Recebido em: 20/08/08
Aprovado em: 01/02/10

Intervalo de confiança e teste de significância *bootstrap* para coeficiente de correlação linear referente à hipótese de um valor não nulo

Giovani Glaucio de Oliveira Costa (UFRRJ, RJ, Brasil) – giovaniglaucio@ufrrj.br
• Av. Governador Roberto Silveira s/n, Jardim da Posse, Nova Iguaçu-RJ, CEP 26.020-740

Resumo: A distribuição amostral do coeficiente de correlação amostral r, sob a hipótese nula de o coeficiente de correlação populacional ñ=0, é simétrica, enquanto que, sob a hipótese nula ñ≠0, é assimétrica. No primeiro caso, se utiliza uma estatística que envolve a distribuição t de Student, e no segundo caso, se recorre a uma alternativa desenvolvida por Fisher, a qual dá origem a uma estatística com distribuição aproximadamente normal, obtida através da transformação da estatística r numa estatística £, que tem distribuição bastante próxima da normal. Como alternativa a este último processo, pode-se criar uma distribuição de amostragem real empírica de estimativas de coeficientes de correlação através de simulações bootstrap e a partir daí construir intervalos de confiança não paramétricos e testar hipóteses para um valor não nulo de ñ, baseada nesta estimação intervalar, sem ter que se preocupar com a normalidade da distribuição de amostragem da estatística r. A proposta deste trabalho é sugerir uma alternativa viável e prática para a construção do intervalo de confiança para ñ ou para o teste de significância de r. De maneira operacional e em tempo hábil, graças ao crescente avanço da informática e a disponibilidade de variados softwares estatístico amigáveis, pode ser realizada com mais freqüência a inferência para o coeficiente de correlação linear nos casos em que o valor do coeficiente de correlação linear populacional testado seja o de não nulidade. Portanto, o presente artigo propõe obter via processo de reamostragem, através da técnica bootstrap, uma distribuição por amostragem real empírica para a estatística r e calcular o seu erro padrão, possibilitando assim construir intervalo de confiança e realizar testes de significância bootstrap para o coeficiente de correlação linear referente à hipótese de um valor não nulo.
Palavras-chaves: coeficiente de correlação linear, intervalo de confiança, teste de significância, hipótese nula de coeficiente de correlação linear diferente de zero, bootstrap.

Abstract: The sampling distribution of the coefficient of sampling correlation r, under the null hypothesis of the coefficient of population correlation ñ=0, is symmetrical, whereas, under the null hypothesis ñ≠0, is anti-symmetrical. In the first case, if it uses a statistics that involves distribution t of Student, and in as in case that, if it appeals to an alternative developed for Fisher, which of the origin to a statistics with approximately normal distribution, gotten through the transformation of statistics r in a statistics £, that it has distribution sufficiently next to the normal one. As alternative to this last process, a distribution of empirical real sampling of estimates of coefficients of correlation through simulation bootstrap and from constructing confidence intervals distribution free there can be created and tests of hypotheses for a not null ñ-value, based in this esteem to intervalar, without having that to be worried about the normality of the distribution of sampling of statistics r. The proposal of this work is to suggest a viable and practical alternative for the construction of confidence intervals for ñ or the test of significance of r. In operational way and in skillful time, thanks to the increasing advance of the computer science and the availability of varied softwares friendly statistician, can be carried through with more frequency the inferential procedure for the coefficient of linear correlation in the cases where the value of the tested coefficient of population linear correlation is of not nullity. Therefore, the present article considers to get way resampling process, through the technique bootstrap, a distribution for empirical real sampling it statistics r and to calculate its error standard, being thus made possible to construct confidence intervals and to carry through significance tests bootstrap for the coefficient of referring linear correlation to the hypothesis of a not null value.
Word-keys: coefficient of linear correlation, confidence intervals, test of significance, null hypothesis of coefficient of different linear correlation of zero, bootstrap.

GEPROS. Gestão da Produção, Operações e Sistemas – Ano 5, nº 2, Abr-Jun/2010, p. 177-186

1. INTRODUÇÃO

O coeficiente de correlação *r* ou *r* (X, Y), introduzido por Karl Pearson, é também denominado correlação momento-produto (DE GROOT e SCHERVISH, 2002). Na população, o coeficiente *ñ* mede a aderência ou a qualidade do ajuste à verdadeira reta, através da qual se procura relacionar as variáveis X e Y ou ainda o grau de relação (linear) existente entre elas (DEGROOT e SCHERVISH, 2002). Já o coeficiente *r*, mede a quantidade de dispersão em torno da equação linear ajustada através do método dos mínimos quadrados, ou grau de relação das variáveis na amostra (DE GROOT e SCHERVISH, 2002). O *r* é, portanto, uma estimativa de *ñ*, medindo os desvios em relação à linha calculada pelo método dos mínimos quadrados (DE GROOT e SCHERVISH, 2002).

É importante notar que a dispersão em torno da reta poderia igualmente ser medida através do desvio padrão, sendo esse último preferido por muitos estatísticos (FONSECA *et. al.*, 1995). Não obstante, o uso do coeficiente de correlação permanece, principalmente devido à vantagem que apresenta decorrente da facilidade de interpretação e de seu intervalo compreender valores em um intervalo com uma escala reduzida, no intervalo de [-1 ; 1] (FONSECA *et. al.*, 1995).

O uso do coeficiente de correlação por pesquisadores como indicador da relação das variáveis X e Y é de uso corrente na estatística (FONSECA *et. al.*, 1995), mas quase sempre à nível descritivo(GIOVANI COSTA, 2011), pois quando se procura testar a significância de *r* para a hipótese nula de que *ñ*≠0, se esbarra na dificuldade técnica da modelagem da distribuição amostral de *r*, que não se pode admitir normalmente distribuída, por ser assimétrica (GIOVANI COSTA, 2011). Neste caso, Fisher sugere a transformação (FONSECA *et. al.*, 1995):

$$£, = ½ \ln \frac{1+ r}{1- r} = 1{,}1513[\log10(1+r) - \log10(1- r)] \quad (1)$$

O procedimento descrito acima equivale a considerar *r* como a tangente hiperbólica de £,. A vantagem dessa transformação está em que os valores de £ têm distribuição bastante próxima da normal, com:

$$E(£,) = ½ \ln \frac{1+ ñ}{1- ñ} \quad (2)$$

$$ó(£,)= \sqrt{} 1/ (n - 3) \quad (3)$$

Esta transformação permite realizar testes de significâncias e construir intervalos de confiança para os coeficientes de correlação, trabalhando-se com £, e usando a curva normal (FONSECA *et. al.*, 1995). Contudo, o referido processo por ser relativamente complexo inibiu o uso corrente e sistemático da inferência para teste de significância referente a um valor não nulo de *ñ*.

Imagina-se, inicialmente, que os n pares de valores observados das variáveis X e Y constituem uma amostra de tamanho n, extraída da população de todos os pares de valores possíveis dessas variáveis (FONSECA *et. al.*, 1995). Em virtude de se estar considerando duas variáveis, a população é denominada bidimensional e sua distribuição, por hipótese, é uma distribuição normal bidimensional (FONSECA *et. al.*, 1995).

Dessa forma, apesar de *r* só descrever os dados da amostra, o interesse se centraliza, via de regra, no parâmetro da população(FONSECA *et. al.*, 1995). Em particular, se desejaria provar a hipótese nula de que não há relação(linear) alguma na população(H0 : $\tilde{n}$=0) ou obter intervalos de confiança para $\tilde{n}$. Neste caso, $\tilde{n}$ representa o coeficiente de correlação de uma população teórica, o qual é estimado a partir do coeficiente de correlação amostral *r* (FONSECA *et. al.*, 1995).

A distribuição amostral de *r*, sob a hipótese de $\tilde{n}$=0, é simétrica, enquanto que, sob a hipótese $\tilde{n}$≠0, é assimétrica (FINSECA *et. al.*, 1995). No primeiro caso, se utiliza uma estatística que envolve a distribuição *t* de Student (STUDENT,1908), e no segundo caso, se recorre a uma alternativa desenvolvida por Fisher, a qual dá origem a uma estatística £, com distribuição aproximadamente normal (FONSECA *et. al.*, 1995). Esta é a alternativa teórica disponível para o caso da distribuição amostral de *r* for assimétrica: um tanto complexa e que por isso na prática se torna inviável.

Como alternativa, pode-se criar uma distribuição de amostragem real empírica de estimativas de coeficientes de correlação a partir de simulações e daí construírem intervalos de confiança não paramétricos e realizar testes de significância baseados nesta estimação intervalar (HAIR *et. al.*, 2005).

A proposta deste trabalho é justamente sugerir uma alternativa viável e prática para o teste de significância de *r* ou a construção do intervalo de confiança para $\tilde{n}$. De maneira simples e rápida pode ser realizada a inferência para o coeficiente de correlação linear e possibilitar que o cálculo de *r*, em casos em que a hipótese nula contenha valor diferente de zero, saia do terreno puramente descritivo.

Este o método para o estudo consiste:

- Gerar uma estimativa empírica, real, da distribuição por amostragem da variável aleatória 'coeficiente de correlação amostral r;
- Aplicar metodologias CIS (*Computer Intensive Statistics*) para obtenção das respectivas distribuições por amostragem;
- Empregar o método CIS *bootstrap*;
- Especificar o erro padrão da estatística 'coeficiente de correlação amostral *r*';
- Com o erro padrão disponível, construir intervalos de confiança e realizar testes de significância para $\tilde{n}$≠0.

A abaixo se relaciona os motivos para aplicação de metodologias CIS:

- Não se conhecem com a precisão necessária os parâmetros característicos teóricos da distribuição por amostragem da variável aleatória 'coeficiente de correlação amostral *r*', quando $\tilde{n}$≠0;
- O estudo da distribuição por amostragem do 'coeficiente de correlação amostral *r*, quando $\tilde{n}$≠0, assume a forma assimétrica, o que dificulta a especificação teórica do erro padrão;
- A distribuição por amostragem fornece um modo direto para conhecer as estimativas do erro padrão;
- O *bootstrap* permite obter, então, de forma experimental e empírica as distribuições por amostragem da estatística em foco;
- O método *bootstrap* permite ladear a insuficiência da teoria da amostragem que se faz sentir quando se calcula o coeficiente de correlação amostral *r* em diversas situações práticas (MURTEIRA, 1990).

GEPROS. Gestão da Produção, Operações e Sistemas – Ano 5, nº 2, Abr-Jun/2010, p. 177-186

O desenvolvimento recente da informática tem permitido que técnicas de reamostragem como o *bootstrap* possam ser operacionalizadas de maneira rápida e precisa (HAIR *et. al.*, 2005). Recorrendo a 'computação pesada' consegue-se solucionar problemas para as quais a teoria da estatística tradicional não encontra solução (HAIR *et. al.*, 2005).

Como produto deste estudo, pretende-se estabelecer um critério alternativo simples e prático para testar a significância do coeficiente de correlação amostral *r*, utilizando a noção de intervalo de confiança *bootstrap*, onde a hipótese da normalidade da referida distribuição por amostragem não precisa ser presumidamente normal.

2. METODOLOGIA DO TRABALHO

Como comentado na introdução, este trabalho objetiva propor um processo inferencial para o coeficiente de correlação. A idéia é utilizar técnicas CIS (*Computer Intensive Statistics*), que cogitam o modelo de densidade de probabilidade e que explica o comportamento aleatório da estatística observada e seus parâmetros característicos (EFRON ,1979).

As técnicas CIS dispõem-se principalmente de dois métodos, o *bootstrap* e o *jackknife* (EFRON, 1979). Este trabalho trata da estimação *bootstrap*.

Através deste artigo é especificado o intervalo de confiança para *r* recorrendo-se ao procedimento *bootstrap*. Com estes resultados, pode-se obter um procedimento computacional, um algoritmo, para a construção de intervalos de confiança e realização de testes de significância para as estimativas obtidas.

Os coeficientes de correlação amostral n são calculados e utilizados com grande freqüência, inclusive quando a hipótese nula envolve valor diferente de zero, mas usualmente a nível descritivo, já que o modelo da distribuição por amostragem exata da variável aleatória *r*, não simétrica, torna inviável fazer acompanhar as estimativas do respectivo erro padrão, para não falar na construção de intervalos de confiança ou na realização de testes de significância, quando a hipótese nula formulada é de que $\tilde{n} \neq 0$.

A opção de se usar as metodologias CIS surge quando não se conhece o desvio padrão teórico das estimativas e/ou quando o modelo de distribuição de probabilidade destas estimativas não se adere à curva normal de probabilidades, ou mesmo quando o desvio-padrão da estimativa é de cálculo muito complexo (EFRON, 1979). Este último pressuposto, quando o modelo de distribuição de probabilidade destas estimativas não se adere à curva normal de probabilidades, é violado (hipótese nula $\tilde{n} \neq 0$) com o coeficiente de correlação amostral r.

Nestes casos, com a aplicação do *bootstrap* é possível obter, de forma expedita, através da computação "pesada", estimativas do desvio padrão da estatística em causa em substituição análise teórica (MURTEIRA, 1990). Com o *bootstrap*, por exemplo, é possível determinar a distribuição por amostragem da estatística e seus parâmetros característicos (HAIR *et. al.*, 2005). O método *bootstrap* permite, então, ladear a insuficiência da teoria da amostragem que se faz sentir em diversos estudos de estimação (MURTEIRA, 1990).

3. RESUMO TEÓRICO DE REAMOSTRAGEM

O tipo de estatística não-paramétrica que foi ensinado no passado desempenhou um importante papel na análise de dados que não são variáveis contínuas (EFRON, 1982), em escala nominal ou ordinal, e, portanto, não podem empregar a distribuição normal de probabilidade para fazer estimativas de parâmetros e de intervalo de confiança. Mas existe uma nova perspectiva sobre estimação não-paramétrica que também se relaciona com estimação de parâmetros e de intervalo de confiança para variáveis quantitativas contínuas.

Com isso, não se tem que assumir que o intervalo de confiança para um parâmetro segue a distribuição normal (EFRON, 1982). Pode-se até mesmo gerar intervalos de confiança para parâmetros como a mediana, o que geralmente é difícil de avaliar com as técnicas de inferência paramétrica tradicionais.

Essa abordagem não-paramétrica é conhecida como reamostragem e tem conquistado apoio como uma alternativa aos métodos clássicos de inferência paramétrica (EFRON, 1982).

A reamostragem descarta a distribuição por amostragem assumida de uma estatística e calcula uma distribuição empírica: A real distribuição da estatística ao longo de centenas ou milhares de amostras (HAIR *et. al.*, 2005).

Com a reamostragem, não se tem que confiar na distribuição assumida nem se tem que ser cuidadoso quanto à violação de uma das suposições inerentes. Pode-se calcular uma real distribuição de estatísticas da amostra e pode-se agora ver onde o 95 ou o 99 percentil estão realmente, acreditando-se que a amostra original seja confiável (HAIR *et. al.*, 2005).

Mas de onde vêm as múltiplas amostras? É necessário reunir amostras separadas, aumentando sensivelmente o custo de coleta de dados? Ao longo dos anos estatísticos desenvolveram diversos procedimentos para criar as múltiplas amostras necessárias para a reamostragem a partir da amostragem original (HAIR *et. al.*, 2005).

Agora uma amostra pode gerar um grande número de outras amostras que podem ser empregadas para gerar a distribuição amostral empírica de uma estatística de interesse (HAIR *et. al.*, 2005).

Reamostragem, portanto, não usa a distribuição de probabilidades assumida, mas ao invés disso ela calcula uma distribuição empírica de estatísticas estimadas (HAIR *et. al.*, 2005). Criando múltiplas amostras da amostra original, a reamostragem agora precisa apenas do poder computacional para estimar um valor de uma estatística para cada amostra (HAIR *et. al.*, 2005). Logo que eles estejam todos calculados, pode-se realizar o teste de normalidade dos valores e até mesmo construir intervalos de confiança e realizar testes de hipóteses (HAIR *et. al.*, 2005).

A reamostragem engloba diversos métodos (HAIR *et. al.*, 2005). Para este trabalho, se estudará e aplicará o *bootstrap*.

Uma diferença chave entre os vários métodos de reamostragem é se as amostras são extraídas com ou sem reposição (HAIR *et. al.*, 2005). A amostragem com reposição obtém uma observação a partir da amostra e então a coloca de volta na amostra para possivelmente ser usada novamente. A amostragem sem reposição obtém observações da amostra, mas uma vez obtidas eles não estão mais disponíveis.

O verdadeiro poder da reamostragem vem de amostragem com reposição (HAIR *et. al.*, 2005). Pesquisas têm mostrado (HAIR *et. al.*, 2005)que esse método fornece estimativas diretas dos intervalos de confiança.

GEPROS. Gestão da Produção, Operações e Sistemas – Ano 5, nº 2, Abr-Jun/2010, p. 177-186

O método *bootstrap* obtém sua amostra via amostragem com reposição da amostra original (EFRON, 1982). A chave é a reposição das observações após a amostragem, o que permite ao pesquisador criar tantas amostras quanto necessárias e jamais se preocupar quanto à duplicação de amostras, exceto quando isso acontecer ao acaso (EFRON, 1982). Cada amostra pode ser analisada independentemente e os resultados compilados ao longo da amostra. Por exemplo, a melhor estimativa da média populacional é exatamente a média de todas as médias estimadas ao longo das amostras (EFRON, 1982).

O intervalo de confiança também pode ser diretamente calculado. As duas abordagens mais simples:

1. Calculam o erro padrão simplesmente como o desvio padrão das estimativas estimadas;
2. Literalmente ordenam as estimativas e definem os valores que contém os 5% extremos (ou 1%) dos valores estimados.

Matematicamente, a obtenção da amostra *bootstrap* e suas estimativa do erro padrão é obtida da seguinte maneira:

Seja uma amostra original e a estatística de interesse abaixo:

$x = \{x_1, x_2, x_3, ..., x_{n-1}, x_n\}$.

(1º) Geram-se as amostras *bootstrap* $x_{(1)}, x_{(2)}, x_{(3)}, ..., x_{(n^*)}$ com reposição de x.

(2º) Calculam-se as estimativas da estatística de interesse:

$_{(b)} = F[x_{(b)}]$, b=1, ..., B

(3º) Calcula-se o erro padrão *bootstrap*, S_{boot}, dado por:

$$\left\{ S_{boot} = \frac{1B}{B-1_{b=1}} . \Sigma [b - (*)] \; 2 \; 1/2 \right\}, \text{ sendo} \qquad (4)$$

$$B_{(*)(b)} = \frac{b=1}{B} \qquad (5)$$

Apesar de procedimentos de reamostragem não serem restritos por quaisquer suposições paramétricas, eles ainda têm certas limitações:

- A amostra deve ser grande o bastante e obtida (a princípio aleatoriamente) de forma a ser representativa da população completa. Técnicas de reamostragem não podem conter quaisquer enviezamentos que traga como conseqüência uma amostra não representativa;
- Métodos paramétricos são melhores em muitos casos para fazer estimativas pontuais. Os procedimentos de reamostragem podem completar as estimativas pontuais de métodos paramétricos fornecendo as estimativas de intervalos de confiança;
- As técnicas de reamostragem não são adequadas para estimar parâmetros que têm um domínio amostral muito estreito, como os valores mínimos e máximos. A reamostragem funciona melhor quando estimativas são suficientes.

4. METODOLOGIA PARA TESTAR SIGNIFICÂNCIA

É importante para algumas pesquisas calcular o intervalo de confiança para valores do coeficiente de correlação e em função dele realizar teste de significância do coeficiente de correlação amostral. O intervalo de confiança mede um intervalo onde deve ficar o parâmetro da população, neste caso o coeficiente de correlação, com certo nível de confiança, por exemplo, 95% ou 99% (GIOVANI COSTA, 2011). Uma vez definido o nível de confiança desejado (1 – á), os elementos do cálculo do intervalo é o coeficiente estimado da amostra *r* e o desvio padrão da estimativa sr e o valor de z que corresponde ao nível de confiança exigido pelo pesquisador (GIOVANI COSTA, 2011).

$$IC(\rho,1-\alpha) = rXY \pm z(1-\alpha)sr \quad (6)$$

O teste de significância consiste em verificar se a hipótese nula de valor não nulo está contida neste intervalo (GIOVANI COSTA, 2011).

O problema é que o desvio padrão de *r* é um termo extremamente complexo para calcular (SAMOHYL, 2003) e uma metodologia mais tratável no momento é a proposta por Fisher(1925). Essencialmente, como já se comentou, o problema é que o coeficiente de correlação não segue a distribuição normal, pois é assimétrica (GIOVANI COSTA, 2011). Fisher(1925) desenvolveu,então, uma expressão que transforma o coeficiente *r* em variável aleatória que segue a normalidade, *z* de Fisher, como colocou-se em parágrafos acima. A proposta deste estudo é colocar a disposição de pesquisadores uma metodologia que não substitua, mas que venha acrescentar conhecimento e atualidade à técnica de testar a significância de *r* quando a hipótese nula for $\rho \neq 0$.

Para realizar o teste de significância mencionado para a hipótese nula de não nulidade, se utilizará da idéia corrente de que o intervalo de confiança pode ser usado imediatamente, sem qualquer outro cálculo para testar qualquer hipótese: o intervalo de confiança pode ser considerado como um conjunto de hipóteses aceitáveis (GIOVANI COSTA, 2011). Qualquer hipótese nula que esteja fora do intervalo de confiança deve ser rejeitada. Por outro lado, qualquer hipótese que esteja dentro do intervalo de confiança deve ser aceita (GIOVANI COSTA, 2011). No entanto, ao invés de se utilizar do intervalo de confiança proposto por Fisher este artigo sugere que se utilize o intervalo de confinação não paramétrico *bootstrap*.

Nesta investigação, a hipótese nula é de que a correlação entre X e Y seja diferente de zero, $\tilde{n} \neq 0$. Portanto, se o valor fixado para $\tilde{n}$ na hipótese nula estiver contido no intervalo de confiança, aceita-se a hipótese nula de que a correlação entre X e Y seja diferente de zero, provavelmente o valor fixado na hipótese nula.

5. ESTUDO DE CASO 1

5.1. Simulação *bootstrap*

Como ilustração do desempenho do *bootstrap*, elaborou-se um exemplo numérico onde se aplica esta à metodologia sugerida.

Seja a população alvo dos municípios da região sudeste do Brasil. Como amostra da população definida, tem-se o subconjunto finito de 91 municípios do estado do Rio de Janeiro. Esta é a informação

amostral da qual se dispõe, amostragem não probabilística inacessibilidade a toda a população (GIOVANI COSTA, 2011). A representatividade da amostra é discutível, mas pode-se notar que existem muitos municípios do Rio com perfil de infra-estrutura e condições sócio-econômicos semelhantes aos de outros municípios da região sudeste.

O objetivo deste estudo de caso é procurar relacionar as variáveis Produto Interno Bruto, o PIB(X), e o Índice de Desenvolvimento Humano, o IDH(Y) ou ainda o grau de relação(linear) existente entre elas. Utilizando os dados de fonte secundária da Organização das Nações Unidas(ONU), Atlas de Desenvolvimento Humano no Brasil(2000), pode-se obter a distribuição de probabilidades empírica do coeficiente de correlação. Esta informação tirada da realidade, amostral, será utilizada para se obter o erro padrão da estimativa, que por sua vez será usado para construção do intervalo de confiança para $\tilde{n}$.

A aplicação do *bootstrap* foi feita de acordo com as etapas descritas na seção 3 para obtenção da amostra *bootstrap*. No *bootstrap*, utilizou-se o procedimento da amostragem com reposição descrita no texto, considerando 2500 simulações.

O intervalo de confiança gerado foi o percentílico com uma confiança de 95%. Sua interpretação é o habitual: o procedimento é tal que em provas repetidas pode-se esperar obter intervalos que incluam o valor fixo de $\tilde{n}$ em 95% das vezes.

A computação das estimativas nas 2500 reamostras foi realizada através do pacote estatístico Stata Versão 8.0.

Tabela 1 – Resultados da simulação da correlação IDH X PIB

Estimativa	Nº de Simulações	r-simulado	Intervalo de Confiança	
			LI	LS
Coeficiente de Correlação Amostral	2500	0,91	0,85	0,93

Fonte: elaboração do autor

5.2. Hipótese nula do problema e teste da significância

O teste de significância será realizado utilizando um nível de significância á = 5%. As hipóteses do problema são:

$$H0 : \tilde{N}\tilde{n} = 0{,}90$$
$$H1: \tilde{N}\tilde{n} \neq 0{,}90$$

Genericamente, pode-se afirmar que quaisquer valores escolhidos para figurar na hipótese nula($H0 : \tilde{N}\tilde{n} = \tilde{n}0$) que estivessem compreendidos pelos limites do intervalo de confiança nos conduziria à aceitação de $H0 : \tilde{N}\tilde{n} = \tilde{n}0$, para o mesmo nível de significância *á*.

No problema, foi estabelecido na hipótese nula $\tilde{n}0 = 0{,}90$, o que se conduz a aceitar $H0 : \tilde{N}\tilde{n} = 0{,}90$, isto é, que a relação linear entre o Produto Interno Bruto e o Índice de Desenvolvimento Humano nos municípios do sudeste é fortemente positiva.

6. ESTUDO DE CASO 2

6.1. Simulação *bootstrap*

O presente estudo de caso utiliza uma amostra de trabalhadores com idade entre 25 e 64 anos, residentes nas áreas urbanas do Brasil, extraída da PNAD, Pesquisa Nacional por Amostra de Domicílios (2002). Os dados correspondem a uma amostra aleatória de 10% das observações da PNAD.

A PNAD é uma pesquisa anual feita pelo IBGE representativa de toda a população brasileira. Em cada ano, são entrevistados em torno de 330.000 indivíduos em cerca de 100.000 domicílios. A amostra utilizada envolveu 10.014 indivíduos da amostra original da PNAD.

O objetivo deste trabalho é estimar a relação entre as variáveis "anos de estudos dos indivíduos"(Y) e o "rendimento mensal" (em R$ de 1999).

Com o *bootstrap*, é possível obter a distribuição por amostragem empírica dos coeficientes de correlação linear da relação acima, seu erro padrão, construir o intervalo de confiança e realizar o teste de significância para H0 : $\tilde{N}\tilde{n} = \tilde{n}0$.

A aplicação do *bootstrap* foi com as especificações análogas ao realizado no estudo de caso 1.

Tabela 2 – Resultados das simulações da correlação anos de estudos *versus* rendimento mensal

Estimativa	Nº de Simulações	r-simulado	Intervalo de Confiança	
			LI	LS
Coeficiente de Correlação Amostral	2500	0,41	0,39	0,44

Fonte: elaboração do autor

6.2. Hipótese nula do problema e teste da significância

O teste de significância será realizado utilizando um nível de significância á= 5%.As hipóteses do problema são:

$$H0 : \tilde{N}\tilde{n} \geq 0{,}60$$
$$H1: \tilde{N}\tilde{n} < 0{,}60$$

No presente problema, foi estabelecido na hipótese nula $\tilde{n} \geq 0{,}60$, o que indica que o teste conduz inevitavelmente à rejeição dessa hipótese, pois tais valores encontram-se fora do intervalo de confiança. A relação entre anos de estudos de indivíduos e o rendimento mensal dos mesmos é menor de 0,60. Existem evidências empíricas de que não existe relação significante entre "anos de estudos dos indivíduos"(Y) e o "rendimento mensal" (em R$ de 1999). Contudo, a PNAD é oriunda de amostragem complexa e não se sabe até que ponto este fator possa influenciar realizações de técnicas que deveriam atender o pressuposto de amostragem aleatória simples.

7. CONCLUSÃO

A distribuição amostral de *r*, em geral, não é simétrica, exceto no caso particular em que *ñ*=0. Esse fato torna impraticável a utilização generalizada de intervalos de confiança, exceto se recorrer-se à variável £, introduzida por Fisher, que tem uma distribuição aproximadamente normal. Conhecida a distribuição £, se estará em condições de transformar a estatística *r* na estatística £, e construir intervalo de confiança na forma habitual, reconvertendo posteriormente os valores de £, em *r*. Esta é a alternativa paramétrica para o teste de significância e intervalo de confiança para *ñ*.

Uma alternativa ao recurso exposto acima é se dispuser ao cálculo do intervalo de confiança *bootstrap*, que geraria o intervalo de confiança de maneira empírica e sem a exigência da normalidade da distribuição amostral de *r* e através dele realizar o teste de significância com *ñ*≠0, isto é, H0: *Ññ* ≠ *ñ*0.

Atualmente a computação "pesada", intensiva, não é mais problema, face ao crescente avanço da informática e a disponibilidade de variados *softwares* estatísticos amigáveis. Com isso, torna-se perfeitamente viável a realização sistemática e habitual de testes para hipótese nula em que o valor do coeficiente de correlação linear populacional testado seja diferente de zero.

8. REFERÊNCIAS BIBLIOGRÁFICAS

COSTA, G. G. O. **Curso de Estatística Básica:**Teoria e Prática. Editora Atlas. São Paulo, 2011.

EFRON, B. Bootstrap Methods: Another Look at the Jackknife, **The Annals of Statistic**, 7, 1-26, 1979.

EFRON, B. **The jackknife, the bootstrap, and other resampling methods**, CBNS 38, SIAM-NSF, 1982.

FISHER, R. A. **Applications of student's distribution**. Metro, 5, 90, 104, 1925.

FONSECA, J. S.; MARTINS, G. A.; TOLEDO, G. L. **Estatística Aplicada**. Editora Atlas. São Paulo, 1995.

GROOT, M. H.; SCHERVISH, M. J. **Probability and Statistics**, 3ed, Addison-weley, New York, 2002.

HAIR, J.F; ANDERSON, R.E; TATHAM, R. L; BLACK, W. C. **Análise Multivariada de Dados**. Bookman, Porto Alegre, 2005.

MURTEIRA, B. J. **Probabilidade e Estatística**. McGraw-Hill. Portugal, 1990.

SAMOHYL, R. W. **Introdução à estatística e métodos de previsão em séries temporais:** teoria aprofundada e prática simplificada, 2003.

STUDENT. **On the probable error of the mean**. Biometrika 6, 1,25, 1908.