UNIVERSITA' DEGLI STUDI DI PADOVA

FACOLTA' DI SCIENZE STATISTICHE

CORSO DI LAUREA IN STATISTICA,

POPOLAZIONE E SOCIETA'

Relazione finale

INTERVALLI DI CONFIDENZA PER DISTRIBUZIONI DISCRETE: UNA VALUTAZIONE TRAMITE SIMULAZIONE

Relatore: PROF.SSA ALESSANDRA SALVAN

Laureando: MARAGONI LORENZO

ANNO ACCADEMICO 2005-06

A Lucio, a Sergio,
e agli sponsor(s)

Indice

Cap. 1

Un'introduzione al problema e la proposta di Agresti e Coull

1.1. Intervalli di confidenza per distribuzioni discrete

Risolvendo problemi di inferenza, è spesso necessario costruire intervalli di confidenza per l'ignoto parametro di una distribuzione *discreta*. In simili circostanze, la normalità asintotica dello stimatore garantita dal teorema limite centrale può non fornire approssimazioni particolarmente accurate, soprattutto in presenza di numerosità campionarie scarse o di particolari valori del parametro all'interno del suo campo di variazione.

Lo spettro di possibilità alternative proposte in letteratura è piuttosto ampio: si va dalle tre versioni asintoticamente equivalenti basate sulla verosimiglianza, a test basati sulla distribuzione esatta della variabile aleatoria osservabile, a possibili semplificazioni che riducano gli effetti dell'asimmetria o della discretezza.

Nella pratica, però, per un parametro scalare $\vartheta \in \Theta \subseteq \Re$, il tipo di intervallo più utilizzato (anche e soprattutto per la sua semplicità di comprensione e di calcolo) è quello che sfrutta la normalità asintotica dello stimatore di massima verosimiglianza $\hat{\vartheta}$,

$$\hat{\vartheta} \sim N(\vartheta, i(\vartheta)^{-1}),$$

dove $i(\vartheta)$ rappresenta l'informazione attesa. Spesso al posto di $i(\vartheta)$ si utilizza l'informazione osservata $j(\hat{\vartheta})$, ottenendo intervalli con livello approssimato $1-\alpha$ di forma

$$\hat{\vartheta} \pm z_{1-\alpha/2} j(\hat{\vartheta})^{-1/2},$$

dove $z_{1-\alpha/2}$ indica il quantile 1-α/2 di una Normale standard, che però in certi casi presentano lacune anche gravi. Sembra allora opportuno analizzare le prestazioni di metodi alternativi, alla ricerca di un criterio più efficace per la costruzione delle regioni.

1.2. Il caso binomiale: la proposta di Agresti e Coull

Diversi autori hanno affrontato la questione per particolari tipi di distribuzioni; in particolare, Agresti e Coull (1998), hanno proposto una soluzione interessante per variabili di tipo binomiale. Questa è un'immediata, ma molto utile dal punto di vista concettuale, semplificazione di un metodo proposto per la prima volta da Wilson (1927), il quale per costruire le regioni di confidenza utilizzò il risultato generale per lo stimatore di massima verosimiglianza, ma sfruttando l'informazione attesa (la varianza sotto l'ipotesi nulla) al posto di quella osservata.

In pratica, data una distribuzione Bi(n,π) e detta $\hat{\pi}$ la stima di massima verosimiglianza di π, l'idea è di sfruttare la normalità asintotica non della quantità

$$Z = \frac{\hat{\pi} - \pi}{\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}}$$

bensì di

$$Z = \frac{\hat{\pi} - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}}.$$

Il vantaggio di questo metodo (Henderson e Meyer, 2001), è che necessita di utilizzare solo *una* approssimazione (quella dovuta al teorema limite centrale), e risparmia di utilizzare la seconda, perché l'errore sotto l'ipotesi nulla non ha naturalmente bisogno di essere stimato. Il problema di questa classe di intervalli è l'impossibilità di scriverli in modo compatto: la formula

$$\hat{\pi} + \frac{z^2_{1-\alpha/2}}{2n} \pm \frac{z_{1-\alpha/2}}{1 + z^2_{1-\alpha/2}/n} \sqrt{\frac{\hat{\pi}(1-\hat{\pi}) + z^2_{1-\alpha/2}/4n}{n}}$$

è graficamente pesante, e intuitivamente meno comprensibile, potendo anche risultare fortemente scentrata rispetto alla stima di massima verosimiglianza per numerosità campionarie basse.

L'innovazione di Agresti e Coull è consistita proprio in questo: nel dare a queste regioni, migliori sul piano del rendimento, una forma più leggibile, e farlo attraverso un'approssimazione molto semplice e condivisibile. L'idea di base è stata: approssimare il 1.96 (il valore del quantile corrispondente al livello di confidenza 0.95 di una Normale standard) a 2. In questo modo si crea una sequenza di semplificazioni che portano ad una formula molto più chiara, dalle prestazioni per costruzione identiche alle precedenti e dal livello di confidenza molto simile (il valore corrispondente al quantile-2 di una Normale standard è 0.9545). Il nuovo metodo viene riassunto dagli stessi autori così: "si utilizzi la formula per gli intervalli standard, ma solo dopo aver aggiunto 2 successi e 2 insuccessi" ai dati di partenza.

Per quanto chi ha dimestichezza con tecniche statistiche più complesse potrebbe ritenere questa approssimazione poco utile, non va sottovalutato il

contributo che questa formula può dare a diffondere l'utilizzo di tecniche più accurate anche presso utenti che hanno meno familiarità con tali tecniche (ad esempio studenti ai primi corsi di statistica).

I risultati ottenuti da questa procedura sono stati molto buoni (in termini di copertura effettiva delle regioni), e costringono a riflettere seriamente sul fatto di continuare a utilizzare per distribuzioni binomiali il tipo di intervalli "standard", che hanno l'unico vantaggio di essere semplici da comprendere e di presentarsi in forma compatta, mentre a livello di prestazioni risultano insoddisfacenti anche per numerosità campionarie sostanziose.

Per inciso, la rilevanza pratica di queste applicazioni è tutt'altro che scarsa. Un esempio recente di applicazione in Italia è descritto in un articolo di Olgiati et al. (2006), dove si analizzano le metodologie utilizzate dall'Aeeg (Autorità per l'energia elettrica e il gas) per garantire l'efficienza e la qualità dell'erogazione di questi servizi. L'Aeeg (istituita nel 1997) ha lo scopo di uniformare gli esercenti dei servizi ai livelli di legge, e in definitiva di garantire agli utenti un servizio di buona qualità. Per la verifica di questi standard sono indispensabili procedure statistiche, come verifiche di ipotesi e costruzione di regioni di confidenza. La mancata applicazione di criteri rigorosi, o l'utilizzo di tecniche non affidabili comporterebbero il non rispetto delle soglie e dunque un servizio di scarsa qualità.

Data l'ampia classe di problemi toccati da queste tecniche, viene allora spontaneo, come suggeriscono d'altra parte gli stessi autori, estendere la ricerca ad altre distribuzioni discrete (come ad esempio la distribuzione di Poisson), cercando soluzioni alternative allo scopo di giungere, eventualmente attraverso approssimazioni, a intervalli con livello di

copertura soddisfacente e con una forma compatta. Confrontando i risultati di questi con quelli raggiungibili con il metodo “standard”, e studiando eventuali cambiamenti che si possono presentare al variare della numerosità campionaria e del valore del parametro, si comprenderà se la procedura standard sia soddisfacente per l’inferenza o se (come nel caso binomiale) il suo utilizzo può fuorviare le conclusioni del ricercatore.

L’obiettivo delle successive analisi è proprio questo: cercare un metodo semplice e al tempo stesso affidabile per costruire regioni di confidenza per una distribuzione di Poisson, valutandone le prestazioni attraverso tecniche di simulazione.

1.3. La valutazione delle procedure tramite simulazione

Per valutare l’affidabilità di un metodo per costruire regioni di confidenza, l’indicatore migliore, e il più utilizzato in letteratura (cfr. anche Agresti e Coull, 1998), è il livello di copertura effettivo dell’intervallo, confrontato con quello previsto nominalmente. La valutazione tramite simulazione consiste nel generare una serie di campioni pseudo-casuali da una distribuzione nota (nel nostro caso, di Poisson con λ noto), costruire per ciascuno di essi l’intervallo di confidenza di livello prefissato e calcolare la percentuale di volte in cui l’intervallo contiene il valore “vero” di del parametro, ovvero quello della distribuzione generatrice dei dati.

Quindi, in teoria, se il metodo fosse perfetto, per un intervallo di livello 1-α, replicando un numero molto elevato di volte un esperimento e costruendo ogni volta un intervallo con lo stesso metodo, si dovrebbe

ottenere che il 100(1-α)% delle volte l'intervallo contiene in effetti il vero ed ignoto valore del parametro che caratterizza la distribuzione. Può accadere però, nella pratica, che il livello di confidenza realmente trovato sia inferiore a quello nominale, ed in modo che non dipende semplicemente dalle oscillazioni dovute alla simulazione, bensì strutturalmente insito nel metodo. Questa è una grave lacuna della tecnica di costruzione, in quanto non ci fornisce la giusta idea di quanto ci si possa "fidare" dell'intervallo, portando in genere a sovrastimare la sua attendibilità.

D'altro canto, è da notare che esistono procedure "conservatrici", ovvero che hanno un livello di copertura effettivo *superiore* a quello nominale. Anche se queste sono senza dubbio preferibili rispetto a quelle che sovrastimano la copertura reale, esse rischiano di essere altrettanto fuorvianti per uno sperimentatore che comunque si "aspetti" un certo livello di confidenza dall'intervallo: il fatto che il livello vero sia "diverso" da quello atteso può avere effetti negativi più gravi dei benefici derivanti dal fatto che sia in un certo senso "migliore" delle aspettative.

In definitiva, si tenderà a considerare "migliori" le tecniche che consentono di ottenere livelli di confidenza che non necessariamente siano molto vicini a 1, quanto che siano più vicini possibile al valore nominale, scelta tra l'altro suggerita e operata anche dagli stessi Agresti e Coull (1998).

Un altro possibile indicatore di bontà del metodo citato ed utilizzato in alcuni testi è la lunghezza (ampiezza) degli intervalli: l'idea di fondo è che sia "migliore" una procedura che generi intervalli di ampiezza in media minore. Questa scelta è in realtà piuttosto discutibile, in quanto l'ampiezza delle regioni può variare in funzione della parametrizzazione del modello. Questa dipendenza ha ripercussioni sull'inferenza: cambiando la scelta del

parametro, le conclusioni potrebbero essere diverse. Allora, si è preferito limitarsi a valutare la probabilità di copertura effettiva.

In tutte le analisi svolte, si è utilizzato un numero di simulazioni pari a 10000, per “ripulire” per quanto possibile i grafici dalle oscillazioni attribuibili alla instabilità dovuta alla simulazione e non a reali andamenti sistematici. Per le simulazioni, si è utilizzato l’ambiente R.

Cap. 2

Il caso Poisson e la soluzione “standard”

2.1. Verosimiglianza e quantità collegate per la distribuzione di Poisson

Sia y realizzazione di una variabile aleatoria Y con distribuzione di Poisson con parametro $\lambda > 0$, ossia con funzione di densità

$$f_Y(y;\lambda) = \frac{e^{-\lambda}\lambda^y}{y!}, \text{ per y = 0, 1, 2, …}$$

Per un campione casuale semplice di numerosità n, se si chiama la media campionaria $\hat{\lambda} = \frac{1}{n}\sum y_i$, la funzione di log-verosimiglianza è:

$$l(\lambda) = \sum_{i=1}^{n} y_i \log\lambda - n\lambda = n\left(\hat{\lambda}\log\lambda - \lambda\right),$$

la funzione di punteggio risulta:

$$l_* = n(\frac{\hat{\lambda}}{\lambda} - 1) = n\frac{\hat{\lambda} - \lambda}{\lambda},$$

l’informazione osservata:

$$j(\lambda) = n\frac{\hat{\lambda}}{\lambda^2},$$

e quella attesa, per la non distorsione dello stimatore $\hat{\lambda}$:

$$i(\lambda) = n\frac{\lambda}{\lambda^2} = \frac{n}{\lambda}.$$

Dunque, per una distribuzione di Poisson i tre test asintoticamente equivalenti hanno le seguenti forme:

1. test di Wald: $r_e(\lambda) = (\hat{\lambda} - \lambda)\sqrt{j(\hat{\lambda})} = (\hat{\lambda} - \lambda)\sqrt{\frac{n}{\hat{\lambda}}} \sim N(0,1)$, dove a $j(\hat{\lambda})$ si possono sostituire indifferentemente $j(\lambda)$ o $i(\lambda)$ (o anche $i(\hat{\lambda})$, che risulta uguale a $j(\hat{\lambda})$);

2. test di Rao (o punteggio): la sua forma generale è $r_u(\lambda) = \frac{l_*(\lambda)}{\sqrt{i(\lambda)}} \sim N(0,1)$, ma nel caso di una Poisson si riconduce a $r_\mu(\lambda) = \frac{l_*(\lambda)}{\sqrt{i(\lambda)}} = \frac{n\frac{\hat{\lambda} - \lambda}{\lambda}}{\sqrt{\frac{n}{\lambda}}} = (\hat{\lambda} - \lambda)\sqrt{\frac{n}{\lambda}}$, e risulta quindi equivalente alla versione di Wald se si utilizza l'informazione attesa;

3. test basati direttamente sul rapporto di verosimiglianza (TRV): $r(\lambda) = \text{sgn}(\hat{\lambda} - \lambda)\sqrt{2(l(\hat{\lambda}) - l(\lambda))} \sim N(0,1)$, che nel caso di Poisson diventa $r(\lambda) = \text{sgn}(\hat{\lambda} - \lambda)\sqrt{2n\left(\hat{\lambda}\log\hat{\lambda} - \hat{\lambda} - \hat{\lambda}\log\lambda + \lambda\right)} =$
$= \text{sgn}(\hat{\lambda} - \lambda)\sqrt{2n\left(\hat{\lambda}\log\frac{\hat{\lambda}}{\lambda} - (\hat{\lambda} - \lambda)\right)}$, non ulteriormente semplificabile.

Nel seguito sarà opportuno tenere conto del fatto che il metodo di Wald che utilizza l'informazione osservata è strutturalmente meno affidabile rispetto al TRV, in quanto dipende dalla parametrizzazione del modello e può includere nell'intervallo valori non appartenenti allo spazio parametrico.

2.2. Intervalli di Wald con varianza stimata

Come prima analisi, si valuta la bontà del metodo basato sul test di Wald con varianza stimata per costruire intervalli per la media della distribuzione di Poisson, per capire se la ricerca di soluzioni alternative sia effettivamente necessaria. Questo metodo sarà nel seguito preso come riferimento e chiamato "standard".

Per un campione casuale semplice di numerosità *n*, $Y_1,...,Y_n$, tratto da una distribuzione di Poisson, lo stimatore di massima verosimiglianza $\hat{\lambda}$ è pari alla media campionaria. Allora, per la normalità asintotica dello stimatore normalizzato, vale che, $\forall \lambda \in \Lambda$, dove $\Lambda = \Re^+$:

$$\Pr_\lambda \{ - z_{1-\alpha/2} \le Z \le z_{1-\alpha/2} \} \doteq 1 - \alpha ,$$

dove $Z = \dfrac{\hat{\lambda} - \lambda}{\sqrt{\hat{\lambda}/n}}$ è appunto lo stimatore normalizzato con la varianza stimata, e $z_{1-\alpha/2}$ indica il quantile 1-α/2 di una Normale standard. Dall'uguaglianza precedente, si ricava allora un intervallo di confidenza del cosiddetto tipo "di Wald", la cui forma è:

$$\hat{\lambda} \pm z_{1-\alpha/2} \sqrt{\hat{\lambda}/n} .$$

Questi intervalli godono della seguente desiderabile proprietà: al divergere di *n* la probabilità che contengano il vero valore del parametro converge a 1 (Brown, Cai e DasGupta, 2001).

2.2.1. Dipendenza dalla numerosità campionaria

Si simula allora da una Poisson per un parametro fissato, ad esempio λ = 5, e si calcola il livello di confidenza effettivo dell'intervallo calcolato nel modo suddetto, al variare di *n* ad esempio tra 1 e 100 (per $n\lambda$ = 500 la

distribuzione di $n\hat{\lambda}$ è sufficientemente bene approssimata dalla Normale, come si vede dalla *Figura 1*, quindi se il metodo è valido i risultati dovrebbero iniziare ad essere soddisfacenti).

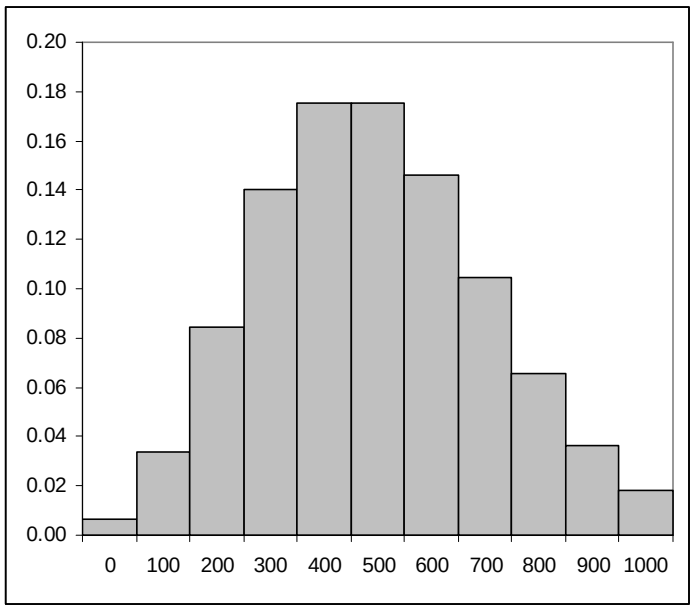

Figura 1: distribuzione di una Poisson(500)

I risultati si riferiscono, come accadrà anche nelle successive analisi, a un livello di confidenza nominale di 0.95; in seguito, si vedrà se i risultati ottenuti cambiano al variare di questa quantità.

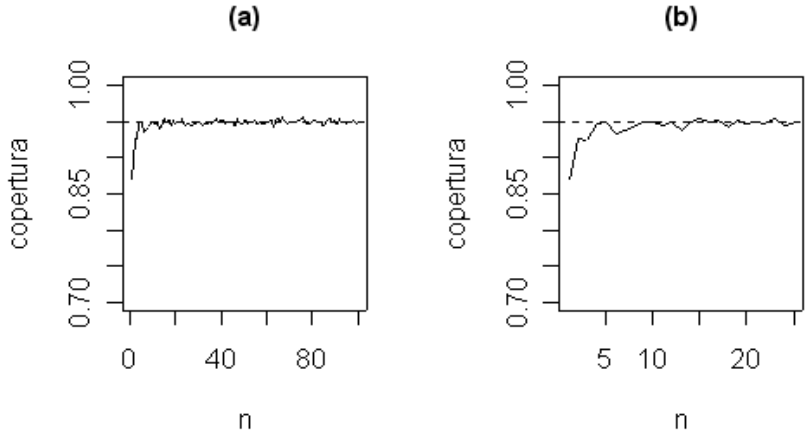

Figura 2: livello di copertura effettivo per λ=5, al variare della numerosità campionaria e ingrandimento dello stesso grafico per valori piccoli di n.

Come si vede dalla *Figura 2*, la copertura sembra piuttosto vicina a 0.95 per valori anche modesti di *n*, e per numerosità superiori a 15-20 i risultati appaiono pienamente soddisfacenti. Dunque, da questa prima analisi, gli intervalli alla Wald non sembrano presentare particolari problemi, soprattutto se ci si limita a considerare campioni non troppo esigui. Si esegue allora la stessa analisi per $\lambda = 1$, ovvero vedendo se le cose cambiano avvicinandosi al "bordo" dello spazio parametrico.

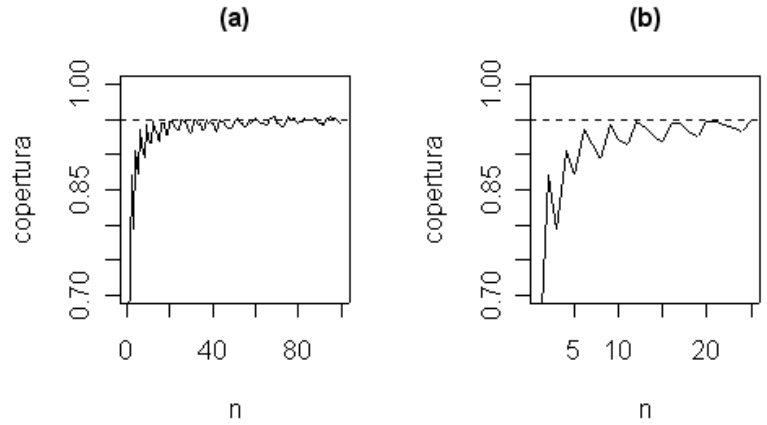

Figura 3: livello di copertura effettivo per λ=1, al variare della numerosità campionaria e ingrandimento dello stesso grafico per valori piccoli di n.

I problemi (si veda la *Figura 3*) cominciano ad essere più gravi: anche per numerosità sostanziose, il livello effettivo di copertura resta quasi sempre al di sotto di quello nominale. Si prova allora un caso ancora più estremo, ovvero $\lambda = 0.1$.

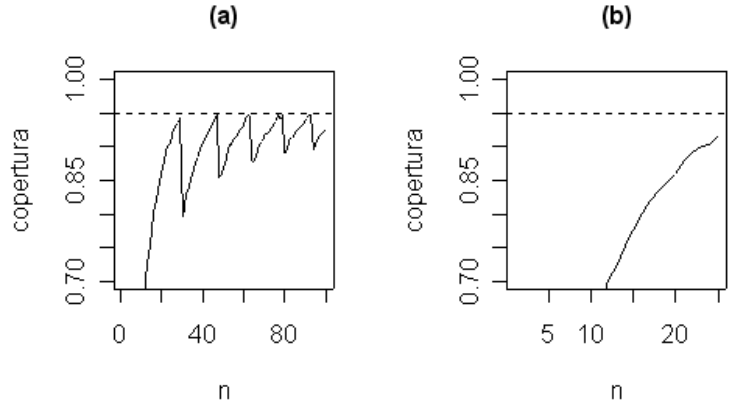

Figura 4: livello di copertura effettivo per λ=0.1, al variare della numerosità campionaria e ingrandimento dello stesso grafico per valori piccoli di n.

Dalla *Figura 4* appare evidente che gli intervalli di confidenza di Wald sono profondamente inadeguati per valori del parametro così piccoli: raggiungono solo raramente la soglia dello 0.95, e presentano delle strane "cadute" nel livello di copertura, che si tenterà di spiegare più avanti.

Comunque, complessivamente sembra emergere un'importante dipendenza del livello di copertura raggiungibile dall'intervallo dalla numerosità campionaria, ma si nota anche una qualche dipendenza dai valori del parametro.

2.2.2. Dipendenza dal valore del parametro

Sembra dunque interessante effettuare l'analisi inversa: tenendo fissata la numerosità si fa variare λ, ad esempio tra 1 e 50 (per $n\lambda$ = 500 la simmetria e quindi l'approssimazione normale è abbastanza buona, si riguardi la *Figura 1*). Le numerosità scelte per l'analisi sono 5, 10, 20, 50.

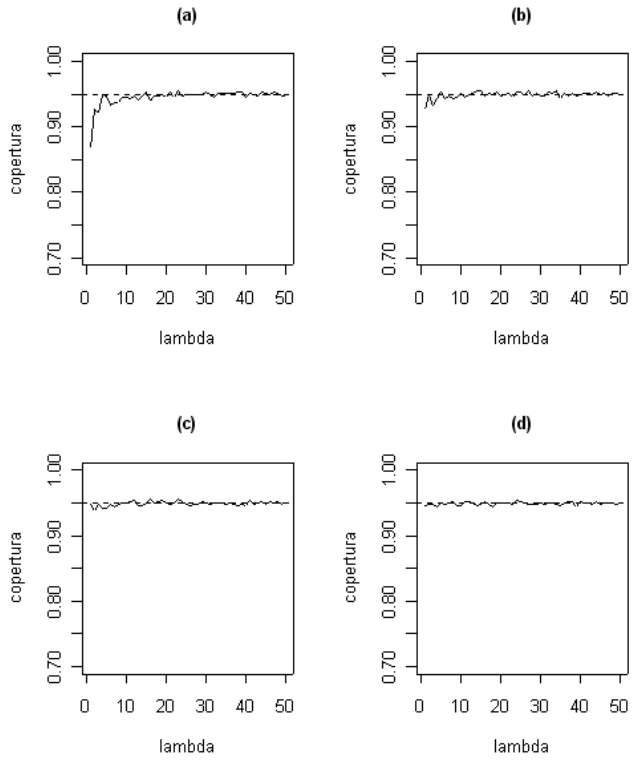

Figura 5: livello di copertura effettivo al variare di λ tra 0 e 50. Le numerosità campionarie sono, nell'ordine definito dalle lettere, 5, 10, 20 e 50.

Guardando la *Figura 5*, si vede come non ci siano particolari problemi già per $n = 10$, anche se rimane qualche dubbio per numerosità molto piccole. E' opportuno allora analizzare nel dettaglio i valori piccoli del parametro, che dovrebbero essere quelli più "problematici": λ viene ristretto all'intervallo (0,5).

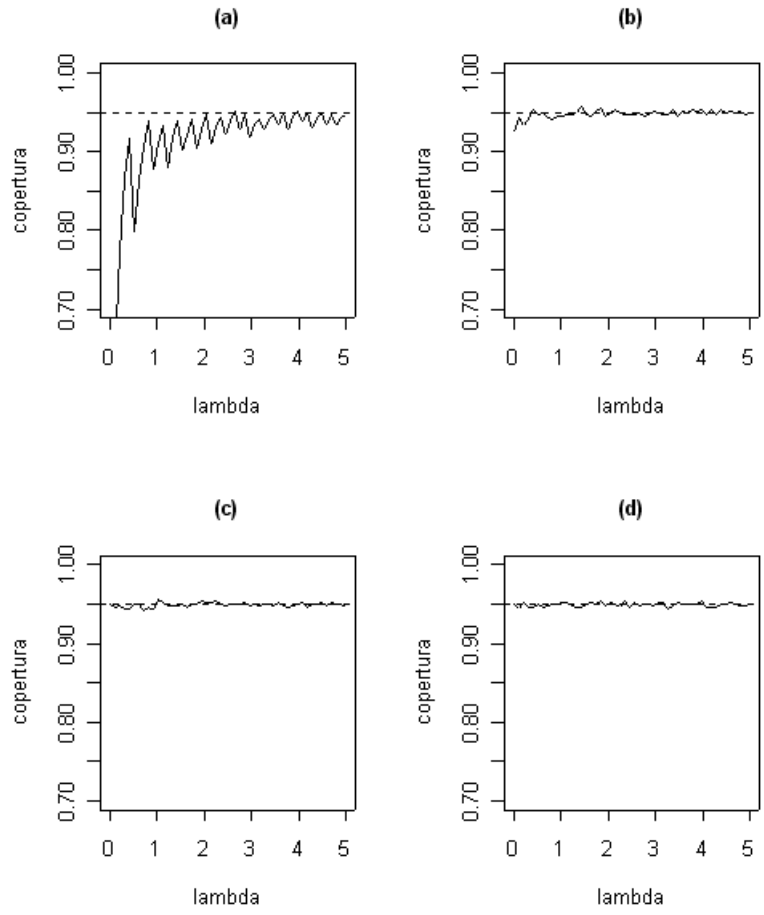

Figura 6: livello di copertura effettivo al variare di λ tra 0 e 5. Le numerosità campionarie sono, nell'ordine definito dalle lettere, 5, 10, 20 e 50.

Dalla *Figura 6* emerge che per $n = 5$ e $n = 10$ il livello di copertura, per quanto vicino alla soglia desiderata, ne rimane quasi sempre al di sotto, anche per valori di λ vicini a 5, mentre per numerosità più elevate l'intervallo ha un comportamento più soddisfacente.

Comunque, le analisi fanno emergere l'esistenza di un'effettiva dipendenza delle prestazioni, oltre che dalla numerosità campionaria, anche dai valori del parametro: in quasi tutto il campo di variazione non si presentano particolari problemi, ma avvicinandoci allo zero la copertura effettiva tende a peggiorare rapidamente.

2.2.3. Dipendenza dal prodotto nλ

Da quanto visto finora emerge il fatto che le prestazioni del metodo risentono in qualche misura sia della numerosità campionaria che del valore di λ, e in particolare si rivelano inadeguate se entrambi i valori sono molto piccoli. Allora può essere interessante studiare il livello di copertura al variare *simultaneo* delle due componenti, metodo frequente di analisi per simulazioni da distribuzioni di Poisson (Barker, 2002). Questa scelta intuitiva ha comunque un fondamento teorico, dal momento che lo stimatore di massima verosimiglianza per un campionamento casuale semplice da una Poisson è, lo si richiama ancora,

$$\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} Y_i \,,$$

da cui si ricava che $n\hat{\lambda} \sim Poisson(n\lambda)$.

Si nota dalla *Figura 7a* che la funzione si stabilizza solo per valori piuttosto elevati del prodotto, e sembra restare comunque leggermente al di sotto della soglia desiderata. Inoltre, per valori particolarmente piccoli (si veda la *Figura 7b*) si hanno cali anche drastici nel livello di copertura.

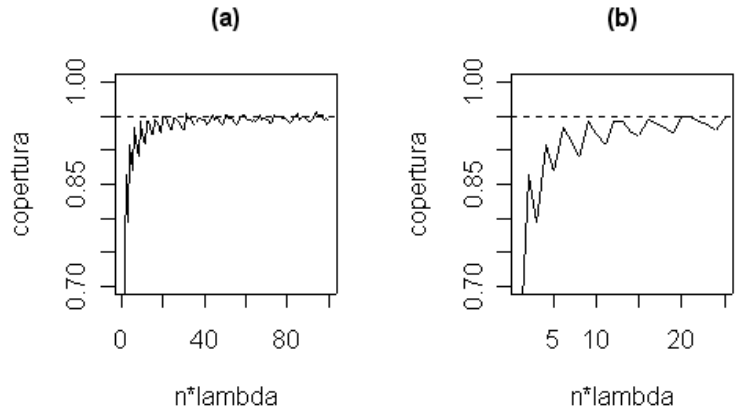

Figura 7: livello di copertura effettivo al variare del prodotto nλ tra 0 e 100 e ingrandimento dello stesso grafico per valori piccoli.

2.2.4. Una prima conclusione

In effetti, come accadeva per il caso binomiale, le prestazioni di questo metodo sono piuttosto insoddisfacenti. Appare allora inevitabile interrogarsi sull'esistenza di alternative migliori. Prima, però, sembra necessario un approfondimento sul fenomeno delle "oscillazioni" riscontrate in molti dei grafici precedenti, che sembrano non essere riconducibili semplicemente alla variabilità dovuta alla simulazione.

2.3. Il fenomeno delle oscillazioni

Questo problema si presenta in molte delle analisi precedenti, spesso lungo tutto il campo di variazione della variabile indipendente, e sia per n fissato che per λ fissato. Ad esempio, dalla *Figura 4a* (qui riproposta per comodità) risulta, contro ogni aspettativa, un livello di confidenza migliore per n intorno a 30 che per n vicino a 80, come emerge dai due "picchi" presentati dalla funzione in corrispondenza di questi due valori, il primo verso l'alto e il secondo verso il basso.

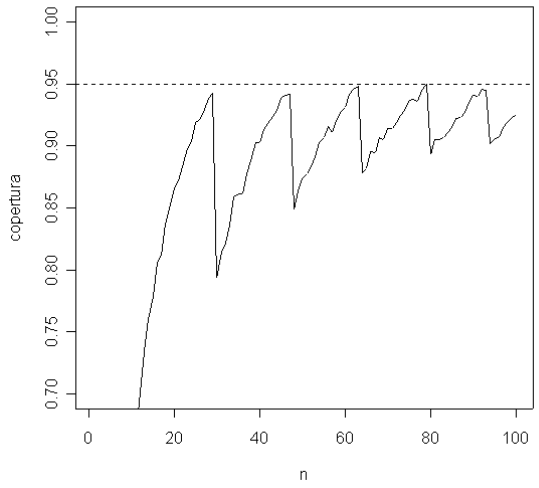

Figura 4a.

Questo spiacevole fenomeno è dovuto al fatto che la distribuzione generatrice dei dati è discreta (si presenta infatti anche nel caso di leggi binomiali), appare sostanzialmente casuale e praticamente ineliminabile (si presenta anche per valori molto elevati di n, o di λ). Nell'analizzare questo fenomeno per distribuzioni binomiali, Brown, Cai e DasGupta (2001) parlano di valori "fortunati" o "sfortunati" per la numerosità e per il parametro, terminologia che rende bene l'idea dell'imprevedibilità del fenomeno: si può notare come valori appartenenti alle due opposte categorie si alternino in continuazione e in modo caotico.

Dal momento che ci attenderemmo che, sostanzialmente, le prestazioni migliorino all'aumentare di n e di λ, il manifestarsi di questo problema non può che essere considerato un indicatore di scarsa attendibilità del metodo, ed è un altro dei motivi per cui sembra opportuno andare alla ricerca di metodi alternativi, più "stabili" di quello standard.

Cap. 3

La ricerca di soluzioni migliori

3.1. Intervalli di Wald con varianza nulla

La prima variazione potenzialmente utile è costruire gli stessi intervalli ma utilizzando l'informazione attesa al posto di quella osservata: con questo secondo approccio, il risultato asintotico da sfruttare è che, $\forall\lambda$,

$$\Pr_\lambda\{-z_{1-\alpha/2} \le Z_0 \le z_{1-\alpha/2}\} \doteq 1-\alpha,$$

dove $Z_0 = \dfrac{\hat{\lambda}-\lambda}{\sqrt{\lambda/n}}$ è lo stimatore normalizzato tramite la varianza nulla invece di quella stimata. La forma dell'intervallo che si ottiene invertendo le due disequazioni è purtroppo, come accadeva nel caso binomiale (Agresti e Coull, 1998), meno compatta e leggibile rispetto a quella molto semplice degli intervalli precedenti:

$$\hat{\lambda} + \frac{(z_{1-\alpha/2})^2}{2n} \pm \frac{(z_{1-\alpha/2})}{2n}\sqrt{(z_{1-\alpha/2})^2 + 4n\hat{\lambda}}\,.$$

Si nota che, a differenza di quello basato sulla varianza stimata, l'intervallo non è centrato su $\hat{\lambda}$; il problema comunque si risolve al divergere di *n*. Si va allora a condurre un'analisi parallela a quella proposta per il metodo precedente, limitandoci però a studiare il livello di copertura in funzione del prodotto $n\lambda$.

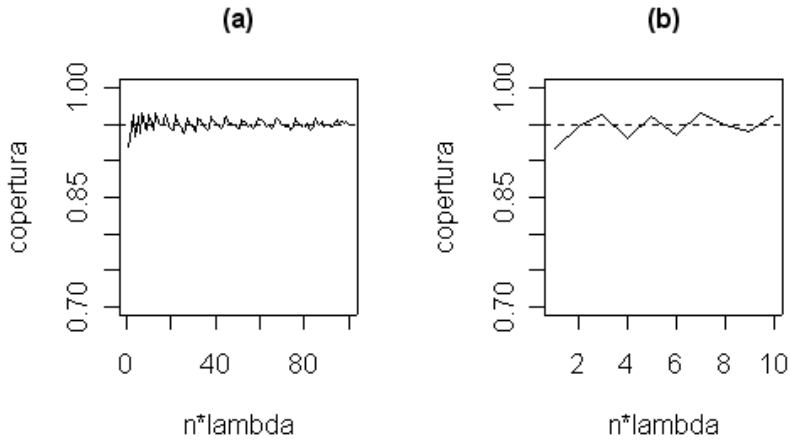

Figura 8: livello di copertura effettivo al variare del prodotto nλ tra 0 e 100 e ingrandimento dello stesso grafico per valori piccoli.

E' evidente, dalla *Figura 8*, come questo metodo sia non solo migliore del precedente (in termini relativi), ma anche molto buono in termini assoluti, anche per valori di λ prossimi allo zero: in particolare le oscillazioni sono molto meno accentuate, ed inoltre sono sostanzialmente "centrate" sulla linea dello 0.95, mentre per il metodo standard restavano sostanzialmente al di sotto di tale soglia. Dunque questo metodo può già sembrare un'alternativa soddisfacente rispetto a quello basato sull'informazione osservata.

Il problema fondamentale della nuova tecnica rimane la scarsa leggibilità della formula, che potrebbe rendere restio ad utilizzarla chi non fosse abituato ad utilizzare software statistici per le proprie analisi. Allora, attraverso una semplice approssimazione, analoga a quella sfruttata da Agresti e Coull, si è tentato di dare alla formula un aspetto più accessibile e che la renda più facile da calcolare. La semplificazione è valida per intervalli di livello 0.95.

3.1.1. Una formulazione più semplice

Per quanto il risultato ottenuto da Agresti e Coull (1998) non sia direttamente generalizzabile, in quanto porta a una formula con

interpretazione pratica immediata in un modo in un certo senso casuale, si è ugualmente cercata un'analoga semplificazione per il caso di Poisson.

Come idea di partenza si è presa la stessa: data la formulazione dell'intervallo basata sulla varianza nulla con livello di copertura 0.95, ovvero

$$\hat{\lambda} + \frac{(1.96)^2}{2n} \pm \frac{(1.96)}{2n}\sqrt{(1.96)^2 + 4n\hat{\lambda}}\,,$$

dove 1.96 è all'incirca il valore del quantile $z_{0.975}$, è stato approssimato 1.96 a 2. Questa semplice approssimazione ne produce altre a cascata, che portano alla forma finale:

$$\frac{\sum_{i=1}^{n} y_i + 2}{n} \pm 2\frac{\sqrt{\sum_{i=1}^{n} y_i + 1}}{n},$$

che ha proprio la struttura di un intervallo di Wald basato sulla varianza stimata: le sole correzioni richieste alla formula standard sono di aggiungere alla somma delle osservazioni 2 nell'espressione della media e 1 nell'espressione della varianza. Questa forma è senz'altro più chiara e leggibile della precedente; ha, per costruzione, prestazioni equivalenti ad essa, e ha un livello di copertura molto vicino all'altra (come visto all'inizio, questo vale circa 0.9545).

3.2. Intervalli basati sul TRV

Il metodo proposto nella sezione precedente appariva soddisfacente. E' senz'altro utile cercare se tra gli altri metodi possibili ce n'è qualcuno che ha prestazioni ancora migliori: si inizia con gli intervalli derivanti dal TRV.

Sapendo che $r(\lambda) = \text{sgn}(\hat{\lambda} - \lambda)\sqrt{2n\left(\hat{\lambda}\log\frac{\hat{\lambda}}{\lambda} - (\hat{\lambda} - \lambda)\right)} \sim N(0,1)$, si può scrivere una regione di confidenza in forma implicita:

$$\hat{\Theta}(y) = \left\{\lambda > 0 : z_{\alpha/2} \le \text{sgn}(\hat{\lambda} - \lambda)\sqrt{2n\left(\hat{\lambda}\log\frac{\hat{\lambda}}{\lambda} - (\hat{\lambda} - \lambda)\right)} \le z_{1-\alpha/2}\right\},$$

e valutarne le prestazioni.

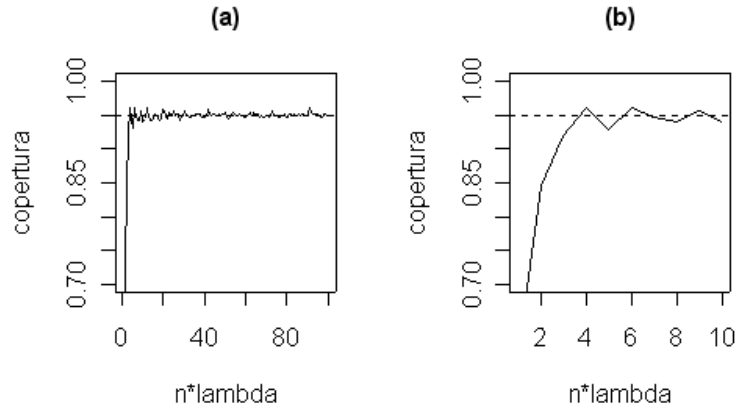

Figura 9: livello di copertura effettivo al variare del prodotto nλ tra 0 e 100 e ingrandimento dello stesso grafico per valori piccoli.

Dalla *Figura 9* il metodo sembra piuttosto affidabile, ma appare peggiore del metodo di Wald basato sulla varianza nulla per valori piccoli del prodotto $n\lambda$.

3.3. Una correzione per l'asimmetria

Dal momento che la distribuzione di Poisson, oltre ad essere discreta, presenta anche un'asimmetria positiva, soprattutto per numerosità basse o valori del parametro piccoli, una strada diversa per migliorare le prestazioni degli intervalli può essere cercare di ridurre questa asimmetria tramite una qualche trasformazione dello stimatore.

Una proposta (Pace e Salvan, 1996) è utilizzare la funzione $g(x) = x^{2/3}$, che è sufficientemente regolare in tutto il suo dominio: allora, se si applica questa trasformazione allo stimatore $\bar{Y}_n$, il quale si ricorda avere distribuzione asintotica $N(\lambda, \frac{\lambda}{n})$, si ottiene che $g(\bar{Y}_n) \sim N\left(g(\lambda), \frac{\lambda}{n}[g'(\lambda)]^2 \right)$.

Dunque lo stimatore trasformato, lo si chiami W_n, si distribuisce così:

$$W_n \sim N\left(\lambda^{2/3}, \frac{4\lambda^{1/3}}{9n} \right).$$

Una regione di confidenza ha allora forma (implicita):

$$\hat{\Theta}(y) = \left\{ \lambda > 0 : z_{\alpha/2} \leq \frac{3\sqrt{n}(\hat{\lambda}^{2/3} - \lambda^{2/3})}{2\lambda^{1/6}} \leq z_{1-\alpha/2} \right\},$$

dove le *z* rappresentano come sempre i quantili di una Normale standard.

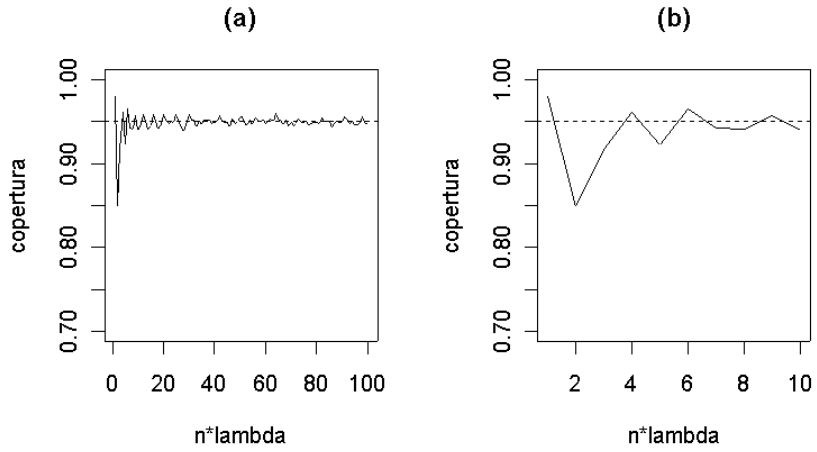

Figura 10: livello di copertura effettivo al variare del prodotto nλ tra 0 e 100 e ingrandimento dello stesso grafico per valori piccoli.

Le prestazioni del metodo, come si evince dalla *Figura 10*, sono abbastanza buone per gran parte dei valori di *nλ*, tranne che per valori piccoli (in realtà, il metodo non è eccellente anche per valori non eccessivamente piccoli).

3.4. Intervalli “esatti”

Un modo sostanzialmente diverso di costruire regioni di confidenza è quello di basarsi direttamente sulla distribuzione esatta delle variabili studiate, senza passare attraverso la funzione di verosimiglianza. Per il caso binomiale, i primi a proporre questo approccio sono stati Clopper e Pearson (1934), ottenendo risultati migliori rispetto all’intervallo “standard”, ma un po’ troppo “conservatori”: gli intervalli presentavano un livello di copertura sistematicamente superiore a quello nominale, il che non sempre è desiderabile, come spiegato in precedenza.

Per la distribuzione di Poisson, chiamata $y = \sum_{i=1}^{n} y_i$, ciò che bisogna fare per ottenere un intervallo di livello 1-α è risolvere in λ le due formule:

$$\sum_{k=0}^{y-1} e^{-1/\lambda} \frac{\lambda^{-k}}{k!} = 1 - \alpha/2$$

e

$$\sum_{k=y}^{n} e^{-1/\lambda} \frac{\lambda^{-k}}{k!} = \alpha/2 .$$

Ciò che si ottiene (Azzalini, 2000) è un intervallo di confidenza di forma:

$$\left(\frac{2n}{c_{2y}}, \frac{2n}{c_{2(y+1)}} \right),$$

dove c_{2y} indica il quantile di livello $\alpha/2$ di una distribuzione X^2_{2y} con $2y$ gradi di libertà e $c_{2(y+1)}$ è il quantile di livello 1-$\alpha/2$ di un $X^2_{2(y+1)}$.

Le prestazioni, al variare del prodotto $n\lambda$, sono illustrate in *Figura 11*, e mostrano essenzialmente due cose:

- la prevedibile “conservatività” del metodo, in quanto la curva è quasi sempre al di sopra della soglia dello 0.95 (in particolare, in diversi casi si è trovato un limite inferiore dell’intervallo praticamente coincidente con lo zero);
- un comportamento soddisfacente (anche se comunque conservatore) anche per valori minimi della media dello stimatore.

Il metodo dunque è senz’altro migliore di quello standard, ma, per i motivi esposti nell’introduzione, si può ritenere meno utile rispetto a quello basato sulla varianza nulla.

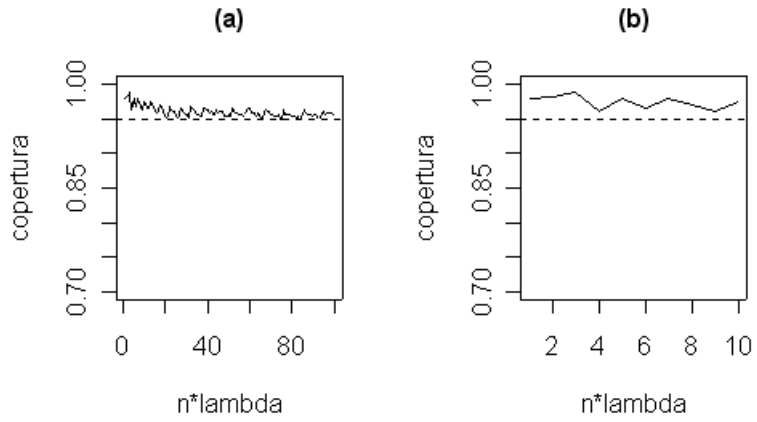

Figura 11: livello di copertura effettivo al variare del prodotto nλ tra 0 e 100 e ingrandimento dello stesso grafico per valori piccoli.

3.5. Effetti del livello di significatività

In tutte le analisi effettuate, si è tenuto fisso il livello di significatività a 0.95. Sembra però doveroso valutare se le prestazioni cambiano al variare di questo fattore, in particolare per quello che sembra il metodo migliore tra quelli valutati: il metodo di Wald con la varianza nulla. Per questa analisi,

per cercare di limitare gli effetti della numerosità e del valore del parametro, si è posto $n\lambda$=100.

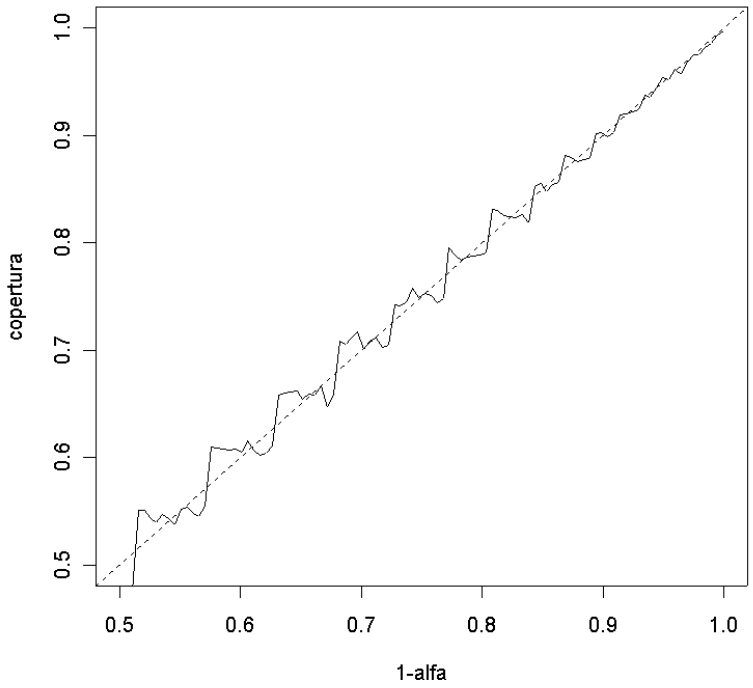

Figura 12: livello di copertura effettivo al variare del livello di copertura teorico tra 0.5050 e 0.9975.

La *Figura 12* valuta la copertura effettiva rispetto a quella prevista: la situazione ottima sarebbe rappresentata dalla bisettrice del grafico. L'andamento osservato non sembra troppo diverso da quello ideale, dunque si conclude che il metodo non peggiora in corrispondenza di particolari livelli di significatività, bensì mantiene sempre buone prestazioni.

Cap. 4

Conclusioni

In definitive, le analisi delle prestazioni delle diverse tecniche per costruire intervalli di confidenza per la media di una Poisson hanno portato a una conclusione generale e a qualche suggerimento specifico:

- gli intervalli di Wald che sfruttano la varianza stimata, a dispetto della loro larga diffusione, sono decisamente poco affidabili. Sembra dunque più opportuno utilizzare una delle altre tecniche proposte, che hanno dimostrato di essere sostanzialmente tutte migliori di quella “standard”;
- in particolare, il metodo in assoluto più attendibile, soprattutto per valori piccoli del prodotto tra valore del parametro e numerosità campionaria, è sembrato quello di Wald basato sulla varianza nulla, che è quindi quello che si consiglia di utilizzare. Inoltre, se si necessita di costruire intervalli di livello vicino a 0.95, si può prendere in considerazione la semplificazione proposta nel sottoparagrafo 3.1.1;
- infine, nei casi in cui possa essere utile un livello di confidenza effettivo superiore a quello richiesto, la tecnica migliore è senz’altro quella “esatta” proposta nel paragrafo 3.4.

Un ulteriore passo nell’analisi potrà essere la valutazione di tecniche analoghe in presenza di osservazioni non identicamente distribuite, ad esempio in modelli di regressione Poisson.

Bibliografia

Agresti, A. e Coull, B.A. (1998). Approximate is better than “exact” for interval estimation of binomial proportions. *The American Statistician*, **52**, 119-126.

Azzalini, A. (2000). *Inferenza Statistica. Una Presentazione basata sul Concetto di Verosimiglianza*. Springer-Verlag, Milano.

Barker, L. (2002). A comparison of nine confidence intervals for a Poisson parameter when the expected number of events is 5. *The American Statistician*, **56**, 85-89.

Brown, L., Cai, T. e DasGupta, A. (2001). Interval estimation for a binomial proportion. *Statistical Science*, **16**, 101-133.

Clopper, C. J., e Pearson, E. S. (1934). The use of confidence intervals for a binomial parameter, *Canadian Journal of Statistics*, **22**, 207-218.

Henderson, M. e Meyer, M. (2001). Exploring the confidence interval for a binomial parameter in a first course in statistical computing. *The American Statistician*, **55**, 337-344.

Olgiati, O., Paglieri, L., Salvati, S. e Secchi, P. (2006). Storia di un caso: intervalli di confidenza per una proporzione per la regolazione della qualità del servizio nel settore energetico nazionale. *Statistica & Società*, **anno IV n. 2**, 22-32.

Pace, L. e Salvan, A. (1996). *Teoria della Statistica: Metodi, Modelli, Approssimazioni Asintotiche*. Cedam, Padova.

Wilson, E.B. (1927). Probabile inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, **22**, 209-212.