UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
FACULDADE DE MEDICINA
PROGRAMA DE PÓS-GRADUAÇÃO EM EPIDEMIOLOGIA

DISSERTAÇÃO DE MESTRADO

ESTIMAÇÃO DO COEFICIENTE DE CORRELAÇÃO DE SPEARMAN PONDERADO

Lidiane Bauer

Orientador: Prof. Dr.Álvaro Vigo

Porto Alegre, abril de 2007.

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
FACULDADE DE MEDICINA
PROGRAMA DE PÓS-GRADUAÇÃO EM EPIDEMIOLOGIA

DISSERTAÇÃO DE MESTRADO

ESTIMAÇÃO DO COEFICIENTE DE CORRELAÇÃO DE SPEARMAN PONDERADO

Lidiane Bauer

Orientador: Prof. Dr.Álvaro Vigo

A apresentação desta dissertação é exigência do Programa de Pós-graduação em Epidemiologia, Universidade Federal do Rio Grande do Sul, para obtenção do título de Mestre.

Porto Alegre, Brasil.
2007

BANCA EXAMINADORA

Profa. Dra. Jandyra Fachel,

Programa de Pós-graduação em Epidemiologia, UFRGS;

Prof. Dr. João Riboldi,

Programa de Pós-graduação em Epidemiologia, UFRGS;

Profa. Dra. Patrícia Klarmann Ziegelmann,

Programa de Pós-graduação em Medicina, UFRGS.

Agradecimentos

Ao Prof. Dr. Álvaro Vigo que mais uma vez enriqueceu minha formação com sua orientação e seus valiosos conselhos que serão fundamentais para toda a minha vida profissional.

Ao suporte financeiro concedido pelo CNPq que foi fundamental ao longo do curso.

À Universidade Federal do Rio Grande do Sul e em especial os professores e funcionários do Programa de Pós Graduação em Epidemiologia.

A todos os meus amigos e colegas que de alguma forma ajudaram na realização deste trabalho e me reconfortaram nas horas difíceis.

E agradeço principalmente às pessoas mais importantes da minha vida, minha família e em especial minha irmã Letícia pela grande ajuda.

SUMÁRIO

Resumo	06
Abstract	07
Lista de Tabelas	08
Lista de Figuras	08
1. APRESENTAÇÃO	09
2. INTRODUÇÃO	10
3. REVISÃO DA LITERATURA	13
4. JUSTIFICATIVA	19
5. OBJETIVOS	19
6. REFERÊNCIAS BIBLIOGRÁFICAS	20
7. ARTIGO	23
8. CONCLUSÕES E CONSIDERAÇÕES FINAIS	51
9. ANEXOS	53
A. Projeto de Pesquisa	54
B. Rotinas Computacionais	63

RESUMO

Para estimar a correlação de duas variáveis que não têm distribuição conjunta normal bivariada, a alternativa mais usual é o coeficiente de correlação de Spearman. Entretanto, quando os dados necessitam de ponderação na análise, como no caso de delineamentos amostrais complexos, não existe método descrito na literatura para estimar essa correlação. Este artigo propõe dois métodos para este cenário e os compara via simulação Monte Carlo. O primeiro método, chamado de método da amostra expandida, consiste em replicar cada observação da amostra em número igual ao seu peso e calcular o coeficiente de Spearman na amostra expandida. No segundo método, o método dos postos, é estimado o coeficiente de correlação de Pearson ponderado nos postos das duas variáveis. Teste de hipóteses tradicional das estimativas produzidas pelos dois métodos também é abordado neste artigo.

Os dois estimadores do coeficiente de Spearman ponderado explorados mostraram desempenhos muito semelhantes, com ausência de viés, pequena variabilidade e mesma eficiência. Entretanto, se recomenda estes métodos quando os dados são medidos em escala.

Este trabalho também explora a estimação pontual do coeficiente de Pearson ponderado e estimação de intervalos de confiança bootstrap, quando a suposição de normalidade bivariada está violada. Sua principal vantagem é evitar potencial influência da expansão da amostra nos postos associados aos valores observados como ocorre com o coeficiente de Spearman.

ABSTRACT

To estimate the correlation of two variables that don't have bivariate normal distribution, the more usual alternative is the Spearman correlation coefficient. However, when the data need of weighting in the analysis like the complex sample surveys, there aren't any methods for estimate this correlation in the literature. This paper proposes two methods for this framework and compares it through the Monte Carlo simulation. The first method which will be called of expanded sample method, consist of replied each observation from sample by its correspondent weight in it. In the second method, called of ranks methods, the ranks of the two variables are calculated, and then are estimated the weighted Pearson correlation coefficient.

This work also explores another solution for making inference to the Pearson coefficient in the presence of weighting and violation of the assumption of normality, the bootstrap confidence interval.

The two estimators proposed showed performance very similar, with or without bias and a little variability. However, a more current proceeding is to estimate the weighted Pearson correlation coefficient and to construct a bootstrap confidence interval, because in this way is unnecessary to know the joint distribution of the two variables. It is important to point out that to Pearson coefficient there is no loss of information in its calculation like in the Spearman coefficient, once in the last one are considerate just the ranks.

Lista de Tabelas

Tabela 1 - Resultados das simulações para os estimadores do coeficiente de Spearman ponderado. 45

Lista de Figuras

Figura 1 **–** Densidades conjuntas suavizadas, marginais e diagrama de dispersão para as populações geradas via simulação. 46
Figura 2 – Histogramas das estimativas pelos dois métodos para dados com distribuição supostamente normal. 47
Figura 3 – Histogramas das estimativas pelos dois métodos para dados com distribuição não normal. 48
Figura 4 – Diagrama de caixas para os erros e erros quadráticos das estimativas.....49
Figura 5 – Histograma das estimativas do coeficiente de correlação de Pearson ponderado entre os valores de interleucina-6 e adiponectina nas 1.000 amostras bootstrap.....50

1. APRESENTAÇÃO

Este trabalho consiste na dissertação de mestrado intitulada “Estimação do coeficiente de correlação de Spearman ponderado”, apresentada ao Programa de Pós-Graduação em Epidemiologia da Universidade Federal do Rio Grande do Sul, em 23 de abril de 2007. O trabalho é apresentado em quatro partes, na ordem que segue:

1. Introdução, Revisão da Literatura e Objetivos
2. Artigo
3. Conclusões e Considerações Finais.
4. Anexos
    A. Projeto de Pesquisa
    B. Rotinas Computacionais

2. INTRODUÇÃO

Em muitos estudos é necessário avaliar a existência de correlação linear entre duas variáveis quantitativas, aqui genericamente representadas por X e Y. Para verificar a existência de tal correlação ou estimar sua magnitude usa-se, freqüentemente, o coeficiente de correlação produto-momento ou apenas coeficiente de correlação de Pearson (r). Este coeficiente foi proposto por Karl Pearson (1857 – 1936) por volta de 1895 e o parâmetro é geralmente denotado por ρ. Nenhuma suposição probabilística é necessária para calcular o coeficiente de Pearson entre duas variáveis quantitativas. Porém, para fazer inferências através do teste de hipóteses tradicional que utiliza distribuição de probabilidade t de Student como distribuição de referência é necessário que a população amostrada tenha distribuição normal bivariada (Zar, 1999). Essa pressuposição muitas vezes não está satisfeita na prática e para contornar este problema uma solução possível é estimar a correlação através do coeficiente de correlação para postos de Spearman (r_S) cujo parâmetro é representado por ρ_S (Zar, 1999; Daniel, 1978).

O coeficiente de correlação de Spearman é a mais antiga estatística baseada em postos e foi introduzida por Spearman em 1904 (Siegel, 1975). Este coeficiente exige que as variáveis supostamente correlacionadas, X e Y, sejam medidas pelo menos em escala ordinal. No caso de variáveis quantitativas com distribuição conjunta diferente da distribuição normal bivariada, esse coeficiente pode ser uma alternativa para substituir o coeficiente de Pearson. Neste caso, a correlação entre X e Y pode ser calculada da mesma forma que o coeficiente de Pearson, porém usando seus postos (Zar, 1999). Para testar a significância estatística desse coeficiente, a distribuição das variáveis não precisa ser

conhecida. Se a amostra for grande, a significância estatística pode ser encontrada através de um teste baseado na distribuição de probabilidade t de Student, como é feito com a estimativa do coeficiente de correlação de Pearson (Daniel, 1978; Siegel, 1975).

Entretanto, o coeficiente de Pearson é uma opção mais atraente para calcular a correlação linear entre duas variáveis quantitativas, pois esse coeficiente usa os valores observados para X e Y enquanto que o coeficiente de Spearman usa seus postos. Ao substituir os valores das variáveis pelos seus postos perde-se a magnitude de cada observação. Quando houverem postos empatados, o procedimento usual é atribuir a média dos postos que lhes caberiam se não tivesse havido empate, e novamente perde-se informação. Para amostras grandes esse problema é ainda pior, pois o número de postos empatados tende a aumentar.

Mesmo quando as variáveis têm distribuição normal bivariada poderá haver outra dificuldade para estimar e fazer inferências sobre a correlação linear, como a presença de pesos amostrais. Isto pode ocorrer, por exemplo, devido a uma amostragem estratificada com tamanho de amostra igual nos estratos. Assim, como o tamanho de cada estrato não é levado em conta, pode ser necessário fazer ponderação na análise. Para o coeficiente de Spearman não existe um método que incorpore a ponderação dos dados. Já para o coeficiente de Pearson existe, mas ainda temos que considerar o efeito dos pesos sobre a distribuição de referência da estatística de teste e na estimação do intervalo de confiança.

Recentemente, com o acelerado desenvolvimento de recursos computacionais, é possível fazer inferências para uma variedade de parâmetros populacionais através de métodos de reamostragem, tais como simulação Monte Carlo, Jackniffe ou Bootstrap. Quando a suposição de normalidade bivariada dos dados não estiver satisfeita estes

métodos podem ser uma alternativa importante para fazer inferênias sobre o coeficiente de correlação (Efron, 1979).

Em particular, o método Bootstrap é uma técnica de reamostragem originalmente desenvolvida para fornecer medidas de vício e variabilidade de estimativas. Posteriormente ela foi estendida para a construção de intervalos de confiança, testes de hipóteses e outras situações inferenciais (Efron & Tibishirani, 1993). A vantagem deste método é que não é necessário conhecer a priori a distribuição de probabilidade do estimador..

Esse trabalho investiga dois procedimentos para estimar e fazer inferências sobre o coeficiente de correlação de Spearman ponderado, com o objetivo de contornar o problema da violação de normalidade bivariada do par (X,Y). Também explora a estimação pontual e por intervalo de confiança bootstrap para o coeficiente de correlação de Pearson ponderado para o mesmo problema.

3. REVISÃO DA LITERATURA

Em diversos estudos epidemiológicos a população que está sendo investigada é constituída por subgrupos ou estratos, podendo ser importante preservar a representatividade dos mesmos no processo de amostragem. Um procedimento usual é utilizar um plano amostral com amostragem estratificada.

Uma amostragem estratificada é um plano amostral em que a população em estudo é dividida naturalmente em k estratos mutuamente excludentes e exaustivos e uma amostra aleatória simples de n_i elementos é retirada de cada estrato i, onde $i = 1,...,k$. O número de elementos amostrados em cada estrato pode ser proporcional ao tamanho do estrato. Assim, a probabilidade de cada elemento da amostra ser selecionado dependerá do estrato ao qual ele pertence e pode ser calculado por n_i / N_i que é a fração amostral no estrato i (Levy & Lemeshow, 1980).

Um método usual de alocação é selecionar o mesmo número de unidades em cada estrato, e então o tamanho de amostra em cada estrato será n/k, onde n é o número total de elementos na amostra e k o número de estratos. Neste caso, para que os resultados das análises possam ser corretamente generalizados para a população, é fundamental corrigir a importância relativa de cada elemento da amostra, mediante a utilização de pesos, de acordo com o estrato ao qual pertence.

Os pesos podem ser definidos como o inverso das frações amostrais de cada estrato N_i / n_i (Deming & Edwards, 1950; Silva & Pessoa, 2002). Dessa forma ocorre uma calibração, possibilitando a obtenção de estimativas não viesadas dos parâmetros. Um

estimador $\hat{\theta}$ é dito não viesado ou não viciado para um parâmetro θ se $E(\hat{\theta}) = \theta$. Ou seja, um estimador é não viciado se o seu valor esperado coincide com o parâmetro de interesse, independente do tamanho de amostra (Cox & Hinkley, 1974).

No Brasil, há estudos com planos complexos que necessitam incorporar pesos amostrais nas estimativas para inferir os resultados para a população. É o caso de estudos que utilizam dados da PNAD (Pesquisa Nacional de Amostra por Domicílio), que são coletados através de um plano de amostragem que inclui todos os aspectos que definem um plano amostral complexo: estratificação das unidades de amostragem, conglomerados, probabilidades desiguais de seleção em um ou mais estágios, e ajustes dos pesos amostrais para calibração com totais populacionais conhecidos. Exemplos são os trabalhos que utilizaram dados da PNAD e pesos amostrais nas análises (Lima-Costa et al., 2002; Almeida et al., 2002; Bahia et al., 2002; Matos et al., 2004).

Os estudos NHANES (*National Health and Nutrition Examination Survey*) são exemplos de estudos epidemiológicos nos quais é fundamental utilizar pesos nas análises para produzir estimativas nacionais sobre estado de saúde e nutrição da população norte-americana (Centers for Disease Control and Prevention (CDC)). O plano amostral do NHANES também apresenta todos os aspectos de um plano amostral complexo.

Em muitos trabalhos que utilizam dados de amostragem complexa é necessário estimar a correlação linear entre duas variáveis quantitativas, e usualmente estimam através da correlação de Pearson ponderado. Por exemplo, dados do NHANES III foram usados para investigar a associação entre proteína C-reativa com níveis de insulina, glicemia e hemoglobina glicada no plasma. Os resultados das correlações de Pearson com ponderação mostraram associações significativas entre proteína C-reativa e insulina, glicemia e

hemoglobina glicada, bem como com outras variáveis tais como idade, educação, consumo de cigarro, índice de massa corporal e atividade física nos períodos de lazer (Wu et al., 2002). Entretanto, diante da possibilidade de haver violações nas suposições do teste de hipóteses tradicional para avaliar a significância do coeficiente de correlação de Pearson, que poderiam alterar as estimativas das probabilidades de significância associadas (valores P), estimaram também o coeficiente de Spearman não ponderado. A decisão foi justificada pelo fato de que o procedimento disponível no programa SAS para incorporar pesos nesse coeficiente não é válido. Estimativas produzidas pelo coeficiente de Spearman sem ponderação e o coeficiente de Pearson ponderado foram muito similares. Esta questão metodológica também surgiu em análises recentes sobre a associação entre marcadores de inflamação sistêmica e adiponectina e o desenvolvimento de diabetes melito tipo 2 estudada no projeto ARIC (Duncan et al, 2003; Duncan et al., 2004).

Outros trabalhos baseados em dados do NHANES que utilizaram o coeficiente de Spearman são (Dixon et al, 2001; Himes et al, 2004; Eisner, 2002) sem, no entanto, explicitar se os pesos foram ou não considerados na análise.

O coeficiente de correlação de Pearson amostral, sem ponderação, é estimado por

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}, \qquad (1)$$

e, incorporando ponderação, é dado por

$$r_w = \frac{\sum_{i=1}^{n} w_i (x_i - \bar{x}_w)(y_i - \bar{y}_w)}{\sqrt{\sum_{i=1}^{n} w_i (x_i - \bar{x}_w)^2 \sum_{i=1}^{n} w_i (y_i - \bar{y}_w)^2}}, \qquad (2)$$

onde $\bar{x}_w = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$, $\bar{y}_w = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i}$ e w_i é um peso da observação i. O teste para as hipóteses $H_0 : \rho = 0$ contra $H_1 : \rho \neq 0$, utilizando a distribuição t de Student como referência, pode ser feito através da estatística

$$t_{n-2} = \frac{r}{\sqrt{(1-r^2)/n-2}} \tag{3}$$

onde r é o valor estimado do coeficiente de correlação de Pearson na presença ou não de pesos amostrais e o valor t_{n-2} é comparado com a distribuição t de Student com $n-2$ graus de liberdade. É importante salientar, entretanto, que o teste é valido se X e Y têm distribuição conjunta normal bivariada.

O coeficiente de correlação por postos de Spearman (r_S) é uma alternativa usual para estimar correlações lineares nas situações em que há violação da suposição de normalidade conjunta para (X,Y).

A correlação de Spearman é uma estatística baseada em postos e foi introduzida por Spearman em 1904 e exige apenas que as variáveis X e Y sejam medidas pelo menos em escala ordinal. No caso de variáveis quantitativas, a correlação entre X e Y pode ser calculada da mesma forma que o coeficiente de Pearson, mas substituindo os valores das duas variáveis pelos seus postos.

Para testar a significância estatística desse coeficiente, a distribuição conjunta das variáveis não precisa ser conhecida. Se a amostra for grande, a significância estatística pode ser avaliada através do teste que utiliza a distribuição de probabilidade t de Student

como distribuição de referência, como é feito com a estimativa do coeficiente de correlação de Pearson (Zar, 1999; Daniel, 1978; Siegel, 1975).

Na literatura não foram encontrados estimadores para o coeficiente de Spearman quando é necessário considerar pesos amostrais, como é visto para o coeficiente de correlação de Pearson (SAS Institute Inc., 1999). Uma alternativa seria determinar os postos associados aos valores observados para as variáveis X e Y, estimar a correlação o coeficiente de correlação de Spearman através da estatística definida na equação (2) e, então, testar a significância utilizando a estatística de teste definida na equação (3). Note que, neste caso, as observações originais foram substituídas pelos correspondentes postos. Outra possível alternativa é expandir a amostra replicando cada observação um número de vezes igual ao peso amostral correspondente, estimar o coeficiente de Spearman da forma usual. A significância do coeficiente estimado pode ser testada mediante a estatística definida na equação (3), corrigindo os graus de liberdade para o tamanho real da amostra, pois, em caso contrário, o nível descritivo amostral (valor P) pode não ser correto devido à expansão artificial da amostra.

Métodos de simulação Monte Carlo podem ser usados para comparar e avaliar a eficiência desses dois procedimentos de estimação e teste para o coeficiente de Spearman ponderado (Gentle, 2002; Manly, 2004). Monte Carlo é um estudo de simulação, onde a partir de um conjunto de dados gerados com uma distribuição de probabilidade especificada, amostras de mesmo tamanho são sorteadas (em geral 10.000) com o objetivo de comparar diferentes procedimentos de estimação e também avaliar suas propriedades (Gentle, 2002).

Outra alternativa para contornar o problema da violação de normalidade bivariada é a utilização do método bootstrap proposto por Efron (Efron, 1979). Este método

é uma alternativa para avaliar a significância estatística do coeficiente de Pearson ponderado ou estimar um intervalo de confiança, e tem a vantagem de não exigir uma distribuição subjacente para o estimador. Foi originalmente desenvolvido para fornecer medidas de viés e de variabilidade de estimadores e posteriormente estendido para a construção de intervalos de confiança, testes de hipóteses e outras situações inferenciais (Efron & Tibishirani, 1993; Gentle, 2002; Manly, 2004).

O método bootstrap nos últimos anos tem sido muito utilizado devido aos avanços computacionais e à incorporação do método em alguns softwares estatísticos usuais. Essa técnica envolve novas amostras (amostras bootstrap) tomadas de uma única amostra observada (amostra original) que será tratada como se esta representasse exatamente toda a população. Desta amostra original de n elementos, B amostras com reposição, também de tamanho são selecionadas aleatoriamente. Em cada uma dessas amostras é recalculado o coeficiente de correlação de Pearson ponderado, gerando sua distribuição amostral.

Para a construção de intervalos de confiança usando bootstrap existem diferentes métodos, como por exemplo, o bootstrap padrão, o bootstrap-t e o método do percentil, podendo ainda considerar correção de viés (Manly, 2004).

4. JUSTIFICATIVA

Para estimar a correlação linear de duas variáveis quantitativas com a presença de pesos amostrais decorrentes, por exemplo, de um plano amostral complexo, é usual calcular o coeficiente de correlação de Pearson ponderado. Porém, para fazer inferências pelo método tradicional é necessário que a distribuição conjunta das variáveis seja normal bivariada. O coeficiente de Spearman não necessita de suposições quanto à distribuição conjunta das variáveis, mas não foi encontrado na literatura um estimador para o coeficiente de Spearman que incorpore os pesos amostrais. Para dados quantitativos, uma alternativa é utilizar o coeficiente de correlação linear de Pearson ponderado para estimar a correlação, construindo intervalos de confiança através de métodos de reamostragem, como o método de bootstrap, que têm a vantagem de ser mais flexível em relação à distribuição de probabilidade do estimador.

5. OBJETIVO

Comparar e eficiência de dois estimadores do coeficiente de Spearman ponderado para estimar a correlação linear entre duas variáveis X e Y através de estudo de simulação Monte Carlo. No mesmo contexto, para variáveis quantitativas, explorar a estimação pontual e por intervalo de confiança para o coeficiente de correlação linear de Pearson ponderado utilizando o método bootstrap.

6. REFERÊNCIAS BIBLIOGRÁFICAS

Almeida MF, Barata RB, Montero CV, Silva ZP. Prevalência de doenças crônicas auto-referidas e utilização de serviços de saúde, PNAD/1998, Brasil. *Ciência e Saúde Coletiva* 2002; 7(4):743-756.

Bahia L, Costa AJL, Fernandes C, Luiz RR, Cavalcanti MLT. Segmentação da demanda dos planos e seguros privados de saúde: uma análise das informações da PNAD/98. *Ciência e Saúde Coletiva* 2002; 7(4):671-686.

Centers for Disease Control and Prevention (CDC). National Center for Health Statistics (NCHS). National Health and Nutrition Examination Survey Data. Hyattsville, MD: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention.

Cox DR, Hinkley DV. *Theoretical Statistics*. London: Chapman and Hall; 1974.

Daniel WW. *Applied nonparametric statistics*. Boston: Houghton-Mifflin; 1978.

Deming WE. *Some Theory of Sampling*. New York: John Wiley & Sons; 1950.

Dixon LB, Winkleby MA, Radimer KL. Dietary Intakes and Serum Nutrients Differ between Adults from Food-Insufficient and Food-Sufficient Families: Third National Health and Nutrition Examination Survey, 1988–1994. *American Society for Nutritional Sciences* 2001; 131:1232-1246.

Duncan BB, Schmidt MI, Pankow JS, Ballantyne CM, Couper D, Vigo A, Hoogeveen R, Heiss G. Low-Grade Systemic Inflammation and the Development of Type 2 Diabetes. The Atherosclerosis Risk in Communities Study. *Diabetes* 2003; vol. 52.

Duncan BB, Schmidt MI, Pankow JS, Bang H, Couper D, Ballantyne CM, Hoogeveen RC, Heiss G. Adiponectin and the Development of Type 2 Diabetes. The Atherosclerosis Risk in Communities Study. *Diabetes* 2004; vol. 53.

Efron B. Bootstrap methods: another look at the jackknife. *Annals of Statistics* 1979; 7: 1-26.

Efron B, Tibshirani RJ. *An Introduction to the Bootstrap*. Chapman and Hall, 1993.

Eisner MD, Environmental Tobacco Smoke Exposure and Pulmonary Function among Adults in NHANES III: Impact on the General Population and Adults with Current Asthma. *Environmental Health Perspectives* 2002; vol. 110 (8).

Gentle JE. *Random Number Generation and Monte Carlo Methods*. George Mason University, 2ª edition; 2002.

Himes JH, Obarzanek E, Baranowski T, Wilson DM, Rochon J, McClanahan BS. Early Sexual Maturation, Body Composition, and Obesity in African-American Girls. *Obesity Research* 2004; vol. 12.

Levy PS, Lemeshow S. *Sampling for health professionals*. Belmont: Lifetime Learning Publications; 1980.

Lima-Costa MF, Barreto SM, Giatti L. A situação socioeconômica afeta igualmente a saúde de idosos e adultos mais jovens no Brasil? Um estudo utilizando dados da Pesquisa Nacional por Amostras de Domicílios – PNAD/98. *Ciência e Saúde Coletiva* 2002; 7(4):813-824.

Manly BFJ. *Randomization, Bootstrap and Monte Carlo Methods in Biology*. New Zeeland: University of Otago, 2ª edition; 2004.

Matos DL, Giatti L, Lima-Costa MF. Fatores sócio-demográficos associados ao uso de serviços odontológicos entre idosos brasileiros:um estudo baseado na Pesquisa

Nacional por Amostras de Domicílios. *Ciência e Saúde Coletiva, Rio de Janeiro* 2004; 20(5):1290-1297.

SAS INSTITUTE INC. (1999). **SAS OnlineDoc®. Version 8**, Cary, NC: SAS Institute Inc.

Siegel S. *Estatística não-paramétrica para as ciências do comportamento*. São Paulo: McGraw-Hill do Brasil; 1975.

Silva PLN, Pessoa DGC, Lila MF. Análise estatística de dados da PNAD: incorporando a estrutura do plano amostral. *Ciência & Saúde Coletiva* 2002; vol.7 (4).

Wu T, Dorn JP, Donahue RP, Sempos CT, Trevisan M. Associations of Serum C-reactive Protein with Fasting Insulin, Glucose, and Glycosylated Hemoglobin**.** *American Journal of Epidemiology* 2002; 155(1).

Zar J. *Biostatistical Analysis*. Upper Saddle River – NJ: Prentice-Hall, 4ª edition; 1999.

7. ARTIGO

Estimação do Coeficiente de Correlação de Spearman Ponderado

Estimation of the Weighted Spearman Correlation

Lidiane Bauer[1,2]
Álvaro Vigo[1,2]

1. Programa de Pós-Graduação em Epidemiologia, Faculdade de Medicina, Universidade Federal do Rio Grande do Sul, Brasil

2. Departamento de Estatística, Instituto de Matemática, Universidade Federal do Rio Grande do Sul, Brasil

Correspondência:

Lidiane Bauer

Departamento de Estatística

Instituto de Matemática, UFRGS

Av. Bento Gonçalves, 9500 – Prédio 43111 – Agronomia

91509-900 – Porto alegre – RS - Brasil

Telefone: +55 51 3308 6225 FAX: +55 51 3308 7301

Email: lidiane.bauer@ufrgs.br

***A ser enviado ao Cadernos de Saúde Pública**

Resumo

Em muitos estudos na área epidemiológica é comum a necessidade de estimar a correlação linear entre duas variáveis quantitativas. Quando os dados necessitam de ponderação na análise, como no caso de delineamentos amostrais complexos, a correlação é estimada usualmente através do coeficiente de Pearson ponderado. Ainda que não haja problemas em estimar a correlação, para testar a significância da correlação pelo método tradicional, é necessário que a distribuição conjunta das variáveis seja normal. O coeficiente de Spearman é a alternativa mais usada quando esta exigência não está garantida. Porém, quando os dados necessitam de ponderação na análise, não existe método descrito na literatura para estimar este coeficiente.

Dois estimadores para este cenário foram comparados via simulação Monte Carlo. O primeiro método consiste em replicar cada elemento da amostra em número igual ao peso correspondente, e então calcular a correlação de Spearman na amostra expandida. No segundo método estima-se o coeficiente de correlação de Pearson ponderado utilizando os postos das duas variáveis, ao invés das observações originais. Um procedimento mais atual e informativo é estimar a correlação de Pearson ponderado e construir intervalo de confiança bootstrap, pois assim é dispensável conhecer a distribuição conjunta das duas variáveis. Outra vantagem é que preserva a informação sobre a relação funcional entre as variáveis usando suas magnitudes e não os postos.

Os dois estimadores para o coeficiente de Spearman ponderado mostraram desempenhos muito semelhantes, com ausência de viés e mesma eficiência. Porém, Para dados quantitativos recomenda-se o coeficiente de Pearson ponderado, estimando o intervalo de confiança via bootstrap, pois preserva a informação original da magnitude das

observações e não sofre a influência dos postos, principalmente no caso da amostra expandida.

PALAVRAS CHAVES: Bootstrap; Correlação; Planos complexos, Pearson; Ponderação; Monte Carlo; Spearman.

Abstract

In many studies in the Epidemiological area there is the interest in knowing the linear correlation between two quantitative variables and the more common is to use the Pearson coefficient. To test this correlation through the traditional method, the assumption of joint normality between these two variables must be satisfied, though there aren't problems in estimating the correlation. The alternative more usual is to estimate by Spearman correlation coefficient. However, when the date need of weighting in the analysis, like the complex sample surveys, there aren't any methods for estimate this coefficient in the literature except for the Pearson coefficient.

This paper proposes two methods for this framework and compares it thought the Monte Carlo simulation. The first method consists of multiplying each element by its correspondent weight in the sample, and then calculating the Spearman's correlation. In the second method the ranks of the two variables are calculated, and then are estimated the weighting Spearman correlation coefficient. As another solution to the problem of violation of the assumption of normality and with weighting, this work explored the punctual estimation and by confidence interval for the weighted Pearson correlation coefficient through the bootstrap method.

The two estimators proposed showed performance very similar, with or without vice and a little variability. However, a more current proceeding is to estimate the weighted Pearson correlation coefficient and to construct a bootstrap confidence interval, because in this way is unnecessary to know the joint distribution of the two variables. It is important to point out that to Pearson coefficient there is no loss of information in its calculation like in the Spearman coefficient, once in the last one are considerate just the ranks.

KEYWORDS: Bootstrap; correlation; weighted; Monte Carlo; Spearman.

1. Introdução

Em muitos estudos é importante avaliar a existência de correlação linear entre duas variáveis quantitativas X e Y, sendo mais comum utilizar o coeficiente de correlação linear de Pearson (r). A significância estatística da estimativa desse coeficiente é usualmente testada utilizando a distribuição t de Student como distribuição de referência, assumindo que as variáveis foram extraídas de uma distribuição normal bivariada. Essa pressuposição muitas vezes não está satisfeita na prática e, para contornar este problema, uma possível alternativa é utilizar o coeficiente de correlação para postos de Spearman (r_S). O coeficiente de correlação de Spearman é a mais antiga estatística baseada em postos e exige que as variáveis sejam medidas pelo menos em escala ordinal e não existem suposições em relação à distribuição conjunta das variáveis [1-3].

Em diversos estudos epidemiológicos a população que está sendo investigada é composta por subgrupos ou estratos, podendo ser importante preservar a representatividade dos mesmos no processo de amostragem. Um procedimento usual é utilizar um plano amostral com amostragem estratificada proporcional ao tamanho dos grupos. Em outra situação, para garantir o poder das comparações ou a estabilidade nos coeficientes de regressão estimados, é comum super-representar grupos minoritários. Assim, muitas vezes é necessário adicionar pesos amostrais nas análises para obter uma estimativa não viesada da correlação populacional. No entanto, para o coeficiente de Spearman não existe um estimador que incorpore a ponderação dos dados. Um estimador para o coeficiente de Pearson ponderado está disponível, mas a validade das inferências depende criticamente da exigência de distribuição normal bivariada estar satisfeita.

A ponderação é um procedimento usual e, muitas vezes, fundamental para que os resultados das análises possam ser corretamente generalizados para a população. Consiste essencialmente em corrigir a importância relativa de cada elemento da amostra, mediante a utilização de pesos, de acordo com o estrato ao qual pertence. Usualmente, os pesos são definidos pelo inverso das frações amostrais de cada estrato[4-5].

No Brasil, há vários exemplos de estudos com planos complexos que necessitam incorporar pesos amostrais para inferir os resultados para a população. É o caso dos dados da PNAD (Pesquisa Nacional de Amostra por Domicílio), que são coletados através de um plano de amostragem complexo que inclui, por exemplo, conglomerados e estratos[6-9].

Os estudos NHANES (*National Health and Nutrition Examination Survey*)[10] são também exemplos de estudos epidemiológicos nos quais é fundamental utilizar pesos nas análises para produzir estimativas nacionais sobre estado de saúde e nutrição da população norte-americana. Por exemplo, dados do NHANES III foram usados para investigar a associação entre proteína C-reativa com níveis de insulina, glicemia e hemoglobina glicada no plasma. Os resultados das correlações de Pearson ponderadas mostraram associações significativas entre proteína C-reativa e insulina, glicemia e hemoglobina glicada, bem como com outras variáveis tais como idade, educação, consumo de cigarro, índice de massa corporal e atividade física nos períodos de lazer. Entretanto, diante da possibilidade de haver violações nas suposições do teste de hipóteses tradicional para avaliar a significância do coeficiente de correlação de Pearson, que poderiam alterar as estimativas das probabilidades de significância associadas (valores P), estimaram também o coeficiente de Spearman não ponderado. A decisão foi justificada pelo fato de que o procedimento disponível no programa SAS para incorporar pesos nesse coeficiente não é válido. Os autores relataram que as estimativas produzidas pelo coeficiente de Spearman sem

ponderação foram muito similares as do coeficiente de Pearson ponderado[11]. Outros trabalhos baseados em dados do NHANES utilizaram o coeficiente de Spearman[12-14] sem, no entanto, explicitar se foi ou não considerado pesos nas análises. Esta questão metodológica também surgiu em análises recentes sobre a associação entre marcadores de inflamação sistêmica e adiponectina com o desenvolvimento de diabete melito tipo 2[15-16].

Apesar da importância, uma detalhada revisão da literatura não mostrou trabalhos metodológicos que apresentem alternativas para a estimação e teste de hipóteses para o coeficiente de Spearman ponderado, como é visto para o coeficiente de correlação de Pearson[17].

O objetivo deste trabalho foi comparar, via simulação Monte Carlo, dois estimadores para o coeficiente de correlação de Spearman ponderado quanto ao viés e precisão. Também são apresentados intervalos de confiança bootstrap para o coeficiente de Pearson ponderado para dados reais que não seguem distribuição normal bivariada.

2. Métodos

Para analisar o comportamento dos estimadores do coeficiente de Spearman ponderado, foram simuladas populações representando o vetor aleatóio (X,Y), com dois estratos de tamanhos diferentes. Na primeira população, dentro de cada estrato (X,Y) tem distribuição conjunta normal bivariada, enquanto que na segunda população a distribuição conjunta não é normal.

A macro *mvn*[18] obtida na página de suporte técnico do SAS foi usada para gerar pares de valores (X,Y) com distribuição normal bivariada dentro dos estratos. Os parâmetros μ_x, μ_y, σ_x, σ_y e σ_{xy}^2, onde μ_x e σ_x representam respectivamente a média e

desvio padrão populacionais do componente X; μ_y e σ_y representam respectivamente a média e o desvio padrão populacionais de Y e σ_{xy}^2 é a covariância populacional entre as variáveis. Assim, por definição, a correlação linear entre X e Y na população é $\rho = \frac{\sigma_{xy}^2}{\sigma_x \times \sigma_y}$. A primeira população considerada é composta por 20.000 elementos com dois estratos ($k = 2$). Com a rotina *mvn* foram gerados 8.000 elementos para o primeiro estrato, especificando os valores dos parâmetros $\mu_{x1} = 10$, $\mu_{y1} = 15$, $\sigma_{x1} = 4$, $\sigma_{y1} = 6$ e $\sigma_{x1y1}^2 = 12$. Para o segundo estrato, foram gerados 12.000 elementos, fixando os valores dos parâmetros $\mu_{x2} = 10$, $\mu_{y2} = 15$, $\sigma_{x2} = 5$, $\sigma_{y2} = 4$ e $\sigma_{x2y2}^2 = 16$. O coeficiente de correlação de Spearman entre X e Y nesta população é igual a 0,644, sendo igual a 0,481 e 0,781 no primeiro e segundo estratos, respectivamente. Note que a população foi gerada a partir de uma mistura de duas distribuições normais bivariadas.

A população com distribuição conjunta não normal foi criada através dos geradores de números aleatórios disponível no programa SAS[17]. As variáveis foram simuladas em duas etapas, para garantir dois estratos distintos na população. No primeiro estrato, com 8.000 elementos, X tem distribuição normal com parâmetros $\mu_x = 15$ e $\sigma_x = \sqrt{5}$ e Y tem distribuição gama com parâmetros $\alpha = 10$ e $\beta = \frac{X}{5}$. No segundo estrato foram gerados 12.000 elementos, onde X tem distribuição normal com parâmetros $\mu_x = 20$ e $\sigma_x = \sqrt{7}$, e Y tem distribuição gama com parâmetros $\alpha = 25$ e $\beta = \frac{X}{10}$. A correlação de Spearman nesta população é 0,746, sendo igual a 0,419 e 0,541 no primeiro e segundo estratos, respectivamente.

Para cada uma destas duas populações, foram selecionadas 10.000 amostras de tamanho 500, com e sem reposição, sendo 250 elementos de estrato. As amostras foram selecionadas usando o procedimento *PROC SURVEYSELECT* disponível no programa SAS[17]. Em cada amostra foi estimado o coeficiente de correlação de Spearman pelos métodos apresentados nas seções 2.1 e 2.2.

Para permitir a comparação dos estimadores, tanto as populações quanto as amostras geradas foram armazenadas em disco, sendo usadas nos dois métodos de estimação. Também foram armazenadas as estimativas obtidas para cada uma das 10.000 amostras, bem como o erro, definido pela diferença entre o valor estimado e o parâmetro populacional, o erro quadrático, o erro absoluto, o valor da estatística de teste e o valor P correspondentes aos testes de significância da correlação. A eficiência dos estimadores foi avaliada através da razão entre os respectivos erros quadráticos médios (EQM).

Para fazer a ponderação no estimador do coeficiente de Spearman, foi calculado o peso de cada elemento da amostra, considerando o tamanho do estrato ao qual pertence e o número de observações selecionadas daquele estrato. Assim, os pesos amostrais para cada estrato foram definidos pelo inverso da fração amostral

$$w_k = \frac{N_k}{n_k}, \tag{2}$$

onde w_k é o peso amostral dos elementos do estrato k, N_k é o número total de elementos no estrato k na população e n_k representa o número de elementos selecionados do estrato k para compor a amostra. Essa definição de peso amostral será mantida em todas as situações exploradas no trabalho.

Os métodos de estimação do coeficiente de Spearman ponderado e do intervalo de confiança bootstrap para o coeficiente de Pearson ponderado são apresentados em detalhes a seguir.

2.1. Método da Amostra Expandida

Este método consiste em replicar cada elemento da amostra w_k vezes, onde w_k é o peso amostral das observações do *k*-ésimo estrato definido na equação (2). Nas amostras simuladas existem dois estratos cujos pesos amostrais são, repectivamente, $w_1 = 8000/250 = 32$ e $w_2 = 12000/250 = 48$. Assim, cada um dos 250 elementos da amostra selecionados do primeiro estrato foi replicado 32 vezes e, similarmente, cada um dos 250 elementos do segundo estrato foi replicado 48 vezes. Esta amostra expandida, com 20.000 elementos, incorpora artificialmente os pesos de cada observação e, assim, pode-se usar os métodos computacionais usuais para estimar a correlação de Spearman. Entretanto, apesar de preservar a importância relativa de cada elemento da amostra na estimativa pontual, os resultados dos testes de significância sobre o coeficiente de correlação de Spearman podem não ser válidos, haja vista que o tamanho da amostra foi artificialmente inflacionado, ocorrendo o mesmo com os graus de liberdade associados à distribuição de referência da estatística de teste.

A estatística de teste é calculada como

$$t_{calc} = \frac{r_{SW}}{\sqrt{(1-(r_{SW})^2)/n-2}} \qquad (3)$$

onde r_{SW} é o valor estimado do coeficiente de Spearman e n o número de pares (X,Y) usados para o cálculo do coeficiente. Como o tamanho da amostra foi artificialmente

expandido, conseqüentemente, os graus de liberdade da distribuição t de Student foram inflacionados. Uma correção no denominador da expressão (3) foi realizada, alterando o tamanho da amostra expandida de 20.000 para o tamanho de amostra original $n = 500$. Para tamanhos de amostra suficientemente grande, o valor observado da estatística definida em (3) pode ser comparada com o valor da distribuição de referência *t* de Student com $n - 2$ graus de liberdade.

Um aspecto importante neste contexto de expansão da amostra é a presença de pesos não inteiros. Uma primeira abordagem seria arredondar para o inteiro mais próximo e fazer a expansão da amostra, mas este procedimento eventualmente poderia produzir viés. Para aumentar a precisão, pode-se primeiro multiplicar por 10 (ou 100, 1000, etc, de acordo com a precisão necessária) e depois arredondar para o inteiro mais próximo. Contudo, para amostras grandes este procedimento poderia exigir grande capacidade de processamento ou de armazenamento de dados, pois a amostra expandida pode ser extremamente grande. Além disso, o número de empates nos postos aumenta drasticamente, podendo influenciar artificialmente a estimativa da correlação de Spearman.

Para avaliar a importância de pesos não inteiros foram geradas as mesmas populações descritas anteriormente, mudando apenas os tamanhos dos estratos, agora com 7.900 e 12.100 elementos. Assim, selecionando amostras de 500 observações (250 em cada estrato) os pesos amostrais foram $w_1 = 7900/250 = 31{,}6$ e $w_2 = 12100/250 = 48{,}4$ respectivamente para o primeiro e o segundo estrato, tendo sido arredondados para 32 e 48.

2.2. Método dos Postos

O método dos postos consiste em utilizar um procedimento similar àquele usado para estimar o coeficiente de correlação de Pearson com ponderação. No contexto simulado o coeficiente de correlação de Pearson ponderado pode ser estimado por

$$r_w = \frac{\sum_{k=1}^{2}\sum_{i=1}^{250} w_k (x_{ik} - \bar{x}_w)(y_{ik} - \bar{y}_w)}{\left[\sum_{k=1}^{2}\sum_{i=1}^{250} w_k (x_{ik} - \bar{x}_w)^2\right]^{1/2}\left[\sum_{k=1}^{2}\sum_{i=1}^{250} w_k (y_{ik} - \bar{y}_w)^2\right]^{1/2}} \qquad (4)$$

onde $\bar{x}_w = \frac{\sum_{k=1}^{2}\sum_{i=1}^{250} w_k x_{ik}}{\sum_{k=1}^{2} n_k w_k}$ e $\bar{y}_w = \frac{\sum_{k=1}^{2}\sum_{i=1}^{250} w_k y_{ik}}{\sum_{k=1}^{2} n_k w_k}$ representam as médias ponderadas das variáveis X e Y, respectivamente, e w_k é o peso amostral das observações do *k*-ésimo estrato.

Para estimar o coeficiente de Spearman ponderado, a idéia é substituir os valores observados (x_i, y_i) na equação (4) pelos correspondentes postos dos valores ordenados, utilizando-se a média dos postos em caso de empates. Para um tamanho de amostra suficientemente grande, o teste de significância do coeficiente pode utilizar como distribuição de referência a distribuição t de Student, como é feito usualmente para o coeficiente de Pearson. Este método de estimação e teste de significância para a correlação de Spearman pode ser executado utilizando procedimentos disponíveis em softwares estatísticos como, por exemplo, o procedimento *PROC CORR* do *SAS* [17].

2.3. Coeficiente de Pearson ponderado via bootstrap

Quando há violação da suposição de normalidade pode não ser válido o resultado do teste de significância do coeficiente de correlação de Pearson (r) que usa a distribuição t de Student como referência. Uma importante alternativa é usar amostras bootstrap para encontrar a distribuição desse estimador, mediante a qual é possível estimar um intervalo de confiança ou testar hipóteses.

O bootstrap é um método de reamostragem, originalmente desenvolvido para fornecer medidas de vício e variabilidade de estimativas[19]. Posteriormente ele foi estendido para a construção de intervalos de confiança, testes de hipóteses e outras situações inferenciais[20]. Sua vantagem está no fato de não ser necessário o conhecimento da distribuição de probabilidade do estimador.

O método bootstrap envolve amostras tomadas de uma única amostra observada que será tratada como se esta representasse exatamente toda a população. Desta amostra original de n elementos são selecionadas aleatoriamente B amostras com reposição, também de tamanho n.

Para ilustrar o método, dados parciais do estudo ARIC (The Atherosclerosis Risk in Communities Study) serão usados para estimar um intervalo de confiança para o coeficiente de correlação linear de Pearson através do método de bootstrap. Este estudo consiste de uma coorte conduzida em quatro comunidades norte-americanas para investigar a etiologia e a história natural da arteriosclerose e fatores de risco de doenças cardiovasculares e diabete. Em cada centro foram selecionados e recrutados aproximadamente 4.000 homens e mulheres com idade entre 45-64 anos da correspondente população. Os 15.792 participantes foram extensivamente examinados na linha de base (1987-89), e convidados para retornar em três visitas clínicas em intervalos de

aproximadamente três anos. A segunda visita ocorreu entre 1990-92, a terceira entre 1993-95 e a quarta entre 1996-98. Os participantes são acompanhados anualmente por telefone para manter contato e para avaliar a condição de saúde da coorte. Os delineamentos do tipo caso-coorte realizados no ARIC permitem usar eficientemente o material biológico congelado para investigar novos fatores de risco, como, por exemplo, os estudos da associação entre inflamação sistêmica e adiponectina com o desenvolvimento de diabete melito tipo 2, publicados recentemente. Um aspecto importante destas análises foi considerar efeito da estratificação e da ponderação, onde os pesos foram definidos pelo inverso da fração amostral dos grupos étnicos de cada centro[15-16.]

Neste trabalho são consideradas apenas as variáveis nível de interleucina-6 (IL6) e adiponectina (ADIPO) observadas nos 668 indivíduos da amostra aleatória da coorte do delineamento do tipo caso-coorte[15-16]. Destes, 314 indivíduos são afro-americanos e 354 são brancos, cujos pesos amostrais são 22,65 e 7,18, respectivamente. As distribuições marginais dessas variáveis são bastante assimétricas, indicando que a suposição de normalidade bivariada não está satisfeita. Para estimar o coeficiente de correlação linear de Pearson ponderado entre estas variáveis, foi considerado o estimador pontual definido na equação (4) e um intervalo com 95% de confiança foi estimado pelo método bootstrap. Para tanto, foram considerados os métodos do bootstrap padrão, com e sem correção de viés, bem como o método do percentil com correção de viés[21]. Estes intervalos foram construídos usando 1.000 amostras bootstrap a partir de uma amostra original com 668 observações.

3. Resultados

A Figura 1 mostra um gráfico de superfície suavizado das densidades conjuntas, os histogramas das marginais e os diagramas de dispersão para as populações geradas usando a macro *mvn*. No contexto em que foram simuladas distribuições normais em cada estrato o gráfico mostra certa similaridade com a normal bivariada.

A Tabela 1 apresenta as médias das 10.000 estimativas pontuais produzidas pelos métodos da amostra expandida e dos postos para amostras com e sem reposição nas duas populações. As médias dos erros, erros absolutos e erros quadráticos também estão apresentados na tabela. O menor erro médio foi igual à -0,000167 na população não normal, para amostragem sem reposição, da amostra expandida e pesos não inteiros, enquanto que o maior erro médio foi de -0,002781 para população normal, amostragem sem reposição, método da amostra expandida e pesos não inteiros.

Os erros quadrático médio (EQM) mostram que os estimadores parecem precisos, haja vista que maior valor observado foi 0,000720 para população normal, para amostragem com reposição, método da amostra expandida e pesos não inteiros. O menor EQM observado foi de 0,000378 na população não normal, usando amostras sem reposição, pelo método dos postos e para pesos inteiros.

Na população com distribuição normal nos estratos a razão entre os EQM variou entre 0,996 e 1,012, sugerindo que os dois estimadores podem ser igualmente eficientes. Quando os estimadores são aplicados à população não normal essa razão também ficou próxima do valor 1, mas variando entre 1,037 e 1,056 e sugerindo que o estimador baseado nos postos pode ser mais eficiente do que o estimador baseado na amostra expandida.

As Figuras 2 e 3 mostram os histogramas das estimativas do coeficiente de correlação de Spearman ponderado. Todas as distribuições parecem simétricas em torno das

verdadeiras correlações. A Figura 4 mostra os diagramas de caixas para os erros e erros quadráticos, sugerindo que as discrepâncias entre os valores estimados e o correspondente parâmetro populacional são pequenas e simétricos em torno do zero, ou seja, que aparentemente não há viés. A maioria dos erros quadráticos está concentrada muito próximo do valor zero, indicando que os estimadores têm boa precisão. Em todas as situações investigadas, a hipótese de que a correlação linear é nula na população foi rejeitada (P<0,001).

Na amostra original, o coeficiente de correlação linear de Pearson ponderado entre os marcadores de inflamação interleucina-6 e adiponectina, foi -0,12092. O intervalo com 95% de confiança pelo método bootstrap padrão sem correção de viés, foi de (-0,1881; -0,0564) e, com correção de viés, forneceu o intervalo (-0,1854; -0,0537). Para o método do percentil com correção de viés o intervalo com 95% de confiança foi (-0,1854 ; -0,0501). A Figura 5 mostra o histograma das estimativas do coeficiente de Pearson ponderado nas 1.000 amostras bootstrap, sugerindo que se ajusta à distribuição normal.

4. Discussão

Em muitos estudos epidemiológicos é comum a necessidade de estimar a correlação linear entre duas variáveis quantitativas na presença de ponderação, especialmente quando planos complexos de amostragem são usados[6-9; 11-16].

Os dois procedimentos para estimar o coeficiente de correlação de Spearman ponderado investigados neste trabalho mostraram-se praticamente equivalentes quanto ao vício e precisão, com pequena vantagem do estimador baseado nos postos na situação em que os dados não têm distribuição normal bivariada. É possível que este pequeno ganho em eficiência seja decorrente do fato de que o método da amostra expandida é influenciado pelo grande número de empates dos postos, haja vista que, como ilustra o caso simulado, o tamanho da amostra original de 500 elementos é artificialmente inflacionado para 20.000 elementos.

É importante salientar também uma limitação natural do coeficiente de Spearman, que apenas avalia uma tendência de crescimento conjunto (correlação positiva) das variáveis ou, em outra situação, na qual o crescimento de uma variável tende a estar associado ao decréscimo da outra variável (correlação negativa). Como isto é captado através dos postos das variáveis, a verdadeira relação funcional (linear, quadrática, logarítmica, etc.) dos valores observados não é, de fato, identificada e quantificada. Por este motivo, certamente é mais informativo utilizar o coeficiente de correlação de Pearson ponderado e construir um intervalo de confiança bootstrap para fazer as inferências. A principal vantagem dos métodos de reamostragem sobre os métodos tradicionais de construção de intervalos de confiança é que não é necessário conhecer a distribuição de probabilidade do estimador [21]. Além disso, o intervalo de confiança pode ser visto como um

teste de hipóteses informal, pois a decisão é equivalente àquela de um teste de hipóteses bilateral. Por exemplo, se o intervalo de confiança bootstrap não contém o valor zero (ou, em outra situação, outro valor de referência), então se pode concluir que a correlação não é nula na população. No exemplo, a estimativa do coeficiente de Pearson ponderado entre os níveis de interleucina-6 e adiponectina foi -0,12092 e nenhum dos intervalos de confiança bootstrap contem o valor de referência zero, sugerindo que estejam correlacionados negativamente na população.

Com base nos resultados das simulações, para estimar correlação linear entre variáveis na presença de ponderação e na presença de violações na suposição de normalidade bivariada, se recomenda a utilização do coeficiente de correlação de Pearson ponderado e intervalo de confiança bootstrap. É importante destacar que métodos de reamostragem, como o bootstrap, podem ser implementados de maneira razoavelmente simples. Um trabalho recente utilizou o intervalo de confiança bootstrap para correlações de Pearson por que não puderam assumir normalidade e a amostra era pequena [22]. Outros trabalhos também são exemplos de estudos na área epidemiológica que utilizaram o método bootstrap em suas análises estatísticas[23-25].

Por outro lado, os estimadores para o coeficiente de Spearman ponderado discutidos neste trabalho são recomendados quando os dados são medidos em escala ordinal.

REFERÊNCIAS

(1) Daniel WW. *Applied nonparametric statistics*. Boston: Houghton-Mifflin; 1978.

(2) Siegel S. *Estatística não-paramétrica para as ciências do comportamento*. São Paulo: McGraw-Hill do Brasil; 1975.

(3) Zar J. *Biostatistical Analysis*. Upper Saddle River – NJ: Prentice-Hall, 4ª edition; 1999.

(4) Deming WE . *Some Theory of Sampling*. New York: John Wiley & Sons; 1950.

(5) Silva PLN, Pessoa DGC, Lila MF. Análise estatística de dados da PNAD: incorporando a estrutura do plano amostral. *Ciência & Saúde Coletiva* 2002; vol.7 (4).

(6) Almeida MF, Barata RB, Montero CV, Silva ZP. Prevalência de doenças crônicas auto-referidas e utilização de serviços de saúde, PNAD/1998, Brasil. *Ciência e Saúde Coletiva* 2002; 7(4):743-756.

(7) Lima-Costa MF, Barreto SM, Giatti L. A situação socioeconômica afeta igualmente a saúde de idosos e adultos mais jovens no Brasil? Um estudo utilizando dados da Pesquisa Nacional por Amostras de Domicílios – PNAD/98. *Ciência e Saúde Coletiva* 2002; 7(4):813-824.

(8) Bahia L, Costa AJL, Fernandes C, Luiz RR, Cavalcanti MLT. Segmentação da demanda dos planos e seguros privados de saúde: uma análise das informações da PNAD/98. *Ciência e Saúde Coletiva* 2002; 7(4):671-686.

(9) Matos DL, Giatti L, Lima-Costa MF. Fatores sócio-demográficos associados ao uso de serviços odontológicos entre idosos brasileiros:um estudo baseado na Pesquisa

Nacional por Amostras de Domicílios. *Ciência e Saúde Coletiva* 2004; 20(5):1290-1297.

(10) Centers for Disease Control and Prevention (CDC). National Center for Health Statistics (NCHS). National Health and Nutrition Examination Survey Data. Hyattsville, MD: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention.

(11) Wu T, Dorn JP, Donahue RP, Sempos CT, Trevisan M. Associations of Serum C-reactive Protein with Fasting Insulin, Glucose, and Glycosylated Hemoglobin**.** *American Journal of Epidemiology* 2002; 155(1).

(12) Dixon LB, Winkleby MA, Radimer KL. Dietary Intakes and Serum Nutrients Differ between Adults from Food-Insufficient and Food-Sufficient Families: Third National Health and Nutrition Examination Survey, 1988–1994. *American Society for Nutritional Sciences* 2001; 131:1232-1246.

(13) Eisner MD. Environmental Tobacco Smoke Exposure and Pulmonary Function among Adults in NHANES III: Impact on the General Population and Adults with Current Asthma. *Environmental Health Perspectives* 2002; vol. 110 (8).

(14) Himes JH, Obarzanek E, Baranowski T, Wilson DM, Rochon J, McClanahan BS. Early Sexual Maturation, Body Composition, and Obesity in African-American Girls. *Obesity Research* 2004; vol. 12.

(15) Duncan BB, Schmidt MI, Pankow JS, Ballantyne CM, Couper D, Vigo A, Hoogeveen R, Heiss G. Low-Grade Systemic Inflammation and the Development of Type 2 Diabetes. The Atherosclerosis Risk in Communities Study. *Diabetes* 2003; vol. 52.

(16) Duncan BB, Schmidt MI, Pankow JS, Bang H, Couper D, Ballantyne CM, Hoogeveen RC, Heiss G. Adiponectin and the Development of Type 2 Diabetes. The Atherosclerosis Risk in Communities Study. *Diabetes* 2004; vol. 53.

(17) SAS INSTITUTE INC. (1999). **SAS OnlineDoc®. Version 8**, Cary, NC: SAS Institute Inc.

(18) SAS INSTITUTE (2005). Data analysis sample programs Disponível em: www.sas.com/techsup/download/stat/mvn.html. [Acessado em novembro de 2005]. Cary, NC: SAS Institute Inc.

(19) Efron B. Bootstrap methods: another look at the jackknife. *Annals of Statistics* 1979; 7: 1-26.

(20) Efron B, Tibshirani RJ. *An Introduction to the Bootstrap*. Chapman and Hall, 1993.

(21) Manly BFJ. *Randomization, Bootstrap and Monte Carlo Methods in Biology*. New Zealand: University of Otago, 2ª edition; 2004.

(22) Lee JS, Ettinger B, Stanczyk FZ, Vittinghoff E, Hanes V, Cauley JA, Chandler W, Settlage J, Beattie MS, Folkerd E, Dowsett M, Grady D, Cummings SR. Comparison of Methods to Measure Low Serum Estradiol Levels in Postmenopausal Women. The *Journal of Clinical Endocrinology & Metabolism* 2006; 91(10):3791-3797.

(23) Oviedo M, Muñoz MP, Dominguez A, Carmona G. Estimated Incidence of Hepatitis A Virus Infection in Catalonia. AEP 2006; vol. 16(11): 812-819. (24) Levy PS, Lemeshow S. *Sampling for health professionals*. Belmont: Lifetime Learning Publications; 1980.

(24) Oliva A, Llabrés M, Fariña JB. Data Analysis of Kinetic Modelling Used in Drug Stability Studies: Isothermal Versus Nonisothermal Assays. *Pharmaceutical Research* 2006; vol. 23(11).

(25) Ndrepepa G, Kastrati A, Mehilli J, Neuman FJ, ten Berg J, Bruskina O, Dotzer F, Seyfarth M, Pache J, Dischinger J, Ulm K, Berger PB, Schömig A. Age-Dependent Effect of Abciximab in Patients With Acute Coronary Syndromes Treated With Percutaneous Coronary Interventions. *American Heart Association* 2006; 114, 2040-2046.

Tabelas e Figuras

Tabela 1 - Resultados das simulações para os estimadores do coeficiente de Spearman ponderado.

	Distribuição Conjunta Normal ρ_S †=0,6440215 (pesos inteiros) ρ_S †=0,645910 (pesos não inteiros)				Distribuição Conjunta Não Normal ρ_S †=0,745767 (pesos inteiros) ρ_S †=0,745429 (pesos não inteiros)			
	Com Reposição		Sem Reposição		Com Reposição		Sem Reposição	
	Amostra expandida	Postos	Amostra expandida	Postos	Amostra expandida	Postos	Amostra expandida	Postos
Pesos Inteiros								
Estimativas Médias	0,642770	0,641997	0,642826	0,642046	0,745010	0,747331	0,745452	0,747840
Erro Médio	-0,001251	-0,002025	-0,001196	-0,001975	-0,000756	0,001564	-0,000315	0,002073
Erro Absoluto Médio	0,021299	0,021326	0,021176	0,021203	0,015963	0,015569	0,015858	0,015531
Erro Quadrático Médio	0,000712	0,000715	0,000710	0,000712	0,000401	0,000379	0,000397	0,000378
Pesos não inteiros								
Estimativas Médias	0,643401	0,644152	0,643129	0,643877	0,745213	0,747387	0,745262	0,747486
Erro Médio	-0,002509	-0,001758	-0,002781	-0,002033	-0,000216	0,001959	-0,000167	0,002058
Erro Absoluto Médio	0,021287	0,021188	0,020892	0,020793	0,015934	0,015600	0,015834	0,015620
Erro Quadrático Médio	0,000720	0,000712	0,000693	0,000685	0,000398	0,000380	0,000396	0,000382

† ρ_S é o valor da correlação de Spearman na população

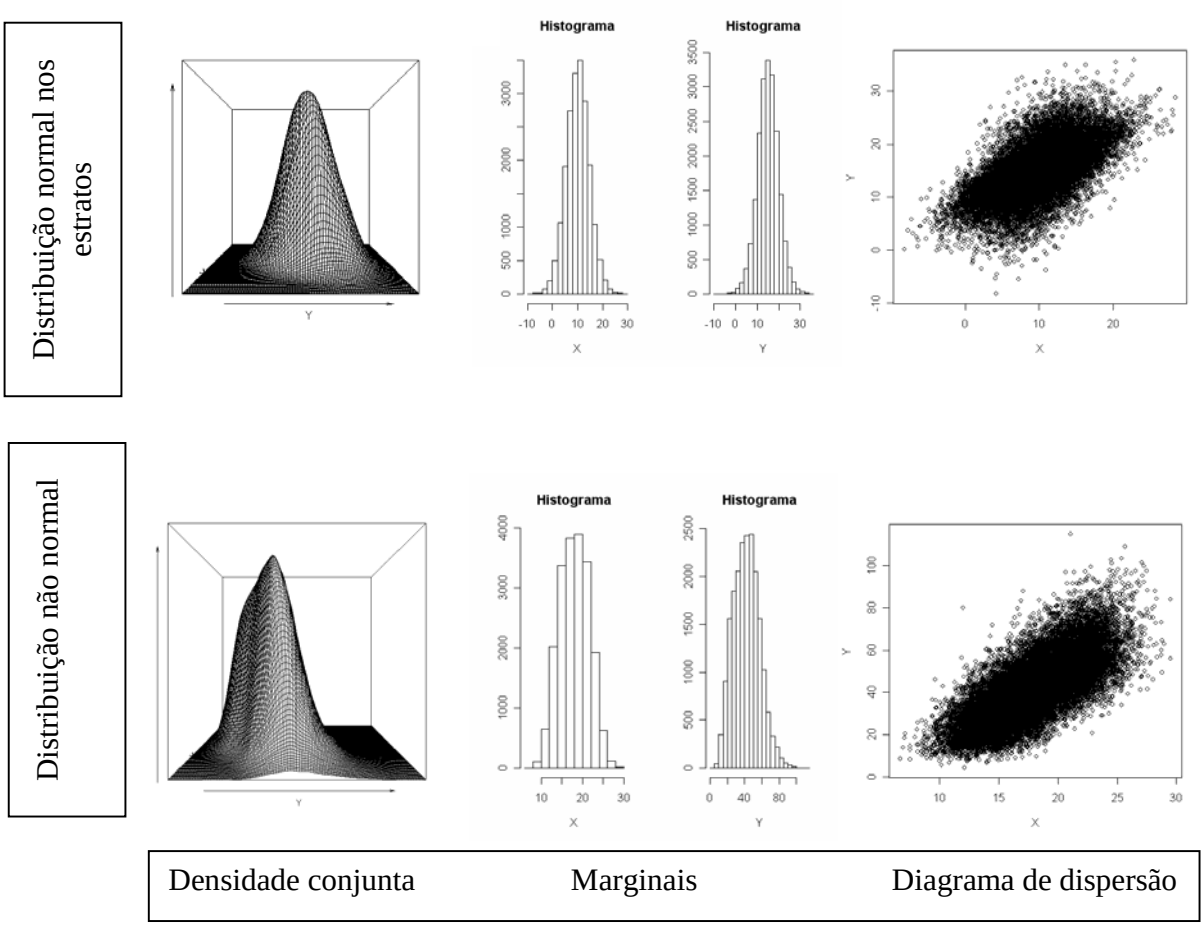

Figura 1 – Densidades conjuntas suavizadas, marginais e diagrama de dispersão para as populações

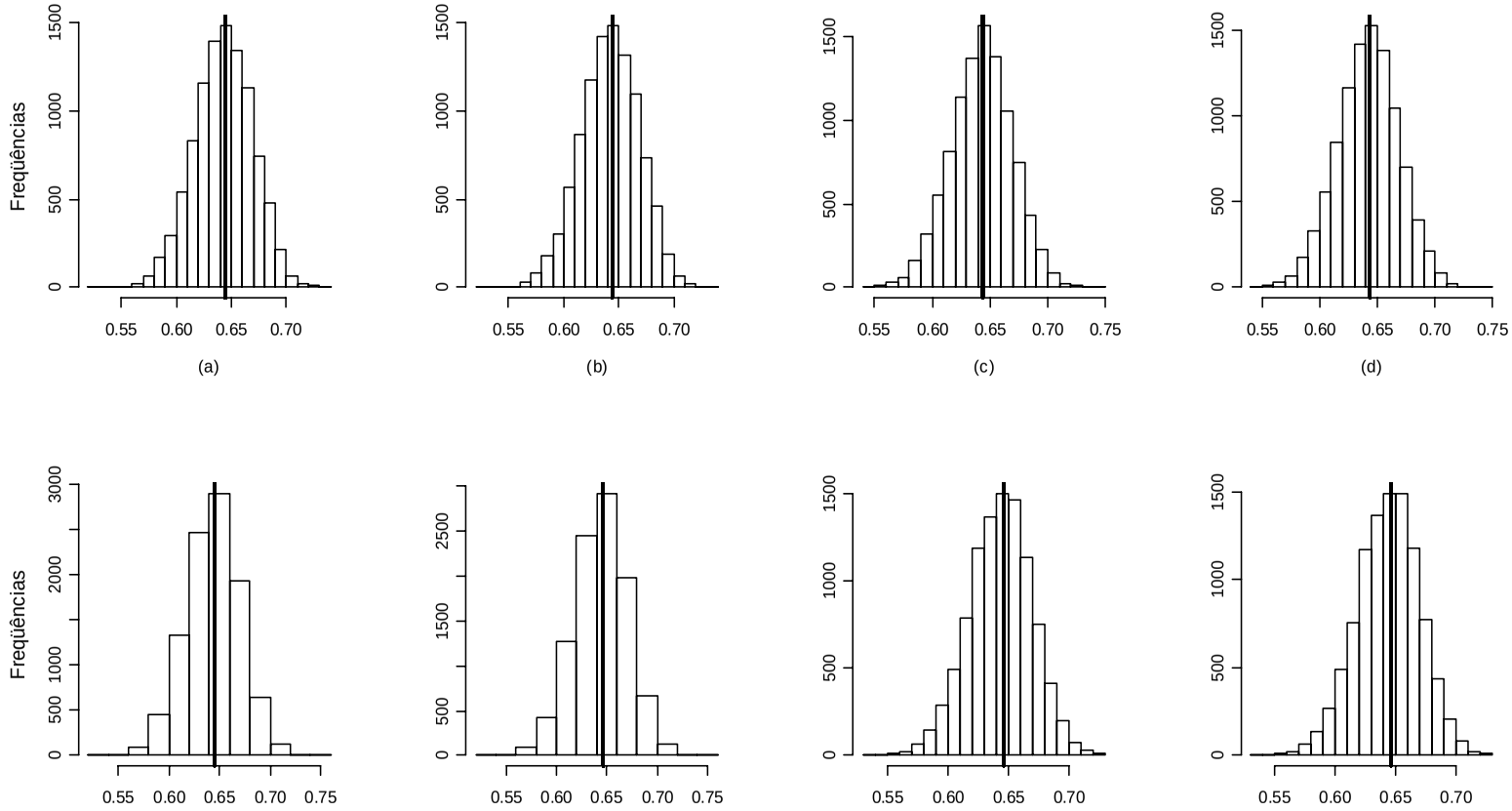

Figura 2 - Histogramas das estimativas da correlação de Spearman para distribuição normal em cada estrato: (a) método da amostra expandida usando amostragem com reposição e pesos inteiros; (b) método dos postos usando amostragem com reposição e pesos inteiros; (c) método da amostra expandida usando amostragem sem reposição e pesos inteiros; (d) método dos postos usando amostragem sem reposição e pesos inteiros; (e) método da amostra expandida usando amostragem com reposição e pesos não inteiros; (f) método dos postos usando amostragem com reposição e pesos não inteiros; (g) método da amostra expandida usando amostragem sem reposição e pesos não inteiros; (h) método dos postos usando amostragem sem reposição e pesos não inteiros.

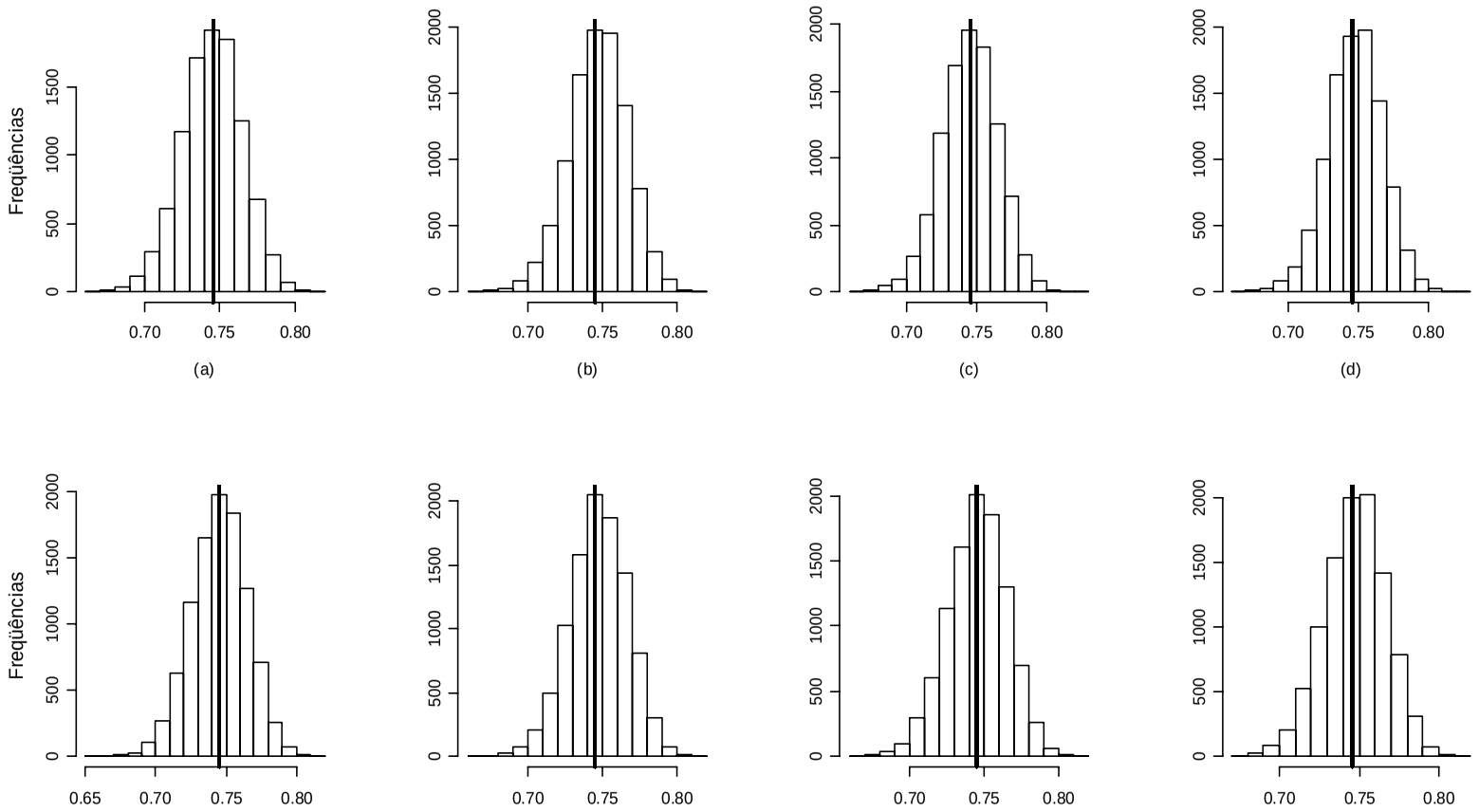

Figura 3 - Histogramas das estimativas da correlação de Spearman considerando distribuição bivariada não normal: (a) método da amostra expandida usando amostragem com reposição e pesos inteiros; (b) método dos postos usando amostragem com reposição e pesos inteiros; (c) método da amostra expandida usando amostragem sem reposição e pesos inteiros; (d) método dos postos usando amostragem sem reposição e pesos inteiros; (e) método da amostra expandida usando amostragem com reposição e pesos não inteiros; (f) método dos postos usando amostragem com reposição e pesos não inteiros; (g) método da amostra expandida usando amostragem sem reposição e pesos não inteiros; (h) método dos postos usando amostragem sem reposição e pesos não inteiros.

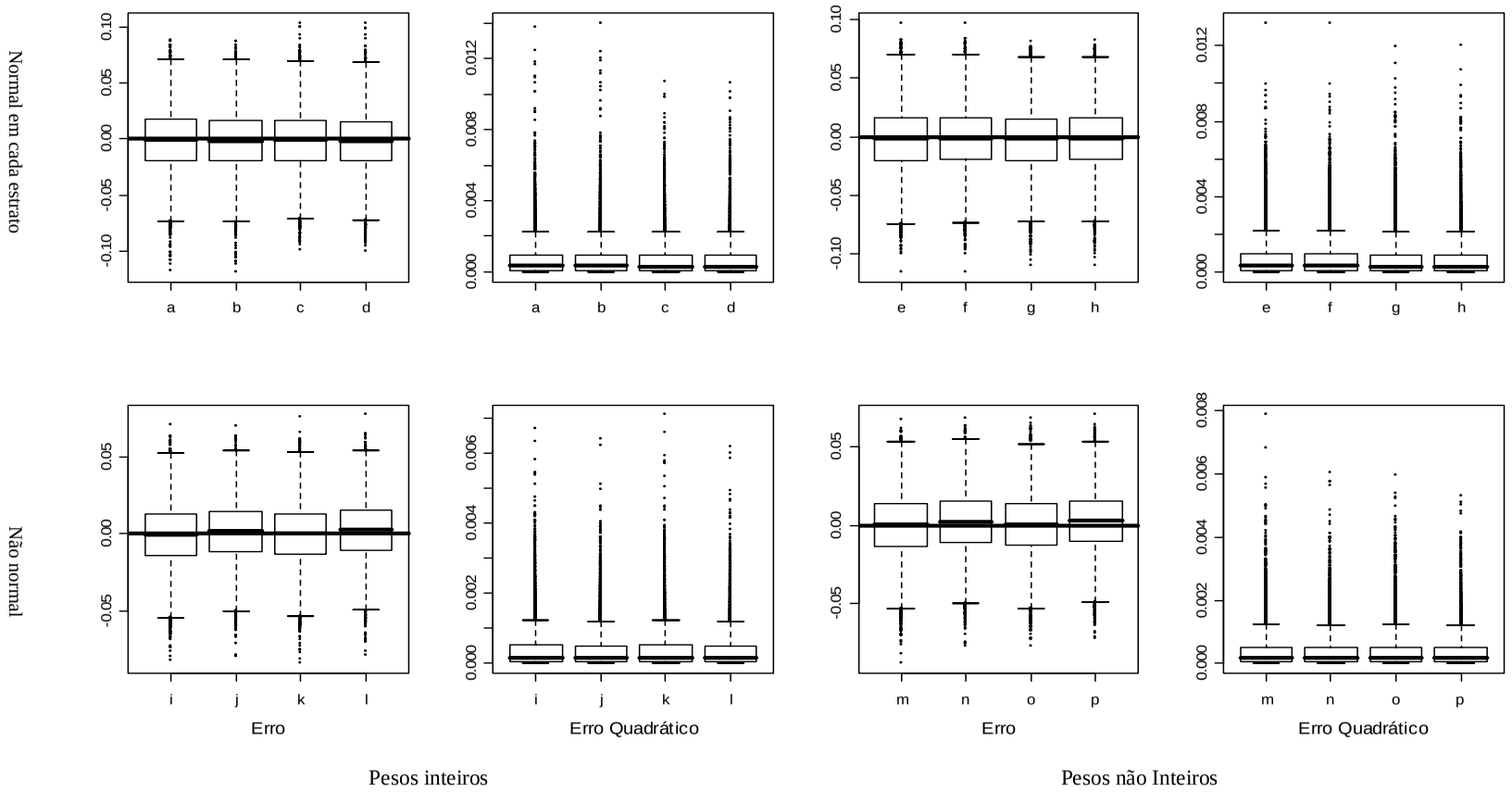

Figura 4 - Diagrama de caixas para os erros e erros quadráticos das estimativas: a) método da amostra expandida usando amostragem com reposição; b) método da amostra expandida usando amostragem sem reposição; c) método dos postos usando amostragem com reposição e pesos inteiros; d) método dos postos usando amostragem sem reposição; e) método da amostra expandida usando amostragem com reposição; f) método da amostra expandida usando amostragem sem reposição; g) método dos postos usando amostragem com reposição; h) método dos postos usando amostragem sem reposição; i) método da amostra expandida usando amostragem com reposição; j) método da amostra expandida usando amostragem sem reposição; l) método dos postos usando amostragem com reposição; m) método dos postos usando amostragem sem reposição; n) método da amostra expandida usando amostragem com reposição; o) método da amostra expandida usando amostragem sem reposição; p) método dos postos usando amostragem com reposição; q) método dos postos usando amostragem sem reposição.

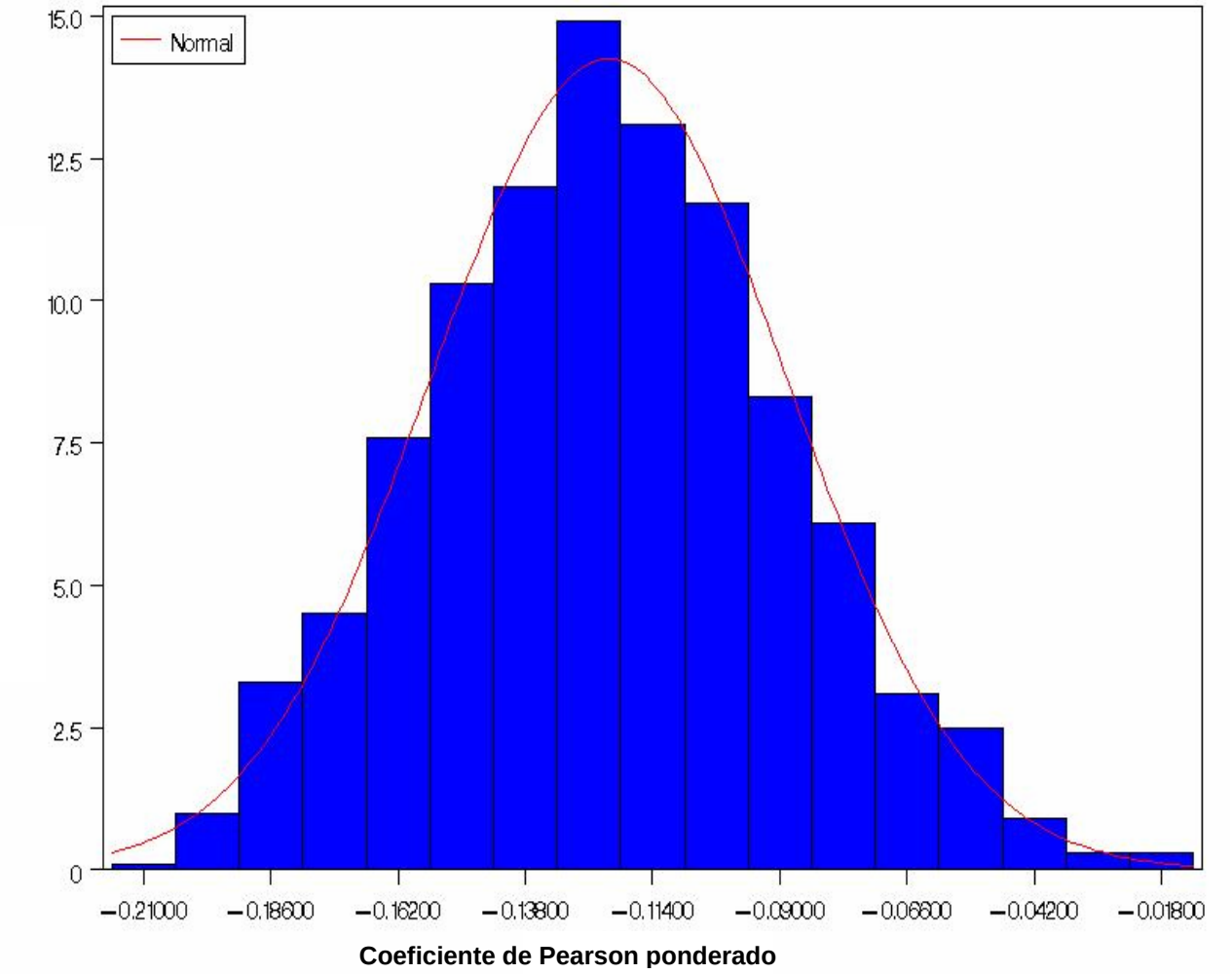

Figura 5 – Histograma das estimativas do coeficiente de correlação de Pearson ponderado entre os valores de interleucina-6 e adiponectina nas 1.000 amostras bootstrap.

8. CONCLUSÕES E CONSIDERAÇÕES FINAIS

Este trabalho investigou dois procedimentos para estimar e testar a correlação linear usando o coeficiente de Spearman na presença de pesos amostrais, como ocorre frequentemente quando se utiliza um delineamento amostral complexo.

Os estimadores do coeficiente de Spearman ponderado mostraram ser não viesados e igualmente eficientes. Os resultados obtidos sugerem que para amostras grandes estes métodos podem ser uma alternativa razoável para estimar a correlação. No entanto, como o coeficiente de Spearman avalia a correlação entre as variáveis através dos postos e não nos valores originais, pode, eventualmente, não captar a verdadeira forma funcional (quadrática, logarítmica, etc.). Outra eventual limitação ocorre quando o número de empates é grande, como ocorre com o método da amostra expandida. Assim, se recomenda a utilização desses métodos quando as variáveis são medidas em escala ordinal.

Também foi ilustrado a estimação da correlação de Pearson ponderado com intervalos de confiança bootstrap na ausência de normalidade. Este método é mais vantajoso, pois não há perda de informação como ocorre com o coeficiente de Spearman, uma vez que as informações originais das variáveis são usados, e não os seus postos. Este método dispensa suposições sobre a distribuição dos dados e do estimador e podem ser implementados de maneira razoavelmente simples.

Esse trabalho não avaliou o desempenho dos estimadores do coeficiente de Spearman ponderado em amostras pequenas, pois a distribuição da estatística de teste exige tamanho de amostra grande. Apesar de ser possível usar o intervalo de confiança como um teste de hipóteses informal para a tomada de decisão, a exploração de testes de hipóteses

para o coeficiente de Pearson ponderado, usando reamostragem, pode ser uma continuidade natural do trabalho.

9. ANEXOS

Anexo A – Projeto de Pesquisa

Anexo B – Rotinas computacionais SAS

1. Rotinas para gerar as populações
    1.1. Macro *mvn*
    1.2. População com distribuição normal bivariada nos estrato
    1.3. População com distribuição conjunta não normal
2. Estimação do coeficiente de Spearman ponderado
    2.1. Método da amostra expandida em amostragem com reposição
    2.2. Método da amostra expandida em amostragem sem reposição
    2.3. Método dos postos em amostragem com reposição
    2.4. Método dos postos em amostragem sem reposição
3. Intervalos de confiança bootstrap para coeficiente de Pearson ponderado

Anexo A – Projeto

Título: Estimação do coeficiente de correlação de Spearman Ponderado

1. Introdução

Em diversos estudos epidemiológicos a população que está sendo investigada é constituída por subgrupos ou estratos, podendo ser importante preservar a representatividade dos mesmos no processo de amostragem. Um procedimento usual é utilizar um plano amostral com amostragem estratificada proporcional ao tamanho dos grupos. No entanto, para garantir o poder das comparações ou a estabilidade nos coeficientes de regressão estimados é comum super-representar grupos minoritários.

Embora seja um procedimento usual, para que os resultados das análises possam ser corretamente generalizados para a população, é fundamental corrigir a importância relativa de cada elemento da amostra, mediante a utilização de pesos, de acordo com o estrato ao qual pertence. Usualmente, os pesos são definidos pelo inverso das frações amostrais de cada estrato (1).

Os estudos NHANES (*National Health and Nutrition Examination Survey*) são exemplos de estudos epidemiológicos nos quais é fundamental utilizar pesos nas análises para produzir estimativas nacionais sobre estado de saúde e nutrição da população norte-americana (2).

Por exemplo, dados do NHANES III foram usados para investigar a associação entre proteína C-reativa com níveis de insulina, glicemia e hemoglobina glicada no plasma. Os resultados das correlações de Pearson ponderadas mostraram associações significativas entre proteína C-reativa e insulina, glicemia e hemoglobina glicada, bem como com outras variáveis tais como idade, educação, consumo de cigarro, índice de massa corporal e atividade física nos períodos de lazer (3). Entretanto, apesar de que possíveis violações nas

suposições do teste de hipóteses da significância do coeficiente de correlação de Pearson poderiam alterar as estimativas das probabilidades de significância associadas (valores P), não puderam estimar correlações de Spearman ponderadas, haja vista que o procedimento disponível no programa SAS não é válido para esta situação. Como alternativa, análises não ponderadas dos postos (ranks) da proteína C-reativa produziram estimativas das correlações de Pearson e Spearman muito similares (3). Esta questão metodológica também surgiu em análises recentes sobre a associação entre marcadores de inflamação sistêmica e adiponectina e o desenvolvimento de diabetes melito tipo 2 no projeto ARIC (4-5).

Outros trabalhos baseados em dados do NHANES que utilizaram o coeficiente de Spearman são (6-8) sem, no entanto, explicitar se os pesos foram ou não considerados na análise.

Apesar da importância do método, uma detalhada revisão da literatura não mostrou trabalhos metodológicos que apresentem alternativas para a estimação e teste de hipóteses para a correlação de Spearman ponderada.

2. Objetivos

Apresentar e comparar a eficiência de diferentes estimadores para o coeficiente de Spearman ponderado mediante estudo de simulação Monte Carlo.

3. Métodos

Na literatura não foram encontrados estimadores para o coeficiente de Spearman quando é necessário considerar pesos na amostra, como é visto para o coeficiente de correlação de Pearson (9). Neste trabalho serão descritas e comparadas duas formas de estimação do coeficiente de correlação de Spearman na presença de ponderação.

Uma alternativa usual de estimação consiste em replicar cada elemento da amostra um número igual ao peso correspondente, ou seja, se o peso de um determinado indivíduo é *k*, então este registro será repetido *k* vezes no banco de dados. Similarmente, este procedimento é realizado para todos os elementos da amostra, produzindo um banco de dados expandido que representa artificialmente os pesos de cada observação. Assim, pode-se usar os métodos computacionais usuais para a estimação da correlação de Spearman. É importante notar, entretanto, que apesar de preservar a importância relativa de cada elemento da amostra na estimativa pontual, os resultados dos testes de significância sobre o coeficiente de correlação de Spearman podem não ser válidos, haja vista que o tamanho da amostra foi artificialmente inflacionado, ocorrendo o mesmo com os graus de liberdade associados à distribuição de referência. Outro aspecto importante é quanto à presença de pesos não inteiros, um aspecto que eventualmente poderia produzir vício na estimação decorrente do arredondamento dos mesmos na etapa de definição do banco de dados multiplicado.

Outra forma de estimação do coeficiente de correlação de Spearman ponderado consiste em utilizar um procedimento similar àquele usado para estimar o coeficiente de Pearson ponderado. Assim, considere que $(X_1,Y_1),(X_2,Y_2),...,(X_n,Y_n)$ representa

observações independente de um fenômeno bivariado (X,Y), cujos pesos correspondentes são $w_1, w_2, ..., w_n$. Então, o coeficiente de correlação de Pearson ponderado pode ser estimado por

$$r_{xy} = \frac{\sum w_i (x_i - \bar{x}_w)(y_i - \bar{y}_w)}{\sqrt{\sum w_i (x_i - \bar{x}_w)^2 \sum w_i (y_i - \bar{y}_w)^2}} \quad (1)$$

onde $\bar{x}_w = \frac{\sum w_i x_i}{\sum w_i}$ e $\bar{y}_w = \frac{\sum w_i y_i}{\sum w_i}$.

Para estimar o coeficiente de Spearman, a idéia é substituir os valores observados (x_i, y_i) na equação (1) pelos correspondentes postos dos valores ordenados, utilizando-se a média dos postos nos casos de empates.

Naturalmente, a adequação deste procedimento, tanto quanto a eficiência quanto ao vício devem ser avaliadas, assim como a distribuição de referência usada nos testes de hipóteses. Para tanto, serão implementadas rotinas no SAS para gerar uma população com N elementos composta por k sub-populações (estratos). Para a distribuição do vetor aleatório (X,Y), serão consideradas a distribuição normal bivariada em cada estrato e não normal. A macro mvn descrita na página de suporte técnico do SAS, será usada para gerar observações de uma população com distribuição normal bivariada com parâmetros μ_{xi}, μ_{yi}, σ_{xi}, σ_{yi} e ρ_i, onde $i = 1,2,...,k$ (10). A população com distribuição não normal será criada com um gerador de números aleatórios disponível em muitos pacotes estatísticos. Com o gerador de números aleatórios, serão criadas as duas variáveis com uma determinada correlação, onde uma delas terá distribuição assimétrica. O coeficiente de

Spearman na população será calculado com estes valores para posteriormente ser comparado com os estimadores mencionados acima.

Através de simulação Monte Carlo (11-12) serão geradas aleatoriamente 10.000 amostras destas populações, para os cenários com e sem reposição, considerando $\frac{n}{k}$ observações cada estrato. Assim, serão 10.000 amostras diferentes contendo n observações que serão guardadas em bancos distintos.

4. Cronograma

Atividades	Abril	Maio	Junho	Julho	Agosto	Setembro	Outubro	Novembro	Dezembro
Aprovação do projeto									
Revisão da literatura									
Desenvolvimento de rotinas computacionais									
Teste das rotinas									
Execução das simulações									
Redação da Dissertação e artigo									
Defesa preliminar									
Sessão pública									

5. Referências

(1) Pedro Luis do Nascimento Silva, Djalma Galvão Carneiro Pessoa, Maurício Franca Lila. *Análise estatística de dados da PNAD: incorporando a estrutura do plano amostral*. Ciência & Saúde Coletiva, vol.7 (4),2002.

(2) Centers for Disease Control and Prevention (CDC). National Center for Health Statistics (NCHS). National Health and Nutrition Examination Survey Data. Hyattsville, MD: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention.

(3) Tiejan Wu, Joan P. Dorn, Richard P. Donahue, Christopher T. Sempos, and Maurizio Trevisan. *Associations of Serum C-reactive Protein with Fasting Insulin, Glucose, and Glycosylated Hemoglobin***.** American Journal of Epidemiology, 155(1), 2002.

(4) Bruce B. Duncan *et al. Adiponectin and the Development of Type 2 Diabetes. The Atherosclerosis Risk in Communities Study*. Diabetes, Vol. 53 (2004).

(5) Bruce B. Duncan *et. al. Low-Grade Systemic Inflammation and the Development of Type 2 Diabetes. The Atherosclerosis Risk in Communities Study*. Diabetes, Vol. 52 (2003).

(6) Lori Beth Dixon, Marilyn A. Winkleby and Kathy L. Radimer. *Dietary Intakes and Serum Nutrients Differ between Adults from Food-Insufficient and Food-Sufficient Families: Third National Health and Nutrition Examination Survey, 1988–1994.* American Society for Nutritional Sciences. 2001.

(7) John H. Himes *et. al. Early Sexual Maturation, Body Composition, and Obesity in African-American Girls.* OBESITY RESEARCH vol. 12, 2004.

(8) Mark D. Eisner. Environmental Tobacco Smoke Exposure and Pulmonary Function among Adults in NHANES III: Impact on the General Population and Adults with Current Asthma. Environmental Health Perspectives, Vol.110 (8), 2002.

(9) SAS INSTITUTE INC. (1999). ***SAS OnlineDoc*®. Version 8**, Cary, NC: SAS Institute Inc.

(10) SAS INSTITUTE (2005). *Data analysis sample programs. Disponível em: www.sas.com/techsup/download/stat/mvn.html*. (Acessado em novembro de 2005), Cary, NC: SAS Institute Inc.

(11) Brian F. J. Manly. Randomization, Bootstrap and Monte Carlo Methods in Biology. University of Otago, New Zealand, 2ª edition, 2004.

(12) James E. Gentle. Random *Number Generation and Monte Carlo Methods*. George Mason University, 2ª edition, 2002.

Anexo B – Rotinas Computacionais (SAS)

1. Rotinas para gerar as populações

1.1 Marcro *mvn*

```
%macro mvn(version,
           varcov=,       /* dataset for variance-covariance matrix */
           means=,        /* dataset for mean vector */
           n=,            /* sample size */
           seed=0,        /* seed for random number generator */
           sample=);      /* output dataset name */


%if &version ne %then %put MVN macro Version 1.0;

 /* Get initial seed value.  If seed<=0, then generate seed from the
    system clock. */

data _null_;
   if &seed le 0 then do;
      seed = int(time());   /* get clock time in integer seconds */
      put seed=;
      call symput('seed',seed);   /* store seed as macro variable */
end;
run;


/* Generate the multivariate normal data in SAS/IML */

    proc iml worksize=100;
use &varcov;            /* read variance-covariance matrix */
read all into cov;
use &means;             /* read means */
read all into mu;
v=nrow(cov);            /* calculate number of variables */
n=&n;
seed = &seed;
l=t(root(cov));         /* calculate cholesky root of cov matrix */
z=normal(j(v,&n,&seed));/* generate nvars*samplesize normals */
x=l*z;                  /* premultiply by cholesky root */
x=repeat(mu,1,&n)+x;    /* add in the means */
tx=t(x);
create &sample from tx;  /* write out sample data to sas dataset */
append from tx;
    quit;

%mend mvn;
```

1.2 População com distribuição normal bivariada nos estrato

```
options ps=58 ls=79 nocenter nodate nonumber formchar='|----|+|---+=|-/\<>*';
title1;
title2;
title3;
footnote;
libname L1 v7 'C:\Documents and Settings\Administrador\Meus documentos\Lidiane\
Mestrado\Dissertação\Simulacoes\Pop normal';

/* Cria matriz de variancias e covarianças para estrato 1*/

data varcov1;
    input x1 x2;
    cards;
    16 12
    12 36
    ;
run;
/* proc print;run; */

/* Cria vetor com as medias de X e Y do estrato 1*/

data means1;
    input x1;
    cards;
    10
    15
    ;
run;
/*proc print;run;*/

/* Cria matriz de variancias e covarianças para estrato 2*/

data varcov2;
    input x1 x2;
    cards;
    25 16
    16 16
    ;
run;
/*proc print;run; */

/* Cria vetor com as medias de X e Y do estrato 2*/

data means2;
    input x2;
    cards;
    10
    15
    ;
run;
/* proc print;run; */

/* Inclui a macro MVN*/

*%include 'c:\vigo\lb\mvn.sas';
%include 'C:\Documents and Settings\Administrador\Meus documentos\Lidiane\
Mestrado\Dissertação\Simulacoes\Pop normal\mvn.sas';
%mvn(varcov=varcov1, means=means1, n=8000 , seed=589, sample=data1);
data DATA1;
    set DATA1;
    ID = _n_;
    STRAT = 1;
    rename COL1=X1;
    rename COL2=X2;
run;
/* proc print; run; */

proc corr data=DATA1;
```

```
    var X1 X2;
run;

%mvn(varcov=varcov2, means=means2, n=12000 , seed=589, sample=data2);
data DATA2;
    set DATA2;
    ID = 8000 + _n_;
    STRAT = 2;
    rename COL1=X1;
    rename COL2=X2;
run;
/* proc print; run; */

proc corr data=DATA2;
    var X1 X2;
run;

proc sort data=Work.Data1 out=WORK._TABLE1_;
   by ID;
run;
proc sort data=Work.Data2 out=WORK._TABLE2_;
   by ID;
run;
data WORK.POP;
   merge WORK._TABLE1_ (in=TABLE1) WORK._TABLE2_ (in=TABLE2) ;
   by ID;
*   if TABLE1 and TABLE2;
run;

***Delete temporary data sets in WORK library******;
proc datasets nolist;
   delete _TABLE1_ _TABLE2_ varcov1 varcov2 means1 means2 data1 data2 ;
run;
quit;

proc print data=POP; run;

data L1.POP;
    set POP;
run;
proc corr data=POP spearman;
    title1 "Correlacao na Populacao";
    var X1 X2;
run;
quit;
```

1.3 População com distribuição conjunta não normal

```
options ps=58 ls=80 nocenter nodate nonumber formchar='|----|+|---+=|-/\<>*';
title1;
title2;
title3;
footnote;
libname L1 V7 'C:\Documents and Settings\Administrador\Meus documentos\Lidiane\Mestrado
\Dissertação\Simulacoes\Não normal';

* Gera a populaçao nao normal;
* Estrato 1 (data set data1) com 8000 elementos. Variavel x tem distribuiçao normal
e y distribuiçao Gama;

data data1;
retain Seed_1 1298573062 Seed_2 447801538;
   do i=1 to 8000;
    ID = _n_;
     x=15+sqrt(5)*rannor(Seed_1);
     y = x/5 * rangam(Seed_2,10);
     strat=1;
    output; end;
   run;

proc print data=data1;
   id i;
   var Seed_1 Seed_2 x y;
run;

*Histograma de x e y;
proc univariate data=data1   noprint;
   var x y;
   histogram / caxes=BLACK cframe=CXF7E1C2 waxis= 1
               cbarline=BLACK cfill=BLUE pfill=SOLID
               vscale=percent hminor=0 vminor=0
               name='HIST';
   run;
symbol;
goptions ftext= ctext= htext=;

proc corr data=data1 spearman;
var x y; run;

* Estrato 2 (data set data2) com 12000 elementos. Variavel x tem distribuiçao normal
e y distribuiçao Gama;

data data2;
retain Seed_1 5688562 Seed_2 879801538;
   do i=1 to 12000;
    ID = 8000+_n_;
    x=20+sqrt(7)*rannor(Seed_1);
    y = x/10 * rangam(Seed_2,25);
    strat=2;
     output; end;
   run;

proc print data=data2;
   id i;
   var Seed_1 Seed_2 x y;
run;

*Histograma de x e y;
proc univariate data=data2   noprint;
   var x y;
   histogram / caxes=BLACK cframe=CXF7E1C2 waxis= 1
               cbarline=BLACK cfill=BLUE pfill=SOLID
               vscale=percent hminor=0 vminor=0
               name='HIST';
   run;
symbol;
```

```
goptions ftext= ctext= htext=;

proc corr data=data2 spearman;
var x y; run;
* Une os dois estratos em um data set chamado POP;

proc sort data=Work.data1 out=WORK._TABLE1_;
   by ID;
run;
proc sort data=Work.data2 out=WORK._TABLE2_;
   by ID;
run;
data WORK.POP;
   merge WORK._TABLE1_ (in=TABLE1) WORK._TABLE2_ (in=TABLE2) ;
   by ID;
*   if TABLE1 and TABLE2;
run;

***Delete temporary data sets in WORK library******;
proc datasets nolist;
   delete _TABLE1_ _TABLE2_ ;
run;
quit;

proc print data=POP; run;

* Salvando data set POP;
data L1.POP;
    set POP;
run;

*Corrrelaçao na populaçao;
proc corr data=POP spearman;
    title1 "Correlacao na Populacao";
    var x y;
    run;
quit;
```

2. Estimação do coeficiente de Spearman ponderado

2.1. Método da amostra expandida em amostragem com reposição

```
* Redirecting and saving the log to the file named "WSpearman.LOG";
proc printto log="C:\Documents and Settings\Administrador\Meus
documentos\Lidiane\Mestrado
\Dissertação\Simulacoes\Pop normal\Com Reposição\Multiplica banco\WSpearman.LOG";
run;
* Redirecting and saving the output to the file named "WSpearman.LST";
proc printto print="C:\Documents and Settings\Administrador\Meus
documentos\Lidiane\Mestrado
\Dissertação\Simulacoes\Pop normal\Com Reposição\Multiplica banco\WSpearman.LST";
run;

options ps=58 ls=79 nocenter nodate nonumber formchar='|----|+|---+=|-/\<>*';
title1;
title2;
title3;
footnote;
libname L1 v7 'C:\Documents and Settings\Administrador\Meus documentos\Lidiane\Mestrado
\Dissertação\Simulacoes\Pop normal';
libname L2 v7 'C:\Documents and Settings\Administrador\Meus documentos\Lidiane\Mestrado
\Dissertação\Simulacoes\Pop normal\Com Reposição\Multiplica banco';

data POP;
    set L1.POP;
    rename x1=X x2=Y;
    TAMANHO = 1;
run;

* Criando variavel com valor da correlacao de Spearman na populacao;
ods output spearmancorr=SPpop(keep=X);
proc corr data=pop spearman;
    title1 "Correlacao na Populacao";
    var X Y;
run;
quit;
data SPpop;
    set SPpop;
    if _n_ = 1 then delete;
    rename X=SPpop;
run;
proc print data=SPpop;
run;
title1;

/* Macro p/ selecionar as a amostras*/
%macro sample(rep);
  %do r=1 %to &rep;

    /* Usando o metodo Probability Proportional to Size with Replacement (PPS_WR)*/
    /* com 250 em cada estrato*/

    proc surveyselect data=POP method=PPS_WR n=250
           seed=195&r out=AMOSTRA&r outhits;
           title1 "Amostragem Estratificada Com Reposicao - Repeticao &r";
        size TAMANHO;  /* artificio p/ que cada estrato tenha o mesmo n */
           strata STRAT;
    run;

    ****** Salvando as amostras;
    data L2.AMOSTRA&r;
        set AMOSTRA&r;
    run;
    ***** Multiplicando o dataset AMOSTRA de acordo com o peso do estrato;
```

```
    data AMOSTRAM&r;
           set AMOSTRA&r;
           if strat=1 then do;
           output; output; output; output; output; output; output; output;
           output; output; output; output; output; output; output; output;
               output; output; output; output; output; output; output; output;
           output; output; output; output; output; output; output; output;
             end;
           else do;
           output; output; output; output; output; output; output; output;
           output; output; output; output; output; output; output; output;
           output; output; output; output; output; output; output; output;
               output; output; output; output; output; output; output; output;
           output; output; output; output; output; output; output; output;
           output; output; output; output; output; output; output; output;
             end;
     run;
     ****Guardando a correlaçao de Spearman com ponderaçao do PROC CORR
            em um dataset chamado spearman;

     ods output spearmancorr=spearman&r(keep=X);
     proc corr data=AMOSTRAM&r spearman;
           title1 "Correlacao na Amostra - Repeticao &r (Dataset Multiplicado)";
           var X Y;
     run;
     quit;
     proc print data=spearman&r;
     run;

    /* Excluindo a linha da correlaçao X com X */
     data spearman&r;
           set spearman&r;
           if _n_ = 1 then delete;
           rename X=WSp;
     run;
%end;
%mend sample;

%sample (10000)

%macro joincoef(rep);
    proc datasets lib=work nolist;
           delete WS;
    run;

    %do r=1 %to &rep;
         **** Une os coeficientes de Spearman calculados para todas as amostras,
                    e os guarda em WS;
         proc append
                   BASE=WS data=spearman&r;
           run;
    %end;

    * Adicionando variavel SPpop no dataset WS;
    title1;
    /*
    title1 "Adicionando variavel SPpop no dataset WS";
    proc print data=WS;
    run;
    */
    data WS;
           set WS;
           one=1;
    run;
    data SPpop;
           set SPpop;
           one=1;
    run;
    data WS (drop=one);
           merge WS SPpop;
```

```
            by one;
    run;

       ***** Calculando a estatistica T e o p-value com o n° de graus de liberdade
correto;
       ***** Calculando o ERRO, EQABS e o EQM2;
       ***** Calculando o intervalo de 95% de confiança para cada Spearman estimado;
       ***** A variavel I indica se o verdadeiro parametro esta contido no intervalo;
    data WS;
          set WS;
             tcalc = WSp/sqrt((1-WSp**2)/(498));
             pvalue = (1-probt(tcalc,498))*2;
             ERRO = WSp-SPpop;
             ERROABS = abs(ERRO);
             ERROQ = ERRO**2;
             Li=Wsp-1.96*sqrt((1-WSp**2)/(498));
             Ls=Wsp+1.96*sqrt((1-WSp**2)/(498));
             I=0;
             if Li<SPpop and Ls>SPpop
             then I=1;
             else I=0;
    run;

%mend joincoef;

%joincoef(10000);

******Calcula os quantis 2.5% e 97.5% da variável WSp (estimativas da correlação de
Spearman);
proc univariate data=WS noprint ;
var WSp;
output out=pctls pctlpts=2.5 97.5 pctlpre=WS_
              pctlname=P25 P975;
run;

******Adiciona os quantis 2.5% e 97.5% no data set WS;
data WS;
           set WS;
           one=1;
    run;
    data work.pctls;
           set work.pctls;
           one=1;
    run;
    data WS (drop=one);
           merge WS work.pctls;
           by one;
    run;

    data WS;
          set WS;
             *****Indica se o IC 95% p/ o coeficiente está entre os quantis 2.5 e 97.5%;
       Ie=0;
             if ws_p25<Li and ws_p975>Ls
             then Ie=1;
             else Ie=0;
             *****Indica se o intervalo baseado nos quantis está contido no IC 95%;
       If=0;
             if ws_p25>Li and ws_p975<Ls
             then If=1;
             else If=0;
             *****Indica se o IC 95% p/ o coeficiente tem o limite superior entre
             os quantis 2.5 e 97.5%;
             I25=0;
             if ws_p25>Li and ws_p25<Ls and ws_p975>Ls
             then I25=1;
             else I25=0;
             *****Indica se o IC 95% p/ o coeficiente tem o limite inferior entre
             os quantis 2.5 e 97.5%;
             I975=0;
             if ws_p25<Li and ws_p975<Ls and ws_p975>Li
```

```
            then I975=1;
            else I975=0;

    run;
proc contents data=ws;run;


title1 "Resumo dos resultados da simulacao";
proc means data=WS maxdec=6 n mean std min max;
    var WSp ERRO ERROABS ERROQ SPpop I Ie If I25 I975;
run;
quit;
      title;
      footnote;
title1 "Diagrama de caixas";
title2 "Estimativas do coef. de correlacao de Spearman";
title3 "Metodo do banco multiplicado";

proc sql;
   create view WORK._TMP_0 as
      select *,1 as _dummy_
      from work.WS
      ;
 *** Box plot ***;
goptions ftext=SWISS ctext=BLACK htext=1 cells;
symbol1 c=BLACK h=1 cells;
axis1 major=none value=none label=none;
proc boxplot data=WORK._TMP_0 ;
   plot (WSp
            )*_DUMMY_
                              / caxis = BLACK
                                cframe = white
                                ctext = BLACK
                                cboxes = BLACK
                                cboxfill = red
                                idcolor = red
                                idsymbol = SQUARE
                                boxstyle = SKELETAL
                                waxis = 1
                                name = 'BOX'
                                description = "Box Plot of WSp"
                                npanel = 15
                                haxis = axis1
   ;
run;
symbol1;
goptions ftext= ctext= htext=;
axis1;

   quit;

title;
      footnote;
title1 "Histograma";
title2 "Estimativas do coef. de correlacao de Spearman";
title3 "Metodo do banco multiplicado";
*** Histogram ***;
goptions ftext=SWISS ctext=BLACK htext=1 cells;
symbol v=SQUARE c=BLUE h=1 cells;
proc univariate data=work.WS   noprint;
   var WSp;
   histogram / caxes=BLACK cframe=CXF7E1C2 waxis= 1
               cbarline=BLACK cfill=CX000080 pfill=SOLID
               vscale=percent hminor=0 vminor=0
               name='HIST'
   ;
run;
symbol;
goptions ftext= ctext= htext=;

data L2.WS;
```

```
    set WS;
run;
proc printto log=LOG;
run;
proc printto print=PRINT;
run;
```

2.2. Método da amostra expandida em amostragem sem reposição

```
* Redirecting and saving the log to the file named "WSpearman.LOG";
*proc printto log="desktop\WSpearman.LOG";

/*proc printto log="c:\vigo\lb\WSpearman.LOG";
run;
*/
proc printto log="C:\Documents and Settings\Administrador\Meus documentos\Lidiane\
Mestrado\Dissertação\Simulacoes\Pop normal\Sem Reposição\Multiplica
banco\WSpearman.LOG";
run;
* Redirecting and saving the output to the file named "WSpearman.LST";

/*proc printto print="c:\vigo\lb\WSpearman.LST";
run;
*/
proc printto print="C:\Documents and Settings\Administrador\Meus documentos\Lidiane\
Mestrado\Dissertação\Simulacoes\Pop normal\Sem Reposição\Multiplica
banco\WSpearman.LST";
run;


options ps=58 ls=79 nocenter nodate nonumber formchar='|----|+|---+=|-/\<>*';
title1;
title2;
title3;
footnote;


libname L1 v7 'C:\Documents and Settings\Administrador\Meus documentos\Lidiane\Mestrado
\Dissertação\Simulacoes\Pop normal';

libname L2 v7 'C:\Documents and Settings\Administrador\Meus documentos\Lidiane\
Mestrado\Dissertação\Simulacoes\Pop normal\Sem Reposição\Multiplica banco';


data POP;
    set L1.POP;
    rename x1=X x2=Y;
    TAMANHO = 1;
run;

* Criando variavel com valor da correlacao de Spearman na populacao;
ods output spearmancorr=SPpop(keep=X);
proc corr data=pop spearman;
     title1 "Correlacao na Populacao";
    var X Y;
run;
quit;
data SPpop;
    set SPpop;
    if _n_ = 1 then delete;
    rename X=SPpop;
run;
proc print data=SPpop;
run;
title1;

/* Macro p/ selecionar as a amostras*/

%macro sample(rep);
  %do r=1 %to &rep;

    /* Usando o metodo Simple Random Sample (SRS)*/
    /* com 250 em cada estrato*/
```

```
    proc surveyselect data=POP method=SRS n=250
           seed=75&r out=AMOSTRA&r;
           title1 "Amostragem Estratificada Sem Reposicao - Repeticao &r";
           strata STRAT;
    run;

    ****** Salvando as amostras;
    data L2.AMOSTRA&r;
        set AMOSTRA&r;
    run;
    ***** Multiplicando o dataset AMOSTRA de acordo com o peso do estrato;

    data AMOSTRAM&r;
           set AMOSTRA&r;
           if strat=1 then do;
           output; output; output; output; output; output; output; output;
           output; output; output; output; output; output; output; output;
        output; output; output; output; output; output; output; output;
           output; output; output; output; output; output; output; output;
             end;
           else do;
           output; output; output; output; output; output; output; output;
           output; output; output; output; output; output; output; output;
           output; output; output; output; output; output; output; output;
        output; output; output; output; output; output; output; output;
           output; output; output; output; output; output; output; output;
           output; output; output; output; output; output; output; output;
             end;
     run;

     ****** Salvando as amostras multiplicadas pelo peso;
     /* data L2.AMOSTRAM&r;
        set AMOSTRAM&r;
     run; */
     ****Guardando a correlaçao de Spearman com ponderaçao do PROC CORR
            em um dataset chamado spearman;


     ods output spearmancorr=spearman&r(keep=X);
     proc corr data=AMOSTRAM&r spearman;
           title1 "Correlacao na Amostra - Repeticao &r (Dataset Multiplicado)";
           var X Y;
     run;
     quit;
     proc print data=spearman&r;
     run;

    /* Excluindo a linha da correlaçao X com X */
     data spearman&r;
           set spearman&r;
           if _n_ = 1 then delete;
           rename X=WSp;
     run;

***** Cria data set temporário (amost)com a 1ª obs. de cada amostra
     para verificar se as amostras são as mesmas p/ o método dos ranks;
data amost&r;
   set amostra&r;
   if _n_>1 then delete;
run;

  %end;
%mend sample;

%sample (10000)


%macro joincoef(rep);
    proc datasets lib=work nolist;
           delete WS;
    run;
```

```
    %do r=1 %to &rep;
        **** Une os coeficientes de Spearman calculados para todas as amostras,
                  e os guarda em WS;
        proc append
                  BASE=WS data=spearman&r;
           run;
    %end;

    * Adicionando variavel SPpop no dataset WS;
    title1;
    /*
    title1 "Adicionando variavel SPpop no dataset WS";
    proc print data=WS;
    run;
    */
    data WS;
           set WS;
           one=1;
    run;
    data SPpop;
           set SPpop;
           one=1;
    run;
    data WS (drop=one);
           merge WS SPpop;
           by one;
    run;

       ***** Calculando a estatistica T e o p-value com o n° de graus de liberdade
correto;
       ***** Calculando o EQM e o EQM2;
       ***** Calculando o intervalo de 95% de confiança para cada Spearman estimado;
       ***** A variavel I indica se o verdadeiro parametro esta contido no intervalo;
    data WS;
          set WS;
             tcalc = WSp/sqrt((1-WSp**2)/(498));
             pvalue = (1-probt(tcalc,498))*2;
             ERRO = WSp-SPpop;
             ERROABS = abs(ERRO);
             ERROQ = ERRO**2;
             Li=Wsp-1.96*sqrt((1-WSp**2)/(498));
             Ls=Wsp+1.96*sqrt((1-WSp**2)/(498));
             I=0;
             if Li<SPpop and Ls>SPpop
             then I=1;
             else I=0;
    run;
%mend joincoef;

%joincoef(10000);

******Calcula os quantis 2.5% e 97.5% da variável WSp (estimativas da correlação de
Spearman);
proc univariate data=WS noprint ;
var WSp;
output out=pctls pctlpts=2.5 97.5 pctlpre=WS_
              pctlname=P25 P975;
run;

******Adiciona os quantis 2.5% e 97.5% no data set WS;
data WS;
           set WS;
           one=1;
    run;
    data work.pctls;
           set work.pctls;
           one=1;
    run;
    data WS (drop=one);
           merge WS work.pctls;
```

```
        by one;
    run;



    data WS;
          set WS;
             *****Indica se o IC 95% p/ o coeficiente está entre os quantis 2.5 e 97.5%;
          Ie=0;
             if ws_p25<Li and ws_p975>Ls
             then Ie=1;
             else Ie=0;
             *****Indica se o intervalo baseado nos quantis está contido no IC 95%;
          If=0;
             if ws_p25>Li and ws_p975<Ls
             then If=1;
             else If=0;
             *****Indica se o IC 95% p/ o coeficiente tem o limite superior entre
             os quantis 2.5 e 97.5%;
             I25=0;
             if ws_p25>Li and ws_p25<Ls and ws_p975>Ls
             then I25=1;
             else I25=0;
             *****Indica se o IC 95% p/ o coeficiente tem o limite inferior entre
             os quantis 2.5 e 97.5%;
             I975=0;
             if ws_p25<Li and ws_p975<Ls and ws_p975>Li
             then I975=1;
             else I975=0;

    run;
proc contents data=ws;run;


title1 "Resumo dos resultados da simulacao";
proc means data=WS maxdec=6 n mean std min max;
    var WSp ERRO ERROABS ERROQ SPpop I Ie If I25 I975;
run;
quit;


      title;
      footnote;
title1 "Diagrama de caixas";
title2 "Estimativas do coef. de correlacao de Spearman";
title3 "Metodo do banco multiplicado";

proc sql;
   create view WORK._TMP_0 as
      select *,1 as _dummy_
      from work.WS
      ;
*** Box plot ***;
goptions ftext=SWISS ctext=BLACK htext=1 cells;
symbol1 c=BLACK h=1 cells;
axis1 major=none value=none label=none;
proc boxplot data=WORK._TMP_0 ;
   plot (WSp
            )*_DUMMY_
                              / caxis = BLACK
                                cframe = white
                                ctext = BLACK
                                cboxes = BLACK
                                cboxfill = red
                                idcolor = red
                                idsymbol = SQUARE
                                boxstyle = SKELETAL
                                waxis = 1
                                name = 'BOX'
                                description = "Box Plot of WSp"
```

```
                              npanel = 15
                              haxis = axis1
   ;
run;
symbol1;
goptions ftext= ctext= htext=;
axis1;

   quit;

title;
      footnote;
title1 "Histograma";
title2 "Estimativas do coef. de correlacao de Spearman";
title3 "Metodo do banco multiplicado";
*** Histogram ***;
goptions ftext=SWISS ctext=BLACK htext=1 cells;
symbol v=SQUARE c=BLUE h=1 cells;
proc univariate data=work.WS   noprint;
   var WSp;
   histogram / caxes=BLACK cframe=CXF7E1C2 waxis= 1
               cbarline=BLACK cfill=CX000080 pfill=SOLID
               vscale=percent hminor=0 vminor=0
               name='HIST'
   ;
run;
symbol;
goptions ftext= ctext= htext=;

*****P/ comparar os valores das amosras c/ os valores das amostras do método Ranks;
proc print data= work.amost1 ;
var x y;
run;

data L2.WS;
    set WS;
run;

proc print data=ws; run;

proc printto log=LOG;
run;
proc printto print=PRINT;
run;
```

2.3. Método dos postos em amostragem com reposição

```
* Redirecting and saving the log to the file named "WSpearman.LOG";
proc printto log="C:\Documents and Settings\Administrador\Meus
documentos\Lidiane\Mestrado
\Dissertação\Simulacoes\Pop normal\Com Reposição\Ranks\WSpearman.LOG";
run;
* Redirecting and saving the output to the file named "WSpearman.LST";
proc printto print="C:\Documents and Settings\Administrador\Meus
documentos\Lidiane\Mestrado
\Dissertação\Simulacoes\Pop normal\Com Reposição\Ranks\WSpearman.LST";
run;

options ps=58 ls=79 nocenter nodate nonumber formchar='|----|+|---+=|-/\<>*';
title1;
title2;
title3;
footnote;
libname L1 v7 'C:\Documents and Settings\Administrador\Meus documentos\Lidiane\Mestrado
\Dissertação\Simulacoes\Pop normal';
libname L2 v7 'C:\Documents and Settings\Administrador\Meus documentos\Lidiane\
Mestrado\Dissertação\Simulacoes\Pop normal\Com Reposição\Ranks';

data POP;
    set L1.POP;
    rename x1=X x2=Y;
    TAMANHO = 1;
run;

* Criando variavel com valor da correlacao de Spearman na populacao;
ods output spearmancorr=SPpop(keep=X);
proc corr data=pop spearman;
     title1 "Correlacao na Populacao";
    var X Y;
run;
quit;
data SPpop;
    set SPpop;
    if _n_ = 1 then delete;
    rename X=SPpop;
run;
proc print data=SPpop;
run;
title1;

/* Macro p/ selecionar as a amostras*/

%macro sample(rep);
  %do r=1 %to &rep;

    /* Usando o metodo Probability Proportional to Size with Replacement (PPS_WR)*/
    /* com 250 em cada estrato*/

    proc surveyselect data=POP method=PPS_WR n=250
           seed=195&r out=AMOSTRA&r outhits;
           title1 "Amostragem Estratificada Com Reposicao - Repeticao &r";
        size TAMANHO;  /* artificio p/ que cada estrato tenha o mesmo n */
           strata STRAT;
    run;

      ****** Salvando as amostras com os ranks;
     data L2.AMOSTRA&r;
         set AMOSTRA&r;
            run;

    ******* Criando os ranks de X e Y;
    proc rank data=AMOSTRA&r out=AMOSTRA&r ties=mean;
      var X;
       ranks RX;
```

```
    run;
    proc rank data=AMOSTRA&r out=AMOSTRA&r ties=mean;
      var Y;
       ranks RY;
    run;

    *** Calcula o coeficiene de Pearson ponderado dos ranks de X e Y;

     ods output pearsoncorr=pearson&r(keep=RX);
     proc corr data=AMOSTRA&r pearson;
           title1 "Correlacao na Amostra - Repeticao &r ";
           var RX RY;
           weight samplingweight;
     run;
     quit;
     proc print data=pearson&r;
     run;

    /* Excluindo a linha da correlaçao RX com RX */
     data pearson&r;
           set pearson&r;
           if _n_ = 1 then delete;
           rename RX=WSp;
     run;

  %end;
%mend sample;

%sample (10000)

%macro joincoef(rep);
    proc datasets lib=work nolist;
           delete WS;
    run;

    %do r=1 %to &rep;
        **** Une os coeficientes de Spearman calculados para todas as amostras,
                  e os guarda em WS;
        proc append
                  BASE=WS data=pearson&r;
           run;
    %end;

    * Adicionando variavel SPpop no dataset WS;
    title1;
    /*
    title1 "Adicionando variavel SPpop no dataset WS";
    proc print data=WS;
    run;
    */
    data WS;
           set WS;
           one=1;
    run;
    data SPpop;
           set SPpop;
           one=1;
    run;
    data WS (drop=one);
           merge WS SPpop;
           by one;
    run;
       ***** Calculando a estatistica T e o p-value com o n° de graus de liberdade
correto;
       ***** Calculando o ERRO, EQABS e o EQM2;
       ***** Calculando o intervalo de 95% de confiança para cada Spearman estimado;
       ***** A variavel I indica se o verdadeiro parametro esta contido no intervalo;
    data WS;
          set WS;
             tcalc = WSp/sqrt((1-WSp**2)/(498));
             pvalue = (1-probt(tcalc,498))*2;
```

```
            ERRO = WSp-SPpop;
            ERROABS = abs(ERRO);
            ERROQ = ERRO**2;
            Li=Wsp-1.96*sqrt((1-WSp**2)/(498));
            Ls=Wsp+1.96*sqrt((1-WSp**2)/(498));
            I=0;
            if Li<SPpop and Ls>SPpop
            then I=1;
            else I=0;
    run;
%mend joincoef;

%joincoef(10000);

******Calcula os quantis 2.5% e 97.5% da variável WSp (estimativas da correlação de
Spearman);
proc univariate data=WS noprint ;
var WSp;
output out=pctls pctlpts=2.5 97.5 pctlpre=WS_
             pctlname=P25 P975;
run;

******Adiciona os quantis 2.5% e 97.5% no data set WS;
data WS;
           set WS;
           one=1;
    run;
    data work.pctls;
           set work.pctls;
           one=1;
    run;
    data WS (drop=one);
           merge WS work.pctls;
           by one;
    run;

data WS;
          set WS;
             *****Indica se o IC 95% p/ o coeficiente está entre os quantis 2.5 e 97.5%;
           Ie=0;
             if ws_p25<Li and ws_p975>Ls
             then Ie=1;
             else Ie=0;
             *****Indica se o intervalo baseado nos quantis está contido no IC 95%;
           If=0;
             if ws_p25>Li and ws_p975<Ls
             then If=1;
             else If=0;
             *****Indica se o IC 95% p/ o coeficiente tem o limite superior entre
             os quantis 2.5 e 97.5%;
             I25=0;
             if ws_p25>Li and ws_p25<Ls and ws_p975>Ls
             then I25=1;
             else I25=0;
             *****Indica se o IC 95% p/ o coeficiente tem o limite inferior entre
             os quantis 2.5 e 97.5%;
             I975=0;
             if ws_p25<Li and ws_p975<Ls and ws_p975>Li
             then I975=1;
             else I975=0;

    run;
proc contents data=ws;run;


title1 "Resumo dos resultados da simulacao";
proc means data=WS maxdec=6 n mean std min max;
    var WSp ERRO ERROABS ERROQ SPpop I Ie If I25 I975;
run;
quit;
```

```
      title;
      footnote;
title1 "Diagrama de caixas";
title2 "Estimativas do coef. de correlacao de Spearman";
title3 "Metodo dos ranks";

proc sql;
   create view WORK._TMP_0 as
      select *,1 as _dummy_
      from work.WS
      ;
*** Box plot ***;
goptions ftext=SWISS ctext=BLACK htext=1 cells;
symbol1 c=BLACK h=1 cells;
axis1 major=none value=none label=none;
proc boxplot data=WORK._TMP_0 ;
   plot (WSp
            )*_DUMMY_
                              / caxis = BLACK
                                cframe = white
                                ctext = BLACK
                                cboxes = BLACK
                                cboxfill = red
                                idcolor = red
                                idsymbol = SQUARE
                                boxstyle = SKELETAL
                                waxis = 1
                                name = 'BOX'
                                description = "Box Plot of WSp"
                                npanel = 15
                                haxis = axis1
   ;
run;
symbol1;
goptions ftext= ctext= htext=;
axis1;
   quit;

title;
      footnote;
title1 "Histograma";
title2 "Estimativas do coef. de correlacao de Spearman";
title3 "Metodo dos ranks";
*** Histogram ***;
goptions ftext=SWISS ctext=BLACK htext=1 cells;
symbol v=SQUARE c=BLUE h=1 cells;
proc univariate data=work.WS   noprint;
   var WSp;
   histogram / caxes=BLACK cframe=CXF7E1C2 waxis= 1
               cbarline=BLACK cfill=CX000080 pfill=SOLID
               vscale=percent hminor=0 vminor=0
               name='HIST'
   ;
run;
symbol;
goptions ftext= ctext= htext=;
data L2.WS;
    set WS;
run;
proc printto log=LOG;
run;
proc printto print=PRINT;
run;
```

2.4. Método dos postos em amostragem sem reposição

```
* Redirecting and saving the log to the file named "WSpearman.LOG";
*proc printto log="desktop\WSpearman.LOG";

/*proc printto log="c:\vigo\lb\WSpearman.LOG";
run;
*/
proc printto log="C:\Documents and Settings\Administrador\Meus
documentos\Lidiane\Mestrado
\Dissertação\Simulacoes\Pop normal\Sem Reposição\Ranks\WSpearman.LOG";
run;
* Redirecting and saving the output to the file named "WSpearman.LST";

/*proc printto print="c:\vigo\lb\WSpearman.LST";
run;
*/
proc printto print="C:\Documents and Settings\Administrador\Meus
documentos\Lidiane\Mestrado
\Dissertação\Simulacoes\Pop normal\Sem Reposição\Ranks\WSpearman.LST";
run;


options ps=58 ls=79 nocenter nodate nonumber formchar='|----|+|---+=|-/\<>*';
title1;
title2;
title3;
footnote;

libname L1 v7 'C:\Documents and Settings\Administrador\Meus documentos\Lidiane\Mestrado
\Dissertação\Simulacoes\Pop normal';


libname L2 v7 'C:\Documents and Settings\Administrador\Meus documentos\Lidiane\
Mestrado\Dissertação\Simulacoes\Pop normal\Sem Reposição\Ranks';

data POP;
    set L1.POP;
    rename x1=X x2=Y;
    TAMANHO = 1;
run;

* Criando variavel com valor da correlacao de Spearman na populacao;
ods output spearmancorr=SPpop(keep=X);
proc corr data=pop spearman;
     title1 "Correlacao na Populacao";
    var X Y;
run;
quit;
data SPpop;
    set SPpop;
    if _n_ = 1 then delete;
    rename X=SPpop;
run;
proc print data=SPpop;
run;
title1;

/* Macro p/ selecionar as a amostras*/

%macro sample(rep);
  %do r=1 %to &rep;

    /* Usando o metodo Simple Random Sample (SRS)*/
    /* com 250 em cada estrato*/

    proc surveyselect data=POP method=SRS n=250
           seed=75&r out=AMOSTRA&r;
           title1 "Amostragem Estratificada Sem Reposicao - Repeticao &r";
```

```
          strata STRAT;
    run;
      ****** Salvando as amostras com os ranks;
    data L2.AMOSTRA&r;
        set AMOSTRA&r;
          run;

    ******* Criando os ranks de X e Y;
    proc rank data=AMOSTRA&r out=AMOSTRA&r ties=mean;
      var X;
      ranks RX;
    run;
    proc rank data=AMOSTRA&r out=AMOSTRA&r ties=mean;
      var Y;
      ranks RY;
    run;


    *** Calcula o coeficiene de Pearson ponderado dos ranks de X e Y;

     ods output pearsoncorr=pearson&r(keep=RX);
     proc corr data=AMOSTRA&r pearson;
           title1 "Correlacao na Amostra - Repeticao &r ";
           var RX RY;
           weight samplingweight;
     run;
     quit;
     proc print data=pearson&r;
     run;

    /* Excluindo a linha da correlaçao RX com RX */
     data pearson&r;
           set pearson&r;
           if _n_ = 1 then delete;
           rename RX=WSp;
     run;

***** Cria data set temporário (amost)com a 1ª obs. de cada amostra
     para verificar se as amostras são as mesmas p/ o método Banco Multiplicado;
data amost&r;
   set amostra&r;
   if _n_>1 then delete;
run;

  %end;
%mend sample;

%sample (10)


%macro joincoef(rep);
    proc datasets lib=work nolist;
          delete WS;
    run;

    %do r=1 %to &rep;
        **** Une os coeficientes de Spearman calculados para todas as amostras,
                  e os guarda em WS;
        proc append
                  BASE=WS data=pearson&r;
          run;
    %end;

    * Adicionando variavel SPpop no dataset WS;
    title1;
    /*
    title1 "Adicionando variavel SPpop no dataset WS";
    proc print data=WS;
    run;
    */
    data WS;
```

```
            set WS;
            one=1;
    run;
    data SPpop;
            set SPpop;
            one=1;
    run;
    data WS (drop=one);
            merge WS SPpop;
            by one;
    run;
       ***** Calculando a estatistica T e o p-value com o n° de graus de liberdade
correto;
       ***** Calculando o ERRO, EQABS e o EQM2;
       ***** Calculando o intervalo de 95% de confiança para cada Spearman estimado;
       ***** A variavel I indica se o verdadeiro parametro esta contido no intervalo;
    data WS;
          set WS;
              tcalc = WSp/sqrt((1-WSp**2)/(498));
              pvalue = (1-probt(tcalc,498))*2;
              ERRO = WSp-SPpop;
              ERROABS = abs(ERRO);
              ERROQ = ERRO**2;
              Li=Wsp-1.96*sqrt((1-WSp**2)/(498));
              Ls=Wsp+1.96*sqrt((1-WSp**2)/(498));
              I=0;
              if Li<SPpop and Ls>SPpop
              then I=1;
              else I=0;
    run;
%mend joincoef;

%joincoef(10);

******Calcula os quantis 2.5% e 97.5% da variável WSp (estimativas da correlação de
Spearman);
proc univariate data=WS noprint ;
var WSp;
output out=pctls pctlpts=2.5 97.5 pctlpre=WS_
             pctlname=P25 P975;
run;

******Adiciona os quantis 2.5% e 97.5% no data set WS;
data WS;
            set WS;
            one=1;
    run;
    data work.pctls;
            set work.pctls;
            one=1;
    run;
    data WS (drop=one);
            merge WS work.pctls;
            by one;
    run;


data WS;
          set WS;
             *****Indica se o IC 95% p/ o coeficiente está entre os quantis 2.5 e 97.5%;
           Ie=0;
             if ws_p25<Li and ws_p975>Ls
             then Ie=1;
             else Ie=0;
             *****Indica se o intervalo baseado nos quantis está contido no IC 95%;
           If=0;
             if ws_p25>Li and ws_p975<Ls
             then If=1;
             else If=0;
             *****Indica se o IC 95% p/ o coeficiente tem o limite superior entre
             os quantis 2.5 e 97.5%;
```

```
            I25=0;
            if ws_p25>Li and ws_p25<Ls and ws_p975>Ls
            then I25=1;
            else I25=0;
            *****Indica se o IC 95% p/ o coeficiente tem o limite inferior entre
            os quantis 2.5 e 97.5%;
            I975=0;
            if ws_p25<Li and ws_p975<Ls and ws_p975>Li
            then I975=1;
            else I975=0;

    run;
proc contents data=ws;run;


title1 "Resumo dos resultados da simulacao";
proc means data=WS maxdec=6 n mean std min max;
    var WSp ERRO ERROABS ERROQ SPpop I Ie If I25 I975;
run;
quit;


      title;
      footnote;
title1 "Diagrama de caixas";
title2 "Estimativas do coef. de correlacao de Spearman";
title3 "Metodo dos ranks";

proc sql;
   create view WORK._TMP_0 as
      select *,1 as _dummy_
      from work.WS
      ;
*** Box plot ***;
goptions ftext=SWISS ctext=BLACK htext=1 cells;
symbol1 c=BLACK h=1 cells;
axis1 major=none value=none label=none;
proc boxplot data=WORK._TMP_0 ;
   plot (WSp
            )*_DUMMY_
                           / caxis = BLACK
                             cframe = white
                             ctext = BLACK
                             cboxes = BLACK
                             cboxfill = red
                             idcolor = red
                             idsymbol = SQUARE
                             boxstyle = SKELETAL
                             waxis = 1
                             name = 'BOX'
                             description = "Box Plot of WSp"
                             npanel = 15
                             haxis = axis1
   ;
run;
symbol1;
goptions ftext= ctext= htext=;
axis1;

   quit;

title;
      footnote;
title1 "Histograma";
title2 "Estimativas do coef. de correlacao de Spearman";
title3 "Metodo dos ranks";
*** Histogram ***;
goptions ftext=SWISS ctext=BLACK htext=1 cells;
symbol v=SQUARE c=BLUE h=1 cells;
proc univariate data=work.WS   noprint;
   var WSp;
```

```
   histogram / caxes=BLACK cframe=CXF7E1C2 waxis= 1
               cbarline=BLACK cfill=CX000080 pfill=SOLID
               vscale=percent hminor=0 vminor=0
               name='HIST'
   ;
run;
symbol;
goptions ftext= ctext= htext=;

*****P/ comparar os valores das amosras c/ os valores das amostras do método Banco
Multiplicado;
proc print data= work.amost1 ;
var x y;
run;

data L2.WS;
    set WS;
run;

proc print data=ws; run;

proc printto log=LOG;
run;
proc printto print=PRINT;
run;
```

3. Intervalos de Confiança Bootstrap para o coeficiente de Pearson ponderado

```
* Redirecting and saving the log file";
proc printto log="c:\vigo\ppgepi\lidiane\simul\boot_person rep1000.LOG";
run;

* Redirecting and saving the output file";
proc printto print="c:\vigo\ppgepi\lidiane\simul\boot_person rep1000.LST";
run;

options ps=58 ls=200 nocenter nodate nonumber formchar='|----|+|---+=|-/\<>*';
title1;
title2;
title3;
footnote;

libname L1 v7 'c:\vigo\ARIC\976';
libname L2 v7 'c:\vigo\ppgepi\lidiane\simul';

data SET1;
    set L1.uc414801 (where=(crs=1));
    keep ID RACE_AA CRS CRSWT ADIPO IL6 ;
run;
/*
proc contents;
run;
*/
proc sort;
    by crswt;
run;
/*
proc print;
    var id race_aa crs crswt;
run;
*/

*** Calcula o coeficiente de Pearson ponderado entre IL6 e ADIPO;
ods output Corr.PearsonCorr=WP_OBS(keep=ADIPO rename=(ADIPO=WP_OBS));
proc corr data=SET1 pearson;
    title1 "Correlacao Observada na Amostra";
    var IL6 ADIPO;
    weight CRSWT;
run;
quit;

/* Excluindo a linha da correlaçao IL6 com IL6*/
data WP_OBS;
    set WP_OBS;
    if _n_ > 1 then delete;
run;
proc print data=WP_OBS;
run;

/*
proc freq data=SET1;
    table RACE_AA;
run;
*/
data BLACK;
    set SET1;
    if RACE_AA = 1;
    *drop IL6 ADIPO;
run;
data WHITE;
    set SET1;
    if RACE_AA = 0;
    *drop IL6 ADIPO;
run;
```

```
* macro para gerar amostras bootstrap;
%macro sample(rep);
  *ods trace on;

  %do r=1 %to &rep;

    /* Usando o metodo Probability Proportional to Size with Replacement (PPS_WR)*/

    proc surveyselect data=BLACK method=URS n=314 seed=195&r out=BLACK&r outhits;
        title1 "Amostragem Estratificada Com Reposicao - Repeticao &r";
    run;
    proc surveyselect data=WHITE method=URS n=354 seed=8888&r out=WHITE&r outhits;
           title1 "Amostragem Estratificada Com Reposicao - Repeticao &r";
    run;

    data L2.BLACK&r;
        set BLACK&r;
    run;
    proc sort data=BLACK&r;
           by ID;
    run;
    /*
    proc print data=BLACK&r;
    run;
    proc freq data=BLACK&r;
           table ID;
    run;
    */
    data L2.WHITE&r;
        set WHITE&r;
    run;
    proc sort data=WHITE&r;
           by ID;
    run;
    /*
    proc print data=WHITE&r;
    run;
    proc freq data=WHITE&r;
           table ID;
    run;
    */
    data AMOSTRA&r;
           merge work.BLACK&r work.WHITE&r;
           by ID;
    run;
    /*
    proc print data=AMOSTRA&r;
    run;
    proc freq data=AMOSTRA&r;
           table ID;
    run;
    */

    *** Calcula o coeficiente de Pearson ponderado entre IL6 e ADIPO;
    ods output Corr.PearsonCorr=PEARSON&r(keep=ADIPO NADIPO rename=(ADIPO=WPearson
NADIPO=N));
    proc corr data=AMOSTRA&r pearson;
           title1 "Correlacao na Amostra - Repeticao &r ";
           var IL6 ADIPO;
           weight CRSWT;
     run;
     quit;

    /* Excluindo a linha da correlaçao IL6 com IL6 */
     data PEARSON&r;
           set PEARSON&r;
           if _n_ > 1 then delete;
     run;
     proc print data=PEARSON&r;
```

```
     run;
%end;
%mend sample;
%sample (1000)

title1;
%macro joinWP(rep);
    %do r=1 %to &rep;
        proc append
                  BASE=WP data=PEARSON&r;
           run;
    %end;

    data WP;
           set WP;
           one=1;
    run;
    data WP_OBS;
           set WP_OBS;
           one=1;
    run;
    data WP (drop=one);
           merge WP WP_OBS;
           by one;
    run;
%mend joinWP;
%joinWP(1000);


*ods trace on;
ods output Means.Summary=IC(keep=WPearson_mean WPearson_stddev);
proc means data=WP mean std;
    var WPearson;
run;
*ods trace off;
/*
proc print data=IC;
run;
*/

* Juntando WPearson_mean e WPearson_stddev ao dataset WP;
data WP;
    set WP;
     one=1;
run;
data IC;
    set IC;
    one=1;
run;
data WP (drop=one);
    merge WP IC;
    by one;
run;
/*
proc print data=WP;
run;
*/


*** Calculando IC 95%;
data WP;
   set WP;
   * IC 95% padrão, isto é, assumindo que WPearson é normalidade - ver Manly
(2004,p.39);

   WPearson95_inf = WPearson_mean - 1.96*WPearson_stddev;
   WPearson95_sup = WPearson_mean + 1.96*WPearson_stddev;

   * IC 95% padrão, com correção do viés - ver Manly (2004,p.41);
   VIES = WPearson_mean - WP_OBS;
   WPearson95V_inf = (WP_OBS - VIES) - 1.96*WPearson_stddev;
```

```
   WPearson95V_sup = (WP_OBS - VIES) + 1.96*WPearson_stddev;

   * IC 95% baseado nos percentis 97.5% e 2.5% a direita da distribuição de WPearson -
ver Manly (2004,p.50);
   if WPearson > .Z
   then if WPearson >= WP_OBS
           then WPearson_DIC = 1;
           else WPearson_DIC = 0;
   else WPearson_DIC = .V;
run;

proc print data=WP(obs=1);
    title1 "Estimativa pontual, desvio padrão e IC95% pelo Bootstrap padrão";
    var Wpearson_mean Wpearson_stddev WPearson95_inf WPearson95_sup;
run;
proc print data=WP(obs=1);
    title1 "Correlacao observada, VIES e IC95% pelo Bootstrap padrão com correção do
vies";
    var WP_OBS VIES WPearson95V_inf WPearson95V_sup;
run;
title1;


*ods trace on;
ods output Freq.Table1.OneWayFreqs=Z0(keep=percent rename=(percent=P0));
proc freq data=WP;
    table WPearson_DIC;
run;
*ods trace off;


data Z0;
    set Z0;
    if _n_ = 1 then delete;
    P0 = P0/100;
    ONE = 1;
run;
/*
proc print data=Z0;
run;
*/
* Juntando datasets WP e Z0;
data WP;
    set WP;
    ONE = 1;
run;

data WP (drop=ONE);
    merge WP Z0;
    by ONE;
run;
/*
proc print data=WP;
run;
*/


*** Calculando IC 95% ;
data WP;
   set WP;

   Z0 = probit(1-P0);
   *INV_Z0 = 1-probnorm(0.25335);
   PU = probnorm(2*Z0 + 1.96);
   PL = probnorm(2*Z0 - 1.96);
run;
data PL_PU;
    set WP;
    ONE = 1;
    keep Z0 PL PU ONE;
    if _n_ > 1 then delete;
```

```
run;
/*
proc print data=PL_PU;
run;
*/

*ods trace on;
ods output Freq.Table1.OneWayFreqs=WPearson_DIST(keep=F_WPearson CumPercent
                                rename=(F_WPearson=Valor_WPearson
CumPercent=WP_CumPercent));
*ods output Freq.Table1.OneWayFreqs;
proc freq data=WP;
    table WPearson;
run;
*ods trace off;

/*
proc print data=WPearson_DIST;
run;
*/
data WPearson_DIST;
    set WPearson_DIST;
    WP_CumPercent = WP_CumPercent / 100;
    ONE = 1;
run;
/*
proc print data=WPearson_DIST;
run;
*/

* Juntando datasets PL_PU e WPearson_DIST;
data WPearson_DIST (drop=ONE);
    merge WPearson_DIST PL_PU;
    by ONE;
run;
proc print data=WPearson_DIST;
run;


data IC_PERC;
    set WPearson_DIST;
    WPearson95perc_INF = .V;
    WPearson95perc_SUP = .V;
    PL_INT = INT(PL*1000);
    WP_CP_INT =INT(WP_CumPercent*1000);
    if WP_CP_INT = PL_INT
     then WPearson95perc_INF = Valor_WPearson;
    PU_INT = INT(PU*1000);
    if WP_CP_INT = PU_INT
     then WPearson95perc_SUP = Valor_WPearson;
run;
proc print data=IC_PERC;
    var Valor_WPearson WP_CumPercent PL PL_INT WP_CP_INT WPearson95perc_INF
                                     PU PU_INT WP_CP_INT WPearson95perc_SUP;
run;
/*
proc print data=IC_PERC;
    var Valor_WPearson WP_CumPercent PL PL_INT WP_CP_INT WPearson95perc_INF
                                     PU PU_INT WP_CP_INT WPearson95perc_SUP;
run;
*/

data IC_PERC95;
    set IC_PERC;
    if WPearson95perc_INF <= .Z & WPearson95perc_SUP <= .Z then delete;
run;
proc print data=IC_PERC95;
    var Valor_WPearson WP_CumPercent PL PL_INT WPearson95perc_INF PU PU_INT
WPearson95perc_SUP;
run;
```

```
data IC_LINF;
    set IC_PERC95;
    ONE = 1;
    if _n_ > 1 then delete;
    keep ONE WPearson95perc_INF;
run;
/*
proc print data=IC_LINF;
run;
*/
data IC_LSUP;
    set IC_PERC95;
    ONE = 1;
    if _n_ < 2 then delete;
    keep ONE WPearson95perc_SUP;
run;
/*
proc print data=IC_LSUP;
run;
*/


* Juntando datasets IC_LINF e IN_SUP;
data IC_PERCENTIL95 (drop=ONE);
    merge IC_LINF IC_LSUP;
    by ONE;
run;
proc print data=IC_PERCENTIL95;
    title1 "IC para Pearson ponderado baseado no percentil 95% via Bootstrap";
run;
title1;

* Juntando datasets IC_PERCENTIL95 com WP;
data IC_PERCENTIL95;
    set IC_PERCENTIL95;
    ONE = 1;
run;
data WP;
    set WP;
    ONE = 1;
run;

data WP (drop=ONE);
    merge WP IC_PERCENTIL95;
    by ONE;
run;
/*
proc print data=WP;
    var WP_OBS WPearson WPearson95perc_INF WPearson95perc_SUP;
run;
title1;
*/

data L2.WP;
     set WP;
run;

proc printto log=LOG;
run;
proc printto print=PRINT;
run;


      title;
      footnote;
*** Histogram ***;
goptions ftext=SWISS ctext=BLACK htext=1 cells;
symbol v=SQUARE c=BLUE h=1 cells;
proc univariate data=WP   noprint;
   var WPearson;
   histogram / caxes=BLACK cframe=WHITE waxis= 1
```

```
              cbarline=BLACK cfill=BLUE pfill=SOLID
              vscale=percent hminor=0 vminor=0
              name='HIST'
              normal( mu=est sigma=est w=1 color=RED
                      noprint )
   ;
   inset normal ;
run;
symbol;
goptions ftext= ctext= htext=;
```