APPLICATION OF BOOTSTRAP METHOD FOR EVALUATING DISCREPANT LEVELS OF RADIUM-226 IN FORAGE PALM (*Opuntia spp*)

Cleomacio Miguel da SILVA[1]
Ademir AMARAL[2]
Romilton dos Santos AMARAL[3]
José Araújo dos SANTOS JÚNIOR[4]
José Wilson VIEIRA[5]

▪ *ABSTRACT: The distribution of natural radionuclides in samples from typically anomalous environments has generally a great significant asymmetry, as a result of outlier. For diminishing statistic fluctuation researchers, in radioecology, commonly use geometric mean or median, once the average has no stability under the effect of outliers. On the other hand, Efron presented a non-parametric method the so-called bootstrap that can be used to decrease the dispersion around the central-tendency value. In this context, the present study had as an objective to evaluate the application of the non-parametric bootstrap method (BM) for determining the average concentration of ^{226}Ra in cultivated forage palms (Opuntia spp.) in soils with uranium anomaly distributed on the dairy milk farms, localized in Pernambuco-Brazil, as well as discussing the utilization of this method in radioecology. The results of ^{226}Ra in samples of forage palm varied from 1,300 to 25,000 mBq.kg^{-1} (dry matter), with arithmetic average of 5,965.86± 5,903.05 mBq/kg. The result obtained for this mean by BM was 5,963.82 ± 1,202.96 mBq.kg^{-1} (dry matter). Thus, the BM permitted reaching a robust arithmetic mean, minimizing the effects of the outliers.*

▪ *KEYWORDS: Bootstrap; radium-226; forage palm*

1 Introduction

In Environmental Statistics the values of the concentrations of contaminants in the environment are considered to have a distribution of probability with tendency to be lognormal, despite the fact that the contaminants distribution in typically anomalous locals has an elevated asymmetry, caused by the outliers effects (OTT, 1994). To reduce

[1] Departamento de Física, Universidade Frederal Rural de Pernambuco – UFRPE, CEP: 52171-900, Recife, PE, Brasil. E-mail: *cleomacio@ig.com.br*
[2] Grupo de Estudos em Radioproteção e Radioecologia (GERAR), Departamento de Energia Nuclear, Universidade Frederal de Pernambuco – UFPE, CEP: 50740-540, Recife, PE, Brasil. E-mail: *amaral@ufpe.br*
[3] Departamento de Energia Nuclear, Universidade Frederal Rural de Pernambuco – UFPE, CEP: 50740-540, Recife, PE, Brasil. E-mail: *romilton@ufpe.br*
[4] Departamento de Química, Universidade Frederal Rural de Pernambuco – UFRPE, CEP: 52171-900, Recife, PE, Brasil. E-mail: *jaraujo@ufpe.br*
[5] Departamento de Física, CEFET-PE, CEP: 50740-540, Recife, PE, Brasil. E-mail: *wilson@yahoo.com.br*

the influence of these latter, radioecologists use geometric mean or median as representative value for a set of data. In the case of geometric average, logarithms, the values obtained by the samples are used to obtain a symmetric curve. In this case, the geometric average shows itself to be more adequate than the arithmetic mean (Toledo and Ovalle, 1983). On the other hand, generally, median is not affected by the anomalous values being the central tendency parameter the most used in statistic analyses of discrepant data (Toledo and Ovalle, 1983). In positive asymmetric probability distribution, the geometric mean and median are always smaller than the arithmetic mean (Singh et al., 1997). However, geometric and a median do not adequately converge for the most representative value of any data set with outlier's values. In the case of the outlier's values, the use of the arithmetic mean is not adequate because the dispersion is very elevated (Toledo and Ovalle, 1983).

In statistics, difficult situations can be seen as problems for a complex analytical solution. The various possible solutions would be the use of a methodology with a great number of calculi, aiming to extrapolate the results from a small group of data. In this context, bootstrap method of resampling, when applied with results obtained from the sample, provides an arithmetic average that is resistant to the fluctuations due to outlier effects. With the systematic use of computational tools, the solution for these cases is obtained by substituting the analytical resolution for the power of computer processing through the resampling bootstrap method. The resampling process consists of sample generation, from experimental samples, whose data taken at random (with repositioning) are used in the formation of each bootstrap sample. Thus, all the results depend directly on the experimental sample. The statistic distribution of interest applied to this kind of sample, conditional to the observed data, is defined as a bootstrap distribution of this statistic (Efron and Tibshirani, 1993).

In studies on the estimation of the concentration of radionuclides distributed in the environment, a repetition of the same experiment, under the same conditions, is not always possible. However, under determined conditions, it is possible to simulate the behavior of some of the measurements taken from a certain sample, aiming to obtain, for example, an average. The bootstrap method tries to carry out what would be desirable in practice: repeat the experimental procedures. Some considerations on the consistency of this method were discussed by Bickel and Freedman (1981). The basic concepts, theoretical properties and applications can be found in Efron and Tibshirani (1993). Taking into account a set of data about ^{226}Ra concentration in forage palms samples (Opuntia spp.), which were cultivated in soils containing anomaly distribution of uranium, this work was designed to evaluate the use of non-parametric bootstrap method for determining the representative concentration value for this radionuclide well as discussing the utilization of this statistical method in radioecology.

2 Materials and methods

2.1 Sampling

Fourteen samples of forage palms (*Opuntia spp.*) were colleted from nine farms located near the mineralization of uranium, around the cities of Pedra and Venturosa, Pernambuco-Brazil, having high occurrence of uranium.

2.2 Determination of ^{226}Ra - experimental data

To quantify ^{226}Ra from forage palms (*Opuntia spp.*) samples, the methodology employed was based on radon emanation classical technique (Goldin, 1961). The Environmental Monitoring Laboratory of the Nuclear Energy Department of the Federal University of Pernambuco is part of the National Intercomparison Program, whose objective is to evaluate the accuracy of the determination of concentrations of ^{226}Ra, by the method used in the work. Through carrying out internal intercomparison an evaluation of the reliability of the method of analysis used was sought, for determining the concentration of ^{226}Ra. Standard samples were prepared, using certified material provided by Institute of Radioprotection Dosimetry (IRD, 1983).

2.3 Statistical analyses by *bootstrap* method

Efron and Tibshirani (1994) presented the basic ideas in the bootstrap method, in the classic scope of inference of statistics, as the following. A bootstrap sample is a sample composed by $x^* = (x_1^*,\ x_2^*,\ \cdots,\ x_n^*)$ that is obtained in a random form with repositioning from the experimental sample $x = (x_1,\ x_2,\ \cdots,\ x_n)$, also designated bootstrap population. Here, the asterisk denotes that x^* is a randomized version, or resampling of x, rather than a new group of actual data. The bootstrap sampling consists of corresponding members of x. For each bootstrap procedure one should carry out a random resampling by sampling with replacement using the n elements from the experimental sample, which will be employed as parent population. Thus, the arithmetic mean $\bar{x}_i^*$ is reached using the follow equation 1. After a number m of resamplings, the arithmetic bootstrap mean $\bar{x}_m^*$ is obtained by equation 2, with standard deviation given from the equation 3.

$$\bar{x}_i^* = \sum_{i=1}^{n} \frac{\bar{x}_{in}^*}{n} \qquad 1$$

$$\bar{x}_m^* = \sum_{i=1}^{m} \frac{\bar{x}_i^*}{m} \qquad 2$$

$$S_m = \sqrt{\frac{1}{m-1} \sum_{i=1}^{m} (\bar{x}_i^* - \bar{x}_m^*)^2} \qquad 3$$

The bootstrap distribution of probability is a result of the sequence bellow:

$$x_1^* = [x_{11}^*,\ x_{12}^*,\ \cdots,\ x_{1n}^*]$$
$$x_2^* = [x_{21}^*,\ x_{22}^*,\ \cdots,\ x_{2n}^*]$$
$$\vdots$$
$$x_m^* = [x_{m1}^*,\ x_{m2}^*,\ \cdots,\ x_{mn}^*]$$

In practice, the bootstrap distribution is built form the Monte-Carlo Method with a number of repetitions, for a sufficiently large "m". In this case, the bootstrap mean approximates the mean of population and the distribution tends to a normal one (Manly, 1997). The convergence is guaranteed by the great numbers law, because, $(x_1^*,\ x_2^*,\ \cdots,\ x_n^*)$ are nothing more than a sample of independent random variables and are identically distributed.

Applying this bootstrap technique for evaluating the experimental investigation data, the following algorithm was designed for determining the concentration of ^{226}Ra in forage palms (*Opuntia spp.*):

(1) By the Monte Carlo method, each experimental datum was selected by using a generator of random numbers, with replacement.
(2) Then, samples of the same size n were obtained from the parent population;
(3) The arithmetic mean was computed for each procedure of resampling;
(4) After m resampling, the bootstrap mean $\bar{x}_m^*$ was obtained and used to estimate the ^{226}Ra average concentration of forage palms (*Opuntia spp.*).
(5) The simulation process was carried out using a program developed in the C++ language. In the present work 65,536 bootstrap iterations were performed.

3 Results and discussion

The values obtained for concentration of ^{226}Ra from experimental measurements varied from 1,300 to 25,000 mBq.kg^{-1} (dry matter), with an arithmetic average of 5,965.86 mBq.kg^{-1}. In the bootstrap sample, the mean concentration varied from 2,440.36 to 14,945.86 mBq.kg^{-1}, with a bootstrap mean of 5,963.82 mBq.kg^{-1}. Table 1 compares the experimental population parameters with the ones obtained by the bootstrap technique.

Tabela 1 - Statistics in the experimental sample and bootstrap

Statistics	Experimental sample (mBq.kg^{-1} in dry matter)	Bootstrap sample (mBq.kg^{-1} in dry matter)
Arithmetic mean	5,965.86	5,963.82
Standard deviation	5,903.05	1,202.96

Observing the table above, the experimental and bootstrap means have no significant difference (according to test LSD-least significant difference). However, the standard deviation (SD) related to experimental arithmetic mean was about five times greater than the bootstrap SD. This suggests that, with the presence of outlier values, the bootstrap arithmetic mean is more stable being representative of the parent population. In other words, in this experimental example, the bootstrap distribution ($\overline{x}^*_m$, S_m) is more adequate to represent the set of data related to the ^{226}Ra concentration for the investigated samples of forage palm than the experimental data. Obviously, without experimental data one cannot build the bootstrap distribution. Figure 1 presents the distribution of the arithmetic means obtained from m samples built using bootstrap replacement procedure.

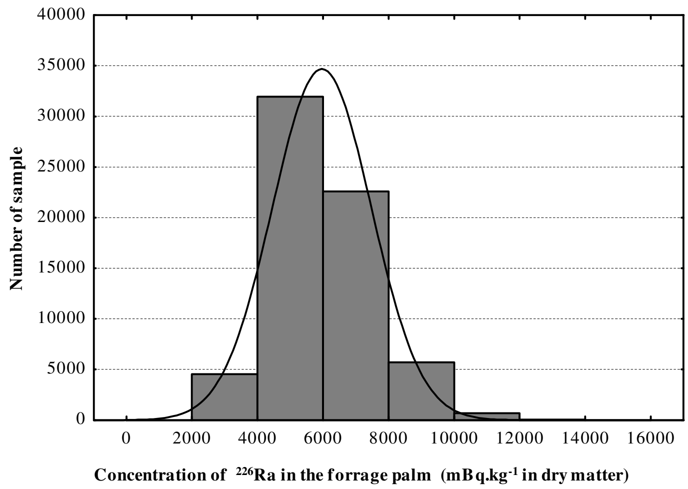

FIGURA 1 - Distribution in the bootstrap concentrations of ^{226}Ra in the forage palm samples.

As predicted by the bootstrap technique, the figure above highlights that the arithmetic distributions concerning ^{226}Ra concentrations in forage palm (*Opuntia spp.*) tend to a normal distribution.

Conclusions

Based on this work, one can infer that bootstrap method is an important tool for determining a robust estimation for arithmetic mean, in particular when the investigator deal with discrepant data and, consequently, the effect of extreme observations must be minimized. This has been the case of studies performed about anomalous presence of radium-226 in the rural region of Pernambuco – Brazil, which was employed on this text.

Acknowledgments

This research was supported by CAPES – Coordenação de Aperfeiçoamento de Pessoal de Nível Superior and CNPq-Brazil – Conselho Nacional de Desenvolvimento Científico e Tecnológico.

SILVA, C. M. da; AMARAL, A.; AMARAL, R. dos S.; SANTOS JÚNIOR, J. A. dos; VIEIRA. J. W. Aplicação do método *bootstrap* na avaliação de níveis discrepantes de rádio-226 em palma forrageira (*Opuntia SPP*). *Rev. Bras. Biom.*, São Paulo, v.25, n.3, p.109-114, 2007.

- *RESUMO*: A distribuição de radionuclídeos naturais em amostras de locais tipicamente anômalos possui elevada assimetria, causada pelos valores outliers. Para atenuar as flutuações estatísticas, os pesquisadores, na radioecologia, utilizam a média geométrica ou a mediana, pois, a média aritmética não possui estabilidade diante dos efeitos causados pelos outliers. Por outro lado, Efron criou o método não paramétrico, denominado de bootstrap, que pode ser utilizado para diminuir a dispersão em torno de uma medida de tendência central. Neste contexto, o presente trabalho teve como objetivo avaliar a aplicação do método bootstrap (MB) não paramétrico para determinar a concentração média de ^{226}Ra em palma forrageira (*Opuntia spp.*) cultivada em solos com anomalia de urânio distribuído nas fazendas produtoras de leite localizadas em Pernambuco-Brasil, com ênfase na discussão da utilização deste método em radioecologia. Os valores de ^{226}Ra nas amostras de palma forrageira variaram de 1.300 a 25.000 mBq.kg^{-1} (matéria seca), com média aritmética de 5.965,86 ± 5.903,05 mBq.kg^{-1}. O valor da média obtido pelo MB foi 5.963,82 ± 1.202,96 mBq.kg^{-1}. Assim, a média gerada pelo MB, constituiu-se uma média aritmética robusta que possuiu estabilidade diante dos efeitos causados pelo outliers.

- *PALAVRAS-CHAVE: Bootstrap*; rádio-226; palma forrageira.

References

BICKEL, P. J.; FREEDMAN, D. A. Some asymptotic theory for the bootstrap, *Ann. Stat*. Hayward, v. 9, p.1196-1217, 1981.

EFRON, B.; TIBSHIRANI, R. J. *An introduction to the bootstrap*. New York: Chapman e Hall, 1993. 436p.

EFRON, B. *The jackknife, the bootstrap and other resampling plans*. Bristol: J.W. Arrowsmith, 1982. 92p.

GOLDIN, A. S. Determination of dissolved radium. *Anal. Chem*., Washington, v. 33, p.406, 1961.

INSTITUTO DE RADIOPROTEÇÃO E DOSIMETRIA. *Manual de procedimentos técnicos do Departamento de Proteção Radiológica Ambiental*. Rio de Janeiro: IRD/CNEN, 1983.

MANLEY, B. F. J. *Randomization, bootstrap and monte carlo methods in biology*. London: Chapman & Hall, 1997. 399p.

OTT, W. R. *Environmental statistics and data analysis*. Florida: Lewis Publishers, 1994. 345p.

SINGH, A. K.; SINGH, A.; ENGELHARDT, M. *The lognormal distribution in environmental applications*. Technology Support Center Issue, 1997.

TOLEDO, G. L.; OVALLE. *Estatística básica*. São Paulo: Atlas, 1983. 459p.

Received in 26.06.2007.

Approved after revised in 22.12.2007.