yule, G. U. (1920). An enquiry into the nature of frequency distributions of multiple happenings, with particular reference to the occurrence of multiple attacks of desease or repeated accidents. *Journal of the Royal Statistical Society, Series* A, *83*, 255-279.

Grogger, J. (1990). The deterrent effect of capital punishment: an analysis of daily homicide counts. *Journal of the American Statistical Association*, *85*, 295-303.

Hall, P. (1994). On the erratic behavior of estimators of *n* in the binomial *n*,*p* distribution. *Journal of the American Statistical Association*, *89*, 344-352.

Husman, J., Hall, B. H. and Griliches, Z. (1984). Econometric models for count data with an application to the patents-R&D relationship. *Econometrica*, *52*, 4, 909-938.

Johnson, N. L. and Kotz, S. Kemp, A. W. (1992). *Univariate Discrete Distributions*. 2nd ed., Wiley, New York.

Kemp, A. W. (1978). "Cluster size probabilities for generalized Poisson distributions," *Communications in Statistics-Theory and Methods*, *7*, 1433-1438.

Lindsay, B. (1986). Exponential family mixture models (with least squares estimators). Annals of Statistics, *14*, 124-137.

Olkin, I., Petkau, A. J. and Zidek, J. V. (1981). A comparison of *n* estimators for the binomial distribution. *Journal of the American Statistical Association*, *76*, 637-642.

Piegorsch, W. W.. (1981). Maximum likelihood estimation for the negative binomial dispersion parameter. *Biometrics*, *46*, 863-867.

Regazzini, E. (1983). Bayesian statistical inference for a nonhomogeneous negative binomial process. Metron, *41*, 1/2, 57-76