ENSINO DA CORRELAÇÃO DE POSTOS NO ENSINO MÉDIO

Antonio Carlos Fonseca Pontes
acfpontes@yahoo.com.br
Universidade Federal do Acre

RESUMO

Dentre os procedimentos estatísticos, um dos mais populares é a correlação linear, em que são estudadas duas variáveis medidas em um único indivíduo concomitantemente. Tal popularidade se justifica em função do possível relacionamento existente entre características num mesmo indivíduo. Entretanto, o coeficiente de correlação linear de Pearson, que é o procedimento mais conhecido para a obtenção desse tipo de relacionamento, nem sempre é adequado, especialmente quando uma ou ambas as variáveis são medidas em escala ordinal. Nessas situações, o coeficiente de correlação de Spearman é mais adequado por levar em consideração a ordem dos dados e não o seu valor intrínseco. Ainda, quando os dados obtidos das variáveis não aderem à distribuição normal devido, por exemplo, à presença de valores discrepantes (*outliers*), o coeficiente de correlação de Spearman é um bom substituto para a verificação do inter-relacionamento das variáveis consideradas. O coeficiente de correlação de Spearman é simples de calcular e de fácil compreensão, especialmente quando o número de pares de dados (ou indivíduos) é pequeno. Neste trabalho busca-se incentivar a introdução de novas metodologias estatísticas no ensino médio, especificamente do coeficiente de correlação de Spearman, fortalecendo e enriquecendo o conhecimento dos alunos e professores do ensino médio e trazendo, ainda que de forma incipiente, um pouco da realidade cotidiana para a sala de aula.

Palavras-chave: estatística não-paramétrica, análise combinatória, coeficiente de correlação de Spearman.

ABSTRACT

Amongst the statistical procedures, one of the most popular is the linear correlation, where two variables measured in a single individual are studied concomitantly. Such popularity if justifies in function of the possible existing relationship between characteristics in one same individual. However, the Pearson linear correlation coefficient, that it is the known procedure more for the attainment of this type of relationship, nor always it is adjusted, especially when one or both variables is measured in ordinal scale. In such situations, the Spearman correlation coefficient is more adequately by taking in consideration the order of the data and not it its intrinsic value. Still, when the gotten data of the variable do not adhere to the normal distribution due, for example, to the presence of outliers, the Spearman correlation coefficient is a good substitute for the verification of the inter-relationship of the considered variables. The Spearman correlation coefficient is simple to calculate and easy to understanding, especially when the number of pairs of data (or individuals) is small. This work search to stimulate the introduction of new statistical methodologies in average education, specifically the Spearman correlation coefficient, fortifying and enriching the knowledge of the students and professors of average education and bringing, despite of incipient form, a little of the daily reality for the classroom.

Key words: nonparametric statistic, combinatorial analysis, Spearman’s correlation coefficient.

1. INTRODUÇÃO

A correlação entre duas variáveis, medidas num mesmo indivíduo, é calculada com o intuito de verificar se existe inter-relacionamento entre essas variáveis. Padronizou-se que tal medida deve estar no intervalo fechado de *-1* a *1*, em que *-1* indica perfeita correlação negativa ou inversa e *1* indica perfeita correlação positiva ou direta. A correlação negativa indica que o crescimento de uma das variáveis implica, em geral, no decrescimento da outra. A correlação positiva indica, em geral, o crescimento ou decrescimento concomitante das duas variáveis consideradas.

Por exemplo, pode-se desejar saber se existe alguma relação entre pares de variáveis como peso e altura de pessoas, população e área de países ou municípios, notas de alunos em disciplinas diferentes, peso e pressão sistólica, idade e níveis de colesterol, dentre outros.

A correlação obtida através do coeficiente de Pearson, que é a medida de correlação mais conhecida, é linear. Assim, nos casos em que a relação entre as variáveis seja não linear (quadrática, cúbica, exponencial, etc.), ela não será medida adequadamente. Nesses casos os dados devem ser transformados para a obtenção da medida adequada. O outro coeficiente de correlação utilizado, o de Spearman, por realizar uma transformação de postos, pode ser utilizado nas situações em que a relação entre os pares de dados não é linear.

2. METODOLOGIA

2.1 Generalidades

Manualmente, ou com o auxílio de ferramentas computacionais, é possível classificar os dados de uma amostra $x_1, x_2,...,x_{n-1}, x_n$ em ordem crescente. Os dados, ordenados dessa forma, formam uma seqüência denotada por $x_{(1)}, x_{(2)},...,x_{(n-1)}, x_{(n)}$, onde os parêntesis no subscrito indicam ordem. De modo formal, dizemos que $x_{(1)}, x_{(2)},...,x_{(n)}$ são as chamadas estatísticas de ordem da amostra $x_1, x_2,...,x_n$ e $x_{(i)}$ indica a *i*-ésima estatística de ordem, ou seja, a *i*-ésima observação ordenada.

Com base nessa ordenação pode-se definir o posto ou *rank* de uma observação. Em geral tem-se que o posto de $x_{(i)}$ é igual a *i*, ou seja, o posto é dado pelo subscrito da estatística de ordem. Obviamente que esta definição refere-se aos postos crescentes. No caso de postos decrescentes, é possível obtê-los através de diferença, ou seja, dados *n* dados ordenados, o posto decrescente da observação que ocupa o *i*-ésima posição quando se consideram os postos crescentes, é dada por

POSTO DECRESCENTE = n – POSTO CRESCENTE + 1 = n – i + 1

Teoricamente, observações empatadas, ou seja, que têm valores iguais, não ocorrem. Na prática, entretanto, tais situações são comuns e nesse caso, valores equivalentes devem ter postos iguais. Uma maneira encontrada para solucionar tais problemas é considerar, para os casos em que haja empates, a média dos postos. Assim, quando duas observações, que teriam supostamente os postos *k* e *k+1* forem iguais, considera-se a média desses postos para ambas. Assim o posto para essas observações seria igual a *[k+k+1]/2 = k+1/2*. Procedimento equivalente é feito se há três ou mais observações empatadas.

A discussão sobre estatísticas de ordem é especialmente útil na definição e aplicação de testes não-paramétricos, em que os dados originais são substituídos por seus respectivos postos. Os testes não-paramétricos são poderosos substitutos dos testes paramétricos, especialmente nos casos em que as amostras são pequenas, naqueles em que a distribuição dos dados não é normal ou ainda quando dados discrepantes (*outliers*) ocorrem. Dentre as técnicas não-paramétricas, o coeficiente de correlação de Spearman (r_s) é uma das mais conhecidas e utilizadas na prática. Esse coeficiente é utilizado em substituição ao coeficiente de correlação de Pearson (r) nos casos em que a binormalidade dos dados não ocorre e ainda em situações envolvendo poucos pares de dados.

2.2 Definições e valores máximo e mínimo

Dadas duas variáveis, *X* e *Y*, cujos valores são X_i e Y_i, $i = 1, 2, ..., n$, pode-se buscar relacionar essas variáveis mediante o uso do coeficiente de correlação. O coeficiente de correlação linear de Spearman, conhecido como o coeficiente de correlação de postos, equivale ao coeficiente de correlação de Pearson adaptado a dados transformados em postos de acordo com a transformação de Wilcoxon. A atribuição de postos, nesse caso, é feita separadamente para cada uma das variáveis. Assim, para a variável *X* atribui-se o posto *1* à menor variável, posto *2* à segunda menor variável e assim por diante, até o posto *n* para a maior variável. O mesmo procedimento é feito para a variável *Y*, independente dos valores da variável *X*. Os empates são tratados como usualmente, ou seja, valores iguais de cada uma das variáveis devem receber o mesmo posto. Assim, se $X_{(k)} = X_{(k+1)}$, ou seja, os valores de ordem *k* e de ordem *k + 1* estão empatados, para ambos atribui-se o posto $(k+k+1)/2 = (2k+1)/2 = k+1/2$. O coeficiente de correlação de postos (Spearman) é dado por

$$r_S = 1 - \frac{6 \times \sum_{i=1}^{n} d_i^2}{n^3 - n}$$

em que $d_i = r_{X_i} - r_{Y_i}$, com r_{X_i} *e* r_{Y_i} variam de *1* a *n*.

O valor máximo para o coeficiente de correlação de Spearman ($r_S = 1$) ocorre quando todos os valores de d_i são nulos, ou seja, quando os postos das variáveis X e Y são iguais para cada um dos indivíduos. O valor mínimo é obtido quando a soma dos quadrados das diferenças é máxima e igual a $\frac{n^3 - n}{3}$ e portando, $r_S = -1$. Tais resultados servem para o professor verificar a veracidade da afirmação de que o coeficiente de correlação de postos varia de *-1* (maior correlação negativa) e *1* (maior correlação positiva).

A correlação negativa ocorre quando há uma inversão dos valores dos postos da variável *Y* em relação à variável *X*. A correlação positiva ocorre se os postos das duas variáveis seguem aproximadamente o mesmo padrão. A obtenção de coeficientes de correlação de postos com valores próximos de zero sugerem a não existência de correlação linear entre as duas variáveis.

O coeficiente de correlação de postos (Spearman) nada mais é que o coeficiente de correlação linear de Pearson aplicado aos postos dos dados, obtidos independentemente para cada variável. Uma diferença que deve ser observada é que os valores *1* ou *-1* para o coeficiente de correlação de Spearman não são raros de ocorrer na prática. No caso do coeficiente de correlação de Pearson, para que ele seja igual a *1* ou *-1*, a variável *Y* deve ser função linear da variável *X*.

2.4 Testes para o coeficiente de correlação de Spearman

Para realizar testes de hipóteses sobre o coeficiente de correlação de postos, em geral utiliza-se o procedimento assintótico aplicado ao coeficiente de correlação de Pearson. Entretanto, tais procedimentos não são adequados quando o número de pares de variáveis é pequeno. Nesses casos, é possível se obter o nível de significância (valor-*p*) exato para o coeficiente obtido ou ainda utilizar testes de permutação aleatórios para a obtenção de valores-*p* aproximados, nos casos em que o número de possíveis permutações é grande.

Uma das variáveis (por exemplo, *X*) pode ser ordenada e fixada, com valores inteiros variando de *1* a *n* (se não houver empates), enquanto que a outra (digamos, *Y*) é permutada. Assim, existem *n!* possíveis de combinações de postos entre *X* e *Y*. Sabe-se que o valor da soma das postos, para qualquer das variáveis, é fixo e igual a *n(n+1)/2*. Assim, fixados *n-1* valores, o *n*-ésimo pode ser obtido por diferença. Assim, é possível diminuir o número de pareamentos possíveis para *(n-1)!*. Para cada permutação, são obtidas as diferenças d_i e seus respectivos quadrados (d_i^2). Sabe-se ainda que a soma das diferenças é nula ($\sum_{i=1}^{n} d_i = 0$) e esse fato pode ser utilizado para a checagem dos resultados. Os outros valores necessários para o cálculo do coeficiente são fixos.

3. EXEMPLOS DE APLICAÇÃO

Em sala de aula, são vários os exemplos em que o coeficiente de correlação pode ser utilizado, abordando situações (dados) cotidianas. Assim, pode-se calcular o coeficiente de correlação existente entre a altura e o peso dos alunos, entre as notas dos discentes em provas de disciplinas diferentes, entre o número de horas de estudo e a nota obtida, entre as idades do pai e a idade da mãe dos alunos, entre as preferências de cada aluno pelas disciplinas ofertadas e suas respectivas notas, dentre outros. Outros exemplos podem ser buscados em situações do dia-a-dia, como, por exemplo, entre o preço de determinados produtos e o número de famílias possuidores daquele tipo de produto, entre a área construída da residência e o número de membros da família, etc. Muitas outras situações podem ser criadas pelos professores, envolvendo outros assuntos que estejam sendo estudados pelos alunos nas diversas disciplinas como, por exemplo, a população e a área ou o IDH (índice de desenvolvimento humano) e a renda per capita de países, nível de renda e incidência de doenças em cidades, número de habitantes e número de eleitores em determinadas cidades, dentre outros. Basta simplesmente tomar duas variáveis que, supostamente, sejam relacionadas para verificar a eficácia desse tipo de coeficiente e exemplificar sua utlização.

Pontes (2003) apresenta as notas médias de cinco juízes para a preocupação ambiental de 27 produtores rurais do Assentamento Sumaré II. Detalhes sobre a maneira de obtenção dessas notas podem ser obtidos no trabalho original. Para exemplificar o método aqui apresentado, foram tomadas as notas dos cinco juízes para os seis primeiros moradores, conforme o Quadro 1.

Quadro 1. Notas e postos de cinco juízes para seis famílias do Assentamento Sumaré II.

	JUIZ 1		JUIZ 2		JUIZ 3		JUIZ 4		JUIZ 5	
CASA	$IREc_{i1}$	Postos	$IREc_{i2}$	Postos	$IREc_{i3}$	Postos	$IREc_{i4}$	Postos	$IREc_{i5}$	Postos
1	3,83	4	2,67	4	2,92	3	2,05	3	3,42	3
2	4,26	5	4,52	5	4,56	6	4,29	5	3,69	5
3	2,12	1	1,50	1	1,69	1	1,53	1	3,24	1
4	3,34	2	2,19	2	2,42	2	2,01	2	3,36	2
5	3,67	3	4,63	6	4,48	5	4,75	6	3,80	6
6	4,37	6	2,59	3	3,11	4	4,00	4	3,63	4

A partir do Quadro 1 são obtidos os dez ($C_{5,2}$) coeficientes de correlação de Spearman (Quadro 2) entre as notas dos cinco juízes. Observa-se que os coeficientes de correlação de Pearson, mais utilizados em trabalhos, não são válidos nos casos em que o número de pares de dados é pequeno ou ainda quando os valores a serem comparados são arbitrários e não resultantes de medidas. No exemplo, as notas tiveram como finalidade a ordenação das famílias em função da sua preocupação ambiental.

Quadro 2. Valores do coeficiente de correlação de Spearman e seus respectivos valores-p.

Pares de Juizes	1 e 2	1 e 3	1 e 4	1 e 5	2 e 3	2 e 4	2 e 5	3 e 4	3 e 5	4 e 5
Coeficiente	0,486	0,714	0,600	0,600	0,886	0,943	0,943	0,943	0,943	1,000
Valor-p	0,329	0,111	0,208	0,208	0,019	0,005	0,005	0,005	0,005	< 0,001

Os valores-p obtidos no quadro referem-se ao teste do coeficiente de correlação linear de Pearson aplicado aos postos dos dados. Esses valores definem se duas variáveis são ou não correlacionadas do ponto de vista estatístico. Em geral, se esse valor for menor que 0,05, considera-se que a correlação entre as variáveis trabalhadas é significativa. Se esses valores não forem obtidos de forma correta, as conclusões não serão válidas.

Observa-se que o valor do coeficiente igual a 1 só ocorre se os postos das variáveis forem todos coincidentes. Isso só ocorre de uma maneira e assim, o valor-p é igual a 1/720 = 0,001389.

Tabela 1. Valores do coeficiente de correlação de postos e valores-p.

R_s	Contagem	Contagem Acumulada	%	% Acumulado
-1,00	1	1	0,83	0,83
-0,90	4	5	3,33	4,17
-0,80	3	8	2,50	6,67
-0,70	6	14	5,00	11,67
-0,60	7	21	5,83	17,50
-0,50	6	27	5,00	22,50
-0,40	4	31	3,33	25,83
-0,30	10	41	8,33	34,17
-0,20	6	47	5,00	39,17
-0,10	10	57	8,33	47,50
0,00	6	63	5,00	52,50
0,10	10	73	8,33	60,83
0,20	6	79	5,00	65,83
0,30	10	89	8,33	74,17
0,40	4	93	3,33	77,50
0,50	6	99	5,00	82,50
0,60	7	106	5,83	88,33
0,70	6	112	5,00	93,33
0,80	3	115	2,50	95,83
0,90	4	119	3,33	99,17
1,00	1	120	0,83	100,00
∑	120		100,00	

O valor 0,943 ocorre se houver inversão dos postos entre valores contíguos (1 e 2, 2 e 3, 3 e 4, 4 e 5, 5 e 6) e apenas essa inversão ocorrer, com os demais pares sendo iguais. Isso ocorre apenas cinco vezes entre as 720 permutações, ou seja, o valor-p nesse caso é

$$Valor - p = \frac{5}{720} + \frac{1}{720} = \frac{6}{720} = 0{,}00833$$

4. CONCLUSÕES

O coeficiente de correlação de Spearman, conhecido como coeficiente de correlação de postos pode ser utilizado no ensino médio, como um elemento de aprendizado dentro do tema análise combinatória. O cálculo desse coeficiente é simples, de fácil entendimento e as permutações necessárias para a obtenção da distribuição nula podem ser obtidas sem grandes esforços. Por outro lado, as discussões que podem ser feitas a partir dos resultados obtidos com dados sociais, econômicos e de outras áreas, especialmente das ciências sociais aplicadas, tendem a enriquecerem as discussões em sala de aula.

Discussões temáticas, com a abordagem de temas sociais nas aulas de matemática podem ser úteis para a transformação social que se busca na formação de cidadãos que tenham interesses e conhecimentos variados. Nenhum conhecimento matemático novo é necessário para que tal finalidade seja alcançada. Assim, o conhecimento básico de conceitos matemáticos do ensino médio é suficiente para a aplicação dos conceitos aqui apresentados.

5. BIBLIOGRAFIA

CHEN, P.Y.; POPOVICH, P.M. Correlation: parametric and nonparametric measures. Thousand Oaks: Sage Publication, Inc. 95p.

KENDALL, M. Rank correlation methods. London: Charles Griffin & Company LTD. 202 p.

PCN Ensino Médio. Ciências da Natureza, Matemática e Suas Tecnologias. Brasília, p.04-11;42-45, 1999.

PONTES, A.C.F. Obtenção dos níveis de significância para os testes de Kruskal-Wallis, Friedman e comparações múltiplas não-paramétricas. Piracicaba, 2000. 140p. Dissertação (M.S.) - Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo.

PONTES, A.C.F.; CORRENTE, J.E. The use of nonparametric contrasts in one-way layouts and random block designs. **Journal of Nonparametric Statistics**, v.17, n.3, p.335-346, 2005.

PONTES, L.O.- Agricultura Familiar: Recuperação e Valoração da Floresta no Assentamento Rural de Sumaré II. Dissertação (MS), ESALQ/USP, Piracicaba, 2003.