



Universidade Estadual da Paraíba
Centro de Ciências e Tecnologia
Departamento de Estatística

Aline Carla da Silva

Utilização da Técnica de Reamostragem Bootstrap em Amostragem Aleatória Simples para os IDHM's do Brasil

Campina Grande
Agosto de 2014.

Aline Carla da Silva

Utilização da Técnica de Reamostragem Bootstrap em Amostragem Aleatória Simples para os IDHM's do Brasil

Trabalho de Conclusão de Curso a ser apresentado como requisito para a conclusão do curso de Bacharelado em Estatística do Departamento de Estatística do Centro de Ciências e Tecnologia da Universidade Estadual da Paraíba para obtenção do título de Bacharel em Estatística.

Orientador:

Kleber Napoleão Nunes de Oliveira Barros

Campina Grande

Agosto de 2014.

É expressamente proibida a comercialização deste documento, tanto na forma impressa como eletrônica. Sua reprodução total ou parcial é permitida exclusivamente para fins acadêmicos e científicos, desde que na reprodução figure a identificação do autor, título, instituição e ano da dissertação.

S586u Silva, Aline Carla da.

Utilização da técnica de reamostragem Bootstrap em amostragem aleatória simples para os IDHM'S do Brasil [manuscrito] / Aline Carla da Silva. - 2014.
33 p.

Digitado.

Trabalho de Conclusão de Curso (Graduação em Estatística) - Universidade Estadual da Paraíba, Centro de Ciências e Tecnologia, 2014.

"Orientação: Prof. Me. Kleber Napoleão Nunes de Oliveira Barros, Departamento de Estatística".

1. Bootstrap. 2. Amostragem Aleat ória Simples. 3. Índice de Desenvolvimento Humano. 4. Teste de Kolmogorov-Smirnov. I. Título. 21. ed. CDD 519.53

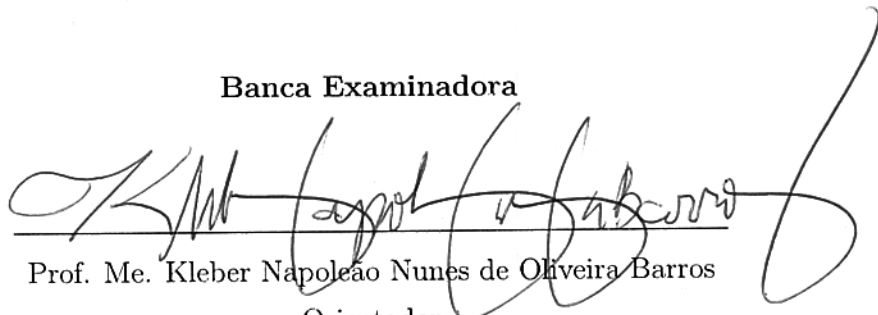
Aline Carla da Silva

Utilização da Técnica de Reamostragem Bootstrap em Amostragem Aleatória Simples para os IDHM's do Brasil

Trabalho de Conclusão de Curso a ser apresentado como requisito para a conclusão do curso de Bacharelado em Estatística do Departamento de Estatística do Centro de Ciências e Tecnologia da Universidade Estadual da Paraíba para obtenção do título de Bacharel em Estatística.

Aprovado em: 15 / 07 / 2014

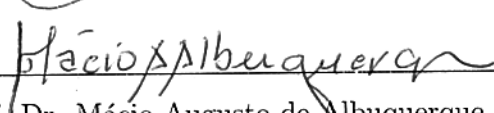
Banca Examinadora



Prof. Me. Kleber Napoleão Nunes de Oliveira Barros
Orientador



Prof. Me. Juárez Fernandes de Oliveira



Prof. Dr. Mácio Augusto de Albuquerque

*Dedico aos meus pais Carlos e
Albanisa, e aos meus irmãos
Ricardo e Kaline.*

Agradecimentos

Agradeço aos professores que me acompanharam durante a graduação, pelo conhecimento adquirido e à Kleber Barros por me orientar na etapa final, apesar dos contratempos.

Agradeço aos amigos de turma Analu Cabral e Edinário Barbosa pela companhia e motivação ao longo dos anos; e à Klecio Lima por tornar a vida universitária mais divertida e incentivar a busca de maiores desafios.

Agradeço aos amigos e companheiros de longa data, Joab Silva, Mariana Melo e Rodrigo Ferreira por continuarem presentes nas ocasiões boas e nem tão boas.

Agradeço ao meu irmão Ricardo Silva; que acompanha, contribui e apoia meu crescimento e minhas escolhas, se fazendo presente nas decisões mais importantes da minha vida.

À Deus, minha família e à todos que fazem parte da minha vida, obrigada.

Campina Grande
Agosto de 2014.

Resumo

A partir dos dados oficiais retirados do site do Programa das Nações Unidas para o Desenvolvimento (PNUD), que diz respeito aos índices de desenvolvimento dos municípios brasileiros nos anos de 2000 e 2010; aplicamos a técnica de reamostragem de Bootstrap e o método da Amostragem Aleatória Simples para estimar os parâmetros de interesse, utilizamos o teste de Kolmogorov-Smirnov para verificação dos pressupostos, em seguida, realizamos também os intervalos de confiança Bootstrap (normal, pivotal e percentil) e o intervalo de confiança frequentista, à nível de comparação. Como a técnica de reamostragem requer um bom desempenho computacional, utilizamos o software RStudio versão 0.98.953. A partir dos resultados se observa que a técnica Bootstrap compete com o intervalo de confiança convencional.

Palavras-chave: Bootstrap, Amostragem Aleatória Simples, Índice de Desenvolvimento Humano, teste de Kolmogorov-Smirnov.

Abstract

From the official data from the United Nations Development Programme for Development Programme (UNDP) site on internet, according to rates of development of Brazilians city councils in the years 2000 and 2010; we apply the technique of bootstrap resampling and The method of Simple Random Sampling to estimate the parameters of interest, used the Kolmogorov-Smirnov test to check the assumptions, Then we also apply bootstrap confidence intervals (normal, pivotal and percentile) and the frequentist confidence interval, for comparison. As the technique resampling requires good computational performance, we use the software RStudio version 0.98.953. From the results, it is observed that the bootstrap technique competes with Conventional confidence interval.

Keywords: Bootstrap, Simple Random Sampling, Human Development Index, Kolmogorv-Smirnov test.

Sumário

1	Introdução	p. 10
2	Revisão de leitura	p. 11
2.1	Amostragem Aleatória Simples	p. 11
2.1.1	Com reposição	p. 11
2.1.2	Sem reposição	p. 12
2.1.3	Estimadores para AASc e AASs	p. 12
2.2	Intervalo de Confiança Frequentista para a média	p. 13
2.3	Estimadores Razão	p. 14
2.4	Técnicas de Bootstrap	p. 15
2.5	Estimadores de Bootstrap	p. 16
2.6	Intervalos de Confiança Bootstrap	p. 16
2.6.1	Intervalo de Confiança Normal	p. 16
2.6.2	Intervalo de Confiança Pivotal	p. 17
2.6.3	Intervalo de Confiança Percentil	p. 18
2.7	Boxplot	p. 18
2.8	Índice de Desenvolvimento Humano - IDH	p. 19
2.9	Cálculo do IDH	p. 20
3	Material e Métodos	p. 22
4	Resultados e Discussões	p. 23
5	Conclusões	p. 29

Referências	p. 30
Apêndice	p. 31
Apêndice A - Códigos R utilizados nas aplicações	p. 31

1 *Introdução*

Bootstrap é uma técnica de reamostragem criada por Bradley Efron (1979), bastante utilizada para estimação do viés, da variância, quantis ou distribuição de amostragem em levantamentos estatísticos, também na construção de intervalos de confiança. A técnica consiste em várias reamostragens do mesmo tamanho da amostra original, estimando e aproximando os parâmetros de interesse; o que exige um certo desempenho computacional.

Foi pensada primeiramente para circunstâncias em que técnicas habituais não são cabíveis, como número de amostras reduzidas onde requer um manuseio mais específico a fim de chegar a uma representatividade, mais fiel possível, da população. Por esse motivo, a técnica de Bootstrap é comumente aplicada a dados originais - cálculos de intervalos de confiança de parâmetros, diminuição do viés em médias e variâncias - e em modelos ajustados, principalmente para otimizá-los (LIERO, 2014).

Tomando como população as razões entre os 5565 índices de desenvolvimento humano municipais (IDHM) dos anos de 2010 e 2000 (PNUD, 2013) - importante ferramenta que mede o desenvolvimento dos países no mundo - objetivamos estimar seus parâmetros de crescimento, que por serem uma razão, quebram a suposição de normalidade, essa quebra pode ser investigada com o auxílio da técnica de Bootstrap.

No capítulo 2, faremos as revisões das técnicas de Amostragem Aleatória Simples (AAS) e Bootstrap, necessárias para o estudo em questão, definiremos os conceitos do Índice de Desenvolvimento Humano e do Índice de Desenvolvimento Humano Municipal, mostrando também, a diferença entre seus cálculos; no capítulo 3 apresentamos a metodologia empregada no estudo. No capítulo 4 mostramos os resultados de todas as aplicações, feitas no R-Studio versão 0.98.953 para sistema operacional Windows. No capítulo 5 apresentamos as devidas conclusões do estudo.

2 *Revisão de leitura*

Primeiramente vamos rever algumas definições da Amostragem Aleatória Simples, para compararmos os resultados após a aplicação das técnicas de Bootstrap.

2.1 Amostragem Aleatória Simples

Sendo o método mais simples e mais importante para selecionar uma amostra a Amostragem Aleatória Simples - AAS, possui algumas vantagens, como a independência entre as unidades sorteadas, que facilita a determinação das propriedades dos estimadores das quantidades populacionais de interesse (Bolfarine; Bussab, 2005). A AAS possui dois casos distintos, a Amostra Aleatória Simples com Reposição (AASc) e Amostra Aleatória Simples sem Reposição (AASs).

2.1.1 Com reposição

A AASc segue os passos:

- Numera-se a população de 1 a N :

$$U = \{1, \dots, N\};$$

- Sorteia-se, com probabilidade igual, uma unidade n_i das N unidades da população;
- Repõe essa unidade n_i na população e sorteia-se outro elemento;
- Repete-se o procedimento até que n unidades tenham sido sorteadas.

Supondo n unidades sorteadas pelo plano AASc, cada tentativa é independente e tem a mesma probabilidade $\frac{n}{N}$ de ser sorteado, uma urna contendo bolas de cores diferentes em que cada cor tem probabilidade $\frac{n}{N}$ de ser sorteada, por exemplo.

A Figura 1 mostra um esquema de AASc, com probabilidade $1/N$ para todas as observações e com reposição das unidades sorteadas .

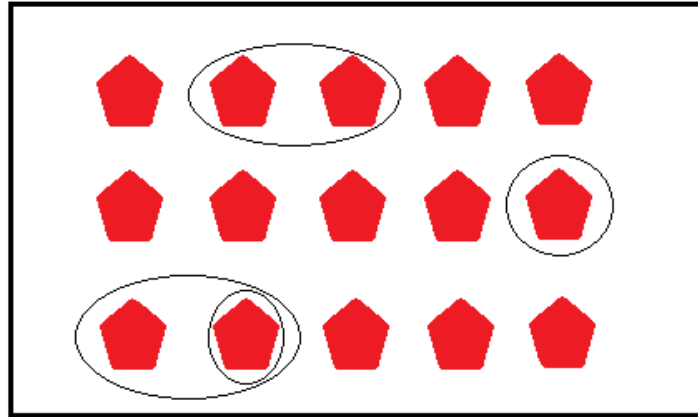


Figura 1: Esquema do método AASc

2.1.2 Sem reposição

Para a Amostra Aleatória Simples sem Reposição (AASs), o procedimento é semelhante ao AASc mas não fazemos a reposição do elemento n_i retirado da população. Dessa forma, cada elemento só aparece uma única vez na amostra com probabilidade $\frac{n!}{N^n}$, como exemplo podemos destacar o bingo em que cada número sorteado aparece uma única vez.

A Figura 2 mostra um esquema AASs, com probabilidade $\frac{n!}{N^n}$ para todas as observações e sem reposição das unidades sorteadas.

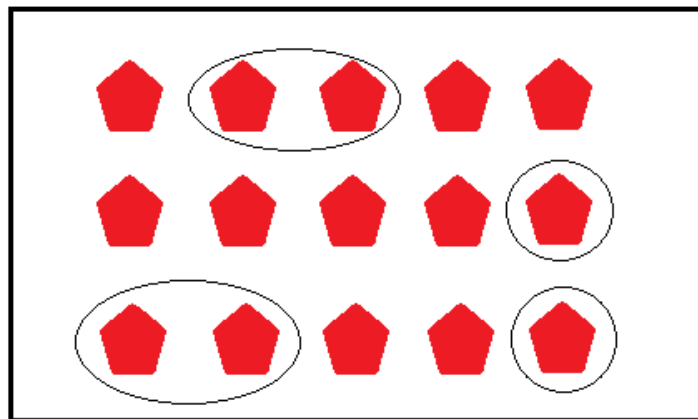


Figura 2: Esquema do método AASs

2.1.3 Estimadores para AASc e AASs

- A média amostral sendo

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n Y_i, \quad (2.1)$$

temos um estimador não viesado da média populacional μ dentro do plano AASc.

Com

$$Var[\bar{y}] = \frac{\sigma^2}{n}. \quad (2.2)$$

Um estimador não viesado para o total populacional é

$$T(s) = N\bar{y}, \quad (2.3)$$

com

$$Var[T] = N^2 \frac{\sigma^2}{n}. \quad (2.4)$$

Para o caso do estimador da variância populacional σ^2 temos,

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{y})^2 \quad (2.5)$$

- Para o plano AASs a média amostral também é dada pela equação (2.1), mas sua variância amostral é dada por:

$$Var[\bar{y}] = (1-f) \frac{S^2}{n}, \quad (2.6)$$

onde $f = \frac{n}{N}$ é denominada fração amostral e $(1-f)$ é o fator de correção para populações finitas. O estimador não viciado para o total populacional segue a equação (2.3) mas sua variância amostral é dada por:

$$Var[T] = N^2(1-f) \frac{S^2}{n}. \quad (2.7)$$

Por fim, temos o estimador não viesado da variância populacional S^2 :

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{y})^2 \quad (2.8)$$

2.2 Intervalo de Confiança Frequentista para a média

Com relação a média populacional, a medida que o tamanho da amostra aumenta, a distribuição de \bar{y} vai se aproximando da distribuição Normal, de acordo com o Teorema Central do Limite (TLC), para n suficientemente grande, temos:

$$P\left(\frac{\bar{y} - \mu}{\sqrt{\sigma^2/n}} \leq z_\alpha\right) \simeq 1 - \alpha, \quad (2.9)$$

onde z_α é um valor $N(0,1)$, de tal forma que a área da densidade da $N(0,1)$ no intervalo $(-z_\alpha; z_\alpha)$ é igual a $1 - \alpha$. Como σ^2 é desconhecido, ele é substituído por seu estimador

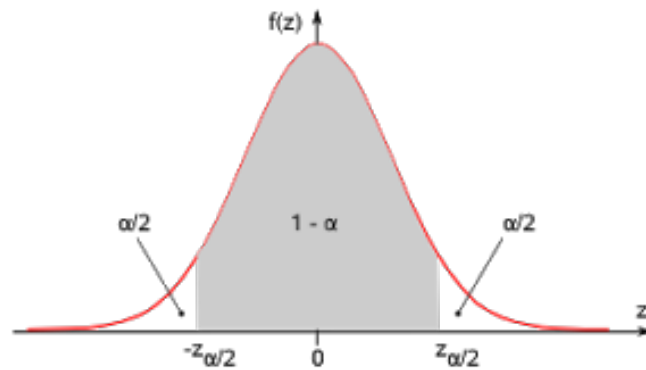
não viciado s^2 , que para n grande é bem próximo de σ^2 . Dessa forma, (2.9) pode ser reescrita como:

$$P\left(\bar{y} - z_\alpha \sqrt{\frac{s^2}{n}} \leq \mu \leq \bar{y} + z_\alpha \sqrt{\frac{s^2}{n}}\right) \simeq 1 - \alpha,$$

e segue que:

$$P\left(\bar{y} - z_\alpha \sqrt{\frac{s^2}{n}}; \bar{y} + z_\alpha \sqrt{\frac{s^2}{n}}\right) \quad (2.10)$$

é um intervalo de confiança para μ com coeficiente aproximadamente igual a $1 - \alpha$.



Fonte: <http://propriedadesdoconcreto.blogspot.com.br>

Figura 3: Intervalo de Confiança em função de z

2.3 Estimadores Razão

Considerando algumas situações em que o elemento i da população finita U , associa-se ao par (X_i, Y_i) , $i = 1, \dots, N$; a variável X é introduzida no problema para melhorar as previsões dos parâmetros. Em casos onde é de interesse a comparação de determinadas quantidades em períodos sucessivos, ou quando o parâmetro é um índice - quociente entre duas variáveis (Bolfarine; Bussab, 2005). Nessas situações pode-se então definir a razão como parâmetro de interesse.

Para utilizar uma variável auxiliar X na estimação de quantidades do tipo razão R , o total τ_Y ou a média μ_Y , utilizamos os seguintes estimadores do tipo razão:

$$\begin{aligned} r &= \hat{R} = \frac{\bar{y}}{\bar{x}}, \\ \hat{\tau}_Y &= T_R = \hat{R}_{\tau_X} = r\tau_X \quad \text{e} \\ \bar{y}_R &= \hat{R}_{\mu_X} = r\mu_X, \end{aligned}$$

respectivamente, onde \bar{x} e \bar{y} são obtidas através do plano amostral AAS.

2.4 Técnicas de Bootstrap

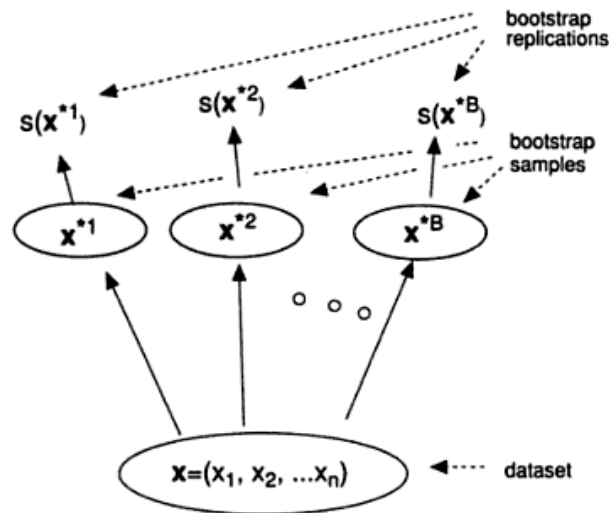
Popularizado na década de 80 devido ao início do uso de computadores para técnicas estatísticas. Bootstrap é muito usado para melhorar propriedades - e conseqüentemente - estimadores e intervalos de confiança para os parâmetros (WASSERMAN, 2004, p.107). Seu principal pressuposto é que a amostra tenha uma boa representação da população desconhecida, pois a amostra observada é tratada como se fosse a população original. Sua ideia básica pode ser resumida em dois passos.

Seja $T_n = g(X_1, \dots, X_n)$, onde T_n é uma função qualquer, suponha que queiramos saber $V_F(T_n)$, variância de T_n (F é uma função de distribuição desconhecida que pode alterar a variância):

- 1 Estimar $V_F(T_n)$ com $V_{\hat{F}_n}(T_n)$
- 2 Aproximar $V_{\hat{F}_n}(T_n)$ usando simulação.

Para $T_n = \bar{X}_n$, temos o passo 1, onde $V_{\hat{F}_n}(T_n) = \hat{\sigma}^2/n$, quando não se tem informações suficientes para estimar $V_F(T_n)$, usa-se o passo 2, simulações.

Observe a figura 4:



Fonte: Efron & Tibshirani, 1993.

Figura 4: Esquema do processo de inicialização de Bootstrap

onde Bootstrap gera n_i amostras independentes de tamanhos iguais a n para a estimativa do erro padrão $s_{boot}(X)$.

2.5 Estimadores de Bootstrap

De acordo com o que foi dito anteriormente, podemos aproximar $V_{\hat{F}_n}(T_n)$ por simulação. A estatística $V_{\hat{F}_n}(T_n)$ implica dizer que T_n é a variância se a distribuição dos dados for \hat{F}_n (WASSERMAN, 2004). Como podemos simular a partir da distribuição F_n quando os dados assumem a distribuição \hat{F}_n ? A resposta é simular X_1^*, \dots, X_n^* de \hat{F}_n e, em seguida, calcular $T_n^* = g(X_1^*, \dots, X_n^*)$. Trata-se de um sorteio da distribuição de T_n .

A idéia pode ser resumida em:

$$\begin{aligned} \text{Sem Bootstrap } F &\implies X_1, \dots, X_n \implies T_n = g(X_1, \dots, X_n) \\ \text{Com Bootstrap } \hat{F}_n &\implies X_1^*, \dots, X_n^* \implies T_n^* = g(X_1^*, \dots, X_n^*) \end{aligned}$$

Como podemos simular X_1^*, \dots, X_n^* de \hat{F}_n ? Observe que F_n coloca peso $1/n$ em cada ponto dos dados X_1, \dots, X_n .

Portanto, cada observação \hat{F}_n é equivalente a um ponto ao acaso a partir do conjunto de dados originais. Assim, para simular $X_1^*, \dots, X_n^* \sim \hat{F}_n$ basta obter n observações com substituição de X_1, \dots, X_n . Em resumo temos:

Estimação da Variância Bootstrap

1. Sorteia-se $X_1^*, \dots, X_n^* \sim \hat{F}_n$
2. Computa-se $T_n^* = g(X_1^*, \dots, X_n^*)$
3. Repete-se os passos 1 e 2, B vezes, para obter $T_{n,1}^*, \dots, T_{n,B}^*$
4. Seja

$$v_{boot} = \frac{1}{B} \sum_{b=1}^B \left(T_{n,b}^* - \frac{1}{B} \sum_{r=1}^B T_{n,r}^* \right)^2 \quad (2.11)$$

2.6 Intervalos de Confiança Bootstrap

Existem várias maneiras de construir intervalos de confiança Bootstrap, aqui discutiremos três deles (WASSERMAN, 2004).

2.6.1 Intervalo de Confiança Normal

É o método mais simples:

$$X_n \pm Z_{\alpha/2} S_{boot} \quad (2.12)$$

onde, S_{boot} é a estimativa de Bootstrap do erro padrão. Este intervalo não é preciso a menos que a distribuição de X_n se aproxime de uma Normal.

2.6.2 Intervalo de Confiança Pivotal

Seja $\theta = X(F)$ e $\theta_n = X(F_n)$ e define o pivô $E_n = \theta_n - \theta$. Seja $\theta_{n,1}^*, \dots, \theta_{n,B}^*$ replicações de Bootstrap de θ_n . Seja $H(e)$ o CDF do pivô:

$$H(e) = P_F(E_n \leq e). \quad (2.13)$$

Definindo $C_n^* = (a, b)$ onde

$$a = \theta_n - H^{-1}\left(1 - \frac{\alpha}{2}\right) \quad e \quad b = \theta_n - H^{-1}\left(\frac{\alpha}{2}\right). \quad (2.14)$$

Segue que

$$\begin{aligned} P(a < \theta < b) &= P(a - \theta_n \leq \theta - \theta_n \leq b - \theta) \\ &= P(\theta_n - b \leq \theta_n - \theta \leq \theta_n - a) \\ &= P(\theta_n - b \leq E_n \leq \theta_n - a) \\ &= H(\theta_n - a) - H(\theta_n - b) \\ &= H\left(H^{-1}\left(1 - \frac{\alpha}{2}\right)\right) - H\left(H^{-1}\left(\frac{\alpha}{2}\right)\right) \\ &= 1 - \frac{\alpha}{2} - \frac{\alpha}{2} \\ &= 1 - \alpha \end{aligned}$$

Assim, C_n^* é um intervalo de confiança $1 - \alpha$ exato para θ . Infelizmente, a e b dependem da distribuição desconhecida H mas podemos obter uma estimativa Bootstrap para H :

$$\widehat{H}(e) = \frac{1}{B} \sum_{b=1}^B I(E_{n,b}^* \leq e) \quad (2.15)$$

onde, $E_{n,b}^* = \hat{\theta}_n^* - \hat{\theta}_n$. Seja e_β^* , com β sendo o quantil amostral de $(E_{n,1}^*, \dots, E_{n,B}^*)$ e seja θ_β^* o quantil amostral β de $(\hat{\theta}_{n,1}^*, \dots, \hat{\theta}_{n,B}^*)$. Note que $e_\beta^* = \theta_\beta^* - \hat{\theta}_n$. Segue que um intervalo de confiança aproximado para $1 - \alpha$ é $C_n = (\hat{a}, \hat{b})$, onde

$$\begin{aligned} \hat{a} &= \hat{\theta}_n - \widehat{H}^{-1}\left(1 - \frac{\alpha}{2}\right) = \hat{\theta}_n - e_{1-\alpha/2}^* = 2\hat{\theta}_n - \theta_{1-\alpha/2}^* \\ \hat{b} &= \hat{\theta}_n - \widehat{H}^{-1}\left(\frac{\alpha}{2}\right) = \hat{\theta}_n - e_{\alpha/2}^* = 2\hat{\theta}_n - \theta_{\alpha/2}^*. \end{aligned}$$

Em resumo, o intervalo de confiança pivotal $1 - \alpha$ de Bootstrap é:

$$C_n = (2\hat{\theta}_n - \theta_{1-\alpha/2}^*; 2\hat{\theta}_n - \theta_{\alpha/2}^*). \quad (2.16)$$

2.6.3 Intervalo de Confiança Percentil

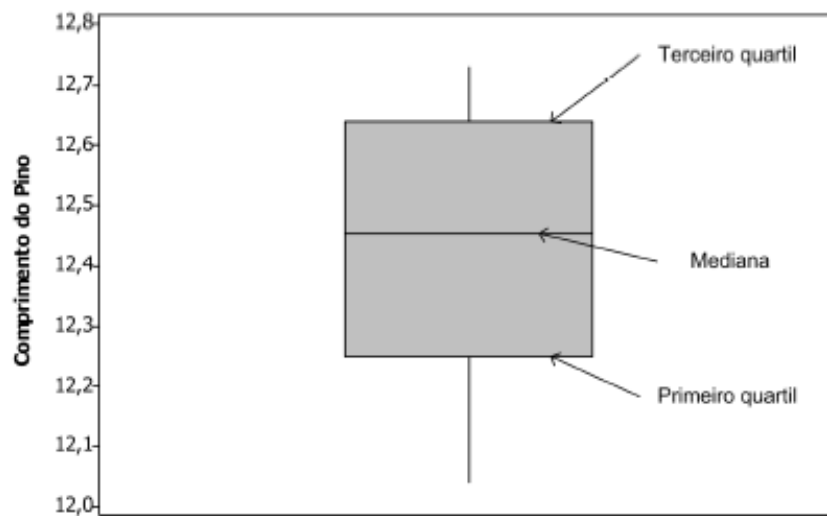
É definido por:

$$C_n = (\theta_{\alpha/2}^*, \theta_{1-\alpha/2}^*) \quad (2.17)$$

em que $\theta_{\alpha/2}^*$ é o quantil amostral $\alpha/2$ e $\theta_{1-\alpha/2}^*$ é o quantil amostral $1 - \alpha/2$.

2.7 Boxplot

O *Boxplot*, muito importante em Estatística pois agrega grande quantidade de informação sobre os dados num único gráfico, possibilitando a leitura de sua variabilidade e a comparação simultânea entre diferentes grupos; é formado pelo 1º quartil (Q_1), mediana (M_d), 3º quartil (Q_3), a distância interquartílica (d_q) definida por $Q_3 - Q_1$ e os limites inferior (l_i) e superior (l_s) definidos por $l_i = Q_1 - 1,5d_q$ e $l_s = Q_3 + 1,5d_q$. Os pontos fora desses limites são considerados valores discrepantes ou *outliers*.



Fonte: <http://www.portalaction.com.br/content/31-boxplot>

Figura 5: Exemplo do gráfico Boxplot

O Boxplot também fornece informações sobre assimetria e dispersão; se a amplitude for consideravelmente maior que a distância interquartílica e a mediana estiver mais próxima de Q_1 do que de Q_3 há fortes indícios de assimetria positiva e de grande dispersão das observações, por exemplo.

2.8 Índice de Desenvolvimento Humano - IDH

Apresentado no primeiro *Relatório de Desenvolvimento Humano do Programa das Nações Unidas para o Desenvolvimento*, em 1990, seu conceito e sua medida foram idealizados pelo economista paquistanês Mahbub ul Haq com colaboração do economista Amartya Sen. Sendo uma alternativa ao Produto Interno Bruto que era a medida de desenvolvimento da época (ATLAS DO DESENVOLVIMENTO HUMANO NO BRASIL, 2013).

Obteve grande repercussão mundial por conseguir unir, em uma única medida, três importantes dimensões da vida humana e ainda ser simples. O **IDH** leva em consideração três requisitos importantes que estão entre os conceitos da expansão da liberdade das pessoas:

- A oportunidade de se levar uma vida longa e saudável - **saúde**

Leva em consideração as oportunidades que as pessoas têm de evitar a morte prematura, e de garantir um ambiente saudável, com acesso à saúde de qualidade, para que possam atingir o padrão mais elevado possível de saúde física e mental.

- Ter acesso ao conhecimento - **educação**

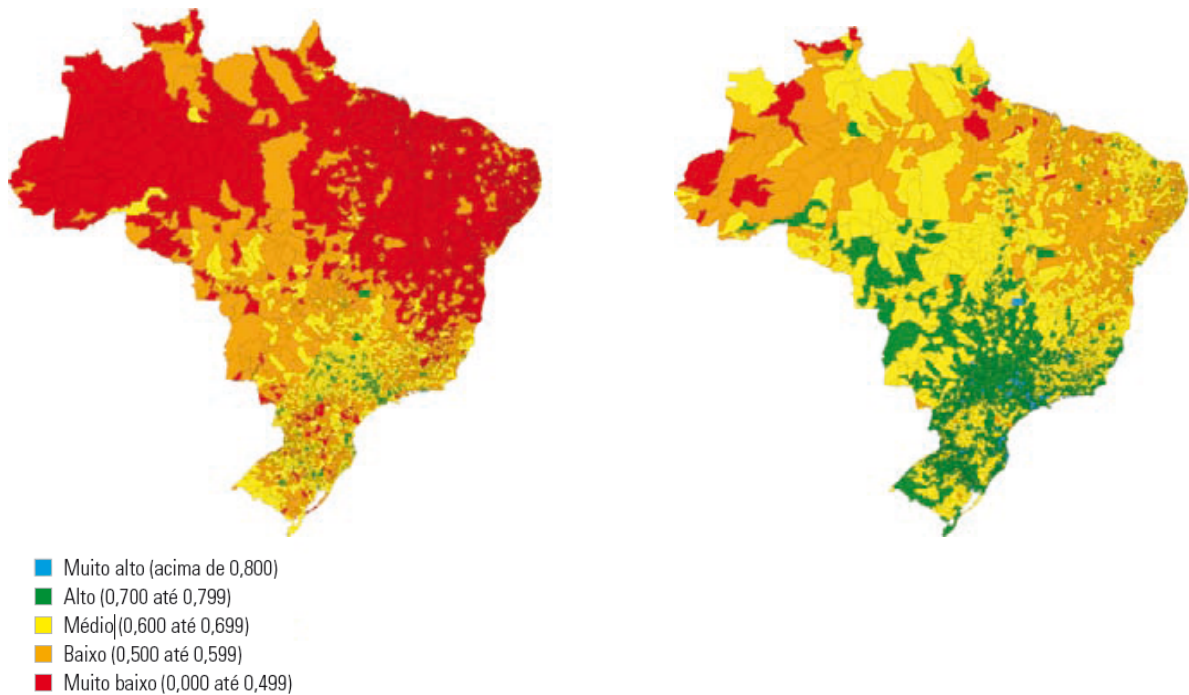
Essencial para o exercício das liberdades individuais, da autonomia e autoestima. A educação expande as habilidades das pessoas em relação a decisão dos seus futuros. Educação constrói dignidade e amplia os horizontes e as perspectivas de vida.

- Poder desfrutar de um padrão de vida digno - **renda**

Fundamental para ter acesso às necessidades básicas (água, comida, moradia) mas também para ir além dessas necessidades e usufruir do exercício da liberdade. Possibilita opções por alternativas disponíveis e sua ausência pode limitar as oportunidades de vida.

No Brasil, tal como em outros países, esse índice é adaptado a realidade dos municípios, realizando assim um IDHM, chamado IDH subnacional (ÍNDICE DE DESENVOLVIMENTO HUMANO MUNICIPAL BRASILEIRO, PNUD, 2013). Tendo como fonte para cálculo os indicadores do Censo Demográfico nacional que garante a unicidade das informações de todos os municípios.

Na figura 6 observa-se o IDHM do Brasil nos anos de 2000 e 2010, respectivamente.



Fonte: Série Atlas do Desenvolvimento Humano no Brasil, 2013

Figura 6: Mapas do IDHM do Brasil - 2000 e 2010

Variando de Muito baixo a Muito alto, podemos perceber o decaimento da faixa Muito baixo, o aumento significativo das faixas Médio e Alto, e o surgimento - ainda que pequeno - da faixa de Muito alto. Nos mapas, pode-se concluir que houve uma melhora do desenvolvimento humano no país na última década.

2.9 Cálculo do IDH

Atualmente os dados são calculados globalmente com uma média geométrica, temos:

$$IDH = \sqrt[3]{EV \times IE \times RN} \quad (2.18)$$

onde:

- EV = esperança de vida ao nascer;
- IE = combinação da média de anos de estudo da população com 25 anos ou mais e a expectativa de anos de estudo.
- RN = Renda Nacional Bruta per capita.

E em relação ao IDHM no Brasil, adaptando-o a cada município e tendo como base os dados dos Censos Demográficos realizados pelo IBGE, temos:

$$IDHM = \sqrt[3]{EV \times IE \times RM} \quad (2.19)$$

onde:

- EV = o número médio de anos que uma pessoa nascida em determinado município viveria a partir do nascimento, mantidos os mesmos padrões de mortalidade.
- IE = a média geométrica entre o percentual de pessoas de 18 anos ou mais de idade com ensino fundamental completo; e a média aritmética do percentual de crianças de 5 a 6 anos frequentando a escola, do percentual de jovens de 11 a 13 anos frequentando os anos finais do ensino fundamental, do percentual de jovens de 15 a 17 anos com ensino fundamental completo e do percentual de jovens de 18 a 20 anos com ensino médio completo; com pesos 1 e 2 respectivamente.
- RM = É a soma da renda de todos os residentes, dividida pelo número de pessoas que moram no município - inclusive crianças e pessoas sem registro de renda ou seja, renda per capita do município.

3 *Material e Métodos*

Como foi dito anteriormente, a aplicação de Bootstrap requer um certo desempenho computacional, por isso, utilizaremos o software RStudio versão 0.98.953 para sistema operacional Windows para analisar os dados. Sendo um software estatístico, ele possui todas as ferramentas necessárias além de ser gratuito e de fácil acesso.

A partir dos dados oficiais dos IDHM's do Brasil referentes aos anos de 2000 e 2010, publicados pelo Programa das Nações Unidas para o Desenvolvimento - PNUD - em 2013; selecionamos tamanhos n_1, \dots, n_{10} de amostras para comparação dos seus resultados em relação ao objetivo em questão.

Cada amostra n_i é repetida $B = 1000$ vezes e tiramos a estimativa da média amostral \hat{y} , a média Bootstrap \hat{y}_{boot} e o desvio padrão amostral e Bootstrap das reamostragens s e s_{boot} , em seguida, se faz os 3 tipos de intervalos de confiança de Bootstrap e o intervalo de confiança convencional (Frequentista) ao nível $\alpha = 0,025$ de significância e comparamos se a verdadeira média se encontra dentro de algum dos intervalos. Esse processo é repetido $nsim=1000$ vezes, e cada vez que a média se encontra dentro do intervalo, conta mais 1, ao fim do processo a soma desses 1's é dividida pelo valor de $nsim$ para obtenção da porcentagem de vezes classificadas corretas; a essa porcentagem, damos o nome de *cobertura* e quanto maior mais eficiente o método.

O algoritmo é repetido $J = 30$ vezes para cada n_i para as 3 variáveis - IDHM 2000, IDHM 2010 e taxa de crescimento R - e assim é possível se obter um grau de incerteza para a taxa de cobertura que pode ser conferida visualmente com o auxílio de um box-plot.

4 *Resultados e Discussões*

Primeiramente, realizamos a análise descritiva da população original, dos IDHM's dos 5565 municípios brasileiros em 2000 e em 2010 e a razão entre eles, para traçar seus parâmetros principais. Os resultados são mostrados na Tabela 1, para o índice de desenvolvimento humano municipal em 2000, em 2010 e a razão entre eles - ou seja, a taxa do crescimento em 10 anos - temos os dados a seguir:

Tabela 1: Parâmetros do IDHM em 2000 e 2010, e sua razão R.

	Mínimo	1° Quartil	Mediana	Média	3° Quartil	Máximo
2000	0,2080	0,4360	0,5330	0,5235	0,6090	0,8200
2010	0,4180	0,5990	0,6650	0,6592	0,7180	0,8620
R	1,044	1,171	1,252	1,286	1,374	2,495

Observa-se um aumento nos valores do IDHM em 2010, o valor Mínimo se aproximou da faixa de IDHM Baixo, uma mudança de faixa para o 1° Quartil saindo de Muito Baixo para Baixo, a Mediana e a Média saíram da faixa de Baixo e passaram para Médio e o 3° Quartil saiu da faixa de Médio para Alto. Graficamente, temos:

A Figura 7 mostra que o índice em 2000, aparentemente, poderia seguir uma distribuição Normal e possuir uma certa simetria, suas observações se concentram na faixa de 0,4 à 0,65.

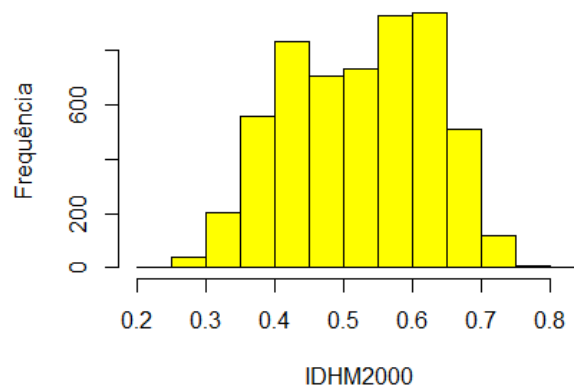


Figura 7: Índice de Desenvolvimento Humano Municipal em 2000

A Figura 8 mostra que o índice em 2010, aparentemente, poderia seguir uma distribuição Normal e possuir uma certa simetria, suas observações se concentram na faixa de 0,55 à 0,75. O que já mostra uma melhora nos índices 10 anos após.

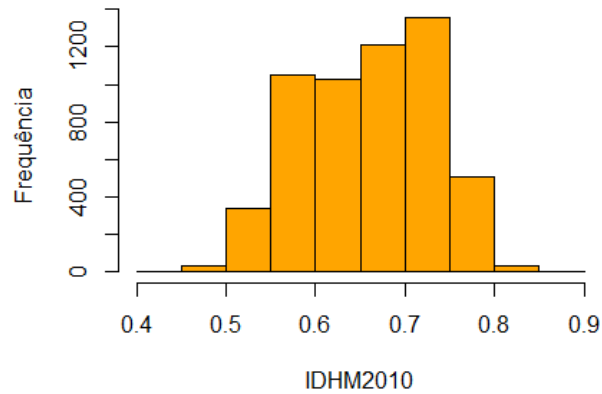


Figura 8: Índice de Desenvolvimento Humano Municipal em 2010

A Figura 9 mostra que o índice da razão R - taxa de crescimento - aparentemente não segue uma distribuição Normal e não possui simetria, suas observações se concentram na faixa de 1,1 à 1,4.

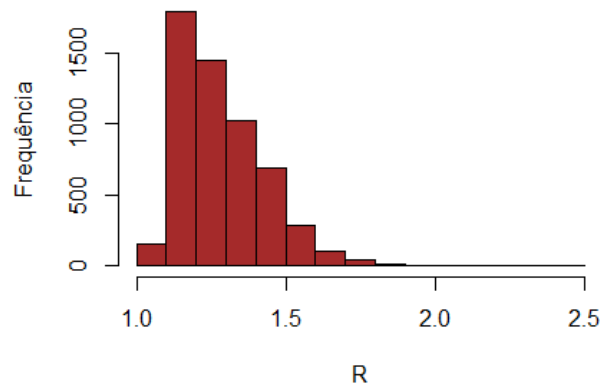


Figura 9: Taxa de crescimento R do IDHM em 10 anos

A partir dos histogramas, observa-se que os IDHM's aparentam ter uma distribuição Normal, mas com o P-valor= $3,314 \times 10^{-11}$ para o IDHM de 2000 e P-valor $< 2,2 \times 10^{-16}$ para o IDHM de 2010, a hipótese de normalidade é rejeitada ao nível $\alpha = 0,01$ de significância pelo teste de Kolmogorov-Smirnov; da mesma forma a taxa de crescimento R, com P-valor $< 2,2 \times 10^{-16}$, não segue uma distribuição Normal - como era de se esperar - por ser uma razão (COCHRAN, 1977).

Continuando com as análises, foi retirada uma amostra piloto de tamanho 20 e foi calculado quanto deveria ser o valor de n para se obter 0,95 de confiança, chegando a conclusão que $n=154$ seria uma quantidade satisfatória, ou seja, com esse número de observações é esperado que o 3º quartil alcance a faixa de 0,95.

Em seguida, foram retiradas amostras de tamanho 5, 10, 20, 30, 50, 70, 100, 120, 150 e 200, com essas amostras foi estimada a média e o desvio para o caso da amostragem aleatória e para o caso Bootstrap em que foi usado $B = 1000$, isto é, a reamostragem foi repetida 1000 vezes para realizar as estimativas Bootstrap. Esse processo para ambos os métodos foi repetido 1000 vezes e todas as vezes que as estimativas caíam dentro do intervalo de confiança com $\alpha = 0,05$ era computado o valor 1, ao fim do processo, os valores computados são divididos por 1000 a fim de obter a porcentagem das vezes que o processo foi classificado como correto; que chamamos de cobertura. Após obter uma estimativa para a taxa de cobertura, se repete o algoritmo todo por mais 29 vezes para se obter o desvio da taxa de cobertura.

Lembrando que temos os intervalos de confiança Normal, Pivotal e Percentil para Bootstrap e o intervalo de confiança Frequentista para a amostragem aleatória simples. Os resultados se encontram na tabela a seguir:

Tabela 2: Coberturas para cada intervalo de confiança

n_i	Normal		Pivotal		Percentil		Frequentista	
	Média	Desvio	Média	Desvio	Média	Desvio	Média	Desvio
5	0,8247	0,0152	0,7987	0,0129	0,8246	0,0129	0,8553	0,0130
10	0,8858	0,0094	0,8715	0,0100	0,8924	0,0089	0,8998	0,0124
20	0,9173	0,0106	0,9092	0,0090	0,9177	0,0086	0,9256	0,0071
30	0,9288	0,0083	0,9233	0,0074	0,9274	0,0083	0,9338	0,0090
50	0,9370	0,0069	0,9317	0,0072	0,9375	0,0075	0,9411	0,0087
70	0,9421	0,0069	0,9353	0,0050	0,9403	0,0070	0,9412	0,0062
100	0,9455	0,0078	0,9411	0,0081	0,9430	0,0066	0,9449	0,0064
120	0,9445	0,0077	0,9459	0,0069	0,9465	0,0057	0,9466	0,0073
150	0,9464	0,0086	0,9457	0,0078	0,9474	0,0071	0,9501	0,0063
200	0,9514	0,0079	0,9500	0,0078	0,9494	0,0053	0,9509	0,0066

Podemos observar que há uma diminuição nos desvios das coberturas à medida que n aumenta.

E seus respectivos boxplots:

Observa-se na Figura 10, cobertura Normal Bootstrap, que com o $n=100$ já há uma inclusão de 0,95 no intervalo; que é um bom resultado. Existindo a presença de *outliers* em $n=50$ e $n=100$, parece haver simetria quando $n=30$.

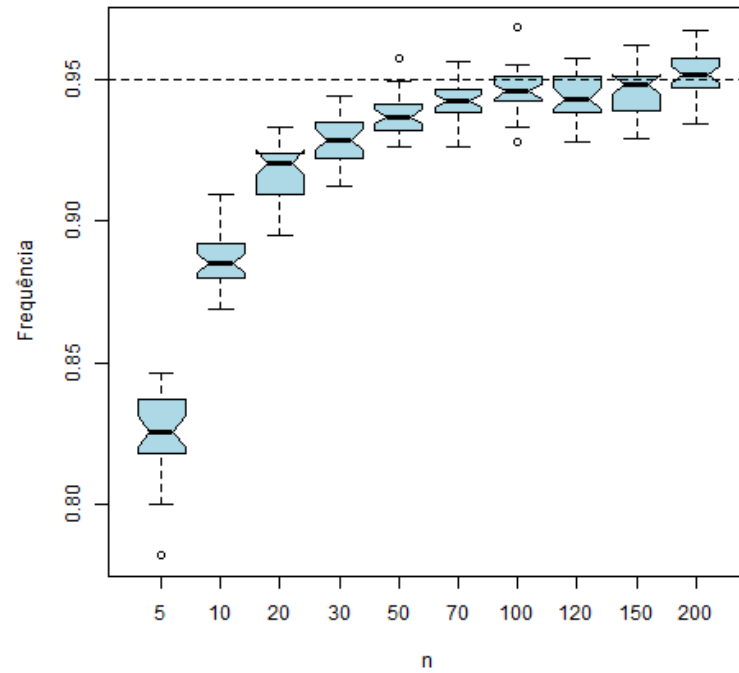


Figura 10: Cobertura Normal

A Figura 11, cobertura Pivotal Bootstrap, há uma inclusão completa do 3° quartil nos 0,95 mas apenas com $n=200$. Observa-se dados discrepantes com $n=10$, $n=30$ e $n=200$.

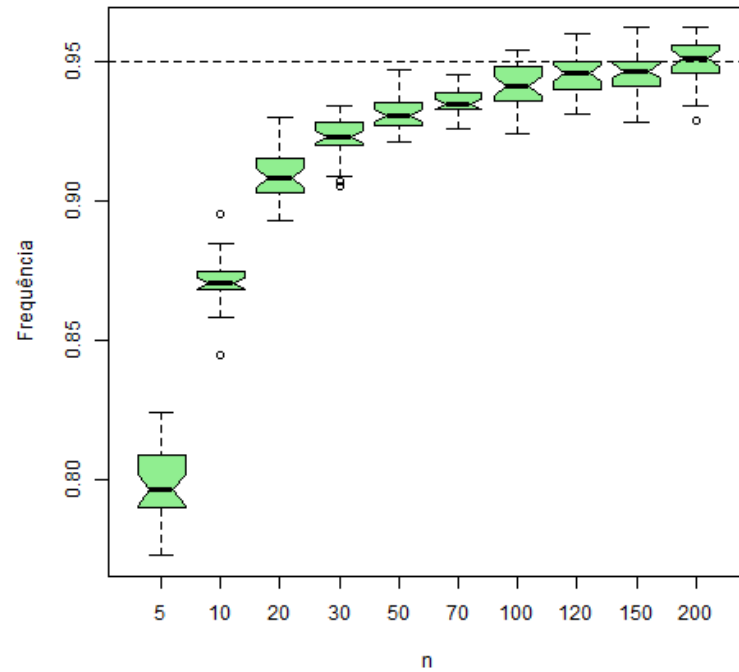


Figura 11: Cobertura Pivotal

Para a cobertura Percentil Bootstrap (Figura 12), a aproximação da faixa de 0,95 começa com $n=120$. Com valores fora do limite inferior em $n=5$ e fora do limite superior em $n=50$.

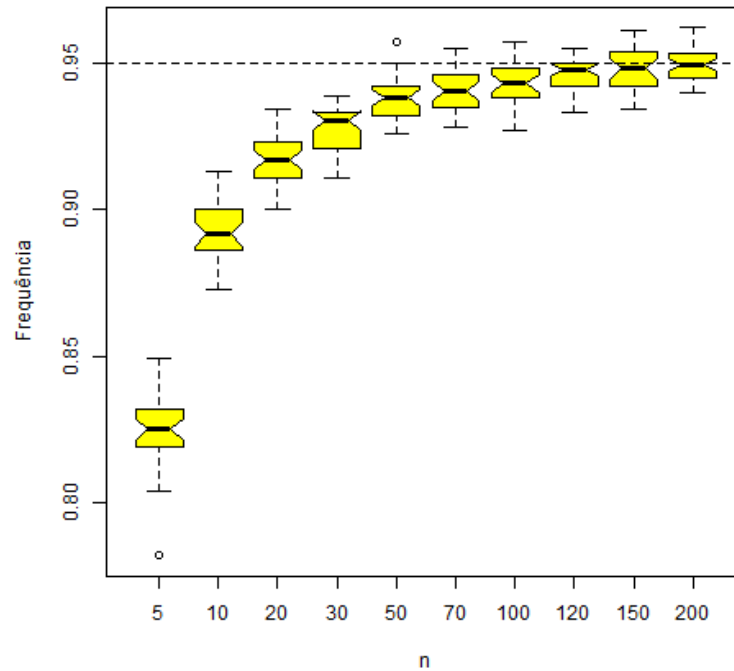


Figura 12: Cobertura Percentil

Na cobertura Frequentista - Figura 13 - a completa inclusão no 3º quartil só foi possível em $n=150$. Observando-se *outliers* nas amostras de tamanho $n=10$, $n=120$ e $n=200$.

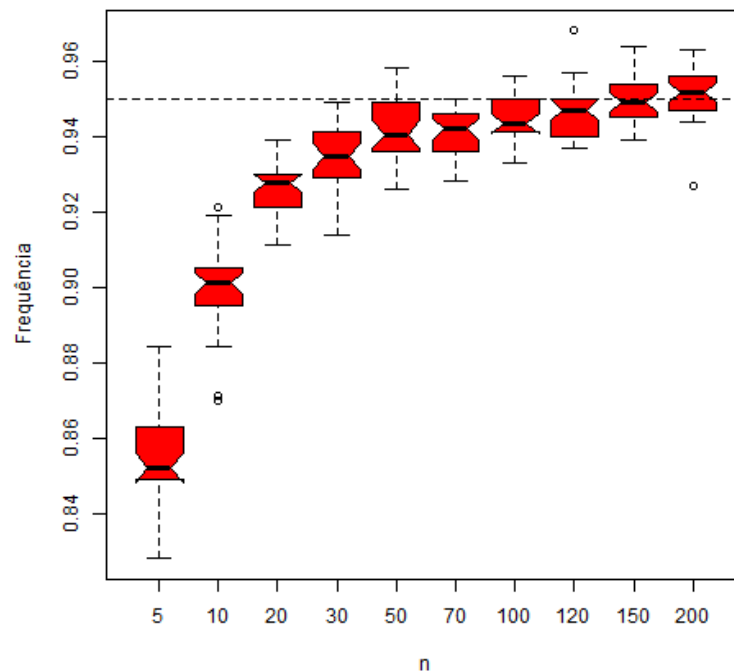


Figura 13: Cobertura Frequentista

Os resultados sugerem que não há diferenças significativas entre os métodos, os gráficos são semelhantes ao convencional mas o que mais se destaca é o Normal Bootstrap.

Tendo visto a eficiência da técnica Bootstrap é fácil imaginá-lo em trabalhos futuros sendo aplicado em outros métodos e modelos. Por não necessitar de muitos pressupostos para estimação de parâmetros, poderia tornar modelos mais complexos, como os Não-Lineares, em modelos mais simples de serem trabalhados. Ou ainda, ser aplicado a estatísticas Não-Paramétricas, onde geralmente se tem amostras pequenas e que não seguem normalidade, os diferentes métodos de cálculo de intervalos de confiança Bootstrap na forma não paramétrica podem ser: o Intervalo de Confiança Bootstrap Percentil das Diferenças, o Intervalo de Confiança Bootstrap t, o Intervalo de Confiança Percentil Corrigido em Relação ao Viés (BCPB) e o Intervalo de Confiança de Correção de Vício Acelerado (BCa).

Exemplos muito práticos são observados em indústrias e fábricas de grande porte, onde o custo para se obter uma amostra é muito alto e se faz necessário o uso de planos amostrais que são otimizados com técnica Bootstrap aplicada na engenharia de produção e controle de qualidade do processo.

5 *Conclusões*

Primeiramente, podemos concluir que houve uma melhora significativa na taxa de crescimento do índice de desenvolvimento humano nos municípios brasileiros entre os anos de 2000 e 2010. Com o uso da teoria da amostragem aleatória simples é possível se concluir que 154 observações seriam suficientes para estimar esse crescimento.

Após as análises feitas com o plano de Amostragem Aleatória Simples e Bootstrap, podemos concluir também, que os resultados obtidos são semelhantes. Mas no caso em questão, temos acesso à população, o que quase nunca acontece na maioria dos problemas que envolvem amostras, e nesses casos, os resultados com Bootstrap podem ser mais interessantes e eficazes.

Também foi observado pelo teste de Kolmogorov-Smirnov, que as distribuições (IDHM 2000, IDHM 2010 e a taxa de crescimento R) não seguiam uma distribuição Normal, ou seja, nem sempre temos um conjunto de dados que satisfazem todos os pressupostos necessários para uma análise estatística. Nestes casos, a reamostragem, como foi constatado, terá um desempenho competitivo.

Levando em consideração a abrangência da técnica de Bootstrap, neste trabalho foi abordada uma pequena parte que diz respeito à estimar parâmetros de interesse - onde essa estimação de parâmetros pode ser estendida para todos os métodos que trabalham com conjuntos de dados e que possuem esse objetivo em comum, como todos os outros tipos de planos amostrais (estratificada, por blocos, conglomerados, etc) ou ainda testes não-paramétricos. E também é aplicada em modelos (sejam eles de regressão, lineares e não-lineares, multivariados, entre outros) com o objetivo de otimizá-los.

Referências

- BOLFARINE, H.; BUSSAB, W. O. **Elementos da Amostragem**. 1°ed. São Paulo: Blucher, 2005.
- COCHRAN, W. G. **Sampling Techniques**. 3°ed. Advisors, 1977.
- EFRON, B; TIBSHIRANI, R. J. **An Introduction to the Bootstrap**. 1°ed. United States of America: Chapman & Hall/CRC, 1993.
- LIERO, H. **An Introduction to the Bootstrap**. University of Potsdam, 2014.
- PNUD. "Atlas do desenvolvimento humano no Brasil", 2013.
- TIBSHIRANI, R. J. et. al. **An Introduction to Statistical Learning: with Applications in R**. 1°ed. New York: Springer, 2013.
- WASSERMAN, L. **All of Statistics: A Concise Course in Statistical Inference**. 1°ed. New York: Spring, 2004.

Apêndice

Apêndice A - Códigos R utilizados nas aplicações

```

setwd('D:IDH' )
idh=read.table('IDHM.txt',head=T)
attach(idh)
idh = idh[-1,]
Y = IDHM2010
Y = IDHM2010/IDHM2000
detach(idh)

summary(IDHM2000)
hist(IDHM2000, col="yellow")
summary(IDHM2010)
hist(IDHM2010, col="orange")
summary(Y)
hist(Y, col="brown")

icfunction = function(Y, B, n, nsim)

c1 = 0; c2 = 0; c3 = 0; c4 = 0
y.boot = 0

for(i in 1:nsim)
y=sample(Y,n)
y.hat = mean(y)
Sy.hat = sd(y)/sqrt(n)

for(b in 1:B) y.boot[b] = mean(sample(y,n,rep=T))
ca = quantile(y.boot, probs = c(.025))
cb = quantile(y.boot, probs = c(.975))
Sy.boot = sd(y.boot)
y.boot = mean(y.boot)

if(y.boot-1.96 Sy.boot < mean(Y) && mean(Y) < y.boot +1.96 Sy.boot) c1 = c1 + 1

```



```

if(2 y.boot - cb < mean(Y) && mean(Y) < 2 y.boot - ca) c2 = c2 + 1
if(ca < mean(Y) && mean(Y) < cb) c3 = c3 + 1
if(y.hat-1.96 Sy.hat < mean(Y) && mean(Y) < y.hat +1.96 Sy.hat) c4 = c4 + 1
return(list(y.hat = y.hat, Sy.hat = Sy.hat, y.boot = y.boot, Sy.boot = Sy.boot, c1 =
c1/nsim, c2 = c2/nsim, c3 = c3/nsim, c4 = c4/nsim))
}

J = 10; I = 30
c1 = c2 = c3 = c4 = matrix(0,I,J)

tamanho = c(5, 10, 20, 30, 50, 70, 100, 120, 150, 200)

ptm <- proc.time()
for(j in 1:J)
for(i in 1:I)
c1[i,j] = icfunction(Y, 1000, tamanho[j], 1000)$ c1
c2[i,j] = icfunction(Y, 1000, tamanho[j], 1000)$ c2
c3[i,j] = icfunction(Y, 1000, tamanho[j], 1000)$ c3
c4[i,j] = icfunction(Y, 1000, tamanho[j], 1000)$ c4

}
proc.time() - ptm

g = as.factor(rep(tamanho, rep(I,J)))
png('coberturanormal.png')
boxplot(split( as.vector(c1), g), col="lightblue", notch=T, xlab='n',
ylab='Frequência')
abline(h = 0.95, lty = 2)
dev.off()

png('cobeturapivotal.png')
boxplot(split( as.vector(c2), g), col="lightgreen", notch=T, xlab='n',
ylab='Frequência')
abline(h = 0.95, lty = 2)
dev.off()

png('cobeturapercentil.png')
boxplot(split( as.vector(c3), g), col="yellow", notch=T, xlab='n',
ylab='Frequência')
abline(h = 0.95, lty = 2)

```

```

dev.off()

png('coberturafreq.png')
boxplot(split( as.vector(c4), g), col="red", notch=T, xlab='n',
ylab='Frequência')
abline(h = 0.95, lty = 2)
dev.off()

A=sample(Y,20)
s=sqrt(var(A))
((1.96 * s) / 0.02)

A=sample(Y,154)
N=(rnorm(5565,mean(IDHM2000),sd(IDHM2000)))
ks.test(N,IDHM2000)
N1=(rnorm(5565,mean(IDHM2010),sd(IDHM2010)))
ks.test(N,IDHM2010)
N2=(rnorm(5565,mean(Y),sd(Y)))
ks.test(N,Y)

```