ALGORITMO "EM" E FAMÍLIA EXPONENCIAL GENERALIZADA: UMA APLICAÇÃO NO EQUILÍBRIO DE HARDY-WEINBERG

Allan Robert da Silva[1], Paulo Sérgio Lucio[2]
Programa de Pós-Graduação em Matemática Aplicada e Estatística - PPGMAE
Universidade Federal do Rio Grande do Norte - UFRN
[1]all_robert02@yahoo.com.br
[2]pslucio@ccet.ufrn.br

RESUMO

O presente trabalho tem o intuito de mostrar uma alternativa de inferir, sobre a verdadeira proporção de freqüências alélicas em uma determinada população mendielana, a partir de uma amostra com dados incompletos. Nesse intuito, será apresentado o algoritmo EM, bastante usado na literatura. Uma forma particular deste algoritmo é obtida quando a distribuição de referência pode ser expressa como pertencente à família exponencial generalizada, caso da população mendielana posteriormente simulada. Após simular a população, será coletada uma amostra, no qual será obtida uma estimativa da verdadeira proporção de freqüências alélicas, em seguida, será empregado um método de reamostragem Bootstrap, a fim de medir a variabilidade das estimativas da amostra e, a partir deste, foram criados intervalos de confiança para o parâmetro de interesse. Os resultados foram bastante significativos, pois os parâmetros da população simulada estiveram sempre contidos nos intervalos obtidos. Todas as simulações foram criadas no software estatístico R.

Palavras-Chave: Simulação, Reamostram, Bootstrap, Intervalo de Confiança.

1- Introdução

O algoritmo EM (Dempster, Laird, e Rubin 1977; McLachlan e Krishnan, 1997) é uma poderosa ferramenta computacional para maximizar a estimativa da verossimilhança com dados incompletos. Chamamos de dados incompletos: falta de dados, elementos desconhecidos, variáveis latentes, tipos de censura de observações, e assim por diante. Uma boa introdução deste algoritmo é apresentada por Flury e Zopper (2000).

Bickel e Doksum(1988) citam uma aplicação buscando estimar a verdadeira proporção de freqüências alélicas em uma população mendielana. O objetivo do nosso trabalho é implementar e discutir esta aplicação usando o software estatístico R. Antes disso, para melhor entendimento desta aplicação, vejamos alguns conceitos básico de biologia.

O *genótipo* de uma pessoa é a sua constituição genética. O *fenótipo* é a expressão observável de um *genótipo* como um caráter morfológico, bioquímico ou molecular. Por exemplo, o *gene* que determina a cor da flor em várias espécies de plantas. Existe um único *gene* controla a cor das pétalas, podendo haver diferentes versões desse mesmo gene. Uma dessas versões pode resultar em pétalas vermelhas, enquanto outra versão originará pétalas brancas.

Alguns organismos são *diplóides* - isto é, têm pares de *cromossomos homólogos*, ou seja, possui a mesma origem embriológica, nas suas *células somáticas*, dominação dada a células do corpo que formem tecidos ou órgãos do corpo, tal como a célula da pele. Assim, existem duas cópias do mesmo *gene* que ocupam um lugar definido no cromossomo. Esse lugar definido é denominado *locus gênico*. Os genes que ocupam o mesmo *locus* em *cromossomos homólogos* são denominados *genes alelos*.

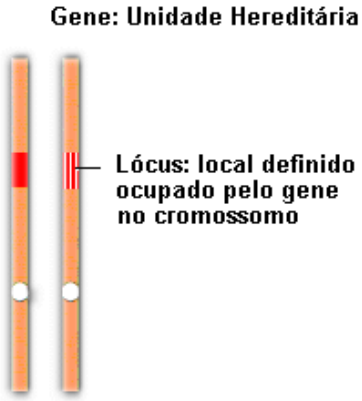

Figura 1(a): Lócus gênico.

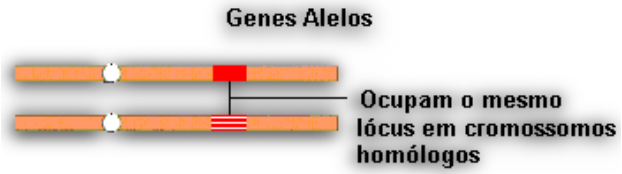

Figura 1(b): Genes alelos.

Os *genes alelos* não são necessariamente idênticos. Quando nas células de um indivíduo os *genes alelos* para um determinado caráter não são idênticos, o indivíduo é denominado *heterozigoto* para o caráter denominado pelo par de genes. Quando os genes alelos são idênticos, o indivíduo é denominado *homozigoto* para aquele caráter.

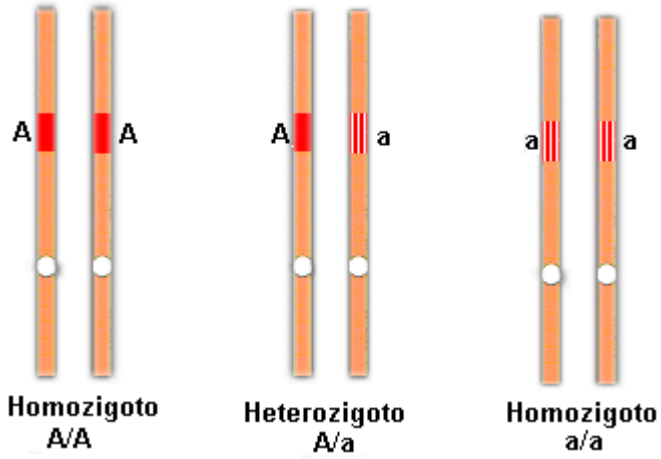

Figura 2: Homozigotos e Heterozigotos.

Como é o comportamento das freqüências alélicas ao longo do tempo?

O equilíbrio de Hardy-Weinberg (também princípio de Hardy-Weinberg, ou lei de Hardy-Weinberg) é a base da genética de populações e afirma que, em uma população mendeliana, dentro de determinadas condições, as freqüências alélicas, raras ou não, permanecerão constantes ao passar das gerações. A demonstração matemática foi realizada independentemente por Godfrey Harold Hardy na Inglaterra e por Wilhelm Weinberg, na Alemanha, em 1908.

A Figura 2, ilustra o caso mais simples de um único locus com dois alelos **A** e **a** com freqüências alélicas de p e q, respectivamente, o princípio H-W prediz que a freqüência genotípica para o homozigoto **AA** será p^2, para o heterozigoto **Aa** será $2pq$ e os outros homozigotos **aa** será de q^2.

A seguir abordaremos a teoria sobre o algoritmo EM e bootstrap e, por fim, serão comentados os resultados das simulações obtidas no software estatístico R.

2- Revisão teórica.

2.1- Descrição do algoritmo EM

Seja X^0 um conjunto de dados observados e X^* um conjunto de dados desconhecidos ou incompletos, e θ um parâmetro de interesse pertencente ao espaço paramétrico. Inicia-se esse algoritmo a partir de θ^0, ou seja, uma estimativa inicial do parâmetro. O algoritmo EM repete as duas etapas seguintes até à convergência.

Passo E (Expectation): calcula-se a esperança da log-verossimilhança com relação ao vetor de variáveis latentes.

$$Q(\theta \mid \theta_k) = E_{X^0|X^*}[\ln L(\theta, X^0 \mid X^*)]$$

Passo M (Maximization): Encontrar um argumento que maximize a esperança encontrada no passo E.

$$\theta_{k+1} = \arg\max Q(\theta \mid \theta_k)$$

Repete-se esse processo iterativo até a uma convergência definida, com isso, se obtém uma seqüência $\{\theta^0, \theta^1, \ldots\}$ que converge para um máximo local dos dados da verossimilhança observada, caso este exista, sob condições bastante gerais (mais detalhes ver Wu, 1983).

2.2- Algoritmo EM e Família Exponencial Generalizada.

Seja $\{P_\theta = \theta \in \Theta\}$ pertence à família exponencial generalizada, ou seja, pode ser escrito da forma: $P(x,\theta) = Exp\{\eta(T(X) - A(\eta)\}h(x)$. Denotaremos por S(X) uma função de X que distingue os valores latentes e os observados e T(X) uma estatística, então:

(a) O algoritmo EM consiste da iteração:

$$A(\theta_{k+1}) = E_{\theta_k}(T(X) \mid S(X) = s)$$

$$\theta_k = \theta_{k+1}$$

Se a solução existe esta é necessariamente única.

(b) Se a seqüência de iterações obtidas é limitada e a equação

$$A(\theta) = E_\theta(T(X) \mid S(X) = s)$$

Tem uma única solução, então há uma convergência para um limite $\hat{\theta}^*$, que é necessariamente é um máximo local de $q(s,t)$.

2.3- Técnica de reamostragem Bootstrap

O objetivo da inferência estatística é descobrir o que é possível ser aprendido sobre uma população, a partir de dados observados em uma amostra aleatória. Neste trabalho será utilizado o método Bootstrap para inferir com que precisão um parâmetro estatístico calculado, a partir dos valores amostrados, estima o correspondente parâmetro populacional. Nesta técnica um dos aspectos importantes é a utilização da amostragem com reposição, garantindo, com isso, a independência estocástica. A série gerada no processo de reamostragem, obtida através de um procedimento de simulação, apresenta características com diferenças, embora pequenas, a cada realização. Realizando duas gerações diferentes, obtêm-se, obviamente, séries diferentes. Portanto, qualquer parâmetro estatístico calculado com base na série gerada obterá incertezas. A avaliação dessas incertezas é o que pretendemos fazer empregando a técnica Bootstrap.

O Bootstrap é um procedimento robusto de simulação estatística para atribuir medidas de precisão a estimativas de parâmetros estatísticos. O atrativo maior deste método é a capacidade de responder problemas estatísticos reais sem o uso de fórmulas complexas. Dois dos objetivos principais do Bootstrap são: estimar o erro-padrão da referida estimativa e criar um intervalo de confiança para o parâmetro.

Esta técnica foi introduzida por Efron(1979), como abordagem ao cálculo do erro-padrão e de intervalos de confiança de parâmetros, no caso em que o número de amostras é reduzido. A idéia é simular uma réplica da experiência, uma leitura mais completa sobre o assunto pode ser vista em EFRON e TIBSHIRANI (1993).

Neste trabalho propomos utilizar a técnica Bootstrap mostrando a sua aplicação à avaliação de incertezas estatísticas de amostragem na análise de freqüências alélicas ao longo do tempo em função do número de simulações.

2.4- Intervalos de Confiança com o Bootstrap:

Uma estimativa dos valores entre os quais pode variar um determinado parâmetro populacional θ é chamado de intervalo de confiança para θ. Uma estimativa de intervalo é freqüentemente mais útil do que apenas uma estimativa pontual. Para criar um intervalo de (1–2α)100% de confiança, onde α é o nível de significância do teste, deve-se seguir os passos:

1- Retirar uma amostra da população de interesse a fim de estimar um parâmetro θ.
2- Fazer um grande número de reamostragens bootstrap, com reposição, de tamanho fixo. Por exemplo, 1000 reamostragens bootstraps de tamanho 200.
3- Em cada bootstrap calcular uma estimativa da estatística de interesse $\hat{\theta}^*$.
4- Ordenar as estatísticas do menor para o maior valor estimado em cada bootstrap.
5- E, finalmente, obter $\left(\hat{\theta}^*_{(q_1)}, \hat{\theta}^*_{(q_2)}\right)$, em que:
   $q_1 = \text{Parte\_inteira}(R\alpha/2)$ e $q_2 = R - q_1 + 1$.

3- Aplicação

Suponha que em uma população do tipo **Hardy-Weinberg** exista apenas um único lócus com alelos **A** e **a**, considerando, $X_i = (\in_{i1}, \in_{i2}, \in_{i3})$, representando três diferentes comportamentos. Consideremos $P_\theta = (X = (1,0,0)) = \theta^2$, $P_\theta = (X = (0,1,0)) = 2\theta(1-\theta)$ e $P_\theta = (X = (0,0,1)) = (1-\theta)^2$. Vamos agora representar essa variável X por S(x), tal que:

$$S_i = X_i, 1 \le i \le m$$
$$S_i = (\in_{i1} + \in_{i2}, \in_{i3})$$

Isso pode acontecer em casos onde obtemos que não se consegue distinguir entre homozigoto **AA** ou heterozigoto **Aa.**

A log-verossimilhança em termos de $S(x)$ é escrita:

$$l_{q,s}(\theta) = \sum_{i=1}^{m} [2 \in_{i1} \log\theta + \in_{i2} \log 2\theta(1-\theta) + 2 \in_{i3} \log(1-\theta)]$$
$$+ \sum_{i=m+1}^{n} [(\in_{i1} + \in_{i2}) \log(1-(1-\theta)^2) + 2 \in_{i3} \log(1-\theta)]$$

Podemos reescrever essa função como pertencente à família exponencial, como a seguir:

$$p(x,\theta) = \exp\{\eta(2N_{1n}(x) + N_{2n}(x)) - A(\eta)\}h(x)$$

Em que:

$$\eta = \log\left(\frac{\theta}{1-\theta}\right),\ h(x) = 2^{N_{2n}(x)},\ A(\eta) = 2n\log(1+e^\eta) \text{ e } N_{jn} = \sum_{i=1}^{n} \in_{ij} (x_i), 1 \le j \le 3.$$

Escrevendo o algoritmo EM, teremos:

$$A'(\eta) = 2n\theta$$

$$E_\theta(2N_{1n} + N_{2n} \mid S) = 2N_{1m} + N_{2m} + E_\theta \sum_{i=m+1}^{n} (2 \in_{i1} + \in_{i2}) \mid \in_{i1} + \in_{i2}, m+1 \le i \le n$$

E, finamente, após algumas simplificações:

$$E_\theta(2N_{1n} + N_{2n} \mid S) = 2N_{1m} + N_{2m} + \frac{2}{2-\hat{\theta}_k} M_n$$

Em que:

$$M_n = \sum_{i=m+1}^{n} (\in_{i1} + \in_{i2}).$$

Daí o algoritmo EM é escrito da seguinte forma:

$$\hat{\theta}_{k+1} = \frac{N_{1m} + N_{2m}/2}{n} + \frac{1}{2-\hat{\theta}_k} \frac{M_n}{n}$$

4- Resultados da simulação e perspectiva de trabalhos futuros:

Foram simuladas cinco populações mendielanas de tamanho 1.000.000 e, para cada uma, foi retirada uma amostra de tamanho 1.000, onde 990 deste são completamente distinguíeis e 10 não o são. Posteriormente, foram realizadas 1.000 reamostragem (Bootstrap), usando somente os dados distinguíveis e, em cada reamostragem, foi usado o algoritmo EM a fim de obter a estimativa do parâmetro. Com isso, foi obtido um vetor de estimativas, no qual foi construído um intervalo de 95% de confiança. Abaixo se encontra uma tabela mostrando os parâmetros e os intervalos simulados.

Tabela 1 - Saída computacional

Parâmetro Populacional (θ)	Intervalos obtidos
0,5	(0,4876; 0,5142)
0,3	(0,2924; 0,3294)
0,1	(0,0907; 0,1118)
0,05	(0,0421; 0,0596)
0,02	(0,0165; 0,0276)

Observa-se que os intervalos obtidos reproduzem com êxito os parâmetros populacionais, o que mostra a eficácia do algoritmo. Vale ressaltar que mesmo tornando o parâmetro populacional mais raro não houve perca de precisão na estimativa dos parâmetros. Essa observação nos faz questionar o quanto uma freqüência alélica deve ser rara para que o algoritmo se torne inviável. A inviabilidade, obviamente, também depende dos tamanhos da população e da amostra retirada. Descobrir em que situações o algoritmo não obtém bons resultados será nosso próximo desafio.

5- Bibliografia utilizada:

BICKEL, Peter J., DOKSUM, Kjell A**.** (1988)**. Mathematical statistics: basic ideas and selected topics**, Vol. I, 2 ed., Prentice Hall, N. Jersey.

DEMPSTER A.P.; LAIRD N.M.; RUBIN D.B.(1977). **Maximum likelihood from incomplete data via the EM algorithm (with discussion)**. J. Roy. Statist. Soc. Ser. B, Vol. 39, No. 1 pp. 1-38.

EFRON, B. (1979). **Bootstrap methods: another look at the jackknife**. The Annals of Statistics, 7, pp. 1-26.

EFRON, B., TIBSHIRANI, R.J.(1993). **An Introduction to the Bootstrap**. New York: Chapman and Hall.

FLURY, Bernard; ZOPPÉ, Alice (2000). **Exercices in EM**. The American Statistician, Vol. 54, n. 3, pp. 207-209.

MCLACHLAN G.J.; KRISHNAN T (1997). **The EM Algorithm and Extensions**. Wiley. New York.

POLLI, D. André(2007). **Um estudo de métodos bayesianos para dados de sobrevivência com omissão de covariáveis**. Dissertação (Mestre em matemática e estatística) - Universidade de São Paulo, São Paulo.

R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.