Uma metodologia para estimar a incidência a partir de prevalências obtidas em *Pesquisas Nacionais por Amostra de Domicílios*: o exemplo da alfabetização

A methodology to estimate incidence from prevalence rates in household surveys: the example of literacy

Greice Maria Silva da Conceição [1]
Maria Tereza Serrano Barbosa [2]

[1] Centro de Estudos de Segurança e Cidadania, Universidade Candido Mendes, Rio de Janeiro, Brasil.
[2] Departamento de Matemática e Estatística, Universidade Federal do Estado do Rio de Janeiro, Rio de Janeiro, Brasil.

Correspondência
G. M. S. Conceição
Centro de Estudos de Segurança e Cidadania, Universidade Candido Mendes.
Rua da Assembléia 10, sala 810, Rio de Janeiro, RJ 20011-901, Brasil.
grconceicao@superig.com.br

Abstract

Age and time-specific incidence rates are highly relevant to epidemiologists. Due to the mathematical relationship between incidence and prevalence, several methods have been developed to obtain incidence based on prevalence. The present study didactically describes a method to estimate age and time incidence, based on: repeated measurements of prevalence for an event; events related to the individual's age; and events with the characteristics of a chronic disease. The method was used to analyze literacy in the Brazilian population, using prevalence by age obtained from the National Household Sample Survey (PNAD) for 1981, 1985, 1989, 1992, 1995, and 2001, and allowed obtaining literacy incidence estimates for ages 5 to 20 years, for the years 1960 through 2001. The method adjusts incidence parameters for age and time with the help of a generalized linear model for categorical data and uses the EM algorithm to obtain non-observed data.

Algorithms; Incidence; Prevalence

Introdução

A avaliação do comportamento de eventos de saúde nas populações tanto pode ser feita a partir dos casos prevalentes, como pode, a cada momento, aferir o número de novos casos do evento – casos incidentes. Devido às características destas duas medidas de freqüência epidemiológica, a prevalência é obtida tendo como base uma amostra transversal e, para que se consiga a incidência, é necessário o acompanhamento, por um longo período de tempo, de uma coorte de indivíduos sem o evento de interesse. Ao estarem essas duas medidas relacionadas matematicamente, alguns métodos foram desenvolvidos para que a partir de uma delas se pudesse chegar à outra. Poder contar com esta alternativa é de extrema importância para a epidemiologia, uma vez que na maioria dos estudos apenas uma destas duas informações encontra-se disponível.

Nesta direção, Keiding [1] utilizou análise de sobrevivência não-paramétrica para estimar incidência e prevalência utilizando dados obtidos de amostras transversais, nos quais existiam apenas informações sobre a idade atual e o *status* da doença, ou ainda informações retrospectivas sobre a idade do início da doença.

Brunet & Struchiner [2] apresentaram um método de reconstrução da taxa de incidência para qualquer idade e ponto no tempo a partir de dados de prevalência disponíveis. O método relaciona, matematicamente, taxas de incidência

e prevalência que podem ser estimadas a partir de amostras, o que não requer conhecimento da distribuição demográfica das subpopulações de interesse.

Já a metodologia apresentada por Satten et al. [3] examina a validade de um estimador de incidência calculado para uma população acompanhada repetidas vezes no tempo, permitindo a comparação entre a incidência calculada por meio de dados transversais com as calculadas utilizando dados longitudinais, levando em consideração que um mesmo indivíduo pode estar mais de uma vez nas amostras utilizadas.

Giampaoli et al. [4] apresentaram uma metodologia matemática para obter estimativas e projeções de incidência e de prevalência usando uma função polinomial de idade e coorte de nascimento e um conjunto de funções cúbicas do ano do diagnóstico. A metodologia emprega como fonte de dados os registros oficiais de população, mortalidade e sobrevivência.

O método em que se apóia este artigo foi proposto por Marschner [5] e estima a taxa de incidência por idade e tempo usando prevalências obtidas de amostras transversais repetidas. A partir de um modelo multiplicativo discretizado no tempo, ele emprega o algoritmo EM para obter os dados não observados.

No Brasil, um bom exemplo para a aplicação desta metodologia é a *Pesquisa Nacional por Amostra de Domicílios* (PNAD), que é repetida anualmente pelo Instituto Brasileiro de Geografia e Estatística (IBGE, http://www.ibge.gov.br). Em 1998, por exemplo, apresentou o suplemento *Acesso e Utilização de Serviços de Saúde* que, entre muitas outras questões ligadas ao setor de saúde, investigou 12 doenças crônicas, diagnosticadas ou percebidas pela população inquirida. Se o tema deste suplemento fosse objeto de realização outras vezes, ele poderia ser utilizado para o cálculo das estimativas da incidência das doenças investigadas a partir das suas prevalências capturadas na população. Isto seria de grande utilidade para subsidiar as ações na área de Saúde Pública, uma vez que estas duas medidas de freqüência epidemiológica se mostram essenciais para o planejamento e a intervenção em serviços de saúde. Assim, o método será descrito de forma tutorial para disseminá-lo entre os pesquisadores da Saúde Pública.

Para a sua aplicação foi necessária a escolha de um evento investigado repetidas vezes e que pudesse estar relacionado à idade do indivíduo e à influência a longo prazo de políticas de planejamento, que tivesse as características de uma doença crônica e não ocorresse de forma homogênea no tempo. Assim, ele teve aqui o objetivo de estimar a incidência da alfabetização na população brasileira por idade e tempo específicos, a partir das prevalências de alfabetizados em cada idade obtidas em seis PNADs.

Material e métodos

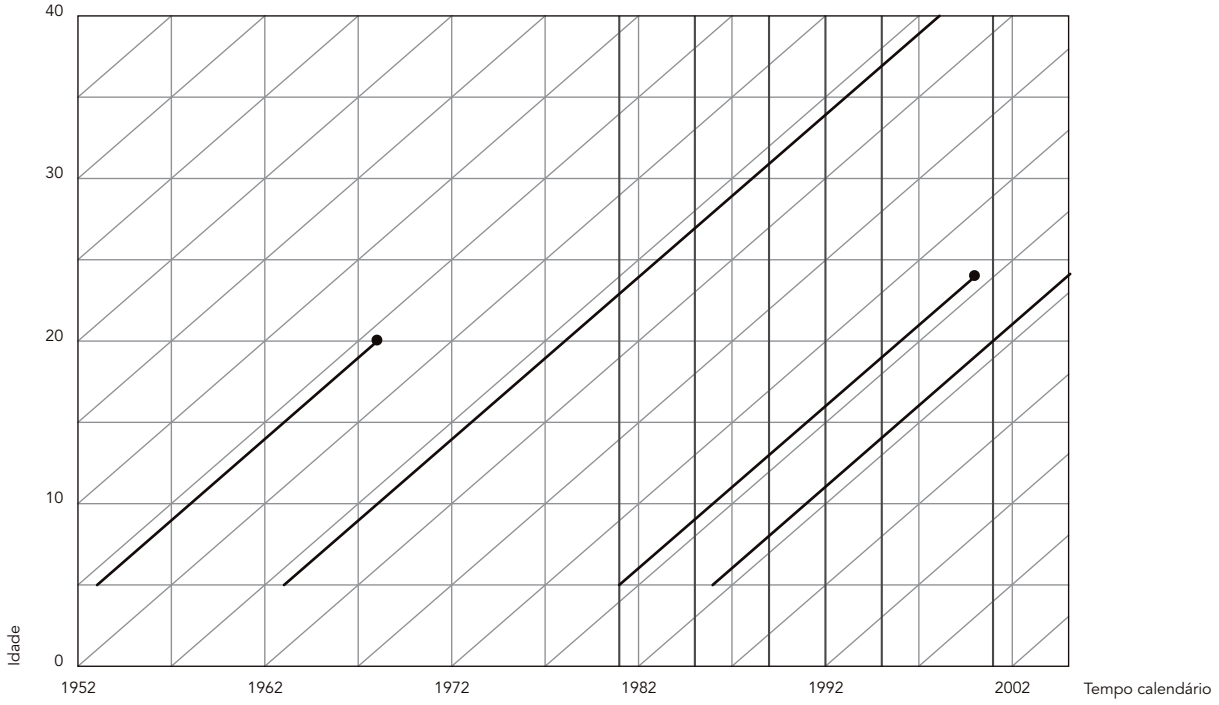
A população-alvo foi constituída por aqueles que pertenciam às coortes de nascimento de 1933 até 1996, ou seja, tinham entre 5 e 68 anos em 2001. A metodologia que será aqui descrita e aplicada teve como objetivo estimar para esta população a incidência de alfabetização em cada idade e em cada ano, tendo como ponto de partida as prevalências obtidas nas amostras das PNADs. As estimativas das prevalências necessárias foram obtidas nas PNADs dos anos de 1981, 1985, 1989, 1992, 1995 e 2001. A base de dados utilizada foi composta pelo ano da pesquisa e pelo número de alfabetizados e não-alfabetizados nas idades entre 5 e 48 anos. Esta faixa compreende os que tinham até 48 anos em 1981, ou seja, os indivíduos que nasceram a partir de 1933. A situação relativa ao evento em questão (alfabetizado ou não-alfabetizado) dos indivíduos pertencentes às coortes foi observada em suas respectivas idades no momento da realização de cada pesquisa; por exemplo, os indivíduos da coorte de nascimento de 1948 tiveram sua situação de alfabetização observada em 1981, 1985, 1989, 1992, 1995 e 2001, quando estavam com 33, 37, 41, 44, 47 e 53 anos, respectivamente. O diagrama de Lexis, apresentado na Figura 1, ilustra possíveis dinâmicas para quatro coortes. Nestas últimas o ponto cheio indica a ocorrência do evento.

O ponto cheio no final dos segmentos de reta indica a ocorrência do evento, e as linhas verticais, as amostras transversais. Assim, o indivíduo da coorte de 1953 que foi alfabetizado em 1973 será captado na amostra de 1981 e contribuirá na prevalência dos que estão alfabetizados aos 28 anos, enquanto um indivíduo da coorte de 1963 que não foi alfabetizado contribuirá para a prevalência dos não-alfabetizados aos 18 anos em 1981 e dos não-alfabetizados aos 38 anos em 2001.

A Figura 2 complementa a dinâmica representada pelo diagrama de Lexis e permite visualizar a idade dos indivíduos de cada coorte nos momentos das PNADs, apresentando também as notações utilizadas para as grandezas observadas ou estimadas em cada tempo ou em cada idade.

Figura 1

Diagrama de Lexis: um exemplo de acompanhamento para quatro coortes.



Casos prevalentes

Ao seguirmos a metodologia de Marschner [5] e a notação proposta por ele, o total de indivíduos na idade *a* em cada amostra *j* será denominada de N_{aj}, e o número de indivíduos de idade *a* alfabetizados antes do tempo *j* de cada amostra será denominado de Z_{aj}. A razão $\frac{Z_{aj}}{N_{aj}}$ estima as prevalências da alfabetização na idade *a* no tempo *j*, tendo como suposição que a idade em que o indivíduo foi alfabetizado não pode ser observada. Ao supor que a probabilidade de estar alfabetizado é a mesma para todos os indivíduos de idade *a* selecionados na amostra da PNAD, o método assume que Z_{aj}, o número de pessoas já alfabetizadas em cada PNAD, tem distribuição binomial com parâmetros (N_{aj}, P_{aj}).

Casos incidentes

Ao considerar a ausência de informação sobre o momento exato da alfabetização como um dado não observado, a metodologia trata o problema através da ótica de dados incompletos. Assim, $X_{at}^{(j)}$ são os casos incidentes de alfabetização na idade *a* no tempo *t* selecionados no tempo *j* de cada PNAD e Y_{at}, o total de indivíduos em risco de serem alfabetizados na idade *a* no tempo *t*. Os indivíduos (não observados) cuja alfabetização ocorreu na idade *a* e no ano *j* de cada PNAD serão denotados por:

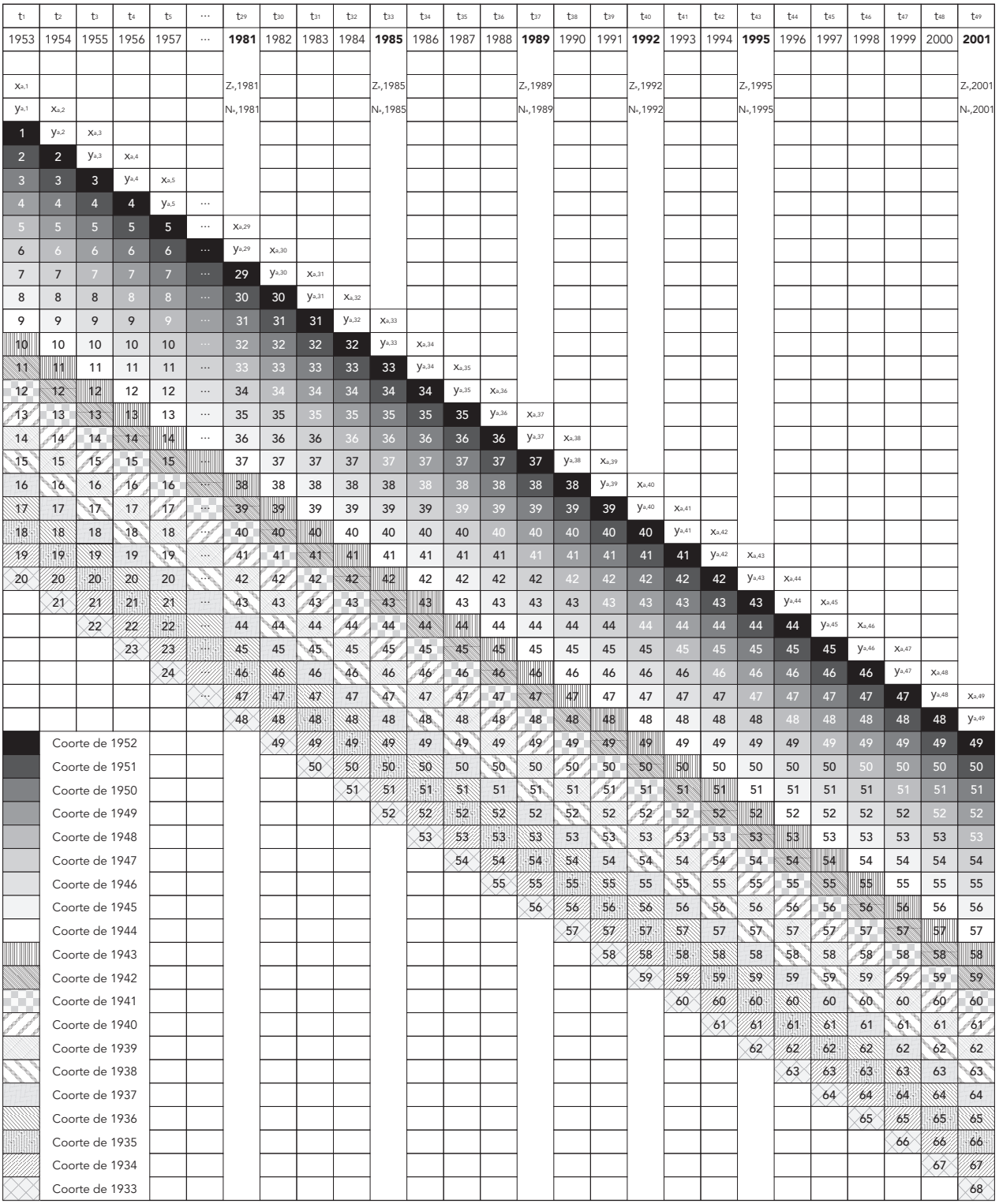
$$X_{at} = \sum_{j} X_{at}^{(j)}$$ (Marschner [5], p. 494),

em que $X_{at}^{(j)}$ também tem distribuição binomial com parâmetros (Y_{at}, P_{at}). Assim, se uma pessoa não tiver sido alfabetizada na idade *a* - 1, a probabilidade de ser alfabetizada na idade *a* será:
$P_{at} = 1 - \exp(\beta_a \lambda_t)$ (Marschner [5], p. 493), em que β_a e λ_t correspondem às taxas de incidência de alfabetização para cada idade *a* e cada tempo *t*.

Já o total Y_{at} corresponderá à diferença entre o total da população estimada na idade *a* no tempo *t* e o número de casos incidentes de alfabetização das idades anteriores à idade *a* no mesmo tempo, sendo obtido por:
$Y_{at} = \sum_{j \in \Gamma_{at}} N_{a+\Gamma_j - t, j} - \sum_{j \in \Gamma_{at}} \sum_{i=1}^{a-1} X_{i,t-a+i}^{(j)}$ (Marschner [5], p. 493).

Figura 2

Coortes acompanhadas, padronização dos tempos, dados estimados e tipo de informação em cada *Pesquisa Nacional por Amostra de Domicílios* (PNAD).



Nota: lê-se: $x_{at} = \hat{x}_{at}$ e $y_{at} = \hat{y}_{at}$

As estimativas dos dados não observados (X_{at} e Y_{at}) serão alcançadas com o auxílio do algoritmo EM [6].

Aqui, o passo E do algoritmo estima $\hat{x}_{at}$ e $\hat{y}_{at}$ para toda idade *a* e tempo *t* e, em seguida, o passo M recebe os dados agora completos $\{X_{at} = \hat{x}_{at}, Y_{at} = \hat{y}_{at}\}$ estimados no passo E. A partir desses dados, ajusta-se um modelo linear generalizado (MLG) com ligação log-log, obtendo-se as novas estimativas dos parâmetros de incidência de alfabetização $\hat{\lambda}^{novo}$ e $\hat{\beta}^{novo}$. O processo continua iterativamente até que a convergência dos parâmetros seja atingida, substituindo em cada iteração os $\hat{\lambda}_t^{antigo} = \hat{\lambda}_t^{novo}$ e $\hat{\beta}_a^{antigo} = \hat{\beta}_a^{novo}$ para cada idade *a* e tempo *t*.

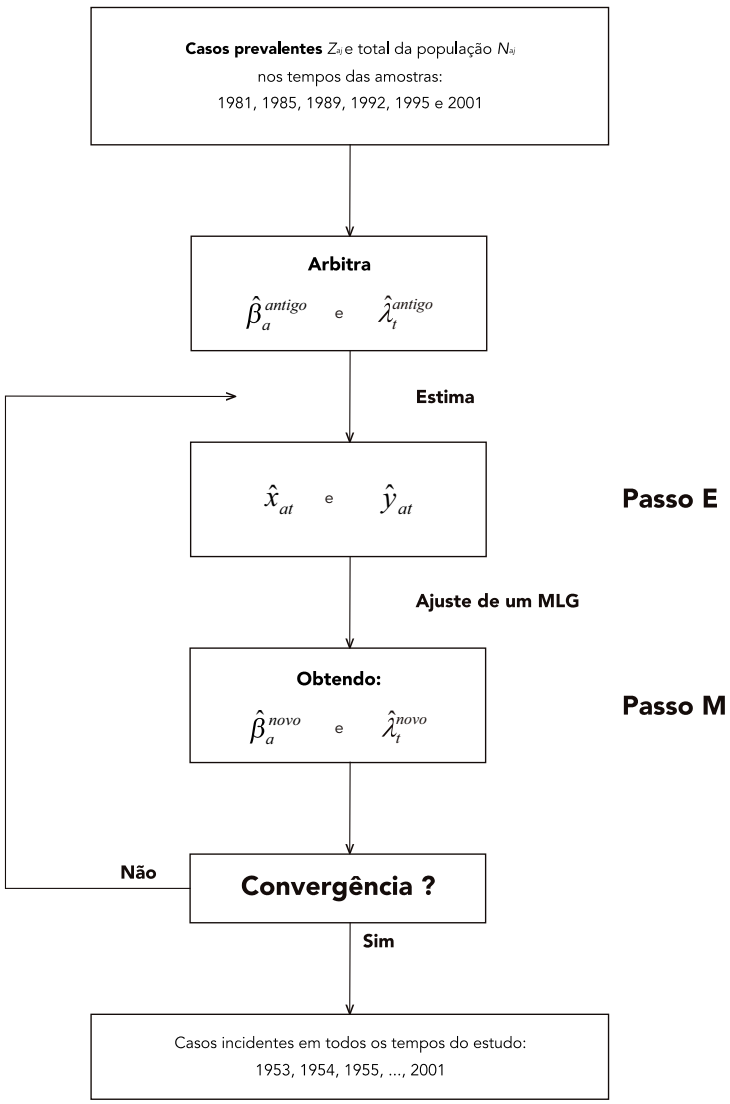
Etapas do processo de estimação

A 1ª etapa consiste na criação do banco de dados com o total de indivíduos em cada amostra *j* e em cada idade *a* (N_{aj}) e o número indivíduos na idade *a* já alfabetizados no tempo *j* de cada amostra (Z_{aj}). Na 2ª etapa são escolhidos os parâmetros iniciais de incidência de alfabetização em cada idade (β_a^{antigo}) e em cada ano (λ_t^{antigo}). Na 3ª etapa é estimado o número de alfabetizados ($\hat{x}_{at}$) e o número de pessoas sob risco de serem alfabetizadas ($\hat{y}_{at}$) – passo E do algoritmo EM, segundo estimadores propostos por Marschner [5]. Na 4ª etapa, com os dados obtidos na 3ª etapa, ajusta-se um MLG para estimar os parâmetros de incidência em cada idade ($\hat{\beta}_a^{novo}$) e em cada ano ($\hat{\lambda}_t^{novo}$). Na 5ª etapa as 3ª, 4ª e 5ª etapas repetem-se até que a convergência seja alcançada. Como critério para a convergência utilizou-se a razão absoluta da diferença da estimativa do parâmetro antigo com o novo, em relação ao parâmetro antigo, até que esta razão apresentasse valor menor ou igual a 0,0001 para todos os parâmetros estimados.

A Figura 3 apresenta o fluxograma com as etapas do processo de estimação que foram executadas por meio do pacote computacional R (The R Foundation for Statistical Computing, Viena, Áustria; http://www.r-project.org) que, além de ser gratuito e possuir uma linguagem de programação relativamente fácil, permitiu a integração das etapas necessárias ao algoritmo EM com aquelas que são fundamentais ao ajuste do MLG a cada iteração do algoritmo.

Figura 3

Etapas do processo de estimação.

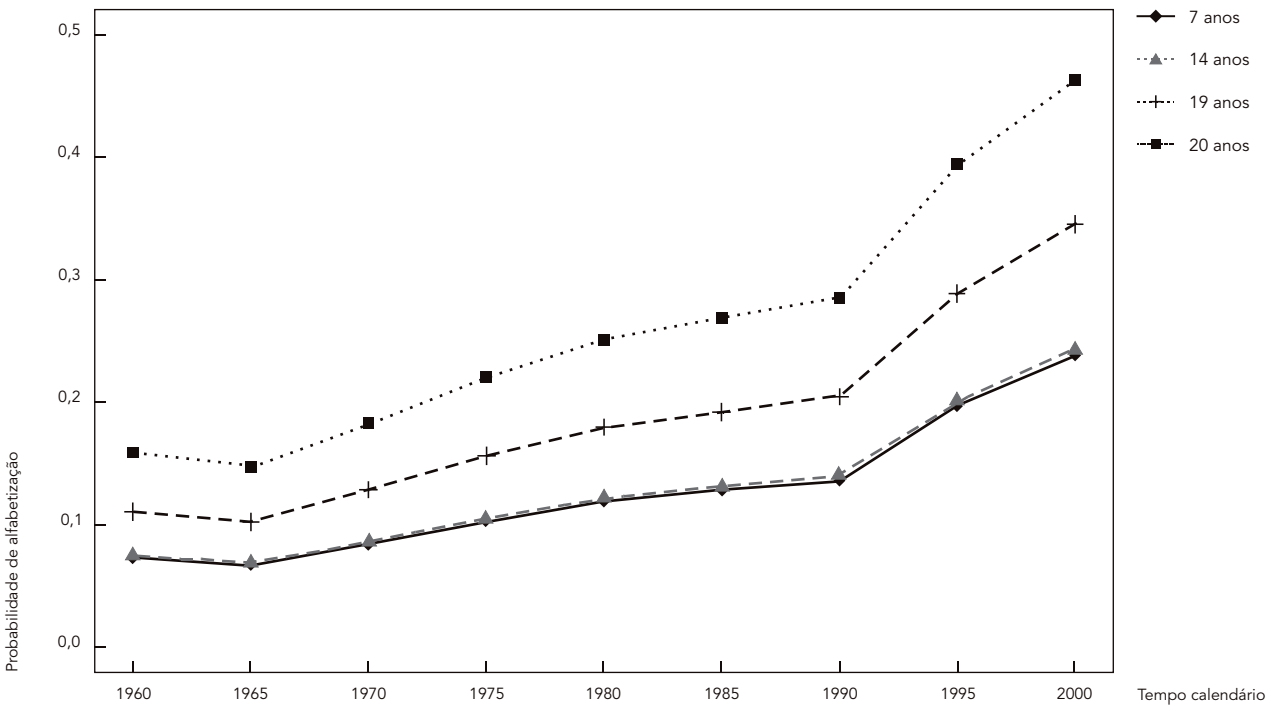


Resultados

A Figura 4 mostra as probabilidades estimadas por idade e tempo para os indivíduos que ainda não estavam alfabetizados aos 7, 14, 19 e 20 anos. Percebe-se que as probabilidades de alfabetização ao longo do tempo, condicionais ao fato de uma pessoa não ter sido alfabetizada até o ano anterior, são sempre crescentes, mas que na fase regular de escolarização (7 a 14 anos) quase não existe diferença entre elas. Ao observar as probabilidades estimadas nas fases posteriores,

Figura 4

Probabilidade de alfabetização por ano segundo idades, 1960/2000.



percebe-se que, mesmo entre idades vizinhas, estas estimativas por idade se afastam. Em 1960, por exemplo, o modelo estimou em 0,16 a probabilidade de alguém ser alfabetizado aos 20 anos, já que não o tinha sido até os 19 anos. Estas probabilidades estimadas se alteram para 0,18; 0,25; 0,29 e 0,46 se consideramos os anos de 1970, 1980, 1990 e 2000, enquanto para os que tinham 19 anos, a probabilidade seria 0,11 em 1960 e, nas demais décadas, 0,13; 0,18; 0,20 e 0,35, respectivamente. Estas diferenças podem ser causadas por mudanças na metodologia de classificação do evento em cada PNAD, mas é possível também que reflitam resultados de políticas públicas aplicadas nessas décadas.

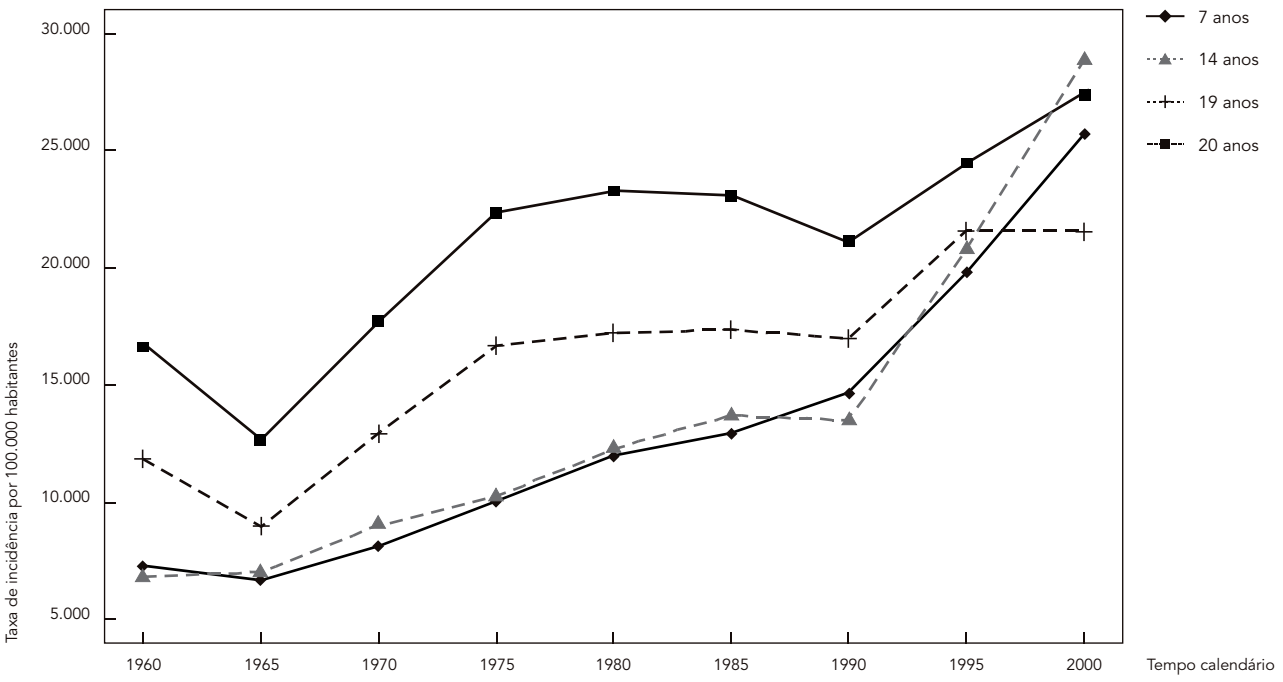
A Figura 5 apresenta as estimativas das taxas de incidência de alfabetização por 100 mil habitantes para as mesmas idades anteriores, ilustrando melhor como a metodologia pode ressaltar particularidades na trajetória de algum evento no tempo em determinadas faixas etárias. Ressalta-se que a instabilidade observada entre os anos de 1960 e 1975 pode ser atribuída também ao fato de a primeira prevalência observada ter sido a do ano de 1981. Ressalta-se que essas diferenças obtidas por esta metodologia desaparecem quando se substituem as dinâmicas de cada coorte, pela informação obtida em apenas um momento do tempo.

Conclusão

As prevalências, apesar de serem de grande importância, não permitem que as mudanças no tempo sejam vistas, ao contrário da incidência, que pode descrever possíveis particularidades específicas tanto no tempo como na idade. A descrição da metodologia ilustra como uma amostra realizada em um momento é passível de ser tratada como um conjunto de subamostras de várias coortes de nascimento, aumentando o seu uso potencial para quem quer planejar, intervir e alterar o comportamento da ocorrência de eventos de saúde. No entanto, os resultados encontrados, assim como os obtidos por qualquer modelo estatístico, devem sempre ser analisados à luz da sua plausibilidade.

Figura 5

Estimativas das taxas de incidência de alfabetização por 100 mil habitantes, 1960/2000.



Por se tratar de modelos que fazem suposições teóricas a respeito das incertezas envolvidas, é necessário sempre buscar melhorar seus resultados, seja através da comparação com outros métodos, tais como os propostos por Brunet & Struchiner [2] ou Keiding [1], ou seja, incluindo métodos de suavização das estimativas dos parâmetros, ou a correção do desenho amostral para estimar a sua variabilidade, que é influenciada por pesos amostrais distintos, segundo Pessoa & Silva [7]. Ao descrever e ao aplicar um dos inúmeros métodos que visam recuperar incidências a partir de prevalência, vislumbra-se que sua divulgação permitirá ampliar as aplicações, tanto na área da saúde pública como em outras áreas do conhecimento, das pesquisas domiciliares realizadas pelo IBGE.

Resumo

Conhecer a incidência de algum evento por idade e tempo é sem dúvida um dos maiores interesses da epidemiologia. Devido à relação matemática que existe entre as medidas de incidência e prevalência, alguns métodos foram desenvolvidos com o objetivo de obter a incidência a partir da prevalência. Este estudo descreve, de forma tutorial, um método que fornece estimativas de incidências por idade e tempo desde que existam prevalências repetidas de um evento, que a ocorrência do mesmo esteja relacionada à idade do indivíduo e que este possua características de uma doença crônica. O estudo foi realizado para a alfabetização da população brasileira, utilizando as prevalências por idade obtidas na Pesquisa Nacional por Amostra de Domicílios *(PNAD) 1981, 1985, 1989, 1992, 1995 e 2001, e permitiu recuperar as estimativas da incidência de alfabetização nas idades entre 5 e 20 anos entre os anos de 1960 a 2000. O método ajusta os parâmetros de incidência para idade e tempo com o auxílio de um modelo linear generalizado para dados categóricos e utiliza o algoritmo EM para obter os dados não observados.*

Algoritmo; Incidência; Prevalência

Colaboradores

G. M. S. Conceição participou da revisão de literatura, elaboração da metodologia, análise dos resultados e redação do artigo final. M. T. S. Barbosa contribuiu na elaboração da metodologia, análise dos resultados e redação do artigo final.

Referências

1. Keiding N. Age-specific and prevalence: a statistical perspective. J R Stat Soc Ser A Stat Soc 1991; 145:371-412.
2. Brunet CR, Struchiner CJ. Rate estimation from prevalence information on a simple epidemiological model for health interventions. Theor Popul Biol 1996; 50:209-26.
3. Satten AG, Janssen RS, Busch MP, Datta S. Validating marker based incidence estimates in repeatedly screened populations. Biometrics 1999; 55: 1224-7.
4. Giampaoli S, Palmieri L, Capocacia R, Pilotto L, Vanuzzo D. Estimanting population-based incidence and prevalence of major coronary events. Int J Epidemiol 2001; 30 Suppl 1:S5-10.
5. Marschner IC. Fitting a multiplicative incidence model to age-and time-specific prevalence data. Biometrics 1996; 52:492-9.
6. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the algorithm (with discussion). Journal of the Royal Statistical Society. Series B (Statistical Methodology) 1977; 39:1-38.
7. Pessoa DGC, Silva PLN. Análise de dados amostrais complexos. São Paulo: Associação Brasileira de Estatística; 1998.

Recebido em 29/Mar/2007
Versão final reapresentada em 21/Set/2007
Aprovado em 25/Out/2007