PROGRAMA DE PÓS-GRADUAÇÃO
EM TECNOLOGIAS ENERGÉTICAS E
NUCLEARES

PROTEN-DEN/UFPE-CRCN-NE/CNEN

Nº 276

DISSERTAÇÃO

DE MESTRADO

ANÁLISE ESTATÍSTICA DE DADOS RADIOECOLOÓGICOS DISCREPANTES USANDO O MÉTODO MONTE CARLO BOOTSTRAP

AUTOR: **ARYKERNE NASCIMENTO CASADO DA SILVA**

RECIFE – PERNAMBUCO – BRASIL

AGOSTO – 2013

UNIVERSIDADE FEDERAL DE PERNAMBUCO
DEPARTAMENTO DE ENERGIA NUCLEAR

Av. Professor Luiz Freire, 3000 - Cidade Universitária
CEP 50740-540 Recife - PE Brasil

ANÁLISE ESTATÍSTICA DE DADOS RADIOECOLÓGICOS DISCREPANTES USANDO O MÉTODO MONTE CARLO BOOTSTRAP

ARYKERNE NASCIMENTO CASADO DA SILVA

ANÁLISE ESTATÍSTICA DE DADOS RADIOECOLÓGICOS DISCREPANTES USANDO O MÉTODO MONTE CARLO BOOTSTRAP

Dissertação submetida ao Programa de Pós-Graduação em Tecnologias Energéticas e Nucleares para obtenção do título de Mestre em Ciências. Área de Concentração: Dosimetria e Instrumentação.

Orientador: Prof. Dr. Romilton dos Santos Amaral (DEN/UFPE)

Coorientador: Prof. Dr. José Wilson Vieira (IFPE, EPP/UPE, DEN/UFPE)

Recife - PE
Agosto - 2013

Catalogação na fonte

Bibliotecário Carlos Moura, CRB-4 / 1502

S586a Silva, Arykerne Nascimento Casado da.

Análise estatística de dados radioecológicos discrepantes usando o método Monte Carlo bootstrap. / Arykerne Nascimento Casado da Silva. - Recife: O Autor, 2013.

66 folhas, il., figs.,tabs.

Orientador: Prof. Dr. Romilton dos Santos Amaral.

Co-orientador: Prof. Dr. José Wilson Vieira.

Dissertação (Mestrado) – Universidade Federal de Pernambuco. CTG. Programa de Pós-Graduação em Tecnologias Energéticas e Nucleares, 2013.

Inclui Referências e Apêndice.

ANÁLISE ESTATÍSTICA DE DADOS RADIOECOLÓGICOS DISCREPANTES USANDO O MÉTODO MONTE CARLO BOOTSTRAP

Arykerne Nascimento Casado da Silva

APROVADA EM: 05.08.2015

ORIENTADOR: Prof. Dr. Ronaldo dos Santos Amaral

CO-ORIENTADOR: Prof. Dr. José Wilson Vieira

COMISSÃO EXAMINADORA:

Prof. Dr. José Araújo dos Santos Júnior – DEN/UFPE

Prof. Dr. Fernando Roberto de Andrade Lima – CRCN-NE/CNEN

Prof. Dr. Dário Costa Primo – DEN/UFPE

Visto e permitida a impressão

Coordenadora(o) do PROTEN/DEN/UFPE

Dedico este trabalho à Lília, Arthur, Laís, Ledice (in memorian) e Nilson (in memorian).

AGRADECIMENTOS

Àqueles que me fazem sobreviver até hoje: Lília, mulher, companheira, cúmplice, amiga e esposa para sempre; Arthur, filho amado, me faz lembrar que posso ser eterno e Laís, filha amada, mostra como eu deveria pensar aos dezoito anos.

Àqueles que me fizeram chegar até aqui: Ledice, mãe dedicada, indicou o caminho e descansa na paz. Nilson, pai saudoso, partiu muito cedo para seu descanso na paz.

Àqueles que me permitiram chegar até aqui: ao Prof. Dr. Romilton dos Santos Amaral pela minha aceitação como orientando, pelas disciplinas ministradas, pelas sugestões no trabalho de pesquisa e por abrir as portas do Grupo de Estudos em Radioecologia; ao Prof. Dr. Cleomácio Miguel da Silva por colocar em poucas palavras o que era necessário fazer; ao Prof. Dr. José Wilson Vieira pelas críticas, sugestões, correções, participação na banca examinadora do primeiro seminário de dissertação, ajuda e tempo dispensados; ao Prof. Dr. José Araújo dos Santos Júnior pelas críticas, sugestões, participações nas bancas examinadoras dos seminários e de defesa da dissertação, além da ajuda dispensada sempre; ao Prof. Dr. André Fellipe Vieira da Cunha pela disciplina ministrada, participações nas bancas examinadoras dos seminários, além das sugestões e críticas; ao Dr. Dário Costa Primo pelas sugestões, críticas e participações nas bancas examinadoras do segundo seminário e de defesa da dissertação e ao Prof. Dr. Fernando Roberto de Andrade Lima pela participação na banca examinadora da dissertação.

Ao Departamento de Energia Nuclear e ao Programa de Pós-Graduação em Tecnologias Energéticas e Nucleares da Universidade Federal de Pernambuco, na figura dos seus docentes, técnico-administrativos e discentes, por permitirem o retorno de um eterno aprendiz.

“Viver!

E não ter a vergonha

De ser feliz

Cantar e cantar e cantar

A beleza de ser

Um eterno aprendiz...”

Gonzaguinha

RESUMO

O método de reamostragem bootstrap vem sendo estudado e utilizado desde 1979. Em Radioecologia existem dificuldades operacionais na obtenção de amostras de campo, levando o pesquisador algumas vezes a trabalhar com um número insuficiente de dados. Além disso, é frequente o surgimento de valores discrepantes nos dados amostrais. Como consequência, a análise estatística das amostras pode requerer a utilização de métodos analíticos paramétricos. O método bootstrap é executado utilizando o poder de processamento dos microcomputadores atuais. As reamostras são obtidas pelo método Monte Carlo através de um sorteio aleatório dos dados reais fornecidos, considerando que eles sejam independentes. Este trabalho foi desenvolvido com o objetivo de aplicar o método bootstrap na análise estatística de dados radioecológicos e auxiliar o pesquisador na obtenção dos estimadores mais adequados para os parâmetros populacionais. O método utiliza o aplicativo R, uma ferramenta interativa poderosa para cálculos estatísticos. A ferramenta permite também programação orientada a objetos através de interpretação. Os dados utilizados para os cálculos foram 14 concentrações do ^{226}Ra na palma forrageira (*Opuntia spp*) que apresentaram variações entre 1.495 a 25.000 mBq.kg^{-1} na matéria seca, com média aritmética simples de 5.965,86 ± 5.903,05 mB.kg^{-1}. A aplicação do método de reamostragem bootstrap com 1.000 iterações, através do algoritmo desenvolvido, obteve um valor médio de 6.012,85 ± 1.597,83 mBq.kg^{-1}. Este, mais representativo que o do conjunto amostral, pois não sofreu influência dos valores discrepantes. Foi possível concluir então que o método bootstrap tem aplicação válida para a análise estatística univariada de dados radioecológicos.

Palavras-chaves: estatística univariada; reamostragem; radioecologia; simulação.

ABSTRACT

The bootstrap resampling method has been studied and used since 1979. There are, in Radioecology, operational difficulties in obtaining samples, leading the researcher sometimes work with insufficient data. Moreover, it is often the appearance of outliers in the sample data. As a consequence, the statistical analysis of samples may require the use of parametric analytical methods. The bootstrap method is performed using the processing power of current computers. The pseudo samples are obtained by Monte Carlo method through a randomly of the actual data provided considering that they are independent. The present work was developed aiming to apply the bootstrap method in statistical radioecological data analysis and to assist the researcher in obtaining the most appropriate estimators for the population parameters. The method uses the software R, a power interactive tool for statistical calculations. The tool also allows object-oriented programming through interpretation. The data used were 14 concentrations of ^{226}Ra in forage palm (*Opuntia spp*) and they showed variations between 1,495 to 25,000 $mBq.kg^{-1}$ in dry matter, with simple arithmetic average 5,965.86 ± 5,903.05 $mBq.kg^{-1}$. The application of the bootstrap resampling procedure with 1,000 iterations through the developed algorithm obtained an average value of 6,0129.61 ± 1,597.83 $mBq.kg^{-1}$. This, more representative of the whole sample, because it was not influenced by the outliers. It was possible to conclude then that the bootstrap method has application valid for univariate statistical analysis of radioecological data.

Keywords: univariate statistic; resampling; radioecology, simulation.

LISTA DE ILUSTRAÇÕES

Página

LISTA DE TABELAS

Página

LISTA DE ABREVIATURAS E SIGLAS

ABN	Artefatos Bélicos Nucleares
BSD	Berkeley Software Distribution
CCN	Ciclo do Combustível Nuclear
CNEN	Comissão Nacional de Energia Nuclear
CRAN	Comprehensive R Archive Network
CRCN	Centro Regional de Ciências Nucleares do Nordeste
DEN	Departamento de Energia Nuclear
IPA	Instituto Agronômico de Pernambuco
IRD	Instituto de Radioproteção e Dosimetria
LPG	Licença Pública Geral
MS	Matéria Seca
PDF	Portable Document Format
NAT^{COS}	Cosmogênico Natural
NAT^{DEC}	Decaimento Natural
NAT^{PRI}	Decaimento Primordial
NUCLEBRAS	Empresas Nucleares Brasileiras S/A
PNI	Programa Nacional de Intercomparação
PROTEN	Programa de Pós-Graduação em Tecnologias Energéticas e Nucleares
RAE	Grupos de Estudos em Radioecologia
RF	Radiofármacos
UFPE	Universidade Federal de Pernambuco

SUMÁRIO

1. INTRODUÇÃO

O ser humano, através da pesquisa científica, busca o desenvolvimento do conhecimento de forma objetiva e estruturada, avançando na procura de soluções para os seus problemas. O pesquisador, para desenvolver a pesquisa, precisa de aprofundamento nos estudos, observação e análise de experimentos que geram conjuntos de dados, estes muitas vezes insuficientes para uma análise estatística confiável.

Uma investigação científica, só pode ser levada a bom termo, através do método científico, que considere os valores experimentais estudados, no âmbito de amostra, representativos da população. É através da inferência estatística que generalizações e conclusões sobre as características de uma população podem ser obtidas, partindo do princípio de que uma amostra pode ser uma representação adequada dessa população. Por esse motivo, surgem os estimadores de medidas de tendência central e de medidas de dispersão, que são utilizados para analisar a representatividade desses dados em relação ao conjunto universo (WILCOX, 2009).

Em Radioecologia, área de estudo que engloba elementos da Química, da Física e da Biologia, e que tem como objetivo avaliar os efeitos das radiações ionizantes sobre as populações e sobre os ecossistemas, as pesquisas científicas dependem basicamente da obtenção de amostras de campo e de uma análise estatística segura sobre os dados obtidos (WHICKER; SCHULTZ, 1982). Os radioecologistas utilizam também, além das amostras de campo, o poder de processamento dos computadores para simular dados de radioisótopos naturais e artificiais (CHERNICK, 2007).

Através de estudos radioecológicos, pesquisadores do Grupo de Estudos em Radioecologia (RAE) do Programa de Pós-Graduação em Tecnologias Energéticas e Nucleares (PROTEN) do Departamento de Energia Nuclear (DEN) da Universidade Federal de Pernambuco (UFPE) e do Centro Regional de Ciências Nucleares do Nordeste (CRCN) da Comissão Nacional de Energia Nuclear (CNEN) analisaram a presença de mineralizações de urânio e tório, precursores dos isótopos ^{226}Ra e ^{228}Ra respectivamente, nos solos de fazendas produtoras de leite dos municípios de Pedra e Venturosa, localizadas na microrregião do Vale do Ipanema na mesorregião Agreste do estado de Pernambuco, região Nordeste do Brasil. Surgiu então a necessidade de se estabelecer um estudo completo do comportamento radioecológico dessa área, avaliando os dados obtidos de amostras de solos e rochas da região,

objetivando verificar a exposição da população da região às radiações ionizantes desses radioisótopos (SANTOS JÚNIOR, 2009; SILVA, 2006).

Estudos anteriores sobre a presença dos radioisótopos naturais em diversas regiões da Terra mostram que a distribuição destes em locais considerados anômalos apresenta uma alta assimetria decorrente do efeito de valores discrepantes (SILVA, 2006; SINGH et al., 1997; OTT, 1994). Na interpretação dos resultados experimentais radioecológicos, na classificação, no estudo da similaridade/dissimilaridade, na proveniência das amostras e na tecnologia de produção são utilizados diversos métodos estatísticos multivariados, como os métodos de agrupamento e o de análise fatorial. Contudo, para que seja viável a utilização dessas técnicas estatísticas, faz-se necessário que o conjunto amostral não possua valores discrepantes e que não existam intervalos com lacunas de valores. Além disso, os fenômenos radioativos são governados por eventos probabilísticos que fogem totalmente ao controle do pesquisador, já que são fenômenos naturais. A dependência desses fatores, que não podem ser controlados no momento de definir a amostragem dos dados radioecológicos, e o surgimento de quantitativos que variam desde valores de fundo até números considerados muito elevados, tornam o uso da análise estatística multivariada inadequada para a determinação de um valor médio representativo dos dados, e esse é um dos principais objetivos do tratamento estatístico dos valores amostrais (SILVA et al., 2012; 2011; SILVA, 2006).

Quando os valores experimentais se apresentam coesos em relação a uma medida de tendência central, não significa que esta seja adequada para representar a população. É suficiente que entre os dados amostrais obtidos apareça um valor discrepante para exigir do pesquisador a aplicação de métodos que ou despreze esse valor, ou repita a amostragem, ou ainda a verifique e corrija o processo de obtenção e tratamento das amostras. O desafio do radioecologogista é encontrar, para suas amostras, uma medida de tendência central que apresente um menor valor possível de dispersão, sem perda de informações. A partir desse valor é possível estimar o erro padrão e, consequentemente, intervalos de confiança. (SILVA et al., 2012; CHERNICK, 2007; UPTON; COOK, 2000; WHICKER; SCHULTZ, 1982).

A coleta de material para estudos radioecológicos, imprescindível para a pesquisa, se caracteriza pela dificuldade operacional, pelo alto custo financeiro e ainda por demandar longos períodos. Essas características fazem com que a repetição da coleta para obtenção de novas amostras se torne um problema para o pesquisador e sua equipe. Dessa forma, o radioecologista depara-se constantemente com a difícil tarefa de fazer uma análise estatística em seu conjunto de dados no qual o número de amostras é insuficiente (SILVA et al., 2012, 2011, 2007).

A utilização de simulação através do uso intensivo de computadores, em conjunto com o método de reamostragem bootstrap, evita que o pesquisador da área de Radioecologia fique limitado às soluções analíticas paramétricas a partir de uma distribuição amostral pequena e, precise extrapolar os resultados obtidos para realizar uma análise estatística segura e confiável. As técnicas de simulação, inclusive o método bootstrap, contornam o problema de obtenção de novas amostras quando a população não está mais disponível nas condições em que a pesquisa foi iniciada (SILVA et al., 2012, 2011, 2007).

Em diversas situações os dados experimentais obtidos a partir de amostragens aleatórias se adequam bem ao perfil da distribuição normal, o que permite grande facilidade no desenvolvimento de toda inferência estatística da pesquisa. Em situações mais gerais, o cálculo dos intervalos de confiança para a média aritmética simples é imediato (WILCOX, 2009). Uma das tarefas do pesquisador é verificar se os dados, discretos ou contínuos, obtidos de seus experimentos seguem um determinado tipo de distribuição de probabilidade (SILVA, 2006).

A utilização do poder de processamento de um microcomputador, utilizando o método bootstrap, permitiu a realização de simulações, com diferentes valores de iterações, para produção de pseudoamostras de concentrações de ^{226}Ra em palma forrageira (*Opuntia spp*), partindo de uma pequena amostra experimental com quatorze valores de concentração. Foram obtidas estimativas representativas da população subjacente estudada, sem necessidade de suposições sobre a distribuição de probabilidade da população, intervenção humana nos cálculos, descarte de valores anômalos e utilização de métodos analíticos. Dessa forma, a determinação de um estimador para um parâmetro de interesse e a avaliação da acurácia desse estimador através do erro padrão, além do cálculo do intervalo de confiança para o parâmetro, puderam ser estudados.

Avaliando o contexto descrito acima, o objetivo do presente trabalho foi aplicar o método bootstrap, através de um algoritmo computacional desenvolvido no aplicativo R, na análise estatística de dados radioecológicos com valores discrepantes presentes, considerando o caso em que as observações são amostras, de uma única variável, selecionadas ao acaso, independentes e provenientes de uma população desconhecida. Para tanto, foram utilizados os valores de concentrações de ^{226}Ra na palma forrageira (*Opuntia spp*), vegetal cultivado em uma área anômala situada no agreste do estado de Pernambuco, nos municípios de Pedra e Venturosa, e, frequentemente, utilizada na alimentação do gado leiteiro dessas regiões.

2. REVISÃO DE LITERATURA

2.1 Distribuição dos radionuclídeos naturais e artificiais na crosta terrestre

A radiação ionizante pode ser produzida naturalmente ou ser proveniente de alguma atividade antropogênica, porém seus efeitos sobre os organismos vivos são os mesmos. A Terra é continuamente submetida à radiação cósmica proveniente do sistema solar e de outros sistemas planetários. Além disso, existe a radiação da própria crosta terrestre definida como primordial e proveniente dos elementos químicos radioativos presentes (COLGAN et al., 2008).

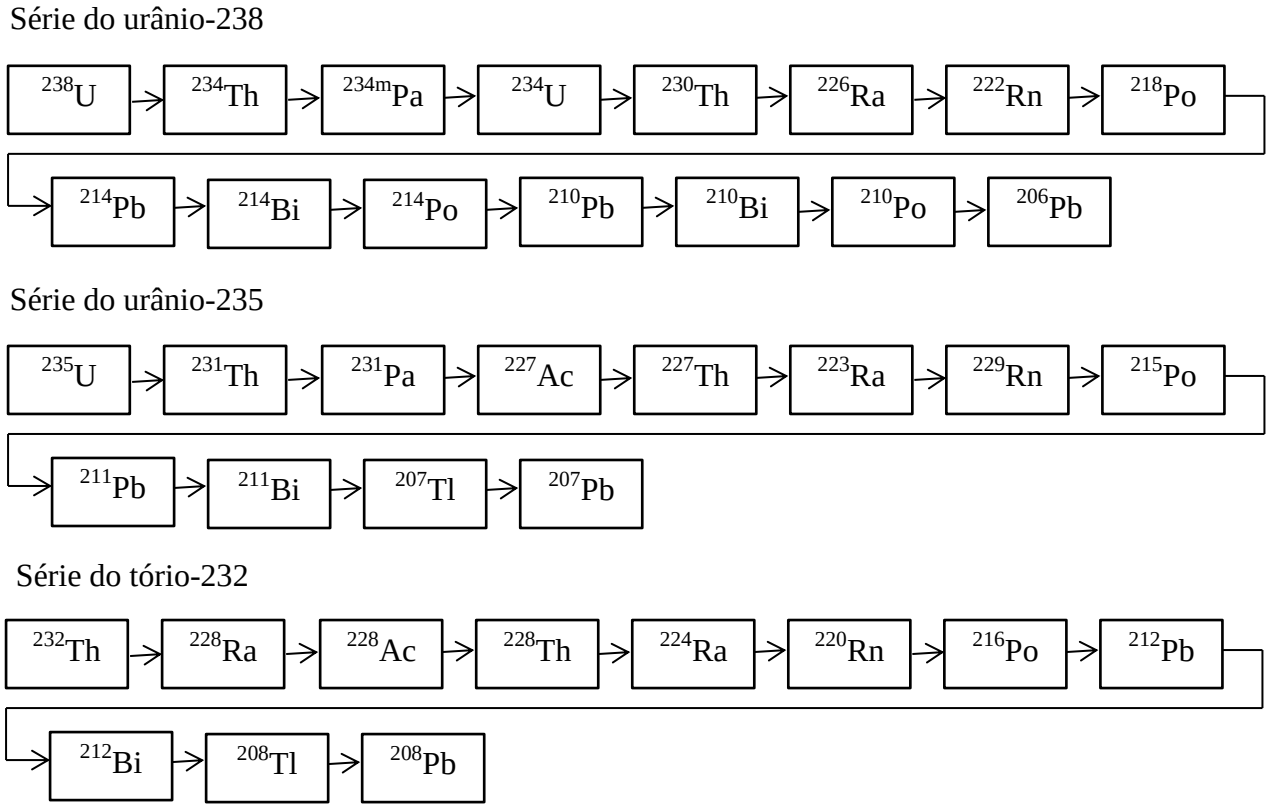
Os radionuclídeos primordiais, que incluem isótopos do urânio, tório, potássio e rubídio, surgiram no período de formação da Terra e estão distribuídos de forma heterogênea na crosta terrestre. O urânio é de fundamental importância nos estudos e pesquisas sobre radioatividade ambiental. Ele está presente naturalmente em rochas e solos, e na crosta terrestre é encontrado em concentrações médias de 3 $mg.kg^{-1}$. Porém, em rochas ricas em fosfato, a concentração do minério de urânio pode alcançar valores de até 40.000 $mg.kg^{-1}$. Os principais isótopos do urânio são o ^{238}U, com 99,3% de ocorrência natural; o ^{235}U, com 0,7% de ocorrência na natureza e o ^{234}U com apenas 0,005% de ocorrência no meio ambiente. O ^{238}U inicia a série radioativa natural do urânio, uma sequência de decaimentos radioativos sucessivos de vários elementos químicos diferentes e que se encerra com a formação do isótopo estável do ^{206}Pb. O ^{235}U inicia a série radioativa dos actinídeos que termina com a formação do isótopo estável do ^{207}Pb (SHAW, 2007; CEMBER, 1996).

O tório, outro elemento de grande importância em Radioecologia, é mais abundante na natureza do que o urânio, sendo que a sua concentração média no solo pode chegar a valores entre 6 e 9 $mg.kg^{-1}$ (SHAW, 2007; CEMBER, 1996; SCHULZ, 1965). O isótopo ^{232}Th é o radionuclídeo inicial da série radioativa natural do tório que termina com a formação do isótopo estável ^{208}Pb (SHAW, 2007; CEMBER, 1996). Em rochas, a concentração de ^{232}Th varia de 1,6 a 20 $mg.kg^{-1}$, com uma média crostal de 10,7 $mg.kg^{-1}$, atingindo concentrações cinco vezes superiores às do urânio (SANTOS JÚNIOR, 2009). Observam-se, através da Figura 1, três diagramas parciais com as séries radioativas naturais de acordo com Cember (1996), Santos Júnior (2009) e Okuno e Yushimura (2010).

Como consequência do decaimento natural dos isótopos das três séries radioativas, são produzidos elementos radioativos filhos de grande importância nos estudos sobre radioecologia como o ^{226}Ra, ^{228}Ra, ^{222}Rn e ^{210}Pb (SHAW, 2007). Segundo Amaral et al. (2005)

aproximadamente 70% de toda radioatividade que incide sobre a população é proveniente de fontes inatas das três séries radioativas naturais. De acordo com Shaw (2007), uma das principais características de um radionuclídeo em termos radioecológicos é o seu tempo de meia vida, período necessário para que uma quantidade de um radionuclídeo perca metade da sua atividade radioativa (OKUNO; YOSHIMURA, 2010; L'ANNUNZIATA, 2003). Sendo assim, é razoável considerar que quanto maior a meia vida do radioisótopo, mais tempo permanecerá nos sistemas bióticos e abióticos, aumentando o impacto nos ecossistemas. O ^{238}U tem tempo de meia vida de 4,47 x 10^9 anos e emite partículas alfa com energias de 4,20 MeV 79% das vezes e 4,15 MeV em 21% das emissões (FIRESTONE et al., 1996 apud SANTOS JÚNIOR, 2009). Além disso, produz radionuclídeos filhos como o ^{226}Ra que apresenta alto risco de provocar danos ao sangue humano (SILVA, 2006).

Figura 1 - Diagramas com as representações parciais das três séries radioativas naturais.



Fonte: Okuno; Yushimura (2010), Santos Júnior (2009), Cember (1996).

Os radionuclídeos ^{40}K, ^{238}U e ^{232}Th, além de vários dos seus produtos radioativos filhos, emitem radiação gama, tornando-se a principal fonte de exposição externa natural desse tipo de radiação para os organismos vivos, inclusive o ser humano (SANTOS JÚNIOR, 2009). Shaw

(2007) afirma que o impacto ambiental de um isótopo radioativo é função das vias ambientais e processos do qual ele toma parte, pois é através desses processos que pode ocorrer uma menor ou maior exposição, interna ou externa, dos organismos à radiação. Através dessa perspectiva é possível considerar a importância dos radionuclídeos naturais de meia vida "curta" nos sistemas bióticos, pois em um curto período metade de toda a sua emissão radioativa afetará o meio. Observam-se, através da Tabela 1, alguns dos elementos radioativos naturais de interesse para os radioecologistas, com o tempo de meia vida em anos e sua respectiva origem.

Tabela 1 - Alguns dos contaminantes radioativos naturais estudados pelos radioecologistas.

Radioisótopo	Tempo de meia vida (anos)	Origem
^{3}H	12,33	Nat^{Cos}, CCN, ABN, RF
^{210}Pb	22,30	Nat^{Dec}
^{222}Rn	$1,048 \times 10^{-2}$	Nat^{Dec}
^{226}Ra	1.600	Nat^{Dec}
^{228}Ra	5,8 anos	Nat^{Dec}
^{230}Th	$7,538 \times 10^{4}$	Nat^{Dec}
^{232}Th	$1,405 \times 10^{10}$	Nat^{Pri}
^{238}U	$4,468 \times 10^{9}$	Nat^{Pri}, CCN

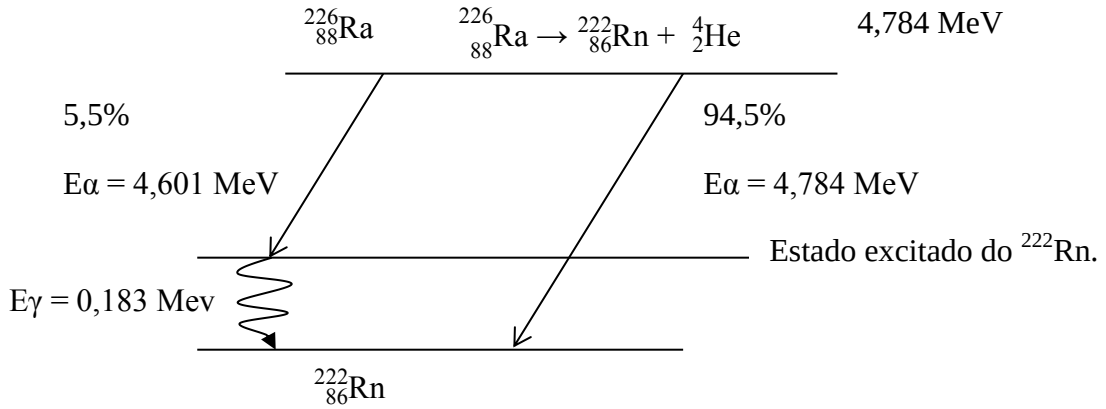
Nat^{dec} – decaimento natural, Nat^{pri} – radionuclídeo primordial, Nat^{cos} – natural cosmogênico, CCN – ciclo do combustível nuclear, ABN - artefatos bélicos nucleares e RF – radiofármacos.
Fonte: SHAW(2007).

O ^{226}Ra é um metal alcalino terroso e um dos produtos do decaimento natural do ^{238}U. Através do seu esquema de decaimento, ilustrado na Figura 2, são exibidas as duas possibilidades de emissão de partícula alfa para o radionuclídeo. O rádio entra no corpo humano através da ingestão de alimentos e água, incorporando-se ao sistema ósseo. Esse radioisótopo e seus produtos de decaimento são responsáveis pela maior fração da dose interna recebida pelo homem, devido às fontes naturais. Quando ingeridos ou inalados, seus produtos de decaimento oferecem alto potencial de risco à saúde dos seres humanos, podendo induzir o surgimento de câncer (EISENBUD; GESELL, 1997 apud SILVA, 2006).

As transformações radioativas ou decaimentos radioativos acontecem sempre através dos seguintes processos: emissão de partículas alfa, emissão isobárica (beta positiva, beta negativa e captura eletrônica) e emissão isomérica (raios gama e conversão interna). Todos esses processos são fenômenos probabilísticos, pois não é possível determinar o momento exato em

que a emissão radioativa ocorre O fenômeno ocorre independentemente da origem do radionuclídeo, se natural ou antropogênico. (CEMBER, 1996).

Figura 2 - Esquema de decaimento alfa do Ra-226.



Fonte: OKUNO; YOSHIMURA (2010).

Diferentes radionuclídeos apresentam diferentes taxas de decaimento radioativo, esses valores são definidos como atividade do elemento. No Sistema Internacional (SI) a unidade da grandeza atividade é o número de decaimentos por segundo, definida como Bequerel. O número de decomposição por segundo dividido pela massa ou pelo volume do material radioativo é definido como atividade específica, uma importante característica que relaciona a radioatividade do radionuclídeo com a massa ou volume da amostra (CEMBER, 1996).

A área de interesse da Radioecologia não se limita ao estudo dos radionuclídeos naturais. Os radioisótopos artificiais, provenientes de atividades antropogênicas, surgiram como resultado das pesquisas voltadas para o desenvolvimento bélico. Em 16 de julho de 1945, ocorreu o primeiro teste de detonação de um artefato atômico no deserto do Novo México nos Estados Unidos. Foi iniciado então, o período da história denominado de “Era Nuclear”. Seguiu-se um período de testes nucleares que, como consequência, lançou no ambiente atmosférico, substâncias radioativas artificiais provenientes da fissão do urânio, ocorrendo, também a ativação de muitos elementos estáveis tanto no ar quanto no solo (SHAW, 2007). As áreas de testes nucleares utilizadas entre as décadas de 50 e 70 do século XX foram selecionadas em locais remotos, contudo existem registros da presença de radionuclídeos antropogênicos, provenientes dessas experiências, a milhares de quilômetros das respectivas explosões (GONZALES; ANDERER, 1989; SHAW, 2007).

Existem cerca de 70 radionuclídeos diferentes originados de fontes naturais (EISENBUD; GESELL, 1997 apud SANTOS JÚNIOR, 2009). Os outros são gerados de atividades antropogênicas que vão desde produtos da fissão do urânio, como ^{137}Cs e ^{90}Sr, até produtos do processo de captura de nêutrons, como o ^{239}Pu proveniente do processo de ativação do ^{238}U (GONZALES; ANDERER, 1989).

Na Tabela 2 estão listados alguns dos contaminantes radioativos artificiais de interesse para os radioecologistas. Da mesma forma que para os radioisótopos naturais, o tempo de meia vida e a atividade específica são grandezas que permitem determinar o impacto dos radionuclídeos artificiais nos ecossistemas.

Tabela 2 - Alguns dos contaminantes radioativos artificiais estudados pelos radioecologistas.

Radioisótopo	Tempo de meia vida (anos)	Origem
^{90}Sr	28,79	CCN, ABN
^{99m}Tc	$6,891 \times 10^{-4}$	RF
^{99}Tc	$2,111 \times 10^{5}$	CCN, ABN
^{129}I	$1,57 \times 10^{7}$	CCN, ABN, RF
^{131}I	$2,197 \times 10^{-2}$	CCN, ABN
^{137}Cs	30,07	CCN, ABN
^{239}Pu	$2,411 \times 10^{4}$	CCN, ABN
^{241}Am	432,20	CCN, ABN

CCN – ciclo do combustível nuclear; ABN – artefato bélico nuclear; RF - radiofármaco.
Fonte: SHAW (2007)

As séries do urânio e do tório, naturalmente presentes nas rochas e no solo, representam, para grande parte da população, a principal fonte de exposição radiológica interna devido à ingestão de água e alimentos contaminados naturalmente por meio de transferência do meio abiótico para o meio biótico. As áreas onde ocorrem acúmulos desses radionuclídeos naturais são definidas como anômalas (SHAW, 2007). Regiões que apresentam rochas e solos com concentrações média de urânio natural da ordem de 2,8 a 3,0 $mg.kg^{-1}$ são classificadas como não anômalas, pois esse é o valor médio de concentração desse radionuclídeo na crosta terrestre (AIETA et al., 1987; TAYLOR, D.; TAYLOR, S., 1997).

Os municípios de Pedra e Venturosa, no Agreste do estado de Pernambuco, região nordeste do Brasil, apresentam valores de concentração máxima de U_3O_8 e de ThO_2 de 22.000

$mg.kg^{-1}$ e 100 $mg.kg^{-1}$, respectivamente. Valores considerados anômalos (COSTA et al., 1977). Os estudos de regiões habitadas e com elevada radioatividade natural são de grande importância já que oferecem oportunidades de observação e avaliação dos possíveis efeitos biológicos da radiação natural no homem e no meio.

Silva (2006), por exemplo, determinou as concentrações de ^{226}Ra e ^{228}Ra na dieta de bovinos de fazendas produtoras de leite na região economicamente identificada como "bacia leiteira" do estado de Pernambuco, no nordeste do Brasil. Esse estudo surgiu do interesse pelas altas concentrações de urânio e tório naturais no solo e nas rochas dessa região. O ^{238}U, ^{235}U e o ^{232}Th produzem no final das suas respectivas séries de decaimento natural isótopos estáveis do chumbo.

Santos Júnior (2009) realizou um estudo radiométrico sobre a concentração do ^{238}U, ^{226}Ra, ^{232}Th e ^{40}K em uma área com mineralizações anômalas de urânio, com o objetivo de identificar níveis elevados de radioatividade natural e estimar a exposição da população.

Alcoforado (2011) estudou a influência de ocorrência de urânio e tório nos níveis de chumbo estável no leite e derivados também na região leiteira do estado de Pernambuco.

Os estudos sobre a radioatividade ambiental permitem a análise do aproveitamento tecnológico do próprio fenômeno. É possível utilizar dados radiométricos como parâmetros auxiliares na identificação dos tipos de solos. Análises bioquímicas e físico-químicas e a verificação dos valores radiométricos das razões K/U e K/ThO de amostras de solos permitem a sua identificação mineralógica. Isso é possível porque os níveis radioativos das rochas podem ser correlacionados com sua idade e forma de ocorrência (NASCIMENTO et al., 2004).

O método de quantificação da erosão de solos através do ^{137}Cs, um radioisótopo artificial proveniente do ciclo do combustível nuclear ou de explosões de artefatos bélicos nucleares, pode ser utilizado para determinar a perda de áreas produtivas para produção agrícola. O processo já está bem difundido em países do hemisfério norte e começa a ser utilizado em alguns estados do sudeste brasileiro (ANTUNES, 2010).

A importância do conhecimento sobre os radioisótopos pode ser avaliada através da frequência dos estudos existentes sobre: o comportamento dos radionuclídeos nos sistemas solo/planta, a radiometria nos ecossistemas florestais tropicais e subtropicais, a radioecologia dos ecossistemas árticos, a ocorrência de radionuclídeos naturais a partir de fontes industriais e a proteção do meio ambiente em relação à exposição à radiação ionizante (SHAW, 2007).

2.2 Análise estatística dos dados radioecológicos

O decaimento radioativo é um fenômeno aleatório. Logo qualquer medição baseada em um radionuclídeo está sujeita a flutuações estatísticas. Estas geram incertezas em todos os procedimentos experimentais radioativos e normalmente podem ser fontes de imprecisão ou erro (KNOLL, 2000). A análise dos fenômenos radioativos naturais e dos dados obtidos a partir deles mostra que existem semelhanças e variabilidades extremas, impossibilitando a aplicação da estatística multivariada (SILVA, 2006). O pesquisador fica restrito à obtenção de medidas de tendência central e de dispersão através da análise de uma única variável estatística para caracterizar os dados experimentais (KNOLL, 2000). Portanto, a utilização de modelos de probabilidade univariada tem grande importância na análise e interpretação dos dados que são utilizados na monitorização do meio ambiente. Para uma determinada variável radiológica ambiental, como a concentração de um radionuclídeo no solo, sua característica probabilística faz com que o início de uma análise seja o exame da distribuição das medições da concentração em um dado local, de forma que o histograma de frequência das concentrações permite uma visão da ocorrência dessa variável (OTT, 1976).

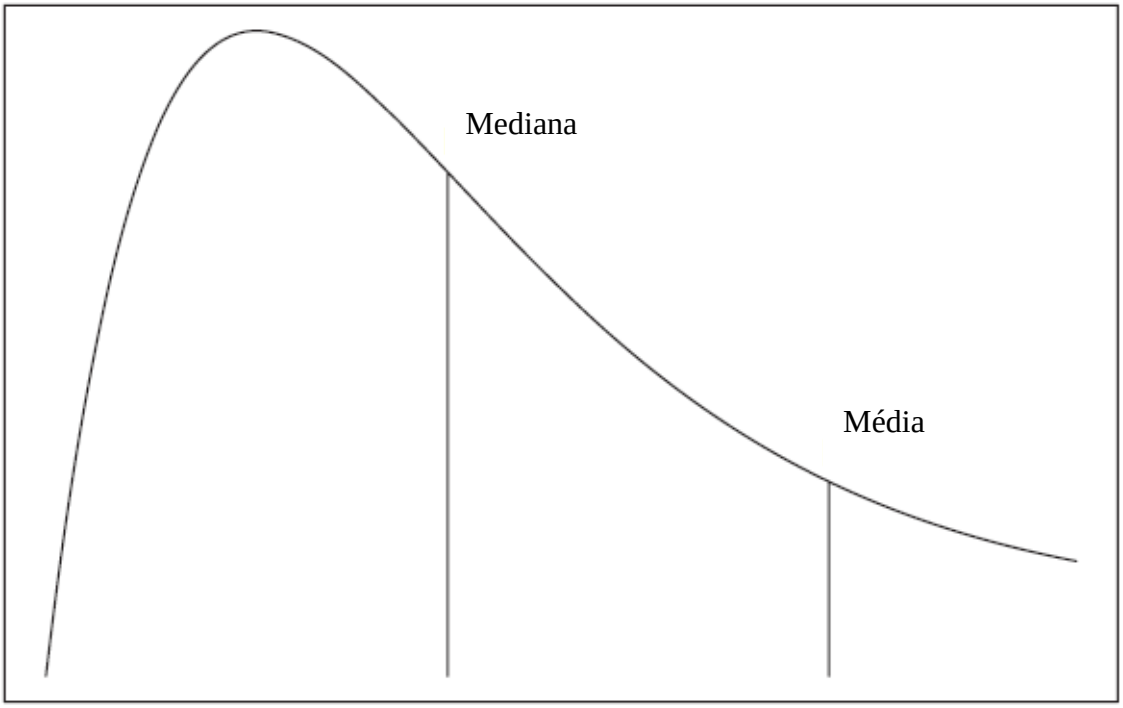
A utilização da média aritmética como melhor representação de um conjunto de dados provenientes de amostras radioecológicas, principalmente de regiões anômalas como as dos municípios de Pedra e Venturosa, é imprópria devido aos valores discrepantes encontrados (SILVA, 2006). Os estatísticos justificam a utilização da mediana como um valor mais apropriado para representar o conjunto de dados já que essa medida não sofre a influência dos valores extremos. Outra técnica sugerida é a utilização da média geométrica dos logaritmos dos valores amostrais, que reduz a influência dos valores discrepantes (WILCOX, 2009).

O grande problema da aplicação da mediana ou da média geométrica dos valores transformados em logaritmos ao conjunto de dados amostrais radioecológicos é que este tende a uma distribuição de probabilidade assimétrica em decorrência da presença de discrepâncias, consequência da elevada dispersão dos dados experimentais obtidos. Sendo assim, os valores dessas duas medidas de tendência central se apresentam sempre menores que a média aritmética simples, não convergindo então para um valor procurado mais representativo do conjunto de dados amostrais (SILVA, 2006). O gráfico de uma distribuição assimétrica à direita está representado na Figura 3 no qual a média aritmética simples e a mediana diferem bastante.

Outro fator importante que limita a utilização da mediana e da média geométrica dos logaritmos na obtenção de um valor procurado de medida central para o conjunto de dados

radioecológicos é o fato de que os valores dessas estatísticas excluem ou discriminam os valores extremos (WILCOX, 2009). Essa característica gera perda de informações durante a análise estatística dos dados (SILVA et al., 2007).

Figura 3 - Distribuição assimétrica à direita com a representação da mediana e da média aritmética simples.



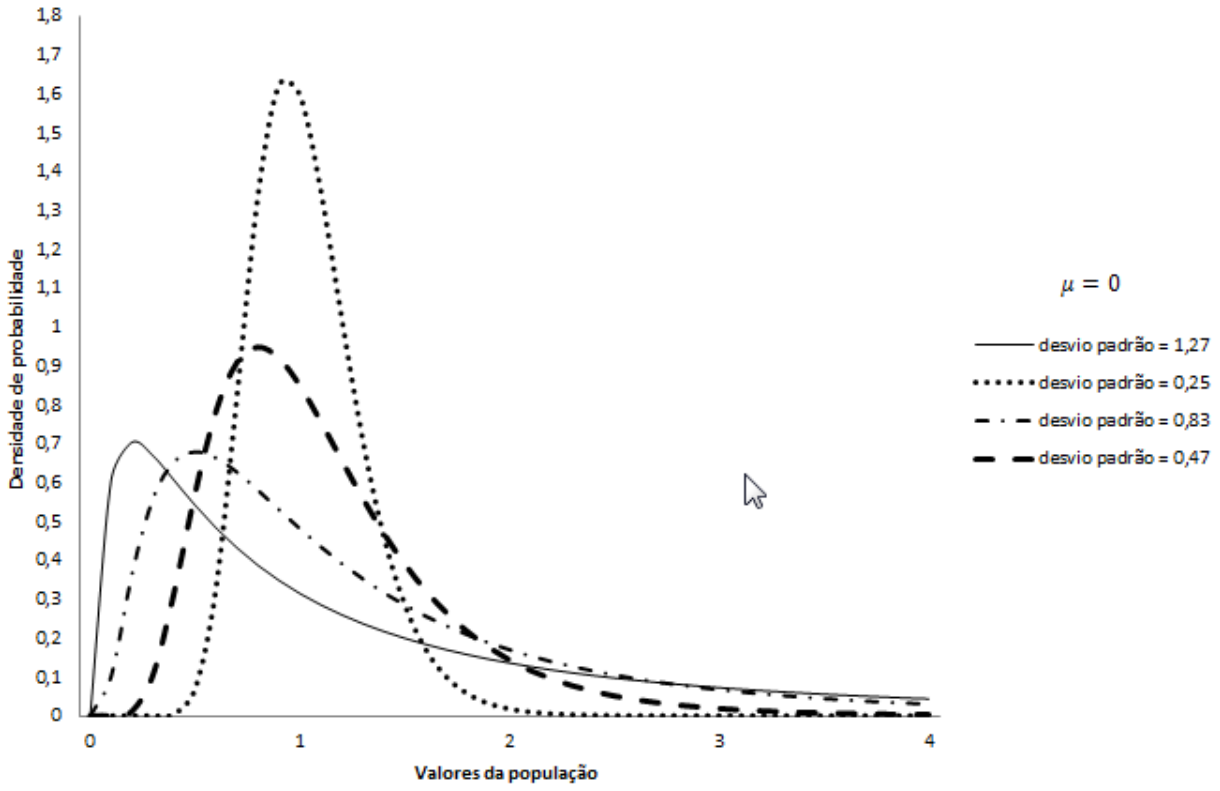
Fonte: WILCOX (2009)

O fato dos dados radioecológicos apresentarem uma distribuição de probabilidade assimétrica faz com que empiricamente a distribuição lognormal se adeque bem para representá-los estatisticamente. A lei do efeito proporcional, teoricamente, parece ser uma justificativa razoável para utilização da distribuição lognormal para um modelo de dispersão dos radionuclídeos no meio ambiente. Assim, uma variável aleatória qualquer deve, em qualquer momento de sua alteração, ter seu valor atual como uma proporção aleatória do valor anterior. A demonstração matemática da lei é complexa, porém já foi demonstrado que a distribuição assintótica dessa variável tem comportamento lognormal, justificando a sua utilização para dados ambientais. Mesmo assim, é o empirismo que tem determinado a utilização da distribuição lognormal como uma das primeiras opções na representação estatística dos dados ambientais (BLACKWOOD, 1992).

Eberhardt e Gilbert (1980) observaram que a distribuição de frequência de dados de elementos transurânicos era sempre fortemente assimétrica, sugerindo o uso da transformação

logarítmica desses dados antes de realizar qualquer análise estatística. Para Dennis e Patil (1988) a distribuição lognormal é a forma ideal de descrever dados de estudos ecológicos e de radionuclídeos no meio ambiente. As características que permitem essa afirmação são: intervalo positivo, assimetria à direita, uma longa calda à direita e a existência de expressões para o cálculo das estimativas dos parâmetros estatísticos. Observam-se na Figura 4 quatro gráficos de distribuições lognormais sobrepostos, com valores de média aritmética simples igual a zero, porém com diferentes desvios padrões. A função de densidade de probabilidade que define a curva da distribuição lognormal está representada na Equação 1.

Figura 4 - Gráficos de quatro distribuições lognormais com média aritmética simples igual a zero e diferentes desvios padrões.



Fonte: BLACKWOOD (1992)

$$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma\,x} e^{\left(-\frac{(\ln x-\mu)^2}{2\sigma^2}\right)} \quad \text{para } x > 0 \qquad (1)$$

2.3 O método de reamostragem bootstrap na análise estatística dos radionuclídeos naturais.

Problemas de otimização e integração surgem frequentemente na inferência estatística. Na realidade, a impossibilidade de calcular analiticamente estimadores associados a um determinado paradigma como, por exemplo, o de máxima verossimilhança, ou da estatística bayesiana, ou ainda do método dos momentos, faz com que soluções através de simulações numéricas geralmente sejam procuradas qualquer que seja a inferência estatística desejada (ROBERT; CASELLA, 2010).

Soluções numéricas, através de simulação, para análise de dados da área de radioecologia, exigem a geração de números aleatórios para criar distribuições que possam ser usadas no cálculo das quantidades desejadas. A possibilidade de gerar grande quantidade de valores para variáveis aleatórias, de acordo com uma dada distribuição, permite a obtenção de resultados assintóticos da inferência clássica, confirmados através da aplicação da Lei dos Grandes Números e do Teorema do Limite Central (ROBERT; CASELLA, 2010).

O termo bootstrap está relacionado a uma situação romanesca de solução quase impossível, mas que de forma inverossímil foi resolvida. A interpretação estatística do termo é passar a ideia de que em situações de alta complexidade, talvez seja possível encontrar soluções simples que aparentemente pareceriam impossíveis de serem executadas. Em Estatística, situações difíceis podem levar o pesquisador a se confrontar com problemas de soluções analíticas complexas (SILVA et al.; 2011). Chernick (2007) afirma que bootstrap é o método de reamostragem para realizar inferência estatística em situações analíticas complexas e Cohen e Cohen (2008) afirmam que o método bootstrap deve ser utilizado quando não se conhece a distribuição subjacente aos dados e quando o número de amostras é pequeno.

A necessidade de fazer uma inferência em amostras pequenas, principalmente quando a população não está disponível para obtenção de novas amostras, é uma situação complicada para qualquer pesquisador. Para essa situação alguns resultados analíticos já foram obtidos através da expansão de Edgeworth, porém a complexidade do método gera grande dificuldade para o pesquisador (HALL, 1992). Com o uso sistematizado de ferramentas computacionais, outra solução para esse caso é obtida substituindo-se a resolução analítica pelo poder de processamento dos computadores através do método de reamostragem bootstrap proposto por Bradley Efron em 1979 (DAVISON; HINKLEY, 1997).

A utilização do método bootstrap exige a produção de pseudoamostras utilizando o processo de escolha aleatória de números inteiros, conhecido como método Monte Carlo.

Dimov (2008) afirma que o método Monte Carlo funciona como um conjunto de métodos de aproximação de solução de problemas de matemática computacional utilizando o processo de sortear números inteiros. O método sempre produz uma aproximação da solução. Logo, a qualidade do processo de aproximação do valor real depende da taxa de convergência do algoritmo utilizado e que pode ser controlada pela função de erro do método. Algoritmos baseados nesse método calculam estimativas estatísticas através de amostragem aleatória de uma variável estudada.

A inferência estatística tem como objetivo estabelecer as propriedades de uma população a partir da análise de dados amostrais aleatórios provenientes da própria população. É possível afirmar que a inferência estatística procura estimar características da distribuição de probabilidade da população através de uma amostra aleatória. O método bootstrap é um procedimento de reamostragem computacional desenvolvido de forma a fornecer uma medida de uma inferência estatística mais segura, como a média aritmética, a mediana, a moda, a variância, o desvio padrão e o intervalo de confiança, utilizando uma amostra de tamanho finito baseada na distribuição empírica dos dados (LUCIO, et al., 2006).

Uma situação de grande dificuldade para o pesquisador da área de radioecologia é o surgimento de valores discrepantes de atividade nas amostras de solo e rocha contendo radionuclídeos naturais. Isso pode ocorrer devido a três importantes fatores: anomalias naturais do terreno, erros na obtenção e no tratamento das amostras, ou ainda erros no cálculo das respectivas incertezas (SILVA, 2006).

Vários métodos, como o da limitação dos pesos estatísticos relativos, o bayesiano modificado e o de Chechev-Egorov, já foram propostos com o objetivo de obter um valor de medida de tendência central mais apropriada para um conjunto de dados discrepantes, porém todos levam em consideração que esses dados são decorrentes de erros nos cálculos dos desvios, e não alteram e nem recalculam o valor de medida de tendência central, apenas modificam os valores das incertezas através da aplicação de fatores apropriados (HELENE; VANIN, 2002).

Helene e Vanin (2002) aplicaram os métodos citados no parágrafo anterior, além do bootstrap, em medições dos tempos de meia-vida e incertezas obtidos experimentalmente com valores discrepantes e observaram que todos apresentaram resultados considerados coerentes e aceitáveis quando aplicados ao conjunto completo de dados experimentais. Contudo, quando aplicado o teste de auto consistência de Rajput e MacMahon (1992), que consiste em analisar em separado dois subconjuntos dos dados amostrais e calcular a média dos resultados obtidos, nos métodos utilizados, apenas o bootstrap apresentou valores consistentes. Demonstrou-se

então que quando aplicados ao conjunto completo de estimativas iniciais já calculadas, todos os métodos apresentaram resultados confiáveis, porém se aplicados a subconjuntos de dados, apenas o bootstrap mostrou-se consistente. A utilização do método bootstrap na análise de dados discrepantes apresentou as seguintes características: redução real de dados, estabilidade diante dos dados discrepantes e obtenção de informações a partir dos dados anômalos sem nenhum pressuposto em relação à distribuição de probabilidade (HELENE; VANIN, 2002).

No caso da análise estatística de dados radioecológicos, as situações difíceis, de que trata o termo bootstrap, encontram-se ainda relacionadas com a distribuição dos valores de concentrações de radionuclídeos naturais existentes no meio ambiente, pois essa distribuição em locais tipicamente anômalos possui elevada assimetria à direita quando plotados em um sistema de eixos ortogonais, fazendo com que a curva de densidade de probabilidade tenha uma longa calda convergindo para o eixo horizontal positivo. A calda a direita pode ainda apresentar picos devido a valores discrepantes existentes (SINGH et al., 1997). É possível que o pesquisador utilize as melhores técnicas de amostragem, com extremo cuidado na obtenção, no tratamento e na análise das amostras. Porém, as amostras podem fornecer um conjunto de dados no qual a concentração dos contaminantes, aferida através das atividades dos radionuclídeos, varia desde valores de fundo até valores considerados anômalos (SILVA, 2007).

O método bootstrap, quando aplicado na reamostragem dos dados originais obtidos, fornece uma média aritmética resistente às flutuações causadas pelos efeitos dos valores anômalos. Neste caso, a reamostragem é utilizada para diminuir a assimetria, acomodando os valores de tal maneira, que a discrepância em torno da média aritmética simples passa a ser a menor possível (EFRON; TIBSHIRANI, 1993).

Existem vários métodos de reamostragem que calculam estimativas a partir de repetidas amostras utilizando o conjunto de dados amostrais originais. Alguns dos mais discutidos são os testes de permutação ou aleatorização, validação cruzada, jackknife e bootstrap. Os dois últimos, jackknife e o boostrap, têm características muito semelhantes e são muito estudados e aplicados. Ambos podem ser utilizados para reduzir a tendência dos estimadores e construir intervalos de confiança para parâmetros como, por exemplo, a média aritmética simples. Os dois métodos tomam a informação da amostra e a reproduz de forma a chegar a distribuições amostrais de interesse. Eles não exigem nenhuma suposição sobre a distribuição estatística para a população subjacente, por isso são definidos como não paramétricos. A diferença entre o jackknife, apresentado por Quenouille (1956), e o bootstrap, introduzido por Efron (1982), é que o primeiro faz n estimativas de um parâmetro sempre excluindo um número de observações

a cada rodada de simulação, enquanto que no bootstrap um número B de amostras de tamanho n, com reposição, é gerado a partir do conjunto de observações iniciais (SINGH et al., 1997).

Vários esquemas diferentes de bootstrap têm sido propostos e, muitos deles, apresentam bom desempenho em uma ampla variedade de situações. O método pode ser implementado tanto na estatística não-paramétrica quanto na paramétrica, dependendo apenas do conhecimento do problema. No caso não-paramétrico, reamostra-se 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 (SILVA, 2006).

O processo de reamostragem consiste em gerar conjuntos de dados a partir da amostra original. Esses são aleatoriamente retirados e utilizados na formação de cada amostra bootstrap. Dessa forma, todo resultado depende diretamente da amostra original. A distribuição da estatística de interesse aplicada aos valores desse tipo de amostragem, condicional aos dados observados, é definida como a distribuição bootstrap dessa estatística (EFRON, 1982). Operacionalmente, o procedimento bootstrap não-paramétrico consiste na reamostragem de mesmo tamanho, com reposição dos dados da amostra original e cálculo da estatística de interesse para cada reamostra (MURTEIRA, 1990).

Efron e Tibshirani (1993) apresentaram as ideias básicas do método bootstrap, no âmbito da inferência clássica da estatística, como se segue. Com $X = (x_1, x_2, \ldots, x_n)$ sendo uma amostra aleatória obtida a partir de uma população com função de distribuição F desconhecida, seja $\hat{\theta}(x_1, x_2, \ldots, x_n)$ um estimador do parâmetro $\theta(F)$ que, como se indica, depende naturalmente de F. Seja $\hat{F}$ a função de distribuição empírica discreta associada à amostra obtida, tal que a cada valor observado x_i, onde i = (1, 2, ..., n), recebe probabilidade de ocorrência (massa probabilística) $\frac{1}{n}$. Então, o valor de $\hat{F}$ é calculado pela Equação 2, onde $\hat{F}_{(n)}(x)$ é o estimador não-paramétrico de máxima verossimilhança de F e $I(x_i \leq x)$ é a função identidade.

$$\hat{F}_{(n)}(x) = \frac{[\sum_{i=1}^{n} I(x_i \leq x)]}{n} \qquad (2)$$

Uma amostra bootstrap é uma amostra $X = (x_1^*, x_2^*, \ldots, x_n^*)$ de tamanho n obtida de forma aleatória e uniforme, com reposição, a partir da amostra original $X = (x_1, x_2, \ldots, x_n)$, também designada população bootstrap. Cada valor da amostra original pode não aparecer ou aparecer várias vezes na amostra bootstrap. Por exemplo, $x_1^* = x_8$, $x_2^* = x_5$, $x_3^* = x_{14}$, $x_4^* =$

$x_8, \ldots, x_n^* = x_5$. A notação com asterisco indica que x^* não é um novo conjunto de dados reais x, mas sim uma versão reamostrada de x. A amostra bootstrap consiste dos correspondentes membros de x, onde: $x_1^* = x_{i_1}^*$, $x_2^* = x_{i_2}^*, \ldots,$ $x_n^* = x_{i_n}^*$. O conjunto $(x_{i_1}^*, x_{i_2}^*, \ldots, x_{i_n}^*)$ representa a i-ésima amostra de tamanho n dos dados originais do conjunto X = (x_1, x_2, ..., x_n).

No método bootstrap, a média amostral calculada é denominada por $\bar{x}_i^*$ e calculada pela Equação 3. A cada procedimento de reamostragem do conjunto original X = (x_1, x_2, ..., x_n), correspondem estimadores, nesse caso as médias amostrais $\bar{x}_1^*, \bar{x}_2^*, \ldots, \bar{x}_i^*$. Assim, o estimador bootstrap da média da população é a média aritmética $\bar{x}_B^*$, calculada através da Equação 4, dos n estimadores $\bar{x}_i^*$. Então, da distribuição $\hat{F}_{(n)}(x)$ obtêm-se B amostras bootstrap de mesmo tamanho n da amostra original, como apresentada na sequência representada pela Equação 5.

$$\bar{x}_i^* = \sum_{i=1}^{n} \frac{\bar{x}_{i_n}^*}{n} \tag{3}$$

$$\bar{x}_B^* = \sum_{i=1}^{B} \frac{\bar{x}_i^*}{B} \tag{4}$$

$$\begin{aligned} x^{*1} &= [x_{1_1}^*, x_{1_2}^*, \ldots, x_{1_n}^*] \\ x^{*2} &= [x_{2_1}^*, x_{2_2}^*, \ldots, x_{2_n}^*] \\ &\vdots \\ x^{*B} &= [x_{B_1}^*, x_{B_2}^*, \ldots, x_{B_n}^*] \end{aligned} \tag{5}$$

Dessa forma, o estimador do desvio padrão da população é calculado pela Equação 6.

$$\hat{s}_B = \sqrt{\frac{1}{(B-1)} \sum_{i=1}^{B} (\bar{x}_i^* - \bar{x}_B^*)^2} \tag{6}$$

Especificamente, $\bar{x}_i^*$ pode ser substituído pelo estimador $\hat{\theta}_i$, para cada procedimento de reamostragem. A média $\bar{x}_B^*$ pode também ser substituída por $\hat{\theta}_B$, que é a média aritmética dos n

estimadores bootstrap. A diferença $\hat{\theta}_B - \hat{\theta}_i$ é o estimador do enviesamento de $\hat{\theta}$. Desse modo, o estimador do erro padrão de $\hat{\theta}$ é calculado pela Equação 7.

$$\hat{s}_B = \sqrt{\frac{1}{(B-1)}\sum_{i=1}^{B}(\hat{\theta}_i - \hat{\theta}_B)^2} \quad (7)$$

A grande vantagem do método bootstrap é que ele pode ser aplicado à praticamente qualquer estatística $\hat{\theta}$, não se limitando apenas à média $\hat{\theta} = \bar{x}$. Isso é muito importante, uma vez que para algumas estatísticas ou não existem fórmulas analíticas ou, quando existem, são complexas e aproximadas para a estimativa dos seus respectivos erros padrões. A reamostragem bootstrap tenta realizar o que seria desejável realizar na prática: repetir os procedimentos experimentais (HELENE; VANIN, 2002).

2.3.1 O algoritmo bootstrap

A técnica de reamostragem bootstrap utiliza o algoritmo Monte Carlo, onde um dispositivo gerador de números aleatórios uniforme seleciona inteiros (1, 2, 3, ... , n) e os relaciona com as posições dos elementos do conjunto original X = (x_1, x_2, ..., x_n) (EFRON, 1982). Na prática, constrói-se a distribuição bootstrap $\hat{F}$ 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 do parâmetro para a estimativa natural do parâmetro $\hat{\theta}_B(B \to \infty) \to \theta(F)$, sendo a construção do algoritmo geralmente simples. Sua convergência está garantida pela Lei dos Grandes Números, pois, os valores $(x_1^*, x_2^*, \ldots, x_n^*)$ nada mais são do que uma amostra de variáveis aleatórias independentes e uniformemente distribuídas com distribuição condicional $\hat{\theta}_B$. Assim, quando B tende a infinito, o estimador $\hat{\theta}_B$ aproxima-se do parâmetro θ (EFRON; TIBSHIRANI, 1993).

A construção do seguinte algoritmo foi sugerido por Silva (2006) utilizando o método Monte Carlo para estimar os parâmetros estatísticos na análise de dados por bootstrap:

(**1**) Da amostra experimental, sorteia-se, utilizando um gerador de números aleatórios uniforme, os n valores com reposição para formar as amostras bootstrap de mesmo tamanho da original.

(**2**) Computa-se a estatística desejada $\hat{\theta}_i$ em cada procedimento de reamostragem.

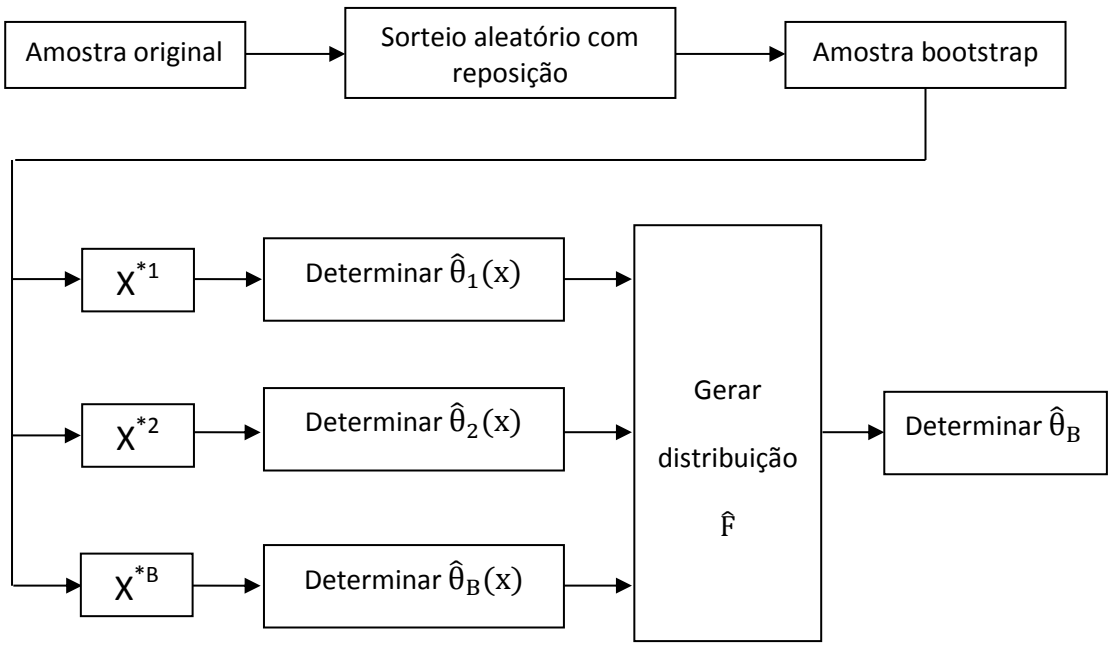
(**3**) Repete-se os passos (**1**) e (**2**) um número B de vezes, obtendo-se, dessa maneira, B valores de $\hat{\theta}_i$.

(**4**) Serão obtidas as B estimativas para formar a distribuição $\hat{F}$.

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

O valor de $\hat{\theta}_B$ pode ser o valor da média aritmética simples ou outra estatística desejada. O procedimento de simulação pode ser realizado utilizando um aplicativo computacional com um gerador de números aleatórios confiável, geralmente implementado nas linguagens de programação mais comuns. A Figura 5 mostra o diagrama de blocos para construir a distribuição bootstrap pelo algoritmo Monte Carlo.

Figura 5 – Diagrama de blocos para implementar o método Monte Carlo Bootstrap.



Fonte: SILVA (2011)

2.3.2 O número de iterações do método bootstrap

O número de iterações B define a quantidade de amostras bootstap que serão geradas e permite verificar a qualidade da estimativa do parâmetro desejado. Em vários estudos esse número foi definido de acordo com estudos experimentais de simulação. Em outras situações, pesquisadores formalizaram expressões teóricas utilizando expansão de Edgeworth, redução de

variância, aproximações lineares e variáveis antitéticas. Porém, qualquer microcomputador atual permite de forma simples e sem um grande custo computacional, a rápida execução de 100.000 ou mais repetições, facilitando a eficiência da simulação e, consequentemente, a definição do número de iterações adequado para o cálculo do parâmetro pretendido (CHERNICK, 2007).

2.4 Aplicativo computacional R

O aplicativo R é um ambiente estatístico computacional de uso livre disponível na rede Internet (BATES et al., 2012). Caracteriza-se por envolver um sistema planejado, completo, dinâmico e coerente, com disponibilidade de armazenar, manipular e representar graficamente dados, além de permitir a realização de cálculos simples até aqueles de alta complexidade.

O programa também possui uma linguagem orientada a objetos que permite desenvolver soluções automatizadas para problemas estatísticos. A execução de programas em R é realizada através de um interpretador, ou seja, após a digitação de uma ou mais linhas de comando e o pressionamento da tecla de entrada do comando do computador, as instruções são imediatamente interpretadas e os comandos são executados.

O aplicativo foi desenvolvido em versões para funcionamento nos computadores que utilizam os sistemas operacionais Linux, Windows® ou MacIntosh®. O objetivo principal do aplicativo R é desenvolver soluções para problemas de estatística aplicada e de estudos científicos. Por esse motivo, o usuário encontrará as técnicas da estatística clássica e também os modernos avanços desenvolvidos através de pesquisas mais recentes (COHEN, Y.; COHEN J., 2008).

O fato de ser um aplicativo livre faz com que qualquer usuário possa utilizar, modificar, desenvolver e submeter novas implementações, que são chamadas de “pacotes”, ao comitê técnico mantenedor. Este é formado por um corpo técnico-científico da Fundação R, organização sem interesse econômico e mantida por centenas de pesquisadores e instituições de desenvolvimento e pesquisa. O comitê avalia as submissões e as disponibiliza através da rede Internet (WILCOX, 2009; BATES et al., 2012).

O aplicativo R não é uma ferramenta comercial de utilização intuitiva, porém já existe um vasto material de apoio tanto para o uso das ferramentas estatísticas, como para a linguagem de programação. O programa é fornecido com uma interface voltada para a digitação através de linhas de comando, o que faz com que o usuário iniciante tenha uma dificuldade maior no

processo de aprendizagem, necessitando de mais tempo para se tornar hábil no programa (DALGAARD, 2002).

O R pode ser considerado um dos melhores programas existentes na área de estatística. O programa é distribuído de forma livre, apresenta todos os métodos da estatística clássica, além dos métodos considerados mais atuais e modernos. Uma das principais características que determinam a qualidade do programa em termos da pesquisa científica é a existência de uma equipe acadêmica, trabalhando em diversas universidades e centros de pesquisas, procurando desenvolver e adicionar constantemente novas técnicas através de rotinas computacionais, chamadas de "pacotes", ao sistema (WILCOX, 2009; BATES et al., 2012).

2.4.1 O pacote boot

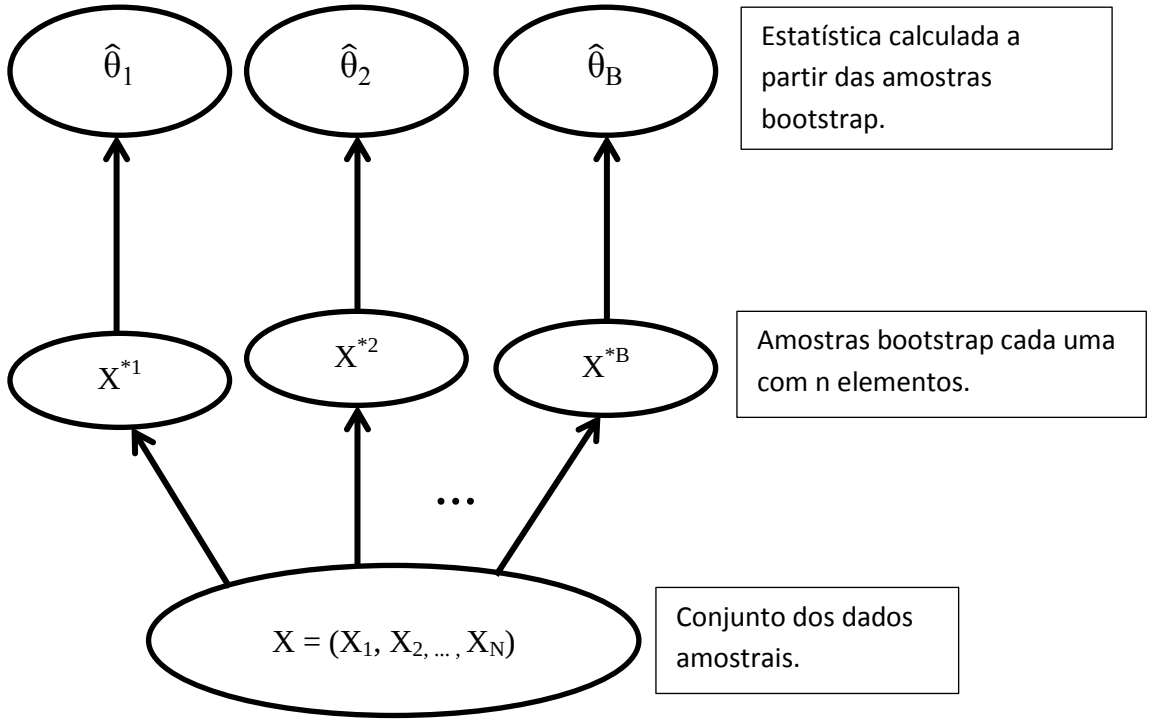
O pacote boot é formado por um conjunto de funções computacionais desenvolvidas por Angelo Canty, professor da Universidade McMaster em Ontário, Canadá, e publicado por Davison e Hinckley (1997) e, posteriormente, adaptado ao aplicativo R por Bryan Ripley, professor de Estatística Aplicada da Universidade de Oxford, Inglaterra. A versão mais recente do código foi disponibilizada em 16 de janeiro de 2012, depende dos pacotes stats e graphics, e funciona na versão 2.14.0 ou superior do aplicativo R. O aplicativo e o pacote boot estão disponíveis no sítio do Comprehensive R Archive Network (CRAN – Rede Abrangente de Arquivos R) sob licença Berkeley Software Distribution (BSD – Distribuição de Programas Berkeley) que permite basicamente o uso e a cópia sem limitações (McKUSICK, 1999).

A principal função do pacote boot recebeu este mesmo nome e utiliza o método bootstrap não paramétrico ou paramétrico para gerar uma estimativa estatística dos valores amostrais informados. Para o bootstrap não paramétrico é possível realizar a reamostragem através dos métodos: bootstrap ordinário, bootstrap balanceado, reamostragem antitética e permutação. A Equação 8 exemplifica a forma de utilizar a função de maneira simples e padrão, onde data é um vetor contendo os dados fornecidos e que serão reamostrados, statistic é uma função que ao ser aplicada ao vetor data retorna um vetor com a amostra bootstrap, R é o número de amostras bootstrap desejadas, as reticências representam argumentos opcionais adicionais da função (DAVISON; HINKLEY, 1997).

$$\text{boot (data, statistic, R, …)} \tag{8}$$

A Figura 6, adaptada de Efron e Tibshirani (1993), resume o funcionamento da função boot. O processo começa com a produção de um grande número de amostras bootstrap independentes definido na função pelo argumento R.

Figura 6 – Esquema representativo da função boot para estimar uma estatística.



Fonte: EFRON; TIBSHIRANI, 1993.

As amostras bootstrap têm o mesmo número de elementos do conjunto X dos dados medidos. A quantidade B de amostras geradas, que corresponde ao número de iterações do método Monte Carlo (CHERNICK, 2007) e ao argumento R, pode ser testada empiricamente para cada estatística estudada. Efron e Tibshirani (1986) sugerem que números típicos para a estimativa do erro padrão da média aritmética simples devem ser definidos entre 50 e 200. Após a geração das amostras, a Figura 6 mostra que para cada uma delas a função boot calcula a estatística $\hat{\theta}$ desejada gerando uma distribuição de valores normalmente denominada de distribuição bootstrap (EFRON; TIBSHIRANI, 1986).

Os valores resultantes obtidos formam, no aplicativo R, um objeto em forma de lista. Este é um conjunto que pode conter vários objetos diferentes como vetores e matrizes. O único item obtido dessa lista, quando a função é utilizada na sua forma mais simples, é a distribuição bootstrap, ou seja, um número B de uma estatística representada pelo argumento “statistic” de

cada reamostra bootstrap. Pela Lei dos Grandes Números e pelo Teorema do Limite Central quanto maior o número de reamostras bootstrap, mais próximo fica o valor do parâmetro desejado da população bootstrap, limitado pelos erros sistemáticos que surgem na geração das pseudoamostras (EFRON; TIBSHIRANI, 1993, 1986; EFRON, 1982).

2.4.2 O gerador de números aleatórios do aplicativo R

A geração de amostras bootstrap é realizada pela escolha aleatória com repetição de n números inteiros com probabilidade $\frac{1}{n}$. Essa ação é executada por um gerador de números aleatórios programado no microprocessador do equipamento utilizado. Porém, máquinas geram sequências que se repetem, ou seja, são sequências previsíveis que são denominadas de sequências de números pseudoaleatórios. Considerando essa informação, é possível afirmar que no nível computacional uma sequência de números é aceita como aleatória, se o algoritmo gerador for distinto e não relacionado com o algoritmo que utilizará os números produzidos (PRESS, et al., 1992).

Define-se números aleatórios como aqueles que estão dentro de um intervalo real, normalmente entre 0 e 1, não são previsíveis e formam uma sequência cuja função de densidade de probabilidade é constante (PARK; MILLER, 1988). Os principais programas de computador para a área científica como o R, o S-Plus, o SAS e o MatLab apresentam funções que produzem números randômicos. Essas funções produzem os números através de um algoritmo gerador que utilizam métodos matemáticos computacionais. O aplicativo R permite a geração de sequências de números aleatórios através da função “runif” que tem a sintaxe apresentado na Equação 9, onde n é o número de elementos da sequência, min é a função que determina o valor mínimo da sequência e max é a função que determina o valor máximo da sequência (COHEN, Y.; COHEN, J., 2008).

runif(n, min, max) (9)

Para gerar a sequência, a função utiliza um valor “semente”, um valor inteiro positivo para iniciar o gerador de números aleatórios, que pode ser controlado através do comando set.seed (definir semente). O valor do argumento semente é um número inteiro e sempre alterado automaticamente pelo próprio R quando uma nova sequência é gerada, fazendo com que o usuário não precise se preocupar com a definição desse argumento (CHAMBERS, 2008).

A função boot, a principal função do pacote de mesmo nome, utiliza o mesmo gerador de números randômicos da função “runif”. O aplicativo R utiliza para gerar os números aleatórios o algoritmo “mersene twister” que foi desenvolvido por Matsumoto e Nishimura (1988). Uma das principais características desse algoritmo é o seu período, tempo em que ocorre a repetição de uma sequência já gerada, da ordem de $2^{19.937}$- 1, além de uma precisão de 32 bits e uma equidistribuição 623-dimensional, e que faz o algoritmo ser considerado um dos mais rápidos e confiáveis (PARK; MILLER, 1988).

3. MATERIAL E MÉTODOS

3.1 Dados utilizados na pesquisa

O grupo de pesquisa RAE do PROTEN do Departamento de Energia Nuclear da UFPE tem como área de estudo principal as atividades dos radionuclídeos naturais de regiões anômalas no meio ambiente.

Dissertações, teses e trabalhos científicos já foram publicados e os dados amostrais obtidos são normalmente atividades específicas de radionuclídeos, em unidades de $mBq.kg^{-1}$ ou $pCi.L^{-1}$, provenientes de diversos tipos de materiais bióticos e abióticos, como: leite, água, solo, rocha, vegetais e alimentos produzidos através de manufatura ou indústria (queijo e leite, por exemplo). Esses dados estão disponíveis em meio digital, normalmente em arquivos no formato de texto ou no formato de planilha eletrônica nos microcomputadores do grupo de pesquisa, ou ainda em tabelas incluídas nas teses, dissertações e em periódicos nacionais e internacionais.

Os dados utilizados neste trabalho, mostrados na Tabela 3, foram obtidos do estudo experimental de Silva (2007), em que 14 (catorze) amostras de palma forrageira (*Opuntia spp*) foram coletadas de 09 (nove) fazendas produtoras de leite localizadas nos municípios de Pedra e Venturosa. As amostras foram analisadas quanto a atividade específica do radionuclídeo ^{226}Ra em $mBq.kg^{-1}$. A área mostrada na Figura 7 já vem sendo objeto de estudos radiométricos desde 1970 por apresentar anomalias de urânio e tório. As fazendas foram identificadas por códigos de F1 a F9. A fazenda F7 foi a área onde a extinta empresa estatal Nuclear Brasileira S/A (NUCLEBRAS) realizou os primeiros estudos sobre a viabilidade de explorar economicamente o minério de urânio encontrado. Análises realizadas em amostras de rochas dessa fazenda encontraram concentrações máximas de U_3O_8 e de ThO_2 de 22.000 $mg.kg^{-1}$ e 100 $mg.kg^{-1}$ respectivamente. (COSTA et al., 1976, 1977).

A seleção das áreas de coleta seguiu o critério da proximidade das fazendas em relação às ocorrências de urânio. As coletas foram realizadas no mês de junho de 2002 (período chuvoso), dezembro de 2003 (período seco) e maio de 2004 (período chuvoso). Em cada local de amostragem foram coletados 5 kg de palma forrageira diretamente no cocho dos bovinos das fazendas escolhidas e armazenados em sacos plásticos.

Os dados obtidos foram provenientes da determinação da atividade específica de ^{226}Ra em $mBq.kg^{-1}$ no material coletado através do método de emanação de ^{222}Rn. Na data das análises o Laboratório de Monitoração Ambiental do DEN-UFPE fazia parte do Programa

Nacional de Intercomparação do Instituto de Radioproteção e Dosimetria (PNI/IRD), que tem como um dos objetivos padronizar, manter e disseminar as grandezas do sistema internacional referentes às medidas das radiações ionizantes (IRD, 2012; SILVA, 2007).

Tabela 3 - Concentração de ^{226}Ra nas amostras de palma das fazendas F1 a F9.

Fazenda	Ano da Coleta	^{226}Ra (mBq.kg^{-1} na MS)*
F1	2002	1.985
F1	2003	1.990
F2	2004	9.300
F3	2003	5.992
F3	2004	2.150
F4	2003	1.495
F5	2003	4.500
F5	2004	5.350
F6	2003	5.060
F6	2004	6.800
F7 (fornecida às vacas)	2004	5.400
F7 (a 30 m da ocorrência)	2004	5.500
F7 (no centro da ocorrência)	2004	25.000
F9	2004	3.000

*MS = matéria seca. Fonte: SILVA (2006).

Duas estações experimentais do Instituto Agronômico de Pernambuco (IPA), localizadas nos municípios de Arcoverde e São Bento do Uma, distantes 40 e 85 km respectivamente das principais ocorrências de urânio e tório, serviram como locais de controle. As fazendas controles foram identificadas como F10 e F11. A Tabela 4 exibe as concentrações do ^{226}Ra das amostras obtidas nas fazendas controles. Esses valores de controle não foram utilizados na pesquisa devido aos valores excessivamente altos. O fato foi explicado pela adubação da área com fertilizantes fosfatados que contêm chumbo, radônio, polônio e outros materiais radioativos.

Figura 7 - Região dos municípios de Pedra e Venturosa na qual estão localizadas as fazendas que apresentam anomalias radioativas em seus terrenos.

Fonte: SILVA (2006)

Tabela 4 - Concentração de ^{226}Ra nas amostras de palma das fazendas controles.

Fazenda	Ano da Coleta	^{226}Ra (mBq.kg^{-1} na MS)*
F10	2002	5.990
F10	2003	6.900
F10	2004	9.900
F11	2002	7.990
F11	2003	1.995

*MS = matéria seca. Fonte: SILVA (2006)

3.2 Desenvolvimento do algoritmo computacional para a simulação bootstrap

O aplicativo R oferece um ambiente de desenvolvimento computacional integrado com uma interface no formato de console que funciona de forma interativa com o usuário. Essa interface não é intuitiva e nem dispõe de recursos visuais. Para obter melhor produtividade na utilização dos recursos de desenvolvimento do algoritmo, foi utilizada a interface gráfica RStudio desenvolvida para facilitar tanto a entrada de dados através de um console, como a criação de linhas de código na linguagem de programação R.

O RStudio é um aplicativo desenvolvido por empresa de mesmo nome e disponibilizado de forma gratuita sob Licença Pública Geral (LPG). O projeto LPG desenvolve e divulga a ideia de colaboração na criação de programas de computador no qual a única exigência é que todo aplicativo sob licença LPG possa ser modificado e distribuído gerando uma nova ferramenta sob licença LPG, ou seja, deverá ser um programa de computador de uso livre para qualquer usuário (McKUSICK, 1999). Foi utilizada a versão RStudio 0.97.48-2012 para o sistema operacional Microsoft Windows®. A instalação do aplicativo requer o programa R na versão 2.11.1 ou superior. O sistema operacional utilizado foi o Microsoft Windows® 7 Profissional de 64 bits com o complemento de correções Service Pack 1.

Para realizar a análise estatística através do método bootstrap utilizando a ferramenta R, através do RStudio, foi necessário criar pequenos roteiros computacionais, com várias linhas de código, que quando executados a partir do R ou do RStudio, faz a simulação de reamostragem bootstrap, os cálculos estatísticos desejados, a criação de gráficos e a gravação dos resultados obtidos. Esses roteiros computacionais são denominados de scripts de programação e neste trabalho foi desenvolvido um script para o cálculo da estatística clássica e para o cálculo da estatística utilizando o método bootstrap. O script foi gravado com o nome de “Projeto Bootstrap” e todas as vezes que foi executado realizou as atividades programadas que foram solicitadas. O código é mostrado no Apêndice A.

3.3 Procedimentos para a análise estatística dos dados

O R tem dois modos básicos de trabalho: o interativo e o programável. O modo interativo utiliza as funções em linhas de comando. As funções são distribuídas em códigos chamados de pacotes que podem ser obtidos livremente pela internet. O modo programável utiliza conceitos

de programação orientada a objetos para o desenvolvimento de scripts e aplicativos (CHAMBERS, 2008).

O script "Projeto Bootstrap" utiliza os dados experimentais de campo, obtidos por Silva (2007), para realizar a análise estatística desejada. Neste trabalho, o autor utilizou uma única variável aleatória na formação do conjunto de dados amostrais independentes e provenientes de uma população desconhecida. Porém, é possível substituir esses dados por qualquer outro conjunto amostral com as mesmas características, ou seja, uma única variável amostral aleatória, proveniente de uma população desconhecida de concentrações de radionuclídeos naturais ou artificiais de um sistema biótico ou abiótico.

3.3.1 Análise estatística das concentrações do ^{226}Ra utilizando a inferência clássica no algoritmo desenvolvido no RStudio

Inicialmente, a atividade específica do ^{226}Ra, obtida da análise das amostras de palma forrageira (*Opuntia spp*) de uma região caracterizada por apresentar anomalias de mineralizações de urânio, foi tratada com as funções da inferência estatística clássica no aplicativo R através do ambiente integrado de desenvolvimento RStudio. Os pacotes base, stats e graphics, todos na versão 2.15.0, apresentam os principais recursos da inferência clássica utilizados nessa etapa.

Os 14 valores de concentração do ^{226}Ra obtidos por Silva (2007) foram digitados como um vetor de valores no script "Projeto Bootstrap" com o nome de dados (Apêndice A). A partir do vetor dados foram calculadas as seguintes estatísticas: valor mínimo, primeiro quartil, mediana, média aritmética simples, terceiro quartil e valor máximo, todas utilizando a função "summary" do pacote base na versão 2.15.0. Utilizando funções específicas do pacote base para cada estatística ou através de fórmulas desenvolvidas no script foram calculados também o desvio padrão para a média aritmética simples, o desvio interquartílico e a amplitude.

A presença de valores atípicos no conjunto de dados amostrais foi verificada através de dois métodos já consolidados na estatística descritiva. O primeiro utiliza a média aritmética simples $\overline{X}$, e o desvio padrão da amostra S através da Equação 6. Nesse método, qualquer valor maior ou igual a 2 será considerado um valor discrepante (WILCOX, 2009) .

$$\left|\frac{X-\overline{X}}{S}\right| \geq 2 \qquad (6)$$

O segundo método é definido como diagrama de caixa ou “boxplot” e utiliza os valores do primeiro e terceiro quartis. De acordo com esse método, um valor do conjunto amostral é considerado anômalo se uma das duas Equações 7 for verdadeira, onde X é um valor amostral, q_1 é o primeiro quartil e q_3 é o terceiro quartil.

$$X < q_1 - 1{,}5(q_3-q_1) \qquad X > q_3 + 1{,}5(q_3-q_1) \tag{7}$$

O método diagrama de caixa permite ainda uma visualização em forma de gráfico destacando os valores dos quartis, da mediana e dos valores discrepantes (WILCOX, 2009).

Um histograma com as frequências das concentrações foi criado e sobreposto com o gráfico da distribuição normal para que a forma dos dados amostrais pudesse ser exibida. Os dados numéricos foram gravados em um arquivo no formato de texto com o nome de “Rádio-226 – Estatística Clássica.txt” e os gráficos foram gravados em um arquivo no formato de “portable document format” ou pdf não editável com o nome “Gráfico Rádio-226 – Estatística Clássica.pdf”. As funções dos pacotes utilizados têm suas sintaxes e argumentos definidos no sistema de ajuda interno do aplicativo R (BATES et al., 2012). Como não existem funções prontas no aplicativo R para o cálculo da amplitude e do desvio interquartílico, essas estimativas foram calculadas através das equações disponibilizadas na literatura da estatística clássica e documentadas no próprio código (WILCOX, 2009; BLACKWOOD, 1992).

3.3.2 Análise estatística das concentrações do ^{226}Ra utilizando o método bootstrap no algoritmo desenvolvido no RStudio

O script “Projeto Bootstrap”, desenvolvido para este trabalho, utilizou o pacote boot e as funções boot e boot.ci desse pacote. A primeira é utilizada para calcular qualquer estimativa estatística desejada pelo método bootstrap; a seguinte, para o cálculo de intervalos de confiança baseado na população de amostras geradas pela função boot. Evidentemente, o intervalo de confiança para a média aritmética simples, só foi calculado após o cálculo da própria média.

Os dados amostrais, o número de reamostras e a estimativa estatística a ser reamostrada foram digitados diretamente no código do programa. As estimativas estatísticas calculadas a partir das amostras geradas foram a média aritmética simples, o desvio padrão e o intervalo de

confiança. Para o cálculo da média aritmética simples foi utilizada a função “mean” do pacote base do aplicativo R.

O número de amostras bootstrap geradas foi definido, de acordo com sugestões de Efron e Tibshirani (1986), para um valor inicial de 100 até um limite máximo de 100.000, com incremento multiplicativo de 10, ou seja, o número de reamostras foi formado pelos elementos do vetor (100, 1.000, 10.000, 100.000). Dessa forma, a primeira simulação gerou 100 amostras bootstrap, a segunda gerou 1.000, a terceira 10.000 e a última 100.000. Após a geração das reamostras, as estimativas média aritmética simples, desvio padrão, intervalos de confiança, primeiro quartil, mediana, terceiro quartil, intervalo interquartílico, valor mínimo, valor máximo e amplitude foram calculadas.

4. RESULTADOS E DISCUSSÃO

4.1 Resultados da análise estatística das concentrações do ^{226}Ra utilizando as funções da estatística clássica no aplicativo R

Os resultados, obtidos a partir dos dados experimentais registrados na Tabela 3, foram gravados em um relatório e armazenado em forma de arquivo de texto com o nome "Ra-226 Estatística Clássica.txt" e estão representados na Tabela 5. O arquivo ocupa apenas 1 kb (quilobyte) de espaço em disco e fica armazenado no diretório de instalação do programa R. Os três primeiros valores calculados: valor mínimo, valor máximo e amplitude apresentam uma visão inicial da extensão e dispersão dos dados das amostras analisadas em termos estatísticos. São valores iniciais e que podem ser facilmente calculados. Apesar de não apresentarem papel fundamental nas conclusões estatísticas obtidas, o elevado valor da amplitude permitiu inferir a necessidade de verificar a existência de valores discrepantes no conjunto.

Tabela 5 - Resultados da análise estatística das concentrações do ^{226}Ra utilizando as funções da estatística clássica no aplicativo R.

Estatísticas calculadas	Estatística clássica para atividade específica do ^{226}Ra (mBq.kg^{-1} na MS)*
Valor mínimo	1.495,00
Valor máximo	25.000,00
Amplitude	23.505,00
Média aritmética simples	5.965,86
Desvio padrão	5.903,05
Primeiro quartil	2.362,50
Mediana	5.205,00
Terceiro quartil	5.869,00
Desvio interquartílico	3.506,50

*MS = matéria seca

A rotina do programa utilizou os métodos de verificação de valores anômalos ou discrepantes através do método clássico, baseado na média aritmética simples e no desvio padrão da amostra, e também, no método do gráfico de caixa ou "boxplot", baseado nos quartis

inferior e superior. Os valores obtidos para o primeiro método estão relacionados na Tabela 6 e apenas o valor 25.000 mBq.Kg^{-1}, da fazenda F7, foi calculado como discrepante ou anômalo em relação ao restante dos dados. É possível observar também que o valor 9.300 mBq.Kg^{-1}, da fazenda F6, na Tabela 6, apesar de não ser indicado como discrepante, é um número consideravelmente alto quando comparado com os outros dados amostrais. Esse método apresentou problemas de mascarar prováveis valores anômalos, pois dependeu da média aritmética simples e do desvio padrão da amostra, ambos fortemente influenciados no cálculo por valores amostrais limítrofes.

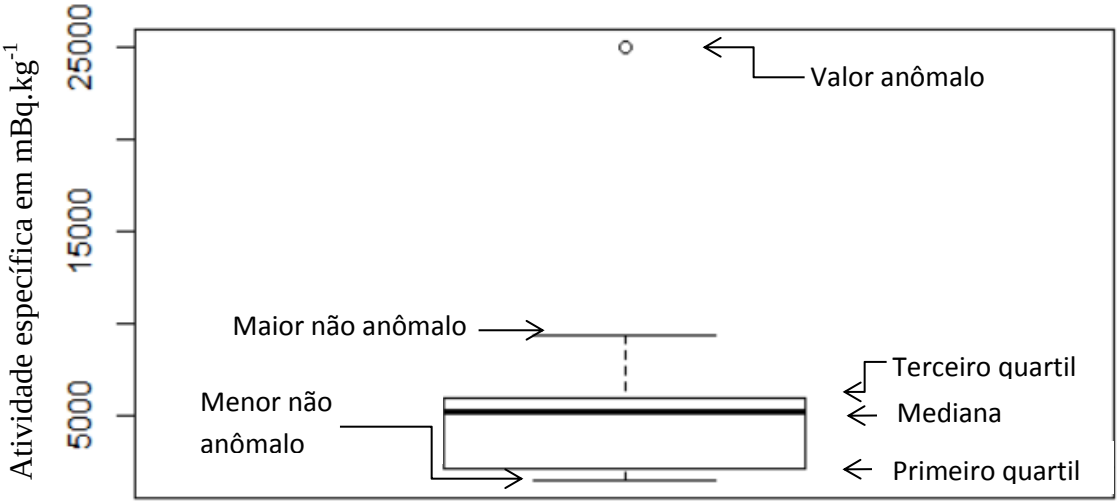
Tabela 6 - Resultados do teste de valor anômalo para os dados amostrais utilizando a média aritmética simples e o desvio padrão da amostra.

Fazenda	Atividade específica (mBq.kg^{-1})	Cálculo $\left\|\frac{X- \bar{X}}{S}\right\| \geq 2$	Resultado
F4	1.495	0,76	Não discrepante
F1	1.985	0,67	Não discrepante
F1	1.990	0,67	Não discrepante
F3	2.150	0,64	Não discrepante
F9	3.000	0,50	Não discrepante
F5	4.500	0,25	Não discrepante
F6	5.060	0,15	Não discrepante
F5	5.350	0,10	Não discrepante
F7	5.400	0,09	Não discrepante
F7	5.500	0,08	Não discrepante
F3	5.992	0,004	Não discrepante
F6	6.800	0,14	Não discrepante
F6	9.300	0,56	Não discrepante
F7	25.000	3,22	Discrepante

A utilização do segundo método através da Equação 7 mostrou que o valor abaixo do qual todos os valores seriam discrepantes foi -2.897,25 mBq.kg^{-1}, valor desconsiderado devido a própria radioatividade natural existente na crosta terrestre, ou seja, não existe radioatividade negativa. Já valores amostrais maiores que 11.128,75 mBq.kg^{-1} foram considerados

discrepantes. Nesse caso, novamente apenas o valor 25.000 mBq.kg^{-1}, da fazenda F7, na Tabela 6, foi considerado uma discrepância em relação ao restante dos dados. O método "boxplot" ainda produziu o relatório em forma de gráfico exibido na Figura 8, gravado no formato pdf (portable document format) com o nome de "boxplot_classico.pdf" no qual ficou destacado o valor encontrado como discrepante. Os resultados obtidos através desse último método são mais representativos do espalhamento dos dados, pois os cálculos dependem apenas dos quartis inferior e superior, não sendo influenciados pelos valores extremos do conjunto amostral.

Figura 8 - Representação do diagrama de caixas para análise da atividade específica do ^{226}Ra em palma forrageira (*Opuntia spp*).

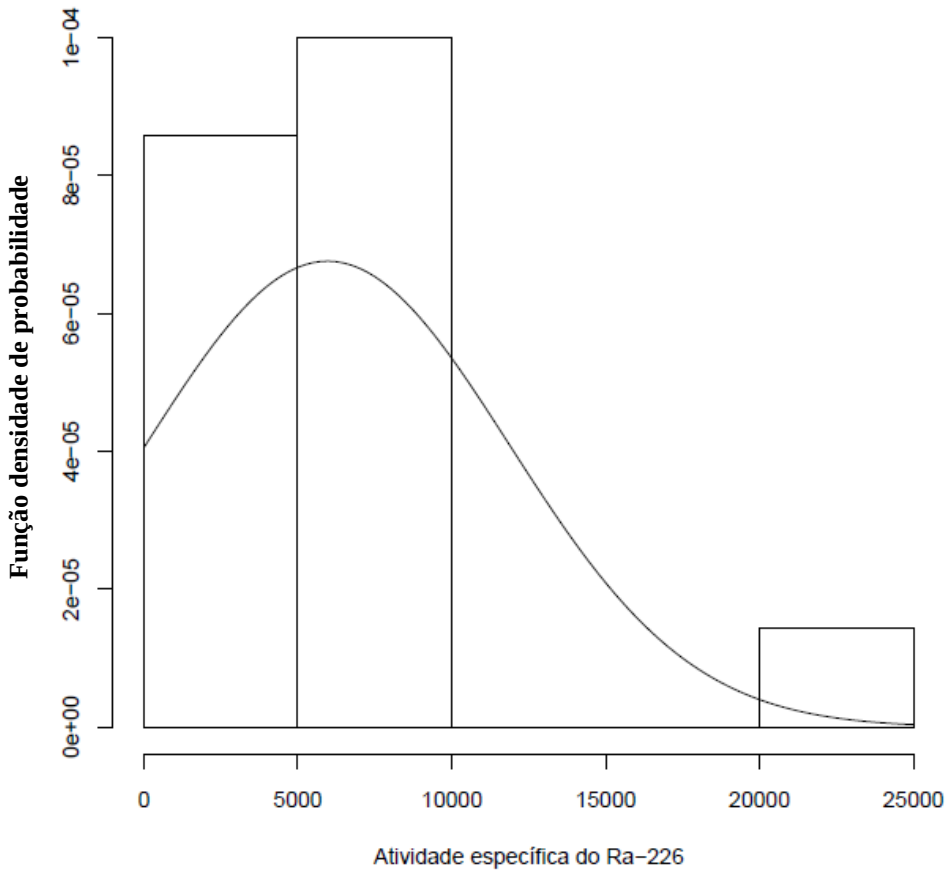


O método "boxplot" permitiu uma visão mais clara do espalhamento dos dados amostrais em relação a uma medida de tendência central do conjunto. O máximo valor não anômalo da amostra foi 9.300 mBq.kg^{-1}, na Tabela 6, porém desse valor até 11.128,75 mBq.kg^{-1} foi calculado um intervalo não considerado anômalo e não representado na amostra. Esse é um grande intervalo vazio que foi calculado pelo método "boxplot". Após esse intervalo foi que ocorreu o único valor discrepante.

O histograma de frequência apresentado na Figura 9, também produzido pelo "script" Projeto Bootstrap apresentou o intervalo de 10.000 mBq.kg^{-1} a 20.000 mBq.kg^{-1} como um grande intervalo vazio da amostra original, o que confirmou uma grande dispersão de dados.

Através do gráfico da figura 9 ficou demonstrado ainda que não é adequado considerar a distribuição normal como uma representação dos dados radioecológicos obtidos. Os valores do histograma que representam os valores radioecológicos não se adequaram a curva normal, ficando muito acima dos valores da curva. Além disso, não existiu simetria no histograma e um grande intervalo vazio apareceu entre 10.000,00 $mBq.kg^{-1}$ e 20.000,00 $mB.kg^{-1}$.

Figura 9 - Gráfico de densidade de probabilidade das amostras de ^{226}Ra em palma forrageira (*Opuntia spp*) sobreposto pela curva da distribuição normal.



O valor da média aritmética simples e do desvio padrão apresentados na Tabela 5 sofreram influência do valor máximo do conjunto de dados amostrais devido à presença de valores discrepantes na amostra. O valor do desvio padrão calculado e colocado na Tabela 5 foi de 5.903,05 $mBq.kg^{-1}$, um valor muito alto e que demonstrou novamente uma grande dispersão dos dados.

A utilização da média aritmética simples como valor mais representativo dos dados amostrais é uma tendência mundial (ARANGO, 2005 apud SANTOS JÚNIOR, 2009) e se adequa bem quando o conjunto de dados se aproxima da distribuição de probabilidade normal. Porém, em amostras provenientes de regiões com anomalias radioativas a distribuição normal não é adequada. O programa plotou a densidade de probabilidade para a amostra e sobrepôs a curva de distribuição normal construída com os valores da média aritmética simples e do desvio padrão da amostra. O resultado, apresentado na Figura 9, demonstrou que a distribuição normal não é uma opção válida para representar a distribuição estudada.

Os cálculos realizados para a estatística clássica foram executados sempre no início do script "Projeto Bootstrap", utilizando operadores e funções do aplicativo R e os resultados obtidos não estão relacionados ao método bootstrap utilizado para simular, a partir das 14 amostras iniciais, as pseudoamostras.

4.2 Resultados da análise estatística das concentrações do ^{226}Ra utilizando o método bootstrap no aplicativo R.

Para as simulações bootstrap, os resultados obtidos estão na Tabela 7 e mostram que a média aritmética simples obtida não variou de forma relevante em relação à média aritmética clássica exibida na Tabela 5. Porém, é possível observar que os desvios calculados através do método bootstrap foram bem menores que o calculado através do método clássico.

A Tabela 7 apresenta também os resultados para os intervalos de confiança bootstrap com 90% e 95% de probabilidade. Para 100 reamostras o valor da média aritmética simples foi de 6.200,49 +/- 1.359,49, ou seja, o valor calculado está no intervalo de 4.841 a 7.559,98. Então, para as 100 amostras a média aritmética simples ficou dentro dos intervalos de confiança bootstrap de 90% e 95%. Para 1.000 reamostras o valor da média aritmética e seu desvio padrão ficaram no intervalo de 4.415,02 a 7.610,68. Novamente o intervalo está dentro dos intervalos de confiança bootstrap nos dois níveis de probabilidade. Para 10.000 reamostras o intervalo da média aritmética com seus respectivos desv $^{(mBq.kg^{-1})}$ 5,46 a 7.456,13, demonstrando mais uma vez a inserção nos intervalos de confiança bootstrap calculados. Finalmente, para 100.000 reamostras, o resultado para média aritmética simples foi de 4.440,63 a 7.480,35, dentro, com as probabilidades de 90 e 95%, dos intervalos de confiança calculados pelo método bootstrap.

Tabela 7 - Resultados das simulações bootstrap para 14 amostras de ^{226}Ra em palma forrageira (*Opuntia spp*).

Nº de reamostras	100	1.000	10.000	100.000
Média aritmética simples	6.200,49	6.012,85	5.940,96	5.960,49
Desvio padrão	1.359,49	1.597,83	1.515,17	1.519,86
Intervalo de confiança 90%	(3.495; 7.698)	(3.291; 8.547)	(3.499, 8.483)	(3.471; 8.471)
Intervalo de confiança 95%	(3.067; 8.396)	(2.787; 9051)	(3.021; 8.960)	(2.992; 8.950)
Valor mínimo	3.476	2.591	2.240	2.280
Valor máximo	10.064	12.271	13.486	14.573
Amplitude	6.588	9.680	11.246	12.293
Primeiro quartil	5.226	4.794	4.769	4.778
Mediana (segundo quartil)	6.067	5.834	5775	5.801
Terceiro quartil	7.222	7.067	6.878	6.909
Desvio interquartílico	1.996	2.273	2.109	2.131

Através do valor calculado da amplitude foi possível perceber que com o aumento do número de reamostras, nas simulações executadas, o espalhamento dos dados aumenta, pois esse valor é calculado utilizando o valor máximo e o valor mínimo. Porém, o desvio interquartílico e a mediana permanecem com valores muito próximos em relação às quatro simulações executadas, mesmo com o aumento dos valores de reamostras, demonstrando que essas estatísticas são imunes aos valores limítrofes gerados nas pseudoamostras.

O método bootstrap não eliminou os valores discrepantes ou anômalos, eles continuaram a surgir, como mostram os diagramas de caixa da Figura 10 para 100 reamostras, da Figura 11 para 1.000 reamostras, da Figura 12 para 10.000 reamostras, da Figura 13 para 100.000 reamostras.

Dessa forma, considerou-se que mesmo existindo valores anômalos ou discrepantes que influenciaram no cálculo da média aritmética simples, o processo de simulação manteve as características dos dados iniciais, de maneira que informações sobre as anomalias radioecológicas não foram perdidas. Essas perdas seriam automáticas se fossem utilizadas técnicas estatísticas que desprezassem os valores anômalos calculados no cálculo da média aritmética simples ou se fosse utilizada a mediana como medida de tendência central.

Figura 10 - Diagrama de caixa das 100 reamostras bootstrap do ^{226}Ra em palma forrageira (*Opuntia spp*).

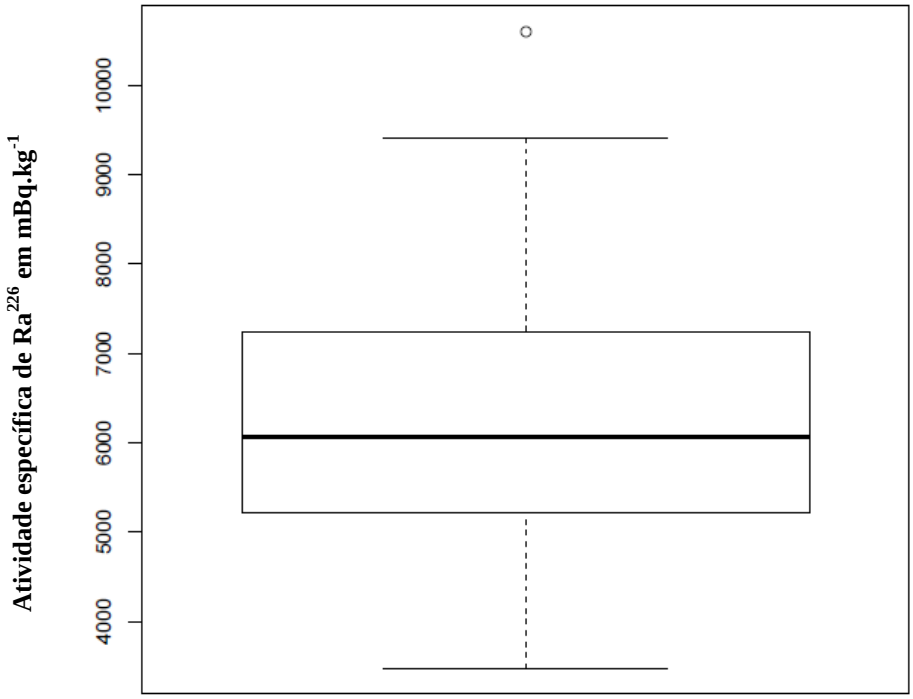


Figura 11 - Diagrama de caixa das 1000 reamostras bootstrap do ^{226}Ra em palma forrageira (*Opuntia spp*).

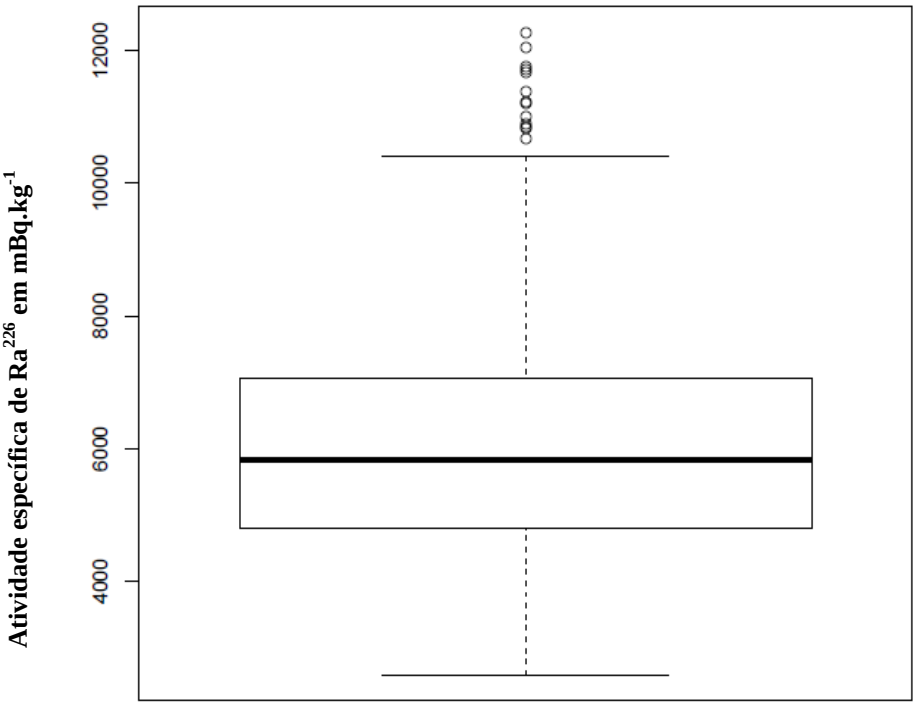


Figura 12 - Diagrama de caixa das 10000 reamostras bootstrap do ^{226}Ra em palma forrageira (*Opuntia spp*).

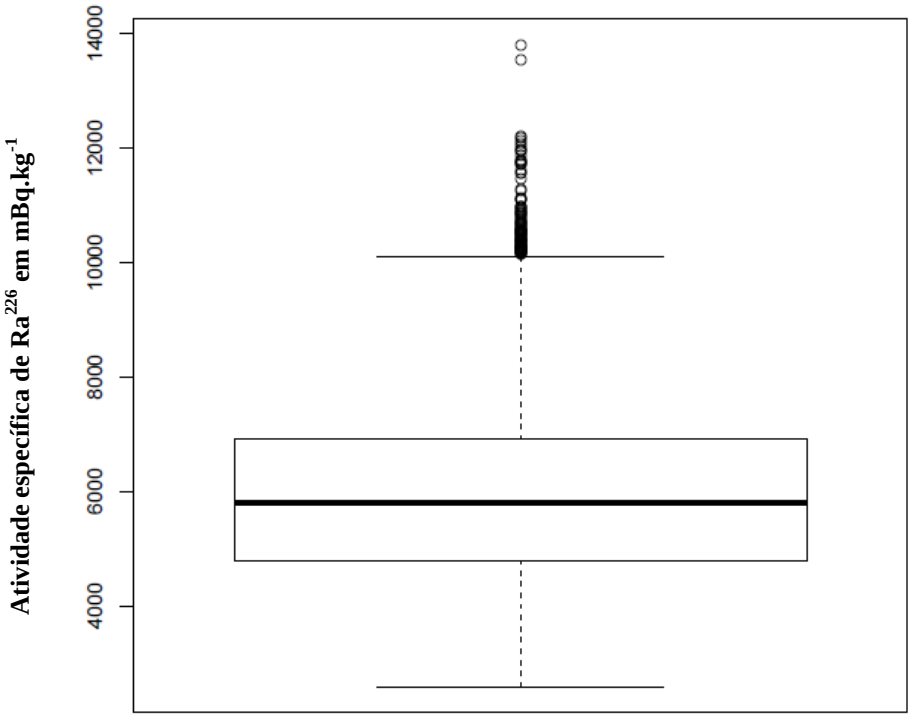
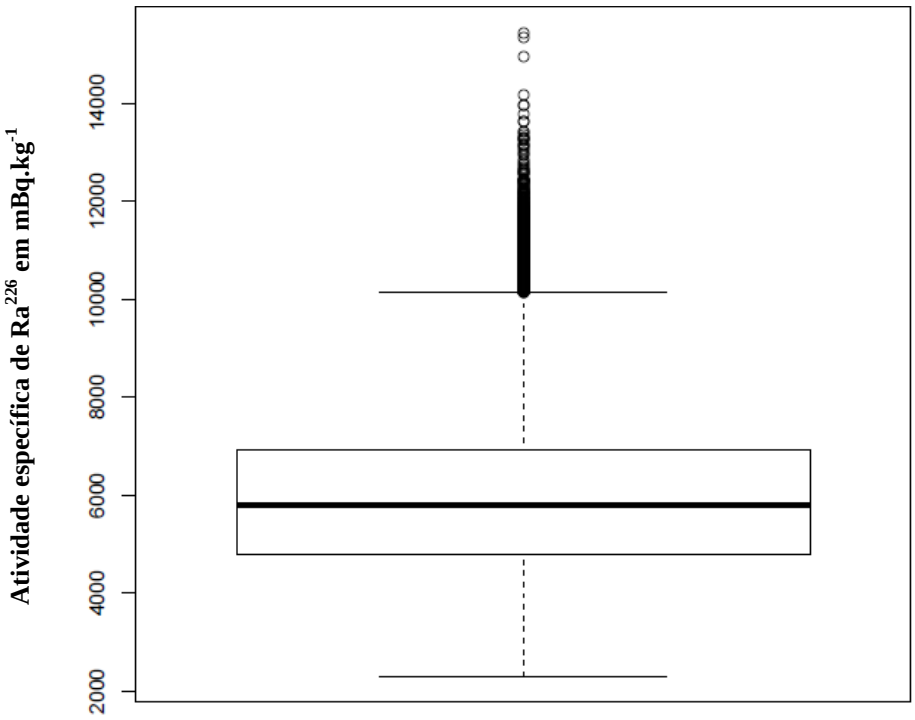
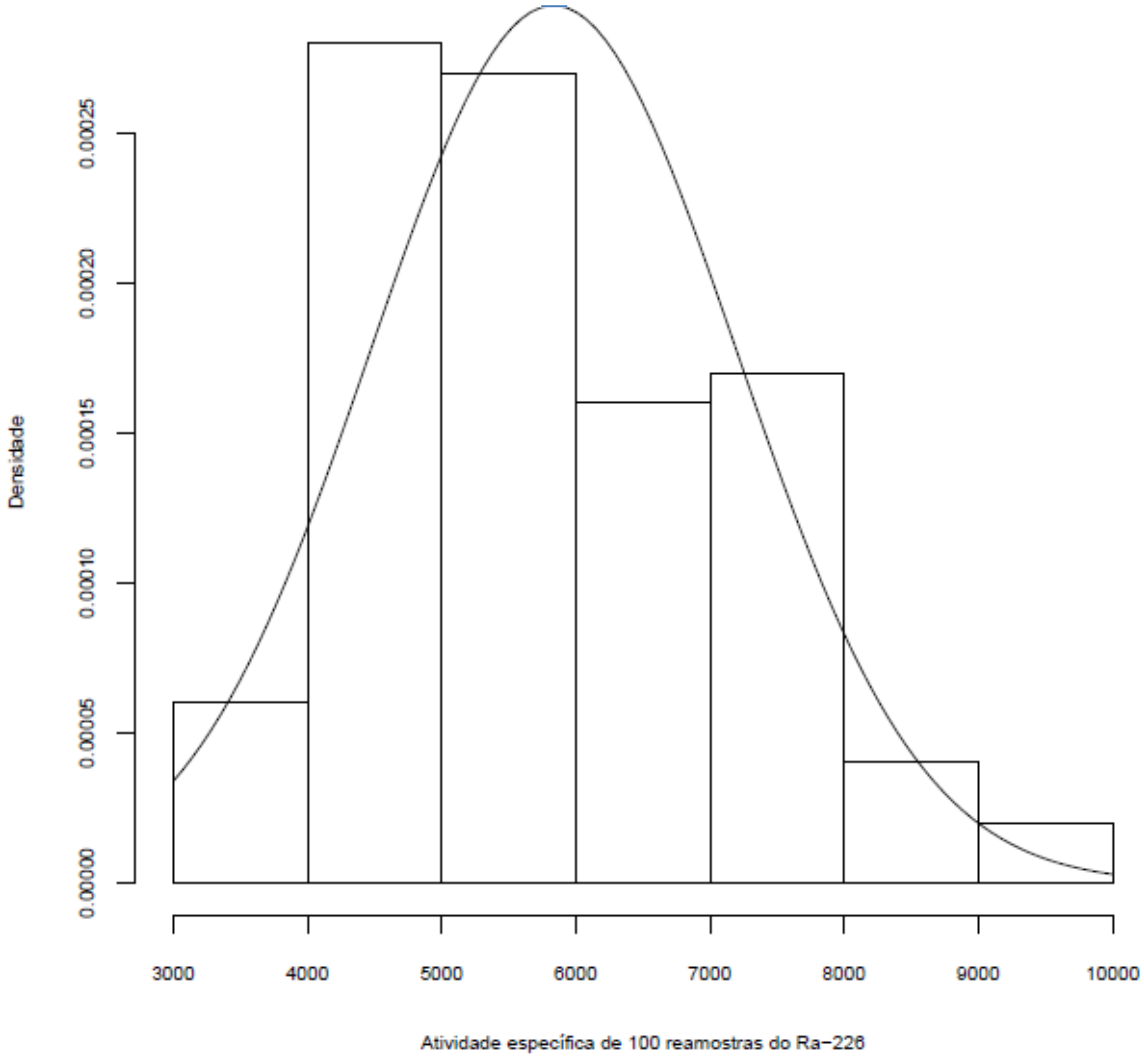


Figura 13 - Diagrama de caixa das 10000 reamostras bootstrap do ^{226}Ra em palma forrageira (*Opuntia spp*).



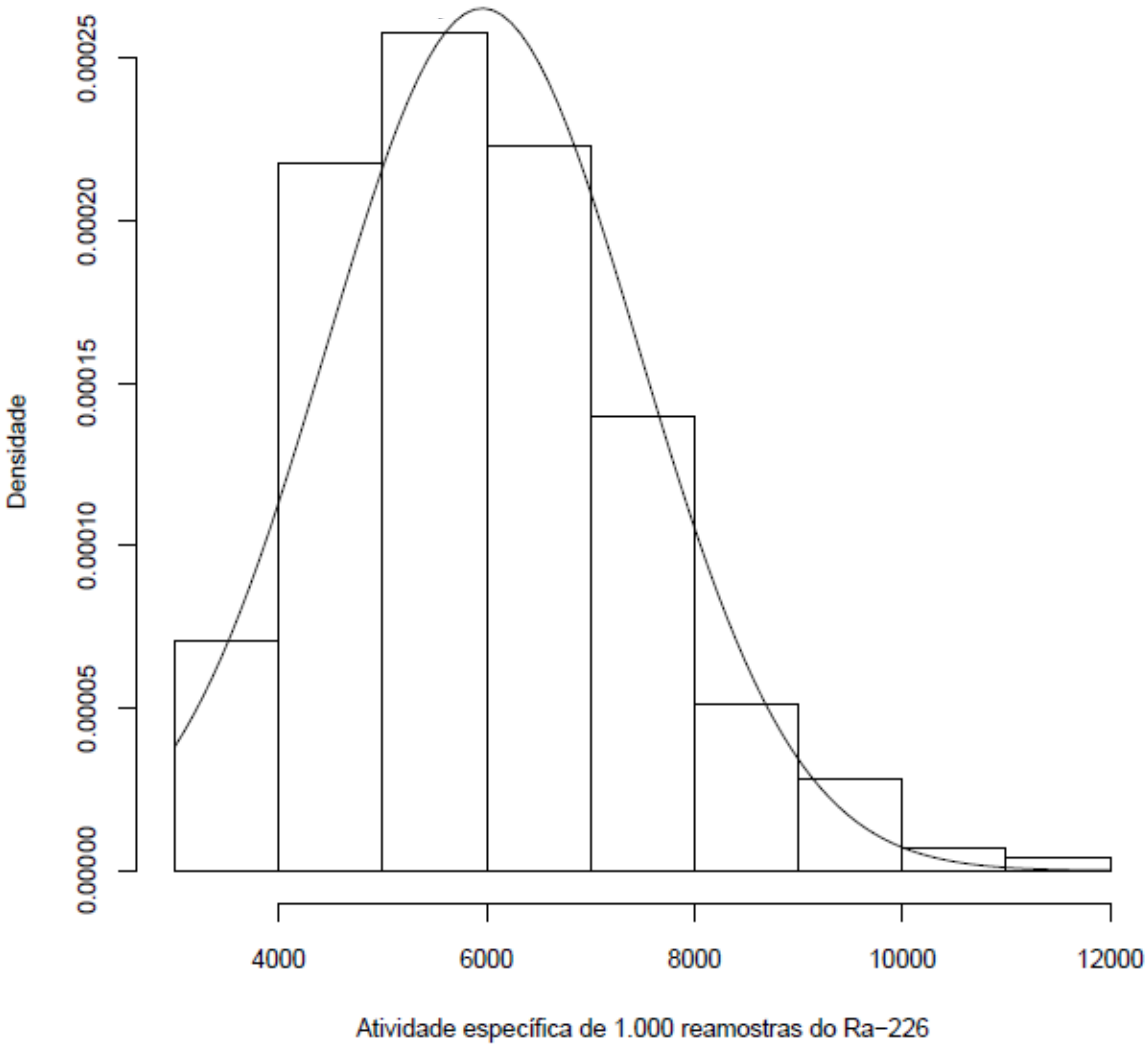
As reamostras obtidas por simulação através do método bootstrap demonstraram ainda que os valores tendem para a distribuição normal. As Figura 14, 12. 13 e 14 exibem, respectivamente, os gráficos de densidade de probabilidade das 100, 1.000, 10.000 e 100.000 reamostras bootstrap sobrepostos pela curva da distribuição normal. Os resultados demonstram a tendência à normalidade das distribuições bootstrap obtidas.

Figura 14 - Gráficos de densidade de probabilidade das reamostras do ^{226}Ra em palma forrageira (*Opuntia spp*) sobreposto pela curva da distribuição normal.



Dessa forma é possível considerar que todas as simulações fizeram com que as medidas de tendência central se aproximassem e se tornassem mais representativas das amostras geradas.

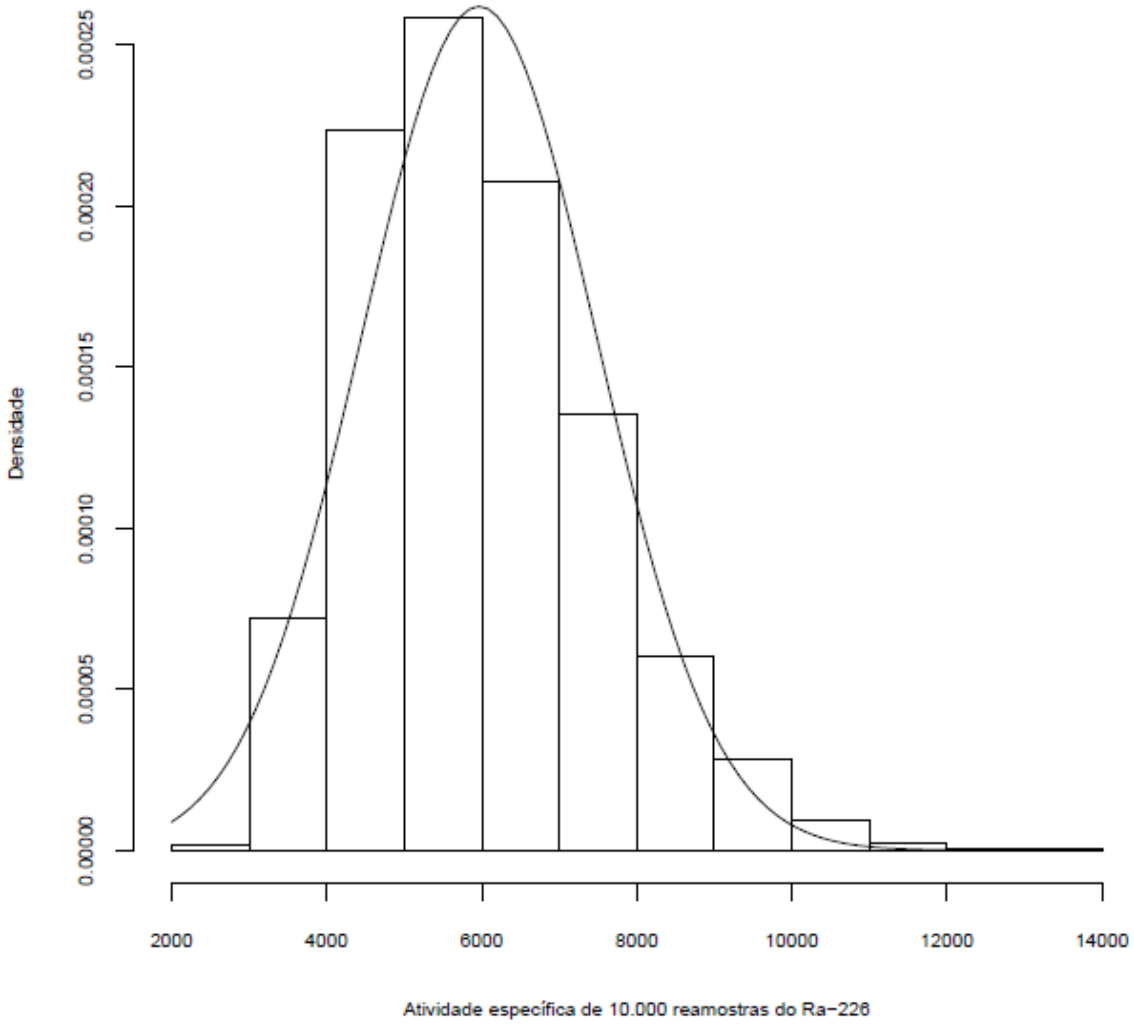
Figura 15 - Gráfico de densidade de probabilidade das 1000 reamostras do ^{226}Ra em palma forrageira (*Opuntia spp*) sobreposto pela curva da distribuição normal.



Foi observado também que não ocorreram espaços descontinuados no eixo horizontal dos gráficos, o que demonstrou que o método bootstrap simula a obtenção de amostras dentro do intervalo das amostras originais sem resultar em ausência de valores.

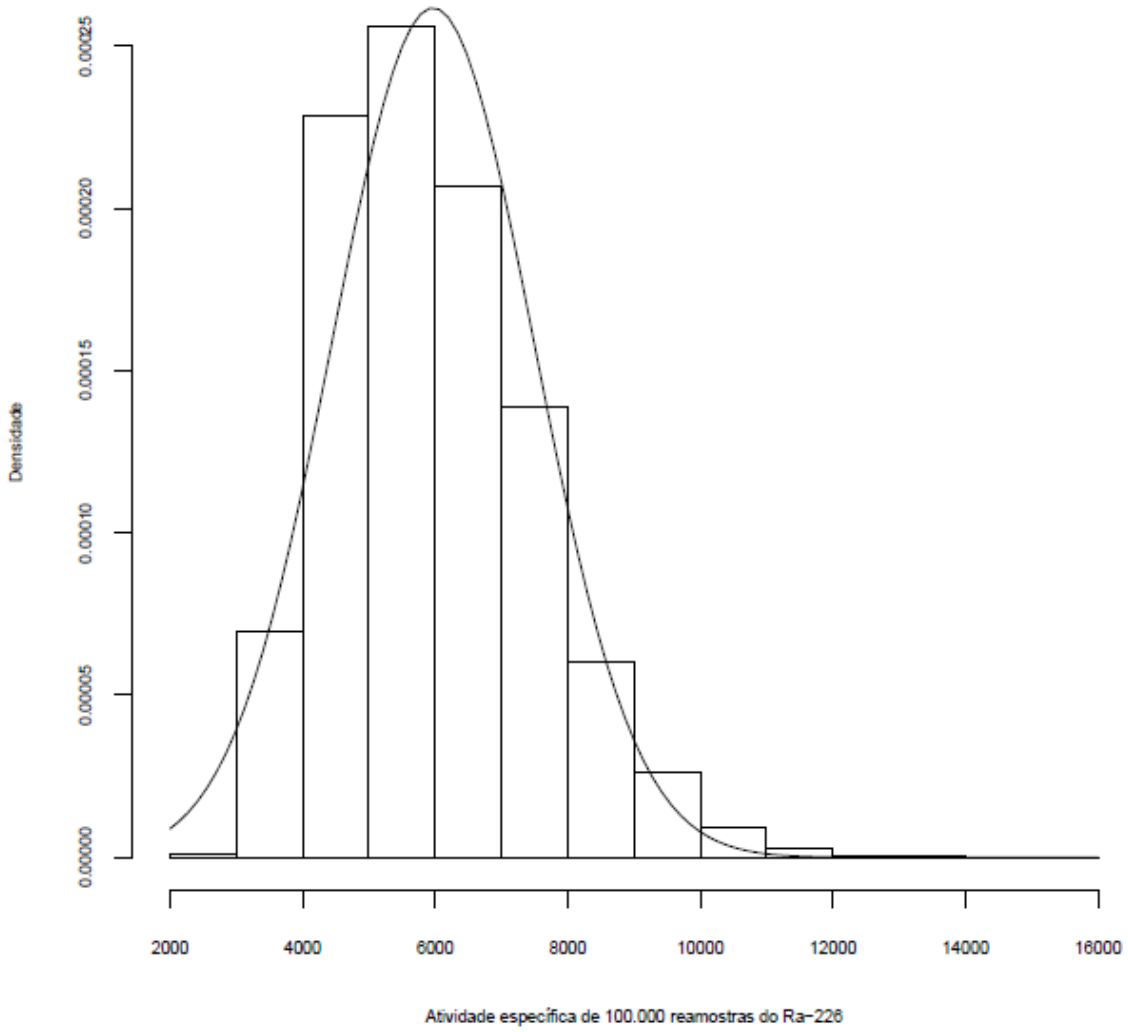
Apesar de ser utilizado apenas para a situação particular de uma amostra de 14 valores de atividade específica em mBq.kg^{-1} do radionuclídeo ^{226}Ra e gerando 100, 1.000, 10.000 e 100.000 reamostras; o "script" Projeto Bootstrap pode ser modificado para outro radionuclídeo, para um número diferente de amostras e qualquer valor de reamostras.

Figura 16 - Gráfico de densidade de probabilidade das 1000 reamostras do 226Ra em palma forrageira (Opuntia spp) sobreposto pela curva da distribuição normal.



Inicialmente, a simplicidade de uso do método bootstrap parece gerar dados sem fundamento. Porém, o que o método faz é calcular as médias aritméticas simples das reamostras de tamanho n, gerando uma população que é utilizada para estimar uma medida de tendência central amostral. Essa população é originada de uma amostra, também de tamanho n, dos dados reais obtidos da pesquisa de campo da população que se deseja estudar.

Figura 17 - Gráfico de densidade de probabilidade das 1000 reamostras do 226Ra em palma forrageira (Opuntia spp) sobreposto pela curva da distribuição normal.



5. CONCLUSÕES

Os resultados obtidos para este trabalho demonstram que para a situação de não normalidade de uma variável aleatória obtida da pesquisa radioecológica, situação considerada muito particular em termos estatísticos, o método bootstrap fornece:

- Uma média aritmética simples mais adequada que a obtida dos dados reais e que pode ser utilizada como uma medida de tendência central mais representativa para um conjunto de valores amostrais de dados radioecológicos provenientes de regiões anômalas devido à redução na dispersão dos valores;
- Uma distribuição de reamostras em que não existem grandes lacunas vazias entre os valores obtidos e que tende a distribuição normal. Além disso, os valores obtidos mostram que os valores discrepantes ou anômalos permanecem na distribuição resultante, mas não tem grande impacto no cálculo da média aritmética simples e, dessa forma, não ocorre perda de informações radioecológicas;
- Intervalos de confiança que garantem que a média da população bootstrap está no mesmo intervalo da média da amostra bootstrap;
- A possibilidade de realizar inferência em pequenas amostras, apesar de não substituir a obtenção de novas amostras.

A utilização do método bootstrap sem a utilização do "script" Projeto Boostrap, desenvolvido para este trabalho, produziria os mesmos resultados, porém demandaria um grande esforço na execução dos cálculos, provocando um período longo de espera e grande probabilidade de geração de erros operacionais devido à participação humana nos cálculos.

O "script" Projeto Bootstrap também permitiu armazenar todos os resultados em arquivos do tipo texto ou pdf, ocupando pouco espaço em disco e de fácil consulta em computadores com sistema operacional Windows.

6. PERSPECTIVAS

O Projeto Bootstrap e a utilização do método bootstrap para inferência de dados radioecológicos discrepantes podem ainda ser aprofundados de forma que as seguintes sugestões de estudos são colocadas:

- Expandir a utilização do método bootstrap para análise de duas ou mais variáveis randômicas provenientes de dados radioecológicos;
- Utilizar o método para execução de testes de hipóteses;
- Desenvolver o “script” Projeto Bootstrap com uma interface gráfica amigável e que possa ser utilizado como um aplicativo completo de código aberto em qualquer microcomputador que funcione com o sistema operacional Windows®;

REFERÊNCIAS BIBLIOGRÁFICAS

AIETA, E. M.; SINGLEY, J. E.; TRUSSEL, A. R.; THORBJARNARSON, K. W.; McGUIRE, M. J. Radionuclides in drinking water: an overview. **Research and Technology**, Denver, v. 79, n. 4, p. 144-152, 1987.

ALCOFORADO, E. S. **Influência de ocorrência de urânio nos níveis de chumbo estável no leite e derivados produzidos no agreste de Pernambuco**. 2011. 66p. Dissertação (Mestrado em Tecnologias Energéticas e Nucleares), Departamento de Energia Nuclear, Universidade Federal de Pernambuco, Recife, 2011.

AMARAL, R. S.; VASCONCELOS, W. E.; BORGES, E; SILVEIRA, S. V.; MAZZILLI, B. P. Intake of uranium and radium-226 due to food crops consumption in the phosphate region of Pernambuco – Brazil. **Journal of Environmental Radioactivity**, v. 82, n. 3, p. 383-393, 2005.

ANTUNES, P. D.; SAMPAIO, E. V. S. B.; FERREIRA Jr.; A. L. G.; GALINDO, I. C. L.; SALCEDO, I. H. Distribuição de ^{137}Cs em três solos representativos do estado de Pernambuco. **Revista Brasileira de Ciência do Solo**, v. 34, n. 3, p. 935-943, 2010.

ARANGO, H. G. **Bioestatística: teórica e computacional.** Rio de Janeiro: GEN, 2 ed., 2005, 423 p.

BATES, D.; CHAMBERS, J.; DALGAARD, P.; FALCON, S.; GENTLEMAN, R.; HORNICK, K.; IACUS, S.; IHAKA, R.; LEISCH, F.; LIGGES, U.; LUMLEY, T.; MAECHLER, M.; MURDOCH, D.; MURRELL, P.; PLUMMER, M.; RIPLEY, B.; SARKAR, D.; LANG, D. T.; TIERNEY, L.; URBANECK, S. The Comprehensive R Archive Network - CRAN. Disponível em: <http://cran.r-project.org>. Acesso em: 11 mai. 2012, 18:32:00.

BLACKWOOD, L. G. The lognormal distribution, environmental data and radiological monitoring. **Environmental Monitoring and Assessment**, v. 21, n. 3, p. 193-210, 1992.

CEMBER, H. **Introduction to Health Physics**. New York: McGraw Hill Professional, 1996. 733 p.

CHAMBERS, J. **Software for data analysis: programming with R.** Springer. New York: Science+Business Media, LLC, 2008. 514 p.

CHERNICK, M. R. **Bootstrap Methods: a guide for practitioners and researchers.** 2nd ed. Newtown: Wiley-Intercience, 2007. 388 p.

COHEN, Y.; COHEN, J. Y. **Statistics and Data with R: an applied approach through examples.** 1st ed. Chichester: John Wiley & Sons Ltd, 2008. 603 p.

COLGAN, P. A.; ORGANO, C.; HONE, C.; FENTON, D. Radiation doses received by the Irish population. RPII 08/01. Dublin: Radiological Protection Institute of Ireland, 2008.

COSTA, A. C.; PEDROSA, I. L.; MENDES, V. A. **Projeto Agreste de Pernambuco**. Convênio DNPM/CPRM. 1976.

COSTA, A. C.; PEDROSA, I. L.; MENDES, V. A. **Projeto Agreste de Pernambuco**. Convênio DNPM/CPRM. Relatório Final, v. 1, 1977.

DALGAARD, P. **Introductory Statistics with R**. New York: Springer, 2002. 284 p.

DAVISON, A. C.; HINKLEY, D. V. **Bootstrap methods and their application**. Cambridge: Cambridge University Press, 1997. 592 p.

DENNIS, B.; PATIL, G. P. Application in Ecology. In: CROW, E. L.; SHIMIZU, K. **Lognormal distributions: theory and applications.** New York: Marcel Dekker, Inc., 1988. chapter 12.

DIMOV, I. T. **Monte Carlo methods for applied scientists**. 1st ed. Singapore: World Scientific Publishing Co. Pte. Ltd., 2008. 308 p.

EBERHARDT, L. L.; GILBERT, R. O. Statistics and sampling in transuranic studies. In: HANSON, W. C. (Ed). **Transuranic elements in the environment**. DOE/TIC-22800 NTIS, 1980. p. 173-186.

EFRON, B.; TIBSHIRANI, R. J. **An introduction to the bootstrap**. 1st ed. Boca Raton: Chapman & Hall/CRC, 1993. 449 p.

EFRON, B.; TIBSHIRANI, R. The bootstrap method for standard errors, confidence intervals, and other measures of statistical accuracy. **Statistical Science**, v. 1, n. 1, p. 1-35, 1986.

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

EISENBUD, M.; GESELL, T. **Environmental radioactivity: from natural, industrial, and**

military sources. New York: Academic Press, 1997. 656 p.

FIRESTONE, R.B.; SHIRLEY, V.S.; CHU, S.Y.F.; BAGLIN, C. M. ZIPKIN, J. **Table of isotopes. U.S.:** Wiley-Interscience. Versão 1.0, CD ROM, 1996, 14.193 p.

GONZALES, A.; ANDERER, J. Radiation versus radiation: nuclear energy in perspective. **Instrumentation Atomic Energy Agency Bulletin**, Viena, v. 31, n. 2, p. 21-31, 1989.

HALL, P. **The bootstrap and Edgeworth expansion**. New York: Springer-Verlag, 1992. 352p.

HELENE, O.; VANIN, V. R. Analysis of discrepant data using a bootstrap procedure. **Nuclear Instruments and Methods in Physics Research A**, v. 481, n. 1-3, p. 626-631, 2002.

INSTITUTO DE RADIOPROTEÇÃO E DOSIMETRIA - IRD. **Radioproteção**. Disponível em: http://www.ird.gov.br/index.php?option=com_content&view=article&id=122&Itemid=1. Acesso em: 20 ago. 2012, 22:40:00.

KNOLL, G. F. **Radiation detection and measurement**. 3rd ed. New York: John Wiley & Sons, Incorporation, 2000. 413p

L'ANNUNZIATA, M. F. **Handbook of radioactivity analysis**. 2nd ed. San Diego: Academic Press, 2003. 1240p.

LUCIO, P. S.; LEANDRO, I. V.; DE PAULA, T. P. Bootstrap aplicado à avaliação de incertezas estatísticas no prognóstico de quantis extremos de precipitação. In: CONGRESSO BRASILEIRO DE METEOROLOGIA, 14, 2006, Florianópolis.

MATSUMOTO, M.; NISHIMURA, T. Mersenne Twister: a 623-dimensionally equidistributed uniform pseudo-random number generator. **ACM Transactions on Modeling and Computer Simulation**, v. 8, n. 1, p.:3-30, 1998.

McKUSICK, M. K. **Open sources: voices from the open source revolution.** 1st ed. O'Reilly, 1999. 280 p.

MURTEIRA, B. J. F. **Probabilidades e estatística**. 2. ed. Lisboa: McGraw-Hill, 1990. 547 p.

NASCIMENTO, C. T. C.; PIRES, A. C. B.; MORAES, R. A. V. Reconhecimento de solos por meio de resistividade elétrica e radiação gama. **Revista Brasileira de Geociências**, v. 34, n. 3, p. 383-392, 2004.

OKUNO, E.; YOSHIMURA, E. **Física das radiações**. São Paulo: Oficina de Textos, 2010. 296 p.

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

OTT, W. R.; MAGE, D. T. A general purpose univariate probability model for environmental data analysis. **Computers and Operations Research**. v. 3, p. 209-216. Pergamon Press, 1976.

PARK, S. K.; MILLER, K. W. Random number generators: good ones are hard to find. **Communications of the ACM**, v. 31, n. 10, p. 1192-1201, 1988.

PRESS, W. H.; TEUKOLSKY, S. A.; VETTERLING, W. T.; FLANNERY, B. P. **Numerical recipes in Fortran: the art of scientific computing.** Cambridge University Press, Cambridge, 1992. 994p.

QUENOUILLE, M. Notes on bias estimation. **Biometrika**, v. 43, n. 3/4, p. 353-360, 1956.

RAJPUT, M. U.; MACMAHON, T. D. Convergence of techniques for the evaluation of discrepant data. **Nuclear Instruments and Methods**. A 312, 289, 1992.

ROBERT, C. P.; CASELLA, G. **Introducing Monte Carlo methods with R**. New York: Springer, 2010. 297 p.

SANTOS JÚNIOR, J. A. **Avaliação radiométrica do U-238, Ra-226, Th-232 e K-40 em uma área anômala do Agreste de Pernambuco**. 2009. 216f. Tese (Doutorado em Tecnologias Energéticas e Nucleares), Departamento de Energia Nuclear, Universidade Federal de Pernambuco, Recife, 2009.

SANTOS JÚNIOR, J. A.; CARDOSO, J. J. R. F.; SILVA, C. M.; SILVEIRA, S. V.; AMARAL, R. S. Determination of radionuclides in environment using gamma-spectrometry. **Journal of Radioanalytical and Nuclear Chemistry**, v. 269, n. 2, p 451-455, 2006.

SHAW, G (Org). **Radioactivity in the terrestrial environment**. 1st ed. Oxford: Elsevier Ltd., 2007. 300 p.

SCHULZ, R. K. Soil chemistry of radionuclides. **Health Physics**. v. 11, n. 12, p. 1317-1324, 1965.

SILVA, C. M.; AMARAL, R. S.; VIEIRA, J. W.; SILVA, A. N. C.; SANTOS JÚNIOR, J. A.S.; ALCOFORADO, E. S. Estimativa de intervalo de confiança para tempo de sobrevida usando desigualdade de Chebyshev via método bootstrap. **Scientia Plena**, v. 8, n. 4, p. 12-23, 2012.

SILVA, C. M.; AMARAL, R. S.; VIEIRA, J. W.; SILVA, A. N. C.; SANTOS JÚNIOR, J. A.S.; ALCOFORADO, E. S. Modelagem de tempo de sobrevida via método bootstrap. **Scientia Plena**, v. 7, n. 10, p. 14-25, 2011.

SILVA, C. M.; AMARAL, A. J.; AMARAL, R. S.; SANTOS JÚNIOR, J. A.; VIEIRA, J. W. Application of bootstrap method for evaluating discrepant levels of radium-226 in forage palm (*Opuntia spp*). **Revista Brasileira de Biometria**, São Paulo, v. 25, n. 3, p. 109-114, 2007.

SILVA, C. M. **Ra-226 e Ra-228 na dieta de bovinos leiteiros do agreste semiárido de Pernambuco e avaliação de risco decorrente do consumo de leite por uma população potencialmente exposta**. 2006. 152f. Tese (Doutorado em Tecnologias Energéticas e Nucleares), Departamento de Energia Nuclear, Universidade Federal de Pernambuco, Recife, 2006.

SINGH, A. K., SINGH, A., ENGELHARDT, M. The lognormal distribution in environmental applications. **Technology Support Center Issue**, EPA/600/R-97/006. 1997.

TAYLOR, D. M.; TAYLOR, S. K. Environmental uranium and human health. **Reviews on Environmental Health**, v. 12, n. 3, p. 147-158, 1997.

UPTON, G.; COOK, I. **Introducing statistics**. 2nd ed. Oxford: Oxford University Press, 2009. 349 p.

WHICKER, F. W.; SCHULTZ, V. **Radioecology: nuclear energy and the environment.** 1st ed. Boca Raton: CRC Press Inc., 1982. 440 p.

WILCOX, R. R. **Basic statistics: understanding conventional methods and modern insights.** Oxford: Oxford University Press, 2009. 341 p.

APÊNDICE A - Código desenvolvido como script com o nome Projeto Bootstrap

As linhas abaixo correspondem ao código de programação desenvolvido na interface gráfica RStudio utilizando os recursos do programa estatístico R para gerar as amostras bootstrap a partir dos dados amostrais originais. Se o código for copiado para o um arquivo novo do RStudio e executado gera os dados discutidos neste trabalho.

```
#Dados do Ra-226 em Palma Forrageira utilizando Monte Carlo-Bootstrap
#Utilizando o pacote boot
#Concentração de Ra-226 em palma forrageira
#A linha abaixo define o números de dígitos no R incluindo as casas decimais e carrega os pacotes.
options(digits=6)
require(boot)
require(graphics)
require(stats)
######################função e dados###################################
#Função que define a estatística ser calculada - no caso média aritmética
media_dados <- function(x,i){
 mean(x[i])
}
#Vetor com os dados da palma forrageira
dados <- c(1985,1990,9300,5992,2150,1495,
        4500,5350,5060,6800,5400,5500,25000,3000)
###################estatística clássica ############################
#calculando valor mínimo, valor máximo, amplitude, média aritmética simples,
#variância, desvio padrão, mediana, primeiro e último quartil,
#desvio interquartílico e média geométrica dos logaritmos dos valores amostrais
estatistica_classica <- summary(dados, digits=6)
amplitude <- max(dados)- min(dados)
#Cálculo da variância
```

```
variancia_classica <-var(dados)
erro_padrao_classico <- sd(dados)
num_dados <-length(dados)
#Desvio interquartilico
#primeiro quartil
quartil1 <- quantile(dados, probs=0.25)
#terceiro quartil
quartil3 <- quantile(dados, probs=0.75)
desvio_quartil <- quartil3-quartil1
#histograma
histograma_classico <- hist(dados, main = " ",
                            xlab = "Atividade específica do Ra-226",
                            ylab="Nº de amostras",
                            cex.axis=0.9,
                            cex.lab=0.9
                           )
#Cálculo dos valores discrepantes
#Método 1
valor_discrepante <- abs(((dados)-mean(dados))/sd(dados))
#Método 2
vd1 <- quartil1 - 1.5*(quartil3-quartil1)
vd2 <- quartil3 - 1.5*(quartil3-quartil1)
############################ bootstrap da média ########################
#Utilização da função boot(data, statistic, R) existem outros argumentos
#data = dados, statistic = media_dados, R é o número de reamostras.
options(digits=6)
resultado_boot <- boot(dados,media_dados,R=100000)
##Sumário do bootstrap
resumo_boot <- summary(resultado_boot$t)
```

```
resumo_boot2 <- sapply(resultado_boot$t, mean, na.rm=TRUE)
#valor_min <- min(resultado_boot$t)
#valor_max
################ bootstrap - intervalo de confiança da média #########
#utilizaçao da função boot.ci(boot.out, conf=0.95, type="all")
#é necessário calcular o bootstrap da estatística desejada primeiro
#resultado_boot é o objeto da classe boot com os valores do cálculo bootstrap
#conf são os intervalos de confiança desejados
#type é o tipo de intervalo de confiança requerido
options(digits=6)
intervalo_boot <- boot.ci(resultado_boot, conf=c(0.90,0.95),
                  type=c("norm", "basic", "perc"))
####################boxplot do bootstrap ############################
grafico_box <- boxplot(resultado_boot$t,ylab ="Atividade específica do Ra-226",
                cex.axis=0.9,
                cex.lab=0.9,
                varwidth=TRUE)
###################Resultados em arquivos############################
#################Em txt #####################################
#Salva o resultado no arquivo Ra-226 Estatística Clássica
sink("c:/Ary/Meeeuuuu/Mestrado/Dissertacao/Defesa/Resultados no R/R-226 Estatistica
Classica.txt")
print(estatistica_classica)
print(amplitude)
print(variancia_classica)
print(erro_padrao_classico)
options(digits=3)
print(media_logs)
options(digits=7)
print(desvio_quartil)
```

```
print("Valor discrepante")

print(vd1);print(vd2)

print("Valor discrepante"); print(dados<vd1)

print("Valor discrepante"); print(dados>vd2)

sink()

#Salva o resultado no arquivo Ra-226 - 100000_1 amostras bootstrap


sink("c:/Ary/Meeeuuuu/Mestrado/Dissertacao/Defesa/Resultados no R/Ra-226 - 100000 amostras bootstrap.txt")

#Imprime a média dos dados originais, o viés entre a população

#e o bootstrap e o desvio padrão do

print(resultado_boot)

#Imprime a média aritmética simples do bootstrap

print(mean(resultado_boot$t))

#Imprime a população bootstrap

print(resultado_boot$t)

#Imprime o intervalo de confiança com probabilidade de 90 e

#95% com normal, básico e percentil

print(intervalo_boot)

#Imprime um sumário

print("Resumo usando summary")

print(resumo_boot)

print("Resumo usando sapply")

print(resumo_boot2)

#Imprime a variância dos dados originais, o viés entre a população

# e o bootstrap e o desvio padrão do bootstrap

#print(resultado_variancia)

#print(var(resultado_variancia$t))

#print(resultado_variancia$t)

sink()
```

```
##############
histograma_classico_relativo<-hist(dados, freq= FALSE, main = " ",
                        xlab = "Atividade específica do Ra-226",
                        ylab="Densidade",
                        cex.axis=0.9,
                        cex.lab=0.9
)
curve(dnorm(x, mean=5965.86, sd=5903.05), add=TRUE)
densidade <- density(dados, main=" ",
                xlab = "Atividade específica do Ra-226",
                ylab="Densidade")
plot(densidade, add=TRUE)
#print(x)
#################Em gráficos pdf ####################################
#Histograma da estatística clássica - em pdf
pdf(file="c:/Ary/Meeeuuuu/Mestrado/Dissertacao/Defesa/Resultados no
R/histograma_classico.pdf")
histograma_classico<-hist(dados, main = " ",
                    xlab = "Atividade específica do Ra-226",
                    ylab="Nº de amostras",
                    cex.axis=0.9,
                    cex.lab=0.9)
dev.off()
#Histograma da estatística clássica - em pdf - frequência relativa
pdf(file="c:/Ary/Meeeuuuu/Mestrado/Dissertacao/Defesa/Resultados no
R/histograma_classico_relativo.pdf")
histograma_classico_relativo<-hist(dados, freq= FALSE, main = " ",
                    xlab = "Atividade específica do Ra-226",
                    ylab="Densidade",
                    cex.axis=0.9,
```

```
                    cex.lab=0.9
)
curve(dnorm(x, mean=5965.86, sd=5903.05), add=TRUE)
#curve(dnorm(x,mean=2,sd=sqrt(4)),col=2,lty=2,lwd=2,add=TRUE)
#hist(rnorm(500,mean=2,sd=sqrt(4)),freq=FALSE)
dev.off()
#histograma da população bootstrap
#Está funcionando
pdf(file="c:/Ary/Meeeuuuu/Mestrado/Dissertacao/Defesa/Resultados no
R/histograma_media_100000.pdf")
histograma_media_100000 <- hist(resultado_boot$t, main = " ",
                    xlab = "Atividade específica do Ra-226",
                    ylab="Nº de amostras",
                    cex.axis=0.6,
                    cex.lab=0.6 )
dev.off()
#boxplot da população bootstrap
pdf(file="c:/Ary/Meeeuuuu/Mestrado/Dissertacao/Defesa/Resultados no
R/boxplot_boot100000.pdf")
boxplot_boot100000 <- boxplot(resultado_boot$t,ylab ="Atividade específica das
      do Ra-226", xlab = "100.000 reamostras", font.axis = "Times New Roman",
                    font.main = "Times New Roman",
                    font.lab = "Times New Roman",
                    cex.axis=0.9,
                    cex.lab=0.9,
                    varwidth=TRUE)
dev.off()
#Densidade de probabilidade do histograma e a curva normal
pdf(file="c:/Ary/Meeeuuuu/Mestrado/Dissertacao/Defesa/Resultados no
R/histograma_boot_relativo100000.pdf")
histograma_boot_relativo100000<-hist(resultado_boot$t, freq= FALSE, main = " ",
```

```
                    xlab = "Atividade específica de 100.000 reamostras do Ra-226",
                    ylab="Densidade",
                    cex.axis=0.6,
                    cex.lab=0.6
)
curve(dnorm(x, mean=mean(resultado_boot$t), sd=apply(resultado_boot$t, 2, sd)),add=TRUE)
dev.off()
```