UNIVERSITÀ DEGLI STUDI ROMA TRE
FACOLTÀ DI SCIENZE M. F. N.

Sintesi di Laurea in Matematica

di

Giuseppe Manlio Luciano

Infezioni HIV, sistema di sorveglianza e codice identificativo

Relatore
Prof. Gianpaolo Scalia Tomba

Correlatore
Dott. Patrizio Pezzotti

Il Relatore

Il Candidato

ANNO ACCADEMICO 2000 - 2001
LUGLIO 2001

Classificazione AMS : 60J05 , 60J22 , 62P10
Parole Chiave : Catene di Markov, Processo di Poisson, Inferenza su dati parziali

Luciano Giuseppe Manlio è nato a Roma il 17 Marzo 1977.
Ha conseguito il Diploma di Maturità scientifica presso il Liceo Scientifico statale "W. Goethe" di Roma nel luglio 1996.
Si è immatricolato al Corso di Laurea in Matematica presso l'Università degli Studi di Roma Tre nell'anno accademico 1996 - 1997.
Ha presentato per la prova di qualificazione all'esame di laurea le seguenti tesine orali :
"Teorema ergodico per catene di Markov e applicazioni"
"Massa inerziale e gravitazionale-fondamenti di relatività generale".

gfggfgfgfd

La principale caratteristica del virus HIV è di essere "invisibile".
Non esistono, infatti, dati certi della sua diffusione, ma solo stime.
Gli unici dati certi afferiscono ai casi di AIDS conclamato, che vengono raccolti, per il nostro Paese, dall'Istituto Superiore di Sanità, organo tecnico scientifico del Ministero della Sanità.

La protezione dell'anonimato nel rispetto della privacy di coloro che si sottopongono al test HIV è una misura cautelare doverosa che deve essere mantenuta; è anche vero che recentemente il Ministero della Sanità sta cercando di risolvere il problema della conoscenza dell'espansione del virus HIV introducendo registri non nominativi (ogni persona un codice) delle persone sieropositive.

L'istituzione di un tale registro (con le cautele del caso dovute alla garanzia reale dell'anonimato delle persone HIV positive) significherebbe un salto in avanti per le strategie di lotta all'AIDS; le diagnosi di AIDS conclamato sono sempre più dilazionate nel tempo, e quindi la registrazione dei casi di AIDS, risultando sottodimensionata, spiega sempre meno l'andamento dell'infezione.

L'invisibilità del virus è inoltre caratterizzata da un altro elemento determinante: la bassa percentuale di persone che si sottopongono al test della ricerca degli anticorpi anti HIV.

L'alone di paura e di diffidenza che da sempre circonda l'AIDS, come patologia in grado di modificare pesantemente le condizioni di vita della persona sotto il profilo fisico, psicologico e sociale, ha suggerito, e continua tuttora a convincere, molte persone a non effettuare il test.

La diagnosi di HIV positivo prima, e di AIDS conclamato poi, nella nostra esperienza corrisponde quasi sempre ad una condanna all'esclusione dal tessuto sociale.

I numerosi casi di discriminazione registratisi in questi anni, soprattutto in ambito scolastico e lavorativo, costituiscono una vera e propria barriera all'assunzione della piena responsabilità, verso se stessi e verso gli altri, rappresentata dal test.

L'HIV e di conseguenza l'AIDS sono dei fantasmi che ci circondano ed essendo "invisibili" e mortali ci spaventano; personalmente l'HIV mi terrorizza e mi sorprende la superficialità con cui molte persone si espongono ad un possibile contagio malgrado le continue campagne di informazione e la conoscenza delle vie di trasmissione del virus.

Noi giovani dobbiamo guardare con attenzione le nuove percentuali della malattia che indicano una crescita del virus negli adolescenti e ognuno di

noi deve cercare nel suo piccolo di impegnarsi nell'ambiente in cui vive in un'opera di sensibilizzazione al problema.

Oltre i problemi strettamente di ordine sanitario è evidente la necessità di capire l'entità del problema e il reale numero di persone sieropositive; lo scopo di questa tesi è proprio nel poter stimare quale sia il numero di persone positive al test nel Lazio e di fornire un modello di stima e di calcolo per gli altri sistemi di sorveglianza dove la necessità di registrare tali notifiche in accordo con il rispetto della privacy dell'individuo non consente di indicare con certezza l'andamento della malattia.

Sentendo personalmente il problema e temendolo, ho accettato con grande interesse la possibilità di discuterlo e approfondirlo sperando nel mio piccolo di proporre un aiuto per chi si impegna quotidianamente per trovare nuove strade per combattere e per limitarne la diffusione.

All'inizio degli anni '80, vennero segnalati ai *Centers for Disease Control* americani alcuni casi di patologie molto rare che si associavano ad una grave deficienza del sistema immunitario in giovani maschi della California.

Nell'arco di pochi mesi sempre più casi di malattie rare associate ad immunodeficienza furono riportati dalla stessa California e da altri stati. Al fine di poter meglio comprendere la causa dell'insorgere di queste patologie e per meglio monitorare l'andamento di queste malattie fu istituita la sorveglianza dei casi della sindrome da immunodeficienza acquisita, divenuta poi tristemente nota come AIDS.

Tutte le nazioni furono successivamente invitate ad istituire dei sistemi di sorveglianza dei casi di AIDS.

L'Italia fu una delle prime ad attivarne uno presso l'*Istituto Superiore di Sanità* (1983), inizialmente su base volontaria e poi divenuto obbligatorio per legge nel 1987.

Solo dalla metà degli anni '80 diventa chiaro che tale sindrome è associata ad una infezione virale fino ad allora sconosciuta alla quale viene dato il nome di virus dell'immunodeficienza umana (*HIV*).

Dal 1985 un test diagnostico commerciale, divenuto disponibile per la diagnosi, ha favorito l'attivazione di sistemi di sorveglianza dell'infezione da HIV anche per persone che ancora non hanno sviluppato l'AIDS.

I sistemi di sorveglianza hanno avuto difficoltà nella loro implementazione in quanto molte persone, temendo una potenziale schedatura degli HIV positivi, si sono opposte all'utilizzo di codici identificativi univoci (cioè basati direttamente su nome e cognome e sui dati anagrafici, anche criptati).

La sorveglianza dei casi di AIDS ha fornito fino ad oggi informazioni estremamente utili sull'epidemia dell'infezioni da HIV in Italia consentendo di conoscere l'andamento temporale e le caratteristiche dei pazienti con AIDS, nonchè di stimare le infezioni da HIV avvenute negli anni passati.

Durante il 1996 è stata osservata in italia un'iniziale diminuzione dei nuovi casi di AIDS (-12%) e dei decessi correllati all'AIDS rispetto al 1995 (-10%), dovuta in gran parte alla maggiore efficacia delle terapie disponibili. La modificazione di parametri essenziali per la costruzione di modelli matematici di back-calculation quali il tempo di incubazione e l'andamento dei casi di AIDS rende difficile la stima dell'epidemia da HIV. In pratica è sempre piu difficile identificare cambiamenti recenti nella diffusione dell'HIV sulla base delle segnalazioni dei casi di AIDS.

Premesso, quindi, che la sorveglianza dei casi di AIDS rimane comunque una preziosa fonte di dati, appare evidente che il monitoraggio delle nuove diagnosi di infezioni da HIV rappresenta il metodo migliore per descrivere le modificazioni in atto nell'epidemia dell'infezione da HIV, nonchè per fornire gli strumenti necessari a pianificare interventi di prevenzione primaria e secondaria.

In alcune regioni e province italiane (Lazio, Friuli Venezia-Giulia, Umbria, Veneto, Modena, Trento) questa esigenza era già sentita negli anni passati portando all'attivazione di sistemi di sorveglianza delle infezioni da HIV che hanno fornito a livello locale dei risultati di grande interesse nel monitoraggio dell'epidemia.

In base a quanto sopra esposto, si sta cercando di attivare un sistema di sorveglianza delle infezioni da HIV in tutte le Regioni ove tale sistema non sia già esistente e viene proposto un coordinamento dei vari sistemi regionali o provinciali di sorveglianza dell'infezione da HIV presso il Centro Operativo AIDS (*COA*) dell'Istituto Superiore di Sanità.

Uno dei problemi principali inerenti un sistema di sorveglianza per le infezioni da HIV è costituito dalla necessità di identificare i casi segnalati ed evitare doppie conte rispettando la riservatezza e l'anonimato degli individui che si sottopongono al test anti-HIV.

Il codice ideale dovrebbe coniugare una bassa probabilità di duplicati ad una quantità minima di dati personali, tali da non permettere di risalire all'identità del soggetto segnalato.

Tuttavia, trovare il giusto equilibrio fra queste due caratteristiche non è semplice: tanto minore è il numero di dati personali inseriti nel codice, tanto

maggiore è la probabilità che due individui diversi vengano identificati con lo stesso codice.

1.1 Metodi di stima del numero di persone differenti presenti in un registro

Presso l'Osservatorio Epidemiologico della regione Lazio(OER) vengono raccolte dal 1985 le notifiche relative ai casi accertati di sieropositività per HIV. Tutti i laboratori di patologia clinica, pubblici e privati e i centri trasfusionali inviano all'OER un documento contenente informazioni cliniche e demografiche della persona che si è sottoposta al test. Le nuove diagnosi vengono individuate dall'archivio delle notifiche attraverso una procedura di linkage che utilizza come chiavi le variabili genere, data e luogo di nascita. Nel momento in cui perviene una notifica con un codice già esistente viene considerata come un caso già accertato e quindi non registrata. Tale procedura porta ovviamente ad un problema di sottostima ed è quindi evidente la necessità di capire quanto questo procedimento influisca sulla stima finale e quindi le possibili modifiche da attuare.

L'attuazione di un determinato codice (ad esempio nel Lazio costituito dal comune di nascita, data di nascita e sesso) porta alla partizione della popolazione in A gruppi ognuno dei quali costituito da G_i individui con $1 \leq \mathrm{i} \leq A$ che condividono lo stesso codice.

Per semplicità consideriamo di avere due individui che condividano lo stesso codice, il primo già notificato con una intensità di testaggio λ_H mentre il secondo non notificato e con una intensità λ_N. Il tempo fino al prossimo test del primo può essere visto come

$$X \sim Exp(\lambda_H)$$

e analogamente per il secondo:

$$Y \sim Exp(\lambda_N)$$

supponendo i due processi indipendenti è facilmente calcolabile:

$$P(X < Y) = \int_0^\infty \int_x^\infty \lambda_H \lambda_N e^{-\lambda_H x - \lambda_N y} \, dxdy = \frac{\lambda_H}{\lambda_H + \lambda_N}$$

che indica la probabilità che, avvenuta una segnalazione, appartenga al primo.

Espandendo il tutto ad un gruppo di G individui dei quali s già notificati, ognuno con una intensità di notifica λ_H ed i restanti $(G - s)$ non notificati con una intensità di notifica λ_N avremo :

$$P_{ss} = \frac{s\lambda_H}{s\lambda_H + (G-s)\lambda_N} = \frac{s}{s + (G-s)\varrho} \quad con \quad \varrho = \frac{\lambda_N}{\lambda_H}$$

che indica la probabilità che avendo già s notificati differenti e arriva solo una nuova segnalazione questa sia di un individuo già notificato.

Introduciamo ora una nuova variabile aleatoria W_k che conta il numero di notifiche necessarie per avere k individui differenti registrati con

$$W_k = \begin{cases} T_1 + T_2 + T_3 + ... + T_k & \text{per } k = 0, 1, ..., G \ , \\ \infty & \text{per } k \geq G + 1 \quad . \end{cases}$$

e con

$$T_k \sim Ge(P_{k-1,k}) = Ge(1 - P_{k-1,k-1});$$

in particolare T_k indicherà le segnalazioni necessarie per passare da $k-1$ *a* k notifiche differenti. Inoltre la probabilità che il numero di notificati differenti X_k dopo k segnalazioni dello stesso codice sia maggiore di z sarà

uguale alla probabilità che il numero di notifiche necessarie per averne $z+1$ differenti sia $\leq k$ cioè:

$$P(X_k > z) = P(W_{z+1} \leq k)$$

e quindi facendo ricorso alla funzione generatrice e ponendo per semplicità di notazioni

$$p_{jj} = p_j \quad q_{\text{J}} = 1 - q_j \quad C_\alpha = \prod_{\substack{j=1 \\ j \neq \alpha}}^{z} \frac{1}{1 - \frac{q_j}{q_\alpha}}$$

avremo che

$$E[X_k] = \sum_{z=1}^{G} P(W_z \leq k) = \sum_{z=1}^{min(k,G)} (1 - \sum_{\alpha=1}^{z} C_\alpha \cdot q_\alpha^k) \tag{1.1}$$

che indica il valore atteso di individui differenti su k segnalazioni dello stesso codice.

Il valore atteso del numero di notificati differenti può essere trovato anche ricorsivamente utilizzando le proprietà delle catene di Markov. La matrice di transizione M associata al nostro processo stocastico sugli stati

$E = (0, 1, 2, 3, 4, ..., G-1, G)$ sarà della forma:

$$M = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & p_{1,1} & 1-p_{1,1} & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & p_{2,2} & 1-p_{2,2} & 0 & \dots & 0 & 0 \\ 0 & 0 & 0 & p_{3,3} & 1-p_{3,3} & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & \dots & p_{G-1,G-1} & 1-p_{G-1,G-1} \\ 0 & 0 & 0 & 0 & 0 & \dots & 0 & 1 \end{pmatrix}$$

quindi essendo $\Pi_0 = (1, 0, 0, ..., 0)$ la distribuzione iniziale avremo:

$$\Pi_k = \Pi_0 \cdot M^k$$

La prima riga della matrice M^k ci darà dunque la densità di probabilità di X_k cioè:

$$w_s^k = M_{0,s}^k \quad con\ k\ \geq 1\ \ ed\ \ 0 \leq s \leq G$$

indicherà la probabilità di avere s notifiche differenti dopo k segnalazioni.

Una volta conosciuta w_s^k avremo inoltre:

$$w_s^{k+1} = P_{s,s} \cdot w_s^k + (1 - P_{s-1,s-1}) \cdot w_{s-1}^k$$

Quindi:

$$E[X_k] = \sum_{s\geq 0}^{G} s \cdot w_s^k \tag{1.2}$$

e

$$Var(X_k) = E[X_k^2] - (E[X_k])^2 = \sum_{s\geq 0}^{G} s^2 \cdot w_s^k - \left(\sum_{s\geq 0}^{G} s \cdot w_s^k\right)^2 \tag{1.3}$$

Un vantaggio di questo metodo, oltre al fatto di fornirci una veloce procedura ricorsiva per calcolare le quantità desiderate, consiste nel poter modificare il valore di ϱ nel tempo, cioè considerare una catena di Markov disomogenea nel tempo.

La (1) e la (2) ci danno il numero di persone differenti notificate su k segnalazioni dello stesso codice. Supponiamo di avere N notifiche e che in realtà il numero di persone differenti notificate sia X e che i codici comportino una partizione della popolazione in A gruppi dei quali P presenti nel registro e cerchiamo di trovare una formula che esprima il numero di persone differenti notificate.

Avremo quindi:

$$X_N = \sum_{s=1}^{P} X_{k_s}(\ G(C_s)\ ,\ \varrho(C_s)\ )$$

con

- C_s=codice essesimo
- k_s=numero di notifiche all'interno dell'essesimo gruppo quindi numero di volte che compare l'essesimo codice
- X_{k_s}=numero reale di persone differenti notificate che condividono l'essesimo codice il quale compare k_s volte

- $G(C_s)$=numero di persone all'interno della popolazione che condividono lo stesso codice essesimo
- $\varrho = \frac{Intensità\ di\ prima\ notifica\ nel\ registro\ del\ Lazio}{Intensità\ di\ rinotifica\ per\ chi\ è\ stato\ notificato\ per\ la\ prima\ volta\ nel\ Lazio}$.

  La dipendenza di ϱ dal codice è dovuta al fatto che all'interno di un codice figura il luogo di nascita di un individuo e quindi λ_N per un individuo nato fuori dal Lazio sarà molto inferiore rispetto al valore per una persona nata nel Lazio, mentre λ_H per un individuo nato fuori dal Lazio ma notificato nella regione, essendo probabile che vi risieda, sarà lo stesso degli abitanti nati nel Lazio.

Quindi, essendo $E_{k_s} = E[X_{k_s}(\ G(C_s)\ ,\ \varrho(C_s)\ )]$ lo stimatore del numero di individui differenti su k segnalazione di uno stesso codice essesimo, lo stimatore $E_N = E[X_N]$ del numero di individui differenti su N segnalazioni generiche sarà della forma:

$$E_N = \sum_{s=1}^{P} E_{k_s}$$

$$Var(E_N) = \sum_{s=1}^{P} Var(E_{k_s})$$

Supponendo che al tempo T_1 abbiamo N_1 notifiche con P codici differenti e al tempo $T_2 \quad N_2 = N_1 + Z$ notifiche con $\tilde{P}$ codici differenti avremo che:

$$E_{N_2} = \sum_{s=1}^{\tilde{P}} E_{k_s+h_s}$$

avendo indicato con k_s il numero di volte che l'essesimo codice compare fino al tempo T_1 e con h_s il numero di volte che l'essesimo codice compare nell'intervallo di tempo $T_2 - T_1$ ed avremo, volendo stimare l'incremento dei nuovi individui notificati durante tale intervallo di tempo:

$$E[N_2 - N_1] = E[N_2] - E[N_1]$$

mentre

$$Var(N_2 - N_1) = Var(N_1) + Var(N_2) - 2 \cdot Cov(N_1, N_2)$$

Per usare il metodo di stima presentato sopra occore avere un'idea dei valori di G e ϱ. Non essendo disponibili sarà necessaria una stima di questi parametri.

1.2 Stima del numero di gruppi distinti presenti nel registro su n notifiche differenti

Supponiamo di avere una popolazione costituita da N individui e di utilizzare un determinato codice (ad esempio nel Lazio sesso, luogo e data di nascita). Ogni persona apparterrà ad un determinato gruppo indotto dal codice utilizzato e avendo scelto n persone a caso tra la popolazione cerchiamo di trovare il numero di gruppi distinti ai quali appartengono gli n individui.
Indicando con:

G = numero dei gruppi indotti dal codice,

G_k = numero dei gruppi con k individui,

$\lambda = \frac{N}{G}$ numero medio di persone per gruppo,

avremo:

$$G = \sum_k G_k \qquad e \qquad N = \sum_k k \cdot G_k$$

Prendendo a caso un individuo della popolazione ed essendo Y una v.a. indicante il numero di persone del gruppo a cui appartiene:

$P(Y = k) = \frac{k \cdot G_k}{N}$

e quindi il valore atteso di Y sarà:

$$E[Y] = \sum_k k^2 \frac{G_k}{N}$$

Sia Z_k con $0 \leq K \leq N$ v.a. che prendendo n persone a caso su una popolazione di N individui indica il numero di gruppi con k persone estratte. Indicando quindi con U_n il numero di gruppi distinti con almeno una persona "estratta" prendendo n persone all'interno della popolazione sarà quindi:

$U_n = G - Z_0 \quad \text{con} \quad 1 \leq \ U_n \leq \ min(n, G)$

Inoltre avremo che:

$$E[z_0] = \sum_k G_k \frac{\binom{N-k}{n}}{\binom{N}{n}} \ \approx \ \sum_k G_k (1 - \tfrac{n}{N})^k \ = \ G \cdot e^{-\lambda \frac{n}{N}}$$

Quindi,

$E[U_n] = G - E[Z_0] = G(1 - e^{-\frac{n}{G}})$

e ponendo $\beta = \frac{n}{G}$

avremo

$E[\frac{U_n}{n}] = \frac{1-e^{-\beta}}{\beta}$

che indica la frazione di gruppi differenti su n individui di una popolazione di G persone.

$\beta = 0.0$	$\beta = 0.1$	$\beta = 0.2$	$\beta = 0.3$	$\beta = 0.4$	$\beta = 0.5$
1.000	0.952	0.906	0.864	0.824	0.787

Tabella 1.1: Frazione di gruppi differenti su n individui di una popolazione di G persone.

CASO COMUNE DI ROMA CON REGISTRO AGGIORNATO AL 31/12/00:

Esistono circa $2 \cdot 365 \cdot 100 \approx 73000 = G$ codici differenti.
Se prendiamo ad esempio 15000 individui, avremo $\beta \approx 0.2$ e circa 13500 codici.
Visto al contrario, dati D codici differenti osservati, avremo:

$n \approx -G \cdot ln(1 - \frac{D}{G})$

e con i dati romani avendo 19588 notifiche di cui 6440 differenti e $G = 73000$ allora $n \approx 6741$ ma considerando il fatto che bisogna porre delle limitazioni dovute all'età possiamo prendere $G \approx 25000$ ottenendo $n \approx 7447$.
Per avere una correzione di circa 2400 individui in accordo con il risultato per Roma con $\varrho \approx \frac{1}{120}$ ci vorrebbe $G \approx 13000$ che potrebbe essere ragionevole considerando quasi solo uomini(eliminando quindi 2 dal denominatore nel calcolo di G) otteneno n ≈ 8891.

1.3 Stima di G

Il numero G di persone che condividono lo stesso codice costituito da comune di nascita, data di nascita e sesso potrebbe essere ottenuto rivolgendosi all'anagrafe dei singoli comuni controllando il registro delle nascite del singolo giorno di un determinato anno. Supponendo che tali registri non siano consultabili o che l'operazione sia troppo laboriosa sarà necessaria una stima del parametro G.

Sia P la popolazione totale di un determinato comune e supponiamo che

le nascite siano uniformemente distribuite nell' anno e che sia equiprobabile nascere maschio o femmina.

In condizioni demograficamente stabili per le quali si dovrebbe avere

$P = (numero\ nati/anno) \cdot (durata\ media\ vita)$

avremo che:

$$G \approx \frac{P}{100 \cdot 2 \cdot 365} = \frac{P}{73000}$$

dove la durata media di 100 anni è un po' esagerata, ma tiene conto della recente diminuzione di natività. Approssimando G con l'intero più vicino avremo per le più grandi città italiane, considerando la popolazione secondo il censimento dell'inizio del 1998:

Numerazione comuni ISTAT	Comune	Popolazione	G
58091	*Roma*	2.653.245	36
15146	*Milano*	1.302.808	18
63049	*Napoli*	1.035.835	14
1272	*Torino*	914.818	13
82053	*Palermo*	688.369	9
10025	*Genova*	647.896	9
37006	*Bologna*	383.761	5
48017	*Firenze*	379.687	5
87015	*Catania*	342.275	5
72006	*Bari*	333.550	5

Tabella 1.2: Le più grandi città italiane con una stima di $G > 1$

1.4 Stima di ϱ

Il registro aggiornato al 31/12/00 è costituito da 37524 notifiche.

Consideriamo tra le notifiche solo le 13049 con codice identificativo paziente indicante le prime tre consonanti del nome e del cognome dell'individuo notificato e tali che siano correttamente indicati il comune di nascita, la data di nascita e il sesso.

Possiamo supporre con buona ragionevolezza che l'insieme costituito dal codice identificativo paziente, comune di nascita, data di nascita, sesso ci consenta di determinare univocamente un individuo. Considerando due notifiche uguali e quindi appartenenti allo stesso individuo se hanno il medesimo codice identificativo, sesso, comune e data di nascita avremo 9461 notifiche differenti tra le 13049 considerate ed in particolare:

Ripetizioni	Totale
1	7081
2	1750
3	328
4	170
5	59
6	38
7	16
8	9
9	7
10	1
11	1
13	1

Tabella 1.3: Delle 13049 notifiche considerate avremo 9461 notifiche differenti delle quali 7081 singole, 1750 doppie, 328 triple. . .

Avremo quindi, poichè il metodo fondamentale di stima di intensità è il quoziente eventi/tempo totale a rischio:

$$\lambda_N = \frac{numero\ di\ persone\ che\ si\ notificano\ per\ la\ prima\ volta}{numero\ persone\ a\ rischio\ di\ prima\ notifica\ per\ intervallo\ di\ tempo} \approx \frac{9461}{1,5 \cdot 10^6 \cdot t}$$

$$\lambda_H = \frac{numero\ di\ persone\ che\ si\ rinotificano}{numero\ persone\ già\ notificate\ per\ \frac{intervallo\ di\ tempo}{2}} \approx$$

$$\approx 2 \cdot \frac{1750 + 2 \cdot 328 + 3 \cdot 170 + ... + 13 \cdot 1}{9461 \cdot t} = \frac{7176}{9461 \cdot t}$$

quindi

$$\varrho = \frac{\lambda_N}{\lambda_H} \approx \frac{8316}{10^6} \approx \frac{1}{120}$$

N. B.

1) $1.5 \cdot 10^6$ indica gli individui effettivamente a rischio(età giusta, ecc. . .). L'intervallo di tempo è essenzialmente tre anni durante i quali le notifiche "complete" sono pervenute.

2) L'intervallo di tempo è stato diviso per due poichè ogni persona è mediamente presente metà intervallo.

È evidente che $E[_{X_k}]$ dipende dai parametri G e ϱ e i seguenti grafici ci mostrano l'andamento di tale valore atteso prendendo in considerazione i valori di G e di ϱ in accordo con le considerazioni fatte successivamente.

fig1

fig2

fig3

1.5 Analisi del registro

Tra le 37.524 notifiche del registro aggiornato al 31/12/00 consideriamo le 37.408 in cui siano correttamente indicati la data di nascita, il comune di nascita e il sesso e concentriamoci sulle notifiche di individui nati nelle città con $G > 1$ e per le quali in caso di segnalazioni plurime dello stesso codice non possiamo dire con certezza di quanti individui differenti si tratti. Delle 37.408 notifiche 15.880 sono costituite da codici differenti e per cui al 31/12/00 l'Osservatorio regionale indicherebbe in 15.880 i casi accertati di sieropositività registrati in Italia.

Procediamo allo studio dettagliato di tali notifiche per i singoli comuni con $G > 1$ in accordo con la tabella 2 e cerchiamo di trovare una stima del numero di notifiche dovute ad individui differenti in accordo con i risultati teorici fino ad ora sviluppati. Iniziamo con le notifiche di individui nati nel comune di Roma. Avremo 19.558 notifiche totali delle quali 6440 differenti ed utilizzando la (1) e la (2) avremo indicando con E_{Roma} il valore atteso del numero dei notificati differenti nati a Roma e V_{Roma} la sua varianza:

$$E_{Roma} = 8854 \qquad V_{Roma} = 1421$$

Con il sistema attuale nel momento in cui perviene una notifica con un codice già esistente viene considerata come un caso già accertato e quindi non registrata e quindi delle 19558 notifiche di individui nati a Roma l'Osservatorio registra 6440 casi differenti mentre la nostra stima ci da 8854 casi differenti con una varianza di 1421 unità.

Studiando per $\varrho = \frac{1}{120}$ il comportamento di Latina che ha circa 112.000 abitanti e $G = 2$ ed è la seconda città del Lazio per numero di abitanti si nota che non si hanno differenze significative tra la nostra stima del numero di individui differenti notificati e quella dell'Osservatorio.

Molte notifiche appartengono ad individui con comune di nascita fuori dal Lazio. La stessa teoria sviluppata fino ad ora vale anche per questi, ma il valore di ϱ deve essere ripensato, poichè

$$\varrho = \frac{\lambda_N}{\lambda_H} \; =$$

$$= \frac{\textit{Intensità di prima notifica nel registro del Lazio}}{\textit{Intensità di rinotifica per chi è stato notificato per la prima volta nel Lazio}}$$

Se un individuo nato fuori dal Lazio viene notificato per la prima volta, è possibile che vi risieda e dunque λ_H sarà come per gli abitanti del Lazio, mentre λ_N sarà molto inferiore poichè si riferisce a molti individui nati nello stesso comune che non risiedono nel Lazio. Prendiamo perciò $\varrho = \frac{1}{350}$ come unico valore per questi calcoli, la cui conclusione è che le correzioni sono minime rispetto alla strategia di contare solo i codici differenti.

Per le più grandi città italiane fuori dal Lazio infatti avremo per $\varrho = \frac{1}{350}$:

Città	G	Numero notifiche	Numero codici differenti	E
Milano	18	384	207	215
Napoli	14	502	253	261.86
Torino	13	111	58	59.72
Palermo	9	211	110	112.13
Genova	9	133	70	71.31
Bologna	5	65	38	38.28
Firenze	5	42	28	28.28
Catania	5	169	73	74
Bari	5	94	59	59.36

Tabella 1.4: Stima del valore atteso per le più grandi città italiane fuori dal Lazio

Per le restanti città con $G < 5$ si può notare che si hanno in media al massimo 10 segnalazioni di uno stesso codice e se consideriamo
G=4 abbiamo:
$\varrho = \frac{1}{350} - - > E_{10} = 1.07$
$\varrho = \frac{1}{120} - - > E_{10} = 1.20$
$\varrho = \frac{1}{40} - - > E_{10} = 1.53$

Quindi su 10 notifiche differenti il valore atteso di individui differenti notificati è abbastanza vicino ad uno e considerando il piccolo numero di notifiche che provengono da tali centri possiamo prendere per buona la stima dell'Osservatorio.

Possiamo dunque concludere, con riserva per il metodo abbastanza rozzo di stima di G e ϱ, che una ragionevole correzione al registro è una aggiunta di 2414 individui ai 15880 già calcolati al 31/12/00 con la strategia del numero di codici differenti e che la correzione derivi interamente dal comune di Roma.

Nel registro inoltre figura un numero abbastanza elevato di individui notificati nati fuori dall'Italia. La probabilità di avere più individui stranieri provenienti dallo stesso continente, nati lo stesso giorno dello stesso anno, dello stesso sesso, sieropositivi e notificati nel Lazio risulta intuitivamente molto bassa quindi in questi casi possiamo considerare buona la stima dell'Osservatorio.

Nella seguenti tabelle sono indicati il numero totale di persone differenti sieropositive stimate e l'incremento annuo con il metodo dell'Osservatorio e con la nostra correzione basandoci sul registro aggiornato al 31/12/00, costituito da 37524 notifiche delle quali 37459 con comune di nascita, data di nascita e sesso correttamente indicati. Sebbene le prime notifiche siano del 1985,essendo esiguo il numero delle stesse fino alla fine del 1988 (solamente 66), nei grafici rappresentanti l'analisi della prevalenza (numero di sieropositivi totali) e dell'incidenza (incremento annuo) i dati partono dal 1989, anno in cui si inizia ad avere un buon numero di notifiche (4231).

Anno	Numero notifiche totali	Osservatorio	E
2000	37459	15880	18294
1999	35490	15313	17595
1998	33902	14510	16702
1997	32355	13736	15833
1996	30348	12859	14831
1995	28085	11862	13695
1994	25294	10740	12388
1993	22555	9617	11093
1992	18981	8323	9841
1991	14126	6765	7677
1990	9195	5119	5654
1989	4231	2919	3106
1988	66	65	65
1987	31	31	31
1986	14	14	14
1985	5	5	5

Tabella 1.5: Prevalenza con metodo dell'Osservatorio e con la nostra correzione

Anno	Numero notifiche annue	Osservatorio	E
2000	1918	563	699
1999	1588	803	893
1998	1547	774	869
1997	2007	877	1002
1996	2290	990	1136
1995	2791	1122	1307
1994	3039	1121	1295
1993	3274	1294	1252
1992	4848	1558	2164
1991	4931	1646	2023
1990	4964	2200	2548
1989	4165	2854	3041
1988	35	34	34
1987	17	17	17
1986	9	9	9
1985	5	5	5

Tabella 1.6: Incidenza con metodo dell'Osservatorio e con la nostra correzione

fig1

fig2

1.6 Commenti ed estensioni

- Può essere interessante chiedersi quale sia stato l'incremento di individui differenti tra due tempi T_1 e T_2.
  Indichiamo con $\widehat{N}(T)$ la stima basata sul registro fino al tempo T, la stima dell'incremento sarà $\widehat{N}(T_2) - \widehat{N}(T1)$, ma la varianza associata di questo numero deve tenere conto della covarianza del nostro processo di stima. Tale covarianza può essere calcolata partendo dal processo fondamentale tratto nel paragrafo 1.

- Nel modello, G si riferisce al numero di persone nate lo stesso giorno, di stesso sesso, nello stesso comune.
  Potrebbe succedere che dopo, diciamo, venti anni, quando il fenomeno di notifica comincia a essere prolungato, che uno o più di questi G individui sia deceduto o emigrato.
  Tali dati potrebbero essere acquisiti dall'anagrafe o stimati con un modello generico.
  In mancanza di informazioni, non proseguiamo su questa via.

- Il codice usato è legato all' anno di nascita e la tendenza a essere notificati varia con l'età; λ_N in un gruppo potrebbe perciò cambiare con il tempo, così come potrebbe cambiare a causa di cambiamenti della tendenza a testarsi, per esempio dovuti a campagne informative o mancanza di tali campagne o altro. λ_N potrebbe invece cambiare con il tempo passato dalla prima notifica, sia per meccanismi naturali (più test iterati all'inizio o verso la fine del periodo d'incubazione) sia per eventuale decesso (precoce) dell'individuo.
  Un'approssimazione a questo tipo di cambiamenti temporali di ϱ può essere ottenuto ricalcolando ϱ ogni volta che arrivi una notifica in un certo G-gruppo, se si usa il metodo algoritmico ($p.8 - 9$).

- È possibile che l'impostazione del modello con intensità per le rinotifiche di un individuo non sia adatto (per esempio, non può descrivere una situazione dove un individuo, durante il primo anno dopo la prima notifica o viene rinotificato zero volte o cinque volte). Bisognerebbe osservare dei pazienti e il loro vero flusso di notifiche per verificare l'aumento di intensità.

- Nel presente lavoro, G è stato stimato grossolanamente per vari comu-

ni, in mancanza dei dati anagrafici esatti.
In teoria, sarebbe interessante incorporare l'incertezza su G nella stima di varianza dello stimatore.
Indichiamo con $\widehat{N}(G)$ una stima ottenuta in un gruppo assumendo il valore G; se G fosse una stima del vero valore γ, diciamo, sarebbe naturale (in stile Bayesiano) ritenere, per esempio, che γ abbia distribuzione di Poisson con media G e ciò porterebbe alla stima pesata $E(\widehat{N}(G) \mid G)$ e al relativo calcolo della varianza approssimata.
Lo stesso si potrebbe dire circa l'incertezza nella stima di ϱ.

Bibliografia

[1] *Sorveglianza delle infezioni da HIV*, documento per la Commissione Nazionale Aids a cura del Centro Operativo AIDS (2000)

[2] Angela Carboni, Daniela D'Ippoliti, Rosa Maria Peano, Carlo A.Perucci, Daniela Porta, Elisabetta Rapiti *Infezioni Hiv ed Aids*, Osservatorio Epidemiologico Regione Lazio (1998).

[3] Marco Bramanti *Calcolo delle probabilità e Statistica*.

[4] Paolo Baldi *Calcolo delle probabilità e Statistica*.

[5] Arnaldo Frigessi *Calcolo delle probabilità*.

[6] *Movimento anagrafico e popolazione residente secondo il sesso, per comune.*, pubblicazione Istat (1998).