COEFICIENTES DE CORRELAÇÃO PARA VARIÁVEIS ORDINAIS E DICOTÔMICAS DERIVADOS DO COEFICIENTE LINEAR DE PEARSON

CORRELATION COEFFICIENT DERIVED FROM PEARSON
LINEAR COEFFICIENT FOR ORDINAL AND DICHOTOMIC VARIABLES

Sachiko Araki Lira

Instituto Paranaense de Desenvolvimento Econômico e Social - IPARDES
Rua Máximo João Kopp, 274 - Bloco 2 - Santa Cândida
CEP: 82630-900 - Curitiba - Paraná
sachiko@onda.com.br

Anselmo Chaves Neto

Universidade Federal do Paraná
Programa de Pós-Graduação em Métodos Numéricos em Engenharia - Centro Politécnico
CEP: 81531-990 - Curitiba - Paraná
anselmo@est.ufpr.br

RESUMO

Devido a grande utilização da Análise de Correlação em diferentes áreas do conhecimento, é importante conhecer os diferentes métodos para a obtenção dos coeficientes de correlação. É comum as situações em que as variáveis não são medidas em nível intervalar, mas em nível ordinal e/ou dicotômica. Apresentam-se no presente trabalho os métodos de coeficientes de correlação derivados do coeficiente linear de Pearson, para situações que envolvem variáveis medidas em nível intervalar, ordinal e dicotômica, quais sejam: coeficiente de correlação ponto bisserial, phi, de Spearman, entre variáveis intervalar e ordinal e entre variáveis ordinal e dicotômica.
Palavras-chave: coeficiente de correlação; coeficiente de correlação para variáveis ordinais e dicotômicas.

ABSTRACT

As the Correlation Analysis is used in several knowledge areas, it is important to be familiar with the different methods existing to obtain correlation coefficients. Usually, we can find situations where variables are not measured by interval level but according to ordinal and/or dichotomic. The present work shows the correlation coefficient methods derived from Pearson linear coefficient and applied to situations involving variables measured by interval, ordinal and dichotomic levels, such as: bi-serial point, phi and Spearman correlation coefficient, between interval and ordinal variables, and between ordinal e dichotomic variables.
Keywords: correlation coefficient; correlation coefficient for ordinal and dichotomic variables

1 - INTRODUÇÃO

A Análise de Correlação é uma ferramenta importante para as diferentes áreas do conhecimento, não somente como resultado final, mas como uma das etapas para a utilização de outras técnicas de análise.

A análise de confiabilidade em sistemas de engenharia tem como objetivo avaliar a probabilidade de não haver falha durante a sua vida útil, atendendo aos objetivos para os quais o sistema foi projetado.

A avaliação da probabilidade de falha é usualmente identificada como a análise de confiabilidade estrutural. Dois métodos analíticos bastante utilizados são: First Order Reliability Method (FORM) e Second Order Reliability Method (SORM). Segundo Haldar e Mahadevan [6], os métodos FORM e SORM assumem implicitamente que as variáveis envolvidas na análise são não correlacionadas. Portanto, deve-se, obter inicialmente as correlações entre as variáveis.

O método usualmente conhecido para medir a correlação entre duas variáveis é o coeficiente de correlação linear de Pearson, também conhecido como coeficiente de correlação do momento produto. Este foi o primeiro método de correlação, introduzido por Karl Pearson em 1897, conforme apresentado em Lira [7].

Uma das suposições para a utilização deste coeficiente é de que as variáveis envolvidas na análise sejam medidas no mínimo em nível intervalar. Entretanto, em muitas situações, não é possível a utilização desse tipo de escala de medida. Foram então desenvolvidos os coeficientes de correlação derivados do coeficiente linear de Pearson para situações que envolvem variáveis medidas em nível ordinal e dicotômica.

2 - OBJETIVO

O objetivo deste trabalho é apresentar uma revisão sobre os diferentes métodos de correlação derivados do coeficiente linear de Pearson, quais sejam: coeficiente de correlação ponto bisserial, coeficiente de correlação phi, coeficiente de correlação de Spearman, coeficiente de correlação entre variável ordinal (rank) e variável intervalar e coeficiente de correlação entre variável ordinal (rank) e variável dicotômica. Estes dois últimos coeficientes são apresentados em Wherry [10].

3 - MÉTODOS PARA A OBTENÇÃO DO COEFICIENTE DE CORRELAÇÃO

3.1. *Coeficiente de correlação linear de Pearson*

O coeficiente de correlação populacional (parâmetro) ρ e sua estimativa amostral $\hat{\rho}$ estão intimamente relacionados com a distribuição normal bivariada, cuja função densidade de probabilidade é dada por:

$$f_{X,Y}(X,Y)=\frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}}\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{X-\mu_X}{\sigma_X}\right)^2-2\rho\left(\frac{X-\mu_X}{\sigma_X}\right)\left(\frac{Y-\mu_Y}{\sigma_Y}\right)+\left(\frac{Y-\mu_Y}{\sigma_Y}\right)^2\right]\right\} \quad (1)$$

$$\text{sendo } \rho_{X,Y}=\rho=\frac{COV(X,Y)}{\sigma_X\sigma_Y}=\frac{\sigma_{X,Y}}{\sigma_X\sigma_Y} \quad (2)$$

o parâmetro populacional

onde:
$COV(X,Y)$ é a covariância entre X e Y
σ_X é o desvio padrão de X
σ_Y é o desvio padrão de Y

O Estimador de Máxima Verossimilhança é dado pela expressão:

$$\hat{\rho}_{X,Y}=\hat{\rho}=\frac{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)\left(Y_i-\overline{Y}\right)}{n\sqrt{\sum_{i=1}^{n}\frac{\left(X_i-\overline{X}\right)^2}{n}}\sqrt{\sum_{i=1}^{n}\frac{\left(Y_i-\overline{Y}\right)^2}{n}}}=\frac{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)\left(Y_i-\overline{Y}\right)}{n\hat{\sigma}_X\hat{\sigma}_Y} \quad (3)$$

onde:
n é o número de observações da amostra
$\overline{X}$ é a média aritmética de X
$\overline{Y}$ é a média aritmética de Y

Este coeficiente é também conhecido como Coeficiente de Correlação do Momento Produto.

Na prática, conforme apresentado em Lira [7], o coeficiente de correlação $\hat{\rho}$ é interpretado como um indicador que descreve a interdependência entre as variáveis X e Y.

Outra forma de interpretar o coeficiente de correlação é em termos de $\hat{\rho}^2=R^2$, denominado coeficiente de determinação ou de explicação. Quando multiplicado por 100, o $\hat{\rho}^2=R^2$ fornece a percentagem da variação em Y (variável dependente), que pode ser explicada pela variação em X (variável independente), ou seja, o quanto de variação é comum às duas variáveis.

Cabe lembrar que o coeficiente de determinação é a relação entre a variação explicada pelo modelo linear ($\hat{Y}=\hat{\alpha}+\hat{\beta}X$, onde $\hat{\alpha}$ e $\hat{\beta}$ são constantes) e a variação total.

A significância do coeficiente de correlação estimado é verificada através de teste de hipóteses. A estatística para testar a hipótese $H_0:\rho=0$ contra $H_1:\rho\neq 0$ tem distribuição t com n - 2 graus de liberdade, ou seja:

$$t=\frac{\hat{\rho}\sqrt{n-2}}{\sqrt{1-\hat{\rho}^2}}\sim t_{n-2} \quad (4)$$

onde:
n é o número de observações da amostra
$\hat{\rho}$ é o coeficiente de correlação linear de Pearson

3.2. *Coeficientes de correlação derivados do coeficiente linear de Pearson*

3.2.1. *Coeficiente de correlação ponto bisserial*

Embora seja usada normalmente como medida de correlação entre escores e itens de testes, a correlação ponto bisserial pode ser empregada em outras situações onde a variável dicotômica pode ser, a título de exemplo, perfeito ou defeituoso, certo ou errado.

O coeficiente de correlação ponto bisserial é derivado do coeficiente de correlação linear de Pearson. Esse método é indicado quando uma das variáveis (Y) é dicotômica e a outra (X), contínua.

Conforme apresentado em Ferguson [4], a correlação ponto bisserial fornece uma medida da relação entre uma variável contínua, e outra variável com duas categorias ou dicotômicas, como perfeito ou defeituoso.

Segundo Ferguson [4], Downie e Heath [3] e Guilford [5], a correlação ponto bisserial é a correlação do momento produto. Se for atribuído 1 para observações de uma categoria e zero para outra, e se for calculado o coeficiente de correlação do momento produto, o resultado será o coeficiente ponto bisserial. Ele é interpretado da mesma forma que $\hat{\rho}$.

O estimador do coeficiente de correlação ponto bisserial foi obtido a partir do estimador do coeficiente de correlação linear de Pearson, conforme apresentado em Guilford [5].

Fazendo $x_i=X_i-\overline{X}$ e $y_i=Y_i-\overline{Y}$, o estimador do coeficiente linear de Pearson é:

$$\hat{\rho}=\frac{\sum_{i=1}^{n}x_iy_i}{\sqrt{\sum_{i=1}^{n}x_i^2\sum_{i=1}^{n}y_i^2}}=\frac{\sum_{i=1}^{n}x_iy_i}{n\sqrt{\frac{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}{n}\frac{\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2}{n}}}=\frac{\sum_{i=1}^{n}x_iy_i}{n\hat{\sigma}_x\hat{\sigma}_y} \quad (5)$$

onde:
n é o número de observações da amostra

$\hat{\sigma}_X$ é o desvio padrão amostral de X

$\hat{\sigma}_Y$ é o desvio padrão amostral de Y

Sendo X uma variável aleatória contínua e Y uma variável aleatória com distribuição de Bernoulli, tem-se, então, que, por conveniência:

$$S_x = \hat{\sigma}_x = \sqrt{\frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{n}} \tag{6}$$

$S_y = \hat{\sigma}_Y = \sqrt{pq}$, onde p = θ e q = (1 - θ) da distribuição de Bernoulli. (7)

Desenvolvendo (5), tem-se:

$$\sum_{i=1}^{n} x_i y_i = \sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right) \tag{8}$$

$$\sum_{i=1}^{n} x_i y_i = \sum_{i=1}^{n}\left[X_i Y_i - X_i\overline{Y} - \overline{X}Y_i + \overline{X}\,\overline{Y}\right]$$

$$\sum_{i=1}^{n} x_i y_i = \sum_{i=1}^{n} X_i Y_i - n\overline{X}\,\overline{Y} \tag{9}$$

Substituindo (9) em (5), tem-se:

$$\hat{\rho} = \frac{\sum_{i=1}^{n} X_i Y_i - n\overline{X}\,\overline{Y}}{nS_x\sqrt{pq}} \quad \text{mas} \quad \sum_{i=1}^{n} X_i Y_i = n_p \times \overline{X}_p$$

e $n\overline{X}\,\overline{Y} = n \times \overline{X} \times p = n_p \times \overline{X}$, então,

$$\hat{\rho} = \frac{n_p \times \overline{X}_p - n_p \times \overline{X}}{nS_x\sqrt{pq}} \tag{10}$$

Dividindo por n, tem-se:

$$\hat{\rho} = \frac{p \times \overline{X}_p - p \times \overline{X}}{S_x\sqrt{pq}} = \frac{\left(\overline{X}_p - \overline{X}\right) \times p}{S_x\sqrt{pq}}$$

Dividindo por $\sqrt{p}$, tem-se:

$$\hat{\rho}_{pb} = \frac{\left(\overline{X}_p - \overline{X}\right)}{S_x}\sqrt{\frac{p}{q}} \tag{11}$$

onde:

$\hat{\rho}_{pb}$ é o coeficiente de correlação ponto bisserial

$\overline{X}_p$ é a média dos valores de X para o grupo superior (grupo cuja variável Y assume valor 1)

$\overline{X}$ é a média total de X da amostra

S_x é o desvio padrão total de X da amostra

p é a proporção de casos do grupo superior (grupo cuja variável Y assume valor 1)

q é a proporção de casos do grupo inferior (grupo cuja variável Y assume valor 1)

De acordo com Wherry [10], a significância do coeficiente estimado é testada pela estatística:

$$t = \frac{\hat{\rho}_{pb}\sqrt{n-2}}{\sqrt{1-\hat{\rho}_{pb}^2}} \sim t_{n-2} \tag{12}$$

onde:

n é o número de observações da amostra

$\hat{\rho}_{pb}$ é o coeficiente de correlação ponto bisserial

3.2.2. *Coeficiente de correlação de Spearman*

Este coeficiente é o mais antigo e também o mais conhecido para calcular o coeficiente de correlação entre variáveis mensuradas em nível ordinal, chamado também de coeficiente de correlação por postos de Spearman, designado "rho" e representado por $\hat{\rho}_s$.

É importante enfatizar, segundo Bunchaft e Kellner [2], que as correlações ordinais não podem ser interpretadas da mesma maneira que para variáveis medidas em nível intervalar. Inicialmente, não mostram necessariamente tendência linear, mas podem ser consideradas como índices de monotonicidade, ou seja, para aumentos positivos da correlação, aumentos no valor de X correspondem a aumentos no valor de Y, e para coeficientes negativos ocorre o oposto. O quadrado do coeficiente de correlação não pode ser interpretado como a proporção da variância comum às duas variáveis.

Seu estimador foi derivado a partir do estimador do coeficiente de correlação linear de Pearson, conforme apresentado em Siegel [8].

Tem-se da expressão (5) que:

$$\hat{\rho} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2}} \tag{13}$$

onde: $x_i = X_i - \overline{X}$ e $y_i = Y_i - \overline{Y}$

Pode-se escrever: $\sum_{i=1}^{n} X_i = \frac{n(n+1)}{2}$, onde

$n = \text{postos=rank} = 1,\ 2,\ 3,\ldots,\ n$ (14)

Os quadrados dos postos são: $1^2,\ 2^2,\ 3^2,\ldots,\ n^2$

Então: $\sum_{i=1}^{n} X_i^2 = \frac{n(n+1)(2n+1)}{6}$ (15)

Assim, $\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 = \sum_{i=1}^{n} X_i^2 - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}$ (16)

$$\sum_{i=1}^{n} x_i^2 = \frac{n(n+1)(2n+1)}{6} - \frac{\left[n(n+1)/2\right]^2}{n}$$

$$\sum_{i=1}^{n} x_i^2 = \frac{n^3 - n}{12} \tag{17}$$

Da mesma forma, obtém-se:

$$\sum_{i=1}^{n} y_i^2 = \frac{n^3 - n}{12} \tag{18}$$

Fazendo a diferença de postos:

$$d_i = x_i - y_i \tag{19}$$

elevando ao quadrado tem-se:

$$d_i^2 = (x_i - y_i)^2 = x_i^2 - 2x_i y_i + y_i^2 \tag{20}$$

fazendo o somatório:

$$\sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} x_i^2 + \sum_{i=1}^{n} y_i^2 - 2\sum_{i=1}^{n} x_i y_i \tag{21}$$

fazendo $\hat{\rho}_s = \dfrac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2}}$, tem-se que

$$\sum_{i=1}^{n} x_i y_i = \hat{\rho}_s \sqrt{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2} \tag{22}$$

substituindo (17), (18) e (22) em (21) tem-se:

$$\sum_{i=1}^{n} d_i^2 = 2\left(\frac{n^3 - n}{12}\right) - 2\hat{\rho}_s \sqrt{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2}$$

Assim, obtém-se:

$$\hat{\rho}_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2 - 1)} \tag{23}$$

onde:

$\hat{\rho}_s$ é o coeficiente de correlação de Spearman

d_i é a diferença entre as ordenações

n é o número de pares de ordenações

Quando a seleção dos elementos que compõem a amostra é feita de forma aleatória, a partir de uma população, é possível determinar se as variáveis em estudo são associadas na população. Ou seja, é possível testar a hipótese de que as duas variáveis estão associadas na população.

Para amostras superiores a 10, segundo Siegel [8], a significância de um valor obtido de $\hat{\rho}_s$ pode ser verificada através de t calculado pelo estimador apresentado a seguir.

$$t = \hat{\rho}_s \sqrt{\frac{n-2}{1 - \hat{\rho}_s^2}} \tag{24}$$

Para $n \geq 10$, a expressão acima tem distribuição t de Student com n-2 graus de liberdade.

Segundo Silveira [9], a relação entre uma escala intervalar e ordinal é de monotonicidade e a transformação monotônica em uma variável causa pouco efeito sobre coeficientes de correlação, razões t e F. Assim, uma variável medida em nível ordinal pode ser tratada como intervalar.

3.2.3. *Coeficiente de correlação phi*

Em algumas situações, as variáveis são medidas em nível nominal ou por categorias discretas e expressas em forma de freqüências. Nesses casos, não é possível a utilização de nenhum dos métodos vistos anteriormente.

O estimador do coeficiente de correlação phi também foi obtido a partir do estimador do coeficiente linear de Pearson, bastando fazer com que a variável X também seja dicotômica e distribuída conforme apresentada a seguir:

		Variável X		
		1	0	TOTAL
Variável 1		a	b	n_p
Y	0	c	d	n_q
TOTAL		$n_{p'}$	$n_{q'}$	n

Tem-se da expressão (11) que o estimador do coeficiente de correlação ponto bisserial é:

$$\hat{\rho}_{bp} = \frac{\left(\overline{X}_p - \overline{X}\right)}{S_x}\sqrt{\frac{p}{q}} \tag{25}$$

$$\text{mas } \overline{X}_p = \frac{a}{n_p} = \frac{a}{a+b} \text{ e } \overline{X}_q = \frac{c}{n_q} = \frac{c}{c+d} \tag{26}$$

$$p = \frac{(a+b)}{n} \text{ e } q = \frac{(c+d)}{n} \tag{27}$$

$$\overline{X} = p\overline{X}_p + q\overline{X}_q = \frac{(a+b)}{n}\frac{a}{(a+b)} + \frac{(c+d)}{n}\frac{c}{(c+d)} = \frac{(a+c)}{n} \tag{28}$$

$$S_x = \sqrt{n_{p'} n_{q'}} = \sqrt{\frac{(a+c)}{n}\frac{(b+d)}{n}} = \frac{1}{n}\sqrt{(a+c)(b+d)} \tag{29}$$

Então, substituindo as expressões (26), (27), (28) e (29) na (25), tem-se:

$$\hat{\phi} = \frac{\frac{a}{(a+b)} - \frac{(a+c)}{n}}{\frac{1}{n}\sqrt{(a+c)(b+d)}}\frac{\sqrt{(a+b)}}{\sqrt{(a+c)}} = \frac{\frac{na - (a+b)(a+c)}{n(a+b)}}{\frac{1}{n}\sqrt{(a+c)(b+d)}}\frac{\sqrt{(a+b)}}{\sqrt{(a+c)}}$$

$$\hat{\phi} = \frac{na - (a+b)(a+c)}{(a+b)\sqrt{(a+c)(b+d)}}\frac{\sqrt{(a+b)}}{\sqrt{(a+c)}}$$

$$\hat{\phi} = \frac{(ad - bc)}{\sqrt{(a+b)(a+c)(b+d)(c+d)}} \tag{30}$$

onde:

$\hat{\phi}$ é o coeficiente de correlação phi

a,b,c,d são as frequências da tabela de contingência

n é a soma das frequências a,b,c e d

O coeficiente de correlação phi está relacionado com χ^2 para a tabela 2x2, dada pela expressão a seguir, como apresentada em Ferguson [4]:

$$\hat{\phi} = \sqrt{\frac{\chi^2}{n}} \text{ ou } \chi^2 = n\phi^2 \tag{31}$$

Por essa razão, pode-se testar a significância de $\hat{\phi}$ calculando o valor de $\chi^2 = n\phi^2$ e comparando com o valor de χ^2, com 1 grau de liberdade [4].

Os valores de $\hat{\phi}$ variam entre -1 e +1. Entretanto, para Bunchaft e Kellner [2] é suficiente que **a** e **d** indiquem ou concordância ou discordância, o mesmo acontecendo com **b** e **c**.

3.2.4. *Coeficiente de correlação entre variáveis dicotômica e ordinal (rank)*

Este coeficiente é utilizado, segundo Wherry [10], quando uma das variáveis (X) é dicotômica e a outra ordinal (rank). O seu estimador também foi obtido a partir do coeficiente de correlação linear de Pearson.

O estimador do coeficiente de correlação linear de Pearson dado pela expressão (3) pode ser reescrito como:

$$\hat{\rho}_{X,Y} = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{n\hat{\sigma}_X\hat{\sigma}_Y} = \frac{\sum_{i=1}^{n} X_i Y_i - \frac{\sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n}}{n\hat{\sigma}_X\hat{\sigma}_Y} \tag{32}$$

Sendo X a variável dicotômica e Y a variável ordinal (rank), então tem-se:

$$\sum_{i=1}^{n} Y_i = \frac{n(n+1)}{2} \text{ onde}$$

$$n = \text{postos=rank} = 1, 2, 3, ..., n \tag{33}$$

Os quadrados dos postos são: 1^2, 2^2, 3^2,..., n^2

$$\text{Então: } \sum_{i=1}^{n} Y_i^2 = \frac{n(n+1)(2n+1)}{6} \tag{34}$$

A média de Y é dada por:

$$\overline{Y} = \frac{\sum_{i=1}^{n} Y_i}{n} = \frac{n+1}{2} \tag{35}$$

E a variância será dada por:

$$\hat{\sigma}_Y = \sqrt{\frac{n^2 - 1}{12}} \tag{36}$$

Sendo X uma variável dicotômica, então:

$$n = n_0 + n_1$$

onde:

n é o número total de observações da amostra

n_0 é o número de observações cuja variável X assume valor zero

n_1 é o número de observações cuja variável X assume valor um

$$\sum_{i=1}^{n} X_i Y_i = \sum_{i=1}^{n_1} Y_i \times 1 + \sum_{i=1}^{n_0} Y_i \times 0 = \sum_{i=1}^{n_1} Y_i \tag{37}$$

$$\sum_{i=1}^{n} X_i = \sum_{i=1}^{n} X_i^2 = n_1 \tag{38}$$

A média e a variância de X serão dadas por:

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{n_1}{n} \tag{39}$$

$$\hat{\sigma}_X = \sqrt{\frac{n_1 - \frac{n^2}{n}}{n}} = \sqrt{\frac{\frac{n \times n_1 - n_1^2}{n}}{n}} = \sqrt{\frac{n_1(n - n_1)}{n^2}} \tag{40}$$

Substituindo (35), (36), (39) e (40) em (32) tem-se:

$$\hat{\rho}_{dr} = \frac{\sum_{i=1}^{n} Y_i - \frac{n_1 n(n+1)}{n}}{\sqrt{\frac{n_1(n-n_1)}{n^2}}\sqrt{\frac{n^2-1}{12}}} = \frac{\frac{2\sum_{i=1}^{n} Y_i - n_1(n+1)}{2n}}{\sqrt{\frac{[n_1(n_1+n_0)]}{n^2}\frac{n^2-1}{12}}}$$

Resultando em:

$$\hat{\rho}_{dr} = \frac{2\sum_{i=1}^{n} Y_i - n_1(n+1)}{\sqrt{\frac{n_1 n_0\left(n^2-1\right)}{3}}} \tag{41}$$

onde:

$\hat{\rho}_{dr}$ é o coeficiente de correlação entre as variáveis X e Y

$\sum_{i=1}^{n} Y_i$ é a soma da variável ordinal Y

n é o número total de observações

n_0 é o número de observações cuja variável X assume valor zero;

n_1 é o número de observações cuja variável X assume valor um.

A significância do coeficiente estimado para amostras com $n \geq 30$, poderá ser obtida através da estatística Z, como segue:

$$Z = \hat{\rho}_{dr}\sqrt{n-1} \qquad (42)$$

3.2.5. *Coeficiente de correlação entre variável ordinal (rank) e intervalar*

Quando tem-se uma variável (X) ordinal e outra (Y) intervalar, é possível estimar o coeficiente através do estimador apresentado em Wherry [10], que também derivou-se do coeficiente linear de Pearson.

Conforme apresentado na expressão (32), o estimador do coeficiente linear de Pearson é:

$$\hat{\rho}_{X,Y} = \frac{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)\left(Y_i-\overline{Y}\right)}{n\hat{\sigma}_X\hat{\sigma}_Y} = \frac{\sum_{i=1}^{n}X_iY_i - \frac{\sum_{i=1}^{n}X_i\sum_{i=1}^{n}Y_i}{n}}{n\hat{\sigma}_X\hat{\sigma}_Y} \qquad (43)$$

Se X é uma variável ordinal (rank), então é possível escrever:

$$\sum_{i=1}^{n}X_i = \frac{n(n+1)}{2} \text{ onde } n = \text{postos= rank } = 1, 2, 3,..., n$$

Os quadrados dos postos são: $1^2,\ 2^2,\ 3^2,...,\ n^2$

$$\text{Então: } \sum_{i=1}^{n}X_i^2 = \frac{n(n+1)(2n+1)}{6} \qquad (44)$$

A média e a variância da variável X serão obtidas por:

$$\overline{X} = \frac{\sum_{i=1}^{n}X_i}{n} = \frac{n(n+1)}{n\times 2} = \frac{n+1}{2} \qquad (45)$$

$$\hat{\sigma}_X = \sqrt{\frac{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}{n}} = \sqrt{\frac{\sum_{i=1}^{n}X_i^2 - \frac{\left(\sum_{i=1}^{n}X_i\right)^2}{n}}{n}} \qquad (46)$$

Tem-se que:

$$\sum_{i=1}^{n}X_i^2 - \frac{\left(\sum_{i=1}^{n}X_i\right)^2}{n} = \frac{n^3-n}{12} \qquad (47)$$

Substituindo (47) em (46) tem-se:

$$\hat{\sigma}_X = \sqrt{\frac{n^3-n}{12\times n}} = \sqrt{\frac{n^2-1}{12}} \qquad (48)$$

Substituindo (45) e (48) em (43) tem-se:

$$\hat{\rho}_{ri} = \frac{\frac{\sum_{i=1}^{n}X_iY_i}{n} - \frac{(n+1)\sum_{i=1}^{n}Y_i}{2n}}{\sqrt{\frac{n^2-1}{12}}\ \hat{\sigma}_Y} = \frac{\frac{\sum_{i=1}^{n}X_iY_i}{n} - \frac{(n+1)\overline{Y}}{2}}{\sqrt{\frac{n^2-1}{12}}\ \hat{\sigma}_Y} \qquad (49)$$

onde:

$\hat{\rho}_{ri}$ é o coeficiente de correlação entre a variável ordinal e intervalar

$\hat{\sigma}_Y$ é o desvio padrão da variável Y

n é o número de observações da amostra

A significância do coeficiente estimado poderá ser obtida através de:

$$t = \hat{\rho}_{ri}\sqrt{\frac{n-2}{1-\hat{\rho}_{ri}^2}} \ \sim\ t_{n-2} \qquad (50)$$

4 - RESULTADOS E DISCUSSÃO

Para a aplicação de diferentes métodos de coeficiente de correlação derivados do coeficiente linear de Pearson, gerou-se diferentes amostras pelo processo de simulação Monte Carlo, utilizando o Statistical Software Analysis (SAS), atendendo às suposições quanto ao nível de mensuração das variáveis envolvidas na análise. Os algoritmos utilizados encontram-se no apêndice.

4.1. *Aplicação do coeficiente de correlação ponto bisserial*

Gerou-se uma amostra aleatória em que a variável X é intervalar e a variável Y é dicotômica. A amostra aleatória e as estatísticas encontram-se nos quadros A.1 e A.2 do apêndice.

O coeficiente de correlação ponto bisserial calculado foi $\hat{\rho}_{pb} = 0{,}76533$. Calculando-se o coeficiente linear de Pearson para as variáveis X e Y, evidentemente obteve-se o mesmo valor, pois trata-se do mesmo coeficiente.

A significância do coeficiente de correlação ponto bisserial quanto do coeficiente linear de Pearson é $\alpha < 0{,}01$, cujo valor de t calculado foi 7,33.

4.2. *Aplicação do coeficiente de correlação de Spearman*

As variáveis X e Y geradas aleatoriamente são ordinais, apresentadas no quadro A.3. Foram calculados os coeficientes de correlação de Spearman e o linear de Pearson, cujo coeficiente estimado foi $\hat{\rho} = \hat{\rho}_S = 0{,}80423$, com $t = 7{,}16$, significativo, portanto, para $\alpha < 0{,}01$.

4.3. *Aplicação do coeficiente de correlação phi*

Gerou-se uma amostra aleatória de variáveis dicotômicas X e Y que se encontra no quadro A.4.

O coeficiente de correlação phi calculado a partir da tabela de contingência apresentada na tabela A.1 é exatamente igual ao coeficiente linear de Pearson, calculado a partir das variáveis dicotômicas X e Y, qual seja $\hat{\rho} = \hat{\phi} = 0{,}78667$.

O teste de significância utilizando a estatística t indica nível de significância $\alpha < 0{,}01$, com $t = 7{,}85$. Utilizando-se

o teste χ^2, a significância também é de $\alpha < 0{,}01$, para $\chi^2 = 31{,}47$.

Para Andenberg [1], quando as variáveis nominais são definidas como dicotômicas, podem ser consideradas nas análises como variáveis medidas em nível intervalar.

Evidentemente, as estatísticas t e χ^2 diferem em seus valores, no entanto, as signficâncias são iguais. Portanto, uma vez definido o nível de significância, aceita-se ou rejeita-se H_0, tanto para a estatística t quanto para χ^2.

4.4. *Aplicação do coeficiente de correlação entre variáveis dicotômica e ordinal*

Tem-se que a variável X é medida em nível ordinal e a variável Y é dicotômica. A amostra aleatória gerada pelo processo de simulação encontra-se no quadro A.5 e as estatísticas da variável X, no quadro A.6.

O coeficiente de correlação $\hat{\rho}_{dr}$ calculado foi 0,86458, exatamente igual ao coeficiente linear de Pearson. A significância do coeficiente $\hat{\rho}$ é $\alpha < 0{,}01$ para $t = 9{,}10$. A estatística Z calculada para testar a significância de $\hat{\rho}_{dr}$ foi 4,66, significativo também para $\alpha < 0{,}01$.

4.5. *Aplicação do coeficiente de correlação entre variáveis intervalar e ordinal*

A amostra aleatória encontra-se no quadro A.7 do Apêndice. O coeficiente de correlação estimado é $\hat{\rho} = \hat{\rho}_{ri} = 0{,}87289$, cuja estatística t é igual a 12,39, significativa portanto para $\alpha < 0{,}01$.

Os coeficientes de correlação ponto bisserial, phi , de Spearman, $\hat{\rho}_{dr}$ e $\hat{\rho}_{ri}$ são derivados do coeficiente linear de Pearson, desenvolvidos para situações em que a suposição de que as variáveis devem ser medidas no nínimo em nível intervalar não são atendidas. No entanto, verificou-se que para cada um dos coeficientes considerando-se o nível de mensuração das variáveis envolvidas, obteve-se a mesma estimativa do coeficiente linear de Pearson, o que indica a possibilidade de utilização deste último coeficiente mesmo em situações que envolvem variáveis ordinais e dicotômicas.

Para os coeficientes de correlação ponto bisserial, de Spearman e $\hat{\rho}_{ri}$, as significâncias dos coeficientes estimados são exatamente iguais a do coeficiente de Pearson, uma vez que a estatística para o cálculo da significância é a mesma, ou seja, t de Student com n-2 graus de liberdade.

Já no caso do coeficiente de correlação phi, utiliza-se a estatística χ^2. Porém, a significância é igual ao da estatística t, para o coeficiente linear de Pearson.

O mesmo ocorre para o coeficiente $\hat{\rho}_{dr}$, cuja estatística utilizada para verificar a significância é Z, que, embora difiram nos valores calculados, as significâncias são iguais.

5 – CONCLUSÃO

Conclui-se portanto que é possível utilizar o coeficiente linear de Pearson para variáveis medidas a nível intervalar, ordinal e dicotômica, tendo as devidas precauções na interpretação, ou seja, o quadrado do coeficiente de correlação não pode ser interpretado como a proporção da variância comum às duas variáveis (R^2), quando envolvem variáveis ordinais e dicotômicas.

Dentre os fatores que afetam o coeficiente linear de Pearson, pode-se citar o tamanho da amostra, principalmente quando é pequeno. Assim, apesar da possibilidade da utilização do coeficiente linear de Pearson, para as variáveis que não são medidas no mínimo em nível intervalar, há que se atentar para a questão do tamanho da amostra, das variáveis envolvidas na análise.

REFERÊNCIAS BIBLIOGRÁFICAS

[1] Andenberg, Michel R.; "Cluster analysis for applications". New York: J. Wiley & Sons, 1958.

[2] Bunchaft, G. and Kellner, S. R. O.; "Estatística sem mistérios". 2.ed. Petrópolis: Vozes, 1999. v.2.

[3] Downie, N. M. and Heath, R. W.; "Basic statistical methods". New York: Harper & Brothers, 1959.

[4] Ferguson, G. A.; "Statistical analysis in psychology and education". 5.ed. New York: McGraw-Hill book, 1981.

[5] Guilford, J. P.; "Fundamental statistics in psychology and education". 4.ed. New York: McGraw-Hill Book, 1950.

[6] Haldar, A. and Mahadevan, S.; " Probability, reliability and statistical methods in engineering design". New York: J. Willey & Sons, 2000.

[7] Lira, S. A.; "Análise de correlação: abordagem teórica e de construção dos coeficientes com aplicações". Curitiba, 2004. 196 p. Dissertação (mestrado). Setores de Ciências Exatas e de Tecnologia, UFPR.

[8] Siegel, Sidney; "Estatística não-paramétrica: para as ciências do comportamento". São Paulo: McGraw-Hill do Brasil, 1975.

[9] Silveira, F. L.; "Estatística paramétrica versus não-paramétrica: um estudo empírico". Scientia, v. 2, n. 2, pp. 115-122, jul-dez 1991.

[10] Wherry, R. J.; "Contributions to correlational analysis". Orlando: Academic Press, 1984.

APÊNDICE

1- ALGORITMOS UTILIZADOS PARA OBTENÇÃO DAS VARIÁVEIS

1.1. *Algoritmo para gerar variáveis normais bivariadas*

```
data normalbi;
  keep x y;
  m1=5; m2=20; v1=2; v2=10; ro=0.80;
    do i=1 to 30; /* tamanho da amostra */
    x=m1+sqrt(v1)*rannor(123);
    y=(m2+ro*(sqrt(v2)/sqrt(v1))*(x-m1))+ sqrt(v2*(1-
       ro**2))*rannor(123);
    output;
    end;
run;
```

1.2. *Algoritmo para gerar variável normal*

```
data normal;
 seed=45;
 n=20;
 do i=1 to n;
   x=rannor(seed);
 output; end;
run;
```

1.3. *Algoritmo para gerar variável Bernoulli*

```
data bernoulli;
seed=45;
n=20;
  do j=1 to n;
  x=ranbin(seed,1,0.4);
  output;  end;
run;
```

2 – AMOSTRAS UTILIZADAS PARA APLICAÇÃO DOS MÉTODOS DE CORRELAÇÃO

QUADRO A.1 - VARIÁVEIS ALEATÓRIAS NORMAL X E BERNOULLI Y

OBS.	X	Y	OBS.	X	Y
1	68,53943	0	21	70,74722	1
2	68,76153	0	22	69,78154	1
3	67,51424	0	23	67,87615	0
4	71,33978	1	24	74,37667	1
5	69,90114	1	25	71,52511	1
6	65,23922	0	26	66,89329	0
7	75,64971	1	27	71,57874	1
8	69,04248	0	28	70,93864	0
9	74,65876	1	29	68,87160	0
10	67,57904	0	30	73,79670	1
11	68,24473	0	31	73,26968	1
12	62,99353	0	32	68,13456	0
13	77,46998	1	33	68,69880	0
14	66,05733	0	34	73,09149	1
15	73,28209	1	35	71,65980	1
16	71,05588	1	36	72,43791	1
17	69,54481	1	37	68,48637	0
18	70,79316	0	38	69,62983	0
19	66,96403	0	39	66,57056	0
20	72,22281	1	40	69,57349	0

QUADRO A.2 - ESTATÍSTICA DESCRITIVA DA VARIÁVEL X SEGUNDO VALORES DA VARIÁVEL Y E TOTAL

Y	FREQ.	MÉDIA	DESVIO PADRÃO
0	21	67,9715	1,7979
1	19	72,4942	2,0555
TOTAL	40	70,1198	2,9510

QUADRO A.3 - VARIÁVEIS ALEATÓRIAS X E Y NORMAIS E TRANSFORMADAS EM ORDINAIS

OBS.	X	Y	OBS.	X	Y	OBS.	X	Y
1	4	1	11	14	11	21	22	21
2	1	2	12	15	12	22	21	22
3	2	3	13	17	13	23	20	23
4	8	4	14	10	14	24	27	24
5	6	5	15	16	15	25	13	25
6	3	6	16	28	16	26	25	26
7	9	7	17	12	17	27	24	27
8	18	8	18	23	18	28	19	28
9	7	9	19	5	19	29	30	29
10	11	10	20	26	20	30	29	30

QUADRO A.4 - VARIÁVEIS ALEATÓRIAS BERNOULLI X E Y

OBS.	X	Y	OBS.	X	Y	OBS.	X	Y	OBS.	X	Y
1	1	1	11	1	1	21	0	0	31	0	0
2	0	1	12	1	1	22	1	1	32	1	1
3	0	0	13	1	0	23	1	1	33	1	1
4	1	1	14	0	0	24	0	0	34	0	0
5	0	0	15	1	1	25	1	1	35	1	1
6	1	1	16	1	1	26	1	1	36	0	1
7	1	1	17	1	1	27	1	1	37	1	1
8	1	0	18	1	1	28	1	1	38	0	0
9	1	1	19	0	0	29	1	1	39	0	0
10	0	0	20	1	1	30	0	0	40	0	0

TABELA A.1 - TABELA DE CONTINGÊNCIA DAS VARIÁVEIS X E Y

Y	X		
	1	0	$\sum$
1	23	2	25
0	2	13	15
$\sum$	25	15	40

QUADRO A.5 - VARIÁVEIS ALEATÓRIAS NORMAL X TRANSFORMADA EM ORDINAL E BERNOULLI Y

OBS.	X	Y	OBS.	X	Y	OBS.	X	Y
1	1	0	11	11	0	21	21	1
2	2	0	12	12	0	22	22	1
3	3	0	13	13	0	23	23	1
4	4	0	14	14	0	24	24	1
5	5	0	15	15	0	25	25	1
6	6	0	16	16	0	26	26	1
7	7	0	17	17	1	27	27	1
8	8	0	18	18	1	28	28	1
9	9	0	19	19	1	29	29	1
10	10	0	20	20	1	30	30	1

QUADRO A.6 - ESTATÍSTICA DESCRITIVA DA VARIÁVEL X SEGUNDO VALORES DA VARIÁVEL Y

Y	$\overline{X}$	S
0	8,5000	4,7610
1	23,5000	4,1833

QUADRO A.7 - VARIÁVEIS ALEATÓRIAS NORMAL X TRANSFORMADA EM ORDINAL E NORMAL Y

OBS	X	Y	OBS	X	Y
1	1	50,1236	26	26	58,1540
2	2	51,1531	27	27	55,4734
3	3	51,1754	28	28	59,6157
4	4	56,7419	29	29	60,1809
5	5	51,1830	30	30	64,0449
6	6	54,5408	31	31	59,9498
7	7	55,5512	32	32	56,5022
8	8	55,2186	33	33	66,6750
9	9	54,5424	34	34	63,8418
10	10	52,2496	35	35	63,0514
11	11	54,8710	36	36	62,3340
12	12	55,9186	37	37	62,6357
13	13	56,1823	38	38	60,7527
14	14	58,7238	39	39	61,2694
15	15	50,9812	40	40	66,5862
16	16	62,0726	41	41	65,3720
17	17	55,7658	42	42	62,2744
18	18	56,9581	43	43	66,4326
19	19	56,4535	44	44	67,7780
20	20	54,9055	45	45	66,9581
21	21	56,3960	46	46	63,9188
22	22	60,1621	47	47	60,6948
23	23	58,0464	48	48	67,7532
24	24	60,8010	49	49	69,7136
25	25	64,1540	50	50	67,1037