Análise de dados discrepantes usando o método bootstrap: O caso da ocorrência anômala radioativa de ^{210}Pb em amostras de solo da região do Agreste Semi-Árido de Pernambuco

Cleomacio Miguel da Silva, José Nildo Tabosa
Empresa Pernambucana de Pesquisa Agropecuária (IPA).
Avenida General San Martin, 1371, San Martin, CEP-50740-540, Recife-PE.
cleomaciomiguel@yahoo.com.br

Romilton dos Santos Amaral, Jorge João Ricardo Ferreira Cardoso, José Araújo dos Santos Júnior
Grupo de Radioecologia. Departamento de Energia Nuclear (DEN-UFPE).
Avenida Professor Luiz Freire, 1000, Cidade Universitária, CEP-50740-540, Recife - PE.

José Wilson Vieira
Centro Federal de Educação Tecnológica - CEFET/PE
Av. Prof. Luiz Freire, 500, Cidade Universitária, 50740-540, Recife – PE.

Emerson Emiliano Gualberto de Farias
Centro Regional de Ciências Nucleares (CRCN). Divisão de Análises Ambientais (DIAMB).
Avenida Professor Luiz Freire, 01, Cidade Universitária, CEP-50740-540, Recife - PE.

Resumo: *Região com elevado nível de radioatividade natural é um local tipicamente anômalo, onde as concentrações dos radionuclídeos podem variar desde o background até valores extremamente elevados. Devido aos efeitos causados pelos valores discrepantes, utiliza-se a mediana como valor mais representativo para qualquer conjunto de dados obtidos de uma região anômala. O bootstrap é um método que procura substituir a análise teórica pelo poder de processamento dos computadores. O presente trabalho teve como objetivo avaliar a aplicação do método "bootstrap" não paramétrico na determinação da mediana das concentrações de ^{210}Pb em amostras de solos provenientes da ocorrência uranífera anômala localizada na cidade de Pedra na região do Agreste Semi-Árido de Pernambuco. Para tanto, realizaram-se simulações computacionais utilizando o método de Monte Carlo, com 60.000 iterações, que resultaram na concentração mediana de 1.573 ± 1.373Bq.kg^{-1}.*

Palavras-chave: ^{210}Pb, dados discrepantes, método bootstrap.

Introdução

O "bootstrap" é uma técnica estatística computacionalmente intensiva que permite a avaliação da variabilidade de estatísticas, com base nos dados de uma única amostra existente. Essa técnica foi introduzida por Efron [1], e, desde então, tem merecido profundo estudo por parte dos estatísticos, não só na parte teórica, como também na aplicada. Porém, sendo um método numérico, a sua operacionalidade somente se tornou viável com o advento e a popularização dos computadores.

O termo "bootstrap" surgiu da frase "to pull oneself up one's bootstrap" retirada de "Adventures of Baron Munchausen" de Rodolph Erich Raspe, século XVII: "The Baron had to the of a deep lake. Just whem it looked like allwas lost, he thought to pick himself up his own bootstrap." Esse texto relata uma situação em que o Barão está afundando em um lago e vendo que tudo estava perdido pensa que conseguirá emergir puxando os cadarços dos próprios sapatos [2]. O sentido estatístico do termo é passar a idéia de que, em situações difíceis, devem-se tentar as mais variadas soluções possíveis.

Na estatística, as situações difíceis podem ser vistas como os problemas de soluções analíticas complexas, e, as variadas soluções possíveis seria a

utilização de uma metodologia com grande quantidade de cálculos, para analisar um pequeno conjunto de dados. A solução para esses casos, com o uso de métodos computacionalmente intensivos, é obtida substituindo-se o poder analítico das expressões teóricas pelo poder de processamento dos computadores.

A idéia principal do método é a amostra "bootstrap", que é retirada da amostra original com reposição. Dessa forma, todo resultado "bootstrap" depende diretamente da amostra original, isto é, os resultados "bootstrap" são robustos para a amostra original. Algumas considerações de regularidade sob as quais esse método é consistente foram discutidas por Bickel e Freedman [3]. Os conceitos básicos, propriedades teóricas e aplicações podem ser encontrados em Efron e Tbshirani [2].

O "bootstrap" pode ser implementado tanto na estatística não-paramétrica quanto na paramétrica, dependendo apenas do conhecimento do problema. No caso não-paramétrico, o método "bootstrap" reamostra os dados com reposição, de acordo com uma distribuição empírica estimada, tendo em vista que, no geral, não se conhece a distribuição subjacente aos dados. No caso paramétrico, quando se tem informação suficiente sobre a forma da distribuição dos dados, a amostra "bootstrap" é formada realizando-se a amostragem diretamente nessa distribuição com os parâmetros desconhecidos substituídos por estimativas paramétricas. A distribuição da estatística de interesse aplicada aos valores da amostra "bootstrap", condicional aos dados observados, é definida como a distribuição "bootstrap" dessa estatística [2]. Operacionalmente, o procedimento "bootstrap" consiste na reamostragem de mesmo tamanho e com reposição dos dados da amostra original, e cálculo da estatística de interesse para cada reamostra "bootstrap" (pseudo-valores) [2]. A técnica de "bootstrap" tenta realizar o que seria desejável realizar na prática, se tal fosse possível: repetir os procedimentos experimentais.

A amostra original

O presente trabalho utilizou o método "bootstrap" não paramétrico, tendo como dados de entrada (amostra original) os valores das concentrações de ^{210}Pb em amostras de solos da principal ocorrência anômala natural de urânio que se encontra localizada numa área da região do Agreste Semi-Árido de Pernambuco. Nesta área, a maior concentração de urânio nas rochas cálcio-silicáticas anfibolíticas foi de 22 kBq.kg^{-1}.

O método bootstrap

Efron e Tbshirani [2] apresentaram as idéias básicas subjacentes ao método de "bootstrap", no âmbito da inferência clássica da estatística, como se segue.

Com $x = (x_1,\ x_2,\ \cdots,\ x_n)$ amostra aleatória obtida a partir de uma população com função de distribuição desconhecida, F, seja, $\hat{\Theta}(x_1,\ x_2,\ \cdots,x_n)$, um estimador do parâmetro $\Theta(F)$ que, como se indica, depende naturalmente de F. Seja $\hat{F}$ a função de distribuição empírica associada à amostra obtida, tal que a cada valor observado x_i, onde $(i = 1, 2, \cdots, n)$, atribui massa probabilística $1/n$. Então, o valor de $\hat{F}$ é dado pela equação 1.

$$\hat{F}_{(n)}(x) = \frac{\left[\sum_{i=1}^{n} I(x_i \le x)\right]}{n} \qquad (1)$$

Onde: $\hat{F}_{(n)}(x)$ é o estimador não-paramétrico de máxima verossimilhança de F e $I(x_i \le x)$ é a função indicadora.

Uma amostra "bootstrap" é uma amostra $x^* = (x_1^*,\ x_2^*,\ \cdots,\ x_n^*)$ obtida de forma aleatória e com reposição a partir da amostra original $x = (x_1,\ x_2,\ \cdots,\ x_n)$, também designada população "bootstrap". A notação com asterisco indica que x^* não é um novo conjunto de dados reais x, mas sim uma versão randomizada, ou reamostrada de x. A amostra "bootstrap" consiste dos correspondentes membros de x, onde: $x_1^* = x_{i1},\ x_2^* = x_{i2},\ \cdots, x_n^* = x_{in}$. O conjunto $(x_{i1}^*, x_{i2}^*,\ \ldots\ , x_{in}^*)$ representa a i-ésima amostra de tamanho n com reposição dos dados originais do conjunto $x = (x_1,\ x_2,\ \cdots,\ x_n)$.

No método "bootstrap", a mediana amostral calculada é denominada por $\tilde{x}$. A cada procedimento de reamostragem do conjunto original $x = (x_1,\ x_2,\ \cdots,\ x_n)$, correspondem estimadores, dados por $\tilde{x}_1,\ \tilde{x}_2,\ \ldots,\ \tilde{x}_n$. O estimador "bootstrap"

da mediana da população é a média aritmética, $\bar{x}_B$, dos n estimadores $\tilde{x}_i$.

Sendo o método "bootstrap" de amostragem com reposição, poderíamos ter, por exemplo: $x_1^*=x_7$, $x_2^*=x_{10}$,$x_3^*=x_2,\cdots,x_B^*=x_3$.

Portanto, o conjunto de dados "bootstrap" é constituído de membros do conjunto de dados original $x=(x_1, x_2, \cdots, x_n)$, onde alguns não aparecem nenhuma vez, outros aparecem uma vez, outros aparecem duas vezes, etc.

Da distribuição $\hat{F}_{(n)}(x)$ tomam-se B amostras "bootstrap" de mesmo tamanho n, como apresentada na seqüência abaixo:

$$\begin{aligned} x_1^* &= \left[x_{11}^*, x_{12}^*, \cdots, x_{1n}^*\right] \\ x_2^* &= \left[x_{21}^*, x_{22}^*, \cdots, x_{2n}^*\right] \\ &\vdots \\ x_B^* &= \left[x_{B1}^*, x_{B2}^*, \cdots, x_{Bn}^*\right] \end{aligned}$$

O estimador "bootstrap" da média da população é a média aritmética, $\bar{x}_B$, dos n estimadores $\tilde{x}_i$. O estimador "bootstrap" do erro padrão é dado pela equação 2.

$$\hat{\sigma}_B=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(\tilde{x}_i-\bar{x}_B)^2} \qquad (2)$$

Especificamente, $\tilde{x}_i$, pode ser substituído pelo estimador $\hat{\Theta}_i$, para cada n amostra "bootstrap". A média "bootstrap" $\bar{x}_B$ pode também ser substituída por $\hat{\Theta}_B$, que é a média aritmética dos n estimadores "bootstrap". A diferença $\hat{\Theta}_B-\hat{\Theta}$ é o estimador do enviesamento de $\hat{\Theta}$. Deste modo, o estimador "bootstrap" do erro padrão de $\hat{\Theta}$ é dado pela equação 3.

$$\hat{\sigma}_B=\sqrt{\frac{1}{B-1}\sum_{i=1}^{B}(\hat{\Theta}_i-\hat{\Theta}_B)^2} \qquad (3)$$

No método "bootstrap", o estimador $\hat{\Theta}$ é normalmente distribuído com média Θ e erro padrão σ. A distribuição "bootstrap" possui forma padronizada dada pela equação 4.

$$Z_{\alpha/2}=\frac{\hat{\Theta}-\Theta}{\sigma} \qquad (4)$$

È importante salientar que a vantagem do método "bootstrap" é que ele pode ser aplicado a, praticamente, qualquer estatística $\hat{\Theta}$, não se limitando apenas à média $\hat{\Theta}=\bar{x}$. Isto é muito importante uma vez que para algumas estatísticas ou não existem fórmulas analíticas ou, quando existem, são difíceis e aproximadas para estimativa dos seus respectivos erros padrões.

O algoritmo bootstrap

O método "bootstrap" passa pelo algoritmo de Monte-Carlo. Nesse caso, um dispositivo gerador de números aleatórios seleciona inteiros $i_1, i_2, \cdots, i_n$, cada um dos quais é igual a algum valor entre 1 e n com probabilidade $1/n$. A amostra bootstrap consiste dos correspondentes membros do conjunto original $x=(x_1, x_2, \cdots, x_n)$ [2]. Na prática constrói-se a distribuição "bootstrap" de F por Monte-Carlo com um número de repetições, B, suficientemente grande. Um indicador do tamanho adequado de B, independente do custo computacional, é a qualidade da convergência da estimativa "bootstrap" do parâmetro para a estimativa natural do parâmetro $\hat{\Theta}_B(B\to\infty)\to\Theta(F)$ [2]. A construção do algoritmo Monte-Carlo para a obtenção da distribuição "bootstrap" das estatísticas usuais é no geral simples. Sua convergência está garantida pela lei dos Grandes Números, pois, $(x_1^*, x_2^*, \cdots, x_n^*)$ nada mais são do que uma amostra de variáveis aleatórias independentes e identicamente distribuídas com distribuição condicional de $\hat{\Theta}_B$. Assim, quando B tende a infinito, a média amostral $\hat{\Theta}_B$ aproxima-se de Θ [1]. O seguinte algoritmo foi construído para calcular a mediana das concentrações de ^{210}Pb nas amostras de solo pelo método de Monte-Carlo:

(**1**) Da amostra original, sorteou-se, utilizando um gerador de números aleatórios com probabilidade

$1/n$, os n valores com reposição para formar as amostras "bootstrap" de mesmo tamanho da original.
(**2**) Computou-se a média aritmética em cada procedimento de reamostragem.
(**3**) Repediu-se o passo (**2**) um número B de vezes obtendo dessa maneira, B valores da estatística em questão.
(**4**) Obtiveram-se as B médias para formar a distribuição $\hat{F}$.

(**5**) Determinou-se o estimador $\hat{\Theta}_B$ da distribuição $\hat{F}$.

O estimador $\hat{\Theta}_B$ foi utilizado para estimar a mediana das concentrações de ^{210}Pb nas amostras de solo. O processo de simulação foi realizado utilizando um programa desenvolvido na linguagem C++.

Resultados e discussão

Na Tabela 1 encontram-se apresentadas os valores das concentrações de ^{210}Pb nas amostras de solos em torno da ocorrência uranífera de Pedra, que variaram de 195 a 86.400 Bq.kg^{-1}.

Tabela 1: Concentração de ^{210}Pb em solos na ocorrência anômala.

Tipo de amostra	**Concentração (Bq.kg^{-1})**
Amostra 01	30.000
Amostra 02	86.400
Amostra 03	1.925
Amostra 04	669
Amostra 05	528
Amostra 06	11.000
Amostra 07	1.284
Amostra 08	441
Amostra 09	459
Amostra 10	195
Amostra 11	306

Como podem ser observado na Tabela 1, houve grande variação nos valores das concentrações de ^{210}Pb nas amostras de solos analisadas. Sendo assim, foi grande a discrepância entre os valores. Os fenômenos radioativos naturais, em cuja análise intervém o método estatístico, bem como os dados estatísticos a eles referentes, caracterizam-se tanto pela sua semelhança, quanto pela sua variabilidade. Desde modo, o cálculo de uma determinada medida de tendência central só se justifica em razão da variabilidade presente no meio ambiente. Entretanto, caso a variação seja grande, como aconteceu com os dados apresentados na Tabela 1, a utilização da média aritmética, como valor mais representativo, seria totalmente inadequado, devido à discrepância entre os valores. Em estudos radioecológicos não existem procedimentos estatísticos utilizados para reduzir os efeitos dos valores anômalos sobre a média aritmética. Devido às flutuações estatísticas causadas pelos valores "outliers" (anômalos), os radioecologistas utilizam a média geométrica ou a mediana como valor mais representativo do conjunto de dados obtidos da amostra. A mediana não é afetada pelos valores anômalos, sendo a medida de tendência central mais utilizada em análises estatísticas de dados discrepantes [4].

Para verificar o comportamento da mediana gerada pelo procedimento "bootstrap", em relação às medidas de tendência central obtidas dos dados experimentais, realizou-se comparações entre medidas de tendência central, como mostra a Tabela 2.

Tabela 2: Mediana experimental e mediana "bootstrap" das concentrações de ^{210}Pb nas amostras de solos.

Estatística	**^{210}Pb (Bq.kg^{-1})**
Mediana experimental	669 ± 5.280
Mediana "bootstrap"	1.573 ± 1.373

Como pode ser observado na Tabela 2, o método "bootstrap" quando foi aplicado na reamostragem dos dados originais obtidos da amostra, forneceu um novo valor para a mediana, que se tornou resistente às flutuações causadas pelos efeitos dos valores discrepantes. Segundo Stuart e Kendall's [5], a variância da mediana é calculada pela equação 5.

$$\sigma^2_{\text{mediana}} = \frac{1}{4nf^2} \qquad (5)$$

Onde: n é o tamanho da amostra e f é a função densidade de probabilidade. Entretanto, para dados discrepantes, a função f não pode ser determinada, e conseqüentemente, não é possível estimar σ_{mediana} utilizando a equação 5. Por outro lado, Toledo e Ovalle [4] afirmaram que o desvio padrão da mediana, conhecido também como desvio quartil ou amplitude semi-interquartílica, é calculado pela equação 6.

$$\sigma_{\text{mediana}} = \frac{Q_3 - Q_1}{2} \qquad (6)$$

Onde: Q_3 e Q_2 são, respectivamente, o terceiro e o primeiro quartil. Entretanto, tomando como base os estudos de Helene e Vanin [6], para dados discrepantes, existem muitas incertezas na estimativa de $\sigma_{mediana}$ quando se utiliza a equação 6. A valor da mediana experimental apresentado na Tabela 2 foi calculado pela equação 6. Neste caso, observa que o desvio padrão foi muito elevado. No entanto, o desvio padrão da mediana "bootstrap" foi significativamente reduzido, como mostra a Tabela 2. Apesar da melhoria significativa no valor da mediana e do seu desvio padrão, a utilização do método "bootstrap" não conseguiu eliminar totalmente os efeitos ambientais intrínsecos causados pela variabilidade das concentrações de ^{210}Pb nos solos da ocorrência uranífera de Venturosa. Apesar disso, o novo valor da mediana calculado pelo método de reamostragem "bootstrap", foi a medida de tendência central mais representativa para o conjunto de dados das concentrações de ^{210}Pb apresentadas na Tabela 1.

Conclusão

O método "bootstrap" foi uma excelente ferramenta na determinação da mediana das concentrações de ^{210}Pb em amostras de solos com valores discrepantes.

O método "bootstrap" pode ser utilizado para determinar a medida de tendência central mais representativa para qualquer conjunto de dados discrepantes.

Agradecimentos

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pela bolsa de estudo concedida.

Referências

[l] B. Efron, The jackknife, the bootstrap and other resampling plans. Bristol: J.W. Arrowsmith, Ltd. 1982. 92 p.

[2] B. Efron, R. J. Tibshirani. An introduction to the bootstrap. New York: Chapman e Hall, 1993. 436 p.

[3] P. J. Bickel, D. A. Freedman. Some asymptotic theory for the bootstrap. Annals Statistics, v. 9, p. 1196-1217. 1981.

[4] G. L. TOLEDO, I. I. OVALLE. Estatística básica. São Paulo: Atlas, 2 ed. 1983. 459 p.

[5] A. Stuart, J. K. O. Kendall's. Advanced theory of statistics. London: Edward Arnold, v.1, 6th, 1994.

[6] O. Helene, V. R. Vanin. Analysis of discrepant data using a bootstrap procedure. Nuclear Instruments and Methods in Physics Research A, v. 481, p. 626-631. 2002.