CORRELAÇÃO (*)

ALFREDO DE OLIVEIRA PEREIRA

Unidade X — Correlação: 1 — Noção de correlação. 2 — Coeficiente "pearsoniano" de correlação e sua interpretação. 3 — Correção de Sheppard. 4 — Previsão estatística; equações de regressão. 5 — Outros coeficientes de correlação. 6 — Precisão e validade dos testes.

1 — NOÇÃO DE CORRELAÇÃO

Correlação é a tendência à variação concomitante, em grandeza e sinal, dos afastamentos de duas séries de dados em relação às respectivas médias.

A correlação pode ser *direta* ou *positiva, inversa* ou *negativa,* ou ainda, *nula.* Diz-se que existe correlação direta entre dois fenômenos, quando, aumentando ou diminuindo um dêles, o outro tem um acréscimo ou uma redução determinada; ao contrário, quando, pelo aumento ou diminuição de um dos fenômenos, temos a redução ou aumento do outro, dizemos que há entre êles correlação inversa ou negativa. A correlação nula indica ausência absoluta de correlação ou dependência entre os dois fenômenos.

2 — COEFICIENTE "PEARSONIANO" DE CORRELAÇÃO E SUA INTERPRETAÇÃO

O *coeficiente de correlação* nada mais é que um simples valor numérico afetado de sinal positivo ou negativo que nos indica a existência ou não de dependência entre duas variáveis e o grau de magnitude da influência, se existente, que uma exerce sôbre a outra. O coeficiente de correlação habitualmente designado pela letra r, varia de $+ 1$ a $- 1$, podendo ser nulo (0):

$r = + 1$	forte correlação direta ou positiva.
$r = 0$	correlação nula.
$r = - 1$	forte correlação inversa ou negativa.

Várias são as expressões matemáticas de determinação do coeficiente de correlação, sendo que a mais comum é:

$$r = \frac{\Sigma (xy)}{n \cdot \sigma_x \cdot \sigma_y} \qquad (38)$$

da autoria de Karl Pearson.

(*) Súmula da 10.ª aula do Curso de Extensão Universitária para Técnicos de Educação e Orientadores Educacionais, organizado pelos Cursos "Magister", sob o patrocínio da Secretaria de Educação e Cultura de São Paulo.

Na fórmula citada, x e y são os desvios dos valores das variáveis em relação às respectivas médias; n, é o número de pares de valores e σ_x e σ_y, os desvios-padrão das variáveis X e Y.

A título de exemplo, vamos calcular o coeficiente "Pearsoniano" de correlação entre dois testes de uma bateria de inteligência, com base nos resultados obtidos por 10 estudantes.

Estudantes	A	B	C	D	E	F	G	H	I	J
Teste I.....	1	2	3	4	6	6	6	7	7	8
Teste II.....	4	3	6	5	4	7	8	7	6	10

O quadro a seguir fornece os elementos necessários à aplicação da fórmula 38:

Teste I (X)	Teste II (Y)	x	y	x^2	y^2	xy
1	4	− 4	− 2	16	4	+ 8
2	3	− 3	− 3	9	9	+ 9
3	6	− 2	0	4	0	0
4	5	− 1	− 1	1	1	+ 1
6	4	+ 1	− 2	1	4	− 2
6	7	+ 1	+ 1	1	1	+ 1
6	8	+ 1	+ 2	1	4	+ 2
7	7	+ 2	+ 1	4	1	+ 2
7	6	+ 2	0	4	0	0
8	10	+ 3	+ 4	9	16	+12
50	60	0	0	50	40	+33

Cálculo das médias:

$$m_x = \frac{50}{10} = 5$$

$$m_y = \frac{60}{10} = 6$$

Cálculo dos desvios-padrão:

$$\sigma_x = \sqrt{\frac{50}{10}} = 2{,}236$$

$$\sigma_y = \sqrt{\frac{40}{10}} = 2{,}000$$

Cálculo do coeficiente "pearsoniano" de correlação:

$$r = \frac{+\ 33}{10 \times 2{,}236 \times 2{,}000} = +\ 0{,}74$$

Tratando-se de uma distribuição de freqüência, o coeficiente "pearsoniano" de correlação poderá ser obtido através da fórmula

$$r = \frac{\frac{\Sigma\,(xy)}{\Sigma f} - C_x \cdot C_y}{\sigma_x' \cdot \sigma_y'} \qquad (39)$$

em que C_x e C_y são os fatôres de correção correspondentes à expressão

$$\frac{\Sigma (df)}{\Sigma f} \text{ e } \sigma_x' \text{ e } \sigma_y'$$

os desvios-padrão das variáveis X e Y não multiplicados pelo intervalo de classe.

Para aplicação da fórmula 39 convém organizar, prèviamente, o *diagrama de dispersão* ("scatter diagram"), que é um quadro de dupla entrada constituído pelas classes das duas variáveis, uma das quais escrita na ordem inversa. Sejam, por exemplo, os resultados colhidos na aplicação simultânea de dois testes X e Y:

Y \ X	0–14	15–29	30–44	45–59	60–74	75–89	f_y
80 — 89					3	4	7
70 — 79				2	5	3	10
60 — 69			1	5	4	2	12
50 — 59			5	7	2		14
40 — 49		9	8	3			20
30 — 39	5	3	7	1			16
20 — 29	1	1	6				8
10 — 19	1	6					7
0 — 9	5	1					6
fx	12	20	27	18	14	9	100

Organizado o quadro, acrescentam-se-lhe na parte lateral direita e na parte inferior esquerda as colunas e linhas d, df, d^2f e Σxy. Em seguida, escolhe-se uma classe de cada variável (geralmente a classe modal) cujo ponto médio representará a respectiva média de origem arbitrária. As freqüências dessas duas classes formam dois eixos perpendiculares entre si, dividindo o quadro de cálculo em *4 quadrantes*. Às subdivisões ou células dos quadrantes atribuem-se desvios em tôrno das médias de origem arbitrária, positivos ou negativos, consoante a natureza do quadrante:

Quadrante I e III ..... desvios positivos
Quadrante II e IV ..... desvios negativos

Multiplicam-se, depois, as freqüências de cada célula pelos respectivos desvios, escrevendo-se os produtos nas próprias células, conforme mostra o quadro-modêlo. Êsses produtos, adicionados algèbricamente, serão transportados, em seguida, para as colunas Σxy das duas variáveis.

X / Y	0–14	15–29	30–44	45–59	60–74	75–89	f_y	d	df_y	d^2f_y	Σxy
80–89	−8	−4	0	4	8 3 24	12 4 48	7	+4	+28	112	+ 72
70–79	−6	−3	0	3 2 6	6 5 30	9 3 27	10	+3	+30	90	+ 63
60–69	−4	−2	0 1 0	2 5 10	4 4 16	6 2 12	12	+ 2	+24	48	+ 38
50–59	−2	−1	0 5 0	1 7 7	2 2 4	3	14	+1	+14	14	+ 11
40–49	0	0 9 0	0 8 0	0 3 0	0	0	20				
30–39	2 5 10	1 3 3	0 7 0	−1 1 −1	−2	−3	16	−1	−16	16	+ 12
20–29	4 1 4	2 1 2	0 6 0	−2	−4	−6	8	−2	−16	32	+ 6
10–19	6 1 6	3 6 18	0	−3	−6	−9	7	−3	−21	63	+ 24
0–9	8 5 40	4 1 4	0	−4	−8	−12	6	−4	−24	96	+ 44
fx	12	20	27	18	14	9	100		+19	471	+270
d	−2	−1		+1	+2	+3					
dfx	−24	−20		+18	+28	+27	+ 29				
d^2fx	48	20		18	56	81	223		Σxy = + 270		
Σxy	+60	+27		+22	+74	+87	+270				

Substituindo os elementos da fórmula 39 pelos valores deduzidos do quadro-modêlo:

$$\Sigma xy = +270$$

$$C_x = \frac{+29}{100} = 0{,}29$$

$$C_y = \frac{+19}{100} = 0{,}19$$

$$\sigma_x = \sqrt{\frac{223}{100} - 0{,}29^2} = 1{,}46$$

$$\sigma_y = \sqrt{\frac{471}{100} - 0{,}19^2} = 2{,}16$$

$$r = \frac{\frac{+270}{100} - 0{,}29 \times 0{,}19}{1{,}46 \times 2{,}16} = +0{,}84$$

Em geral, o valor absoluto de r indica:

Menos de 0,20 — ligeira correlação; interdependência desprezível.

De 0,20 a 0,40 — correlação baixa; relação definida, porém pequena.

De 0,40 a 0,70 — correlação moderada; relação substancial.

De 0,70 a 0,90 — alta correlação.

De 0,90 a 1,00 — correlação absoluta.

O coeficiente é *significativo* quando o seu valor absoluto é superior ao triplo de seu êrro-padrão, expresso pela fórmula:

$$\epsilon_r = \frac{1 - r^2}{\sqrt{n}} \qquad (40)$$

No exemplo dado, o êrro-padrão de $r = 0{,}84$ é:

$$= \epsilon_r \frac{1 - 0{,}84^2}{\sqrt{100}} = -0{,}02944$$

cujo triplo é 0,08832. Êste valor é consideràvelmente menor que 0,84; logo, o coeficiente obtido é altamente significativo, isto é, há mais de 99 probabilidades em 100 de que em outras amostras randômicas do mesmo universo não se apresentem coeficientes que defiram do encontrado senão de $\pm$ 0,08832.

3 — CORREÇÃO DE SHEPPARD

O grupamento simultâneo das duas variáveis em classes altera de modo significativo, como é compreensível, o valor do coeficiente que se obteria sem tabulação. Êste fato se verifica especialmente quando o número de classes é reduzido, como no exemplo dado; daí, a correção proposta por Sheppard, que consiste em subtrair da soma dos quadrados dos desvios de x e y a quantidade $\frac{n-1}{12}$.

Essa correção sempre deverá ser feita, independentemente do número de pares de observações. Levando em conta essas considerações, recomendamos a aplicação da seguinte fórmula em que se contém, para o cálculo do coeficiente "pearsoniano" de correlação, a aludida correção de Sheppard:

$$r = \frac{\Sigma (xy) - \frac{\Sigma df_x \cdot \Sigma df_y}{\Sigma f}}{\sqrt{\left(\Sigma d^2 fx - \frac{(\Sigma dfx)^2}{\Sigma f} - \frac{\Sigma f - 1}{12}\right)\left(\Sigma d^2 f_y - \frac{(\Sigma dfy)^2}{\Sigma f} - \frac{\Sigma f - 1}{12}\right)}} \qquad (41)$$

O emprêgo dessa fórmula corretiva dá para r o valor + 0,86.

4 — PREVISÃO ESTATÍSTICA. EQUAÇÕES DE REGRESSÃO

O coeficiente de correlação existente entre duas variáveis que guardem entre si uma relação de causa e efeito, além de indicar a natureza e a magnitude da relação, poderá ser utilizado, igualmente, como valor prognóstico de um acontecimento, o que importa em dizer que permite fazer a *previsão* dos valores de uma variável dependente de outra.

Essa previsão é feita pelas *equações de regressão*:

$$x = r \frac{\sigma_x}{\sigma_y} y \qquad (42$$

$$y = r \frac{\sigma_y}{\sigma_x} x \qquad (43)$$

em que x e y são desvios a partir das médias de X e de Y. Assim, conhecido um dado valor da variável X, o coeficiente de correlação entre as duas variáveis e os desvios-padrão, será possível fazer uma estimativa do correspondente valor da variável Y.

No exemplo citado no item 2 desta lição encontramos os seguintes valores:

$\sigma_x = 2{,}235 \qquad m_x = 5$

$r = 0{,}74$

$\sigma_y = 2{,}000 \qquad m_y = 6$

Substituindo os elementos das fórmulas 42 e 43 por êsses valores:

$$x = 0{,}74 \frac{2{,}235}{2{,}000} y$$

ou $x = 0{,}83\, y$

$$y = 0{,}74 \frac{2{,}000}{2{,}235} x$$

ou $y = 0{,}66\, x$

Como foi dito anteriormente, x e y são afastamentos a partir das médias de X e de Y; logo as equações de regressão são, a rigor:

$$X - m_x = 0{,}83(Y - m_y)$$
$$Y - m_y = 0{,}66(X - m_x)$$

Sendo $m_x = 5$ e $m_y = 6$:

$$X - 5 = 0{,}83(Y - 6)$$
$$Y - 6 = 0{,}66(X - 5)$$

Simplificando:

$$X = 0{,}83\, Y + 0{,}02$$
$$Y = 0{,}66\, X + 2{,}70$$

Com essas equações é possível prever um resultado de X em função de Y, e vice-versa. Assim, ao valor 10 da variável Y no problema examinado corresponderá provàvelmente o valor 8,32 da variável X:

$$X = 0{,}83 \times 10 + 0{,}02 = 8{,}32$$

5 — OUTROS COEFICIENTES DE CORRELAÇÃO

Nem sempre é possível calcular o coeficiente "pearsoniano" de correlação, dada a natureza dos valores ou a forma especial com

que se apresentam. Muitas vêzes, as variáveis são representadas por valores de posição ("ranks") não se conhecendo os valores absolutos originais; algumas vêzes, uma das variáveis, ou as duas, se apresentam sob a forma de uma classificação com duas alternativas (dicotomia) mùtuamente exclusivas. Em todos êsses casos empregam-se outros métodos de correlação, objetivando a determinação, embora aproximada, do coeficiente "pearsoniano" de correlação. Êsses métodos permitem a determinação dos coeficientes de correlação *ordinal* e *bisserial*, cujas fórmulas de cálculo são:

$$\rho = 1 - \frac{6 \Sigma (D^2)}{n (n^2 - 1)} \qquad (44)$$

$$r_{bis} = \frac{m_2 - m_1}{\sigma_t} \cdot \frac{pq}{y} \qquad (45)$$

em que os símbolos têm a seguinte significação:

ϱ — coeficiente ordinal de Spearman;
D — diferença numérica entre as posições ou "ranks";
n — número de pares de valores;
r_{bis} — coeficiente bisserial de correlação;
m_1 — média de distribuição correspondente aos valores mais baixos da variável dicotômica;
m_2*
σ_t — desvio-padrão da distribuição total;
p — proporção das freqüências de mais alto valor da variável dicotômica em relação ao total;
$q = 1 - p$;
y — ordenada da curva normal que divide a área total em partes proporcionais a p e q.

6 — PRECISÃO E VALIDADE DOS TESTES

Diz-se que os resultados de um teste são *precisos* ou *constantes* quando as discrepâncias verificadas em duas ou mais aplicações são consideradas acidentais, isto é, devidas ao acaso. Para se provar a constância ou precisão de um teste emprega-se a técnica das correlações:

a) entre os resultados de duas aplicações sucessivas do mesmo teste ao mesmo grupo e em condições idênticas (método da prova repetida);

b) entre os resultados das aplicações de dois testes equivalentes num intervalo determinado de tempo (método das formas alternadas);

c) entre os resultados da aplicação sucessiva das duas metades do mesmo teste (método do seccionamento).

O terceiro método ("split-half method"), que é o mais usado, consiste em dividir o teste em duas metades, obtendo-se duas notas para cada examinando. As duas metades do teste constituem formas alternadas, porém, dadas simultâneamente, não havendo, portanto, qualquer intervalo de tempo entre elas.

(*) Média da distribuição correspondente aos valores mais altos da variável dicotômica.

Um processo prático de dividir o teste em duas metades é formar uma com os *itens pares*, e outra, com os *itens ímpares* da prova. A correlação será calculada, conseqüentemente, entre os resultados numa e noutra metade, resultando daí a *intercorrelação* entre as duas partes da prova. Admitida a hipótese de que a organização dos itens seja feita "randômicamente", essa divisão ao meio será tão significativa quanto qualquer outra.

A única falha que apresenta o método é que o coeficiente de constância obtido varia com a extensão do teste. Um meio-teste não merece tanta confiança como um teste inteiro. Êsse inconveniente poderá ser atenuado ou eliminado com o emprêgo da *fórmula específica de Spearman-Brown*:

$$r_{11} = \frac{2_r 1/2 \quad 1/2}{1 + r_{1/2 \quad 1/2}} \qquad (46)$$

em que r_{11} é o coeficiente de constância ou precisão do teste e $r^{1/2\ 1/2}$ a intercorrelação existente entre as duas metades do teste, a qual se calcula pelo processo comum, já estudado.

Validade de um teste é a correspondência entre a habilidade reacional medida pelo teste e essa mesma habilidade reacional objetivamente definida e medida por outros meios. Decorre dêsse fato que um teste é *válido* quando mede o que pretende medir.

Para calcular o coeficiente de validade de um teste, determina-se o coeficiente "pearsoniano" de correlação existente entre os resultados obtidos por um grupo de indivíduos e os "scores" obtidos pelo mesmo grupo de indivíduos ou por outro, estatìsticamente idêntico, num teste similar cuja validade seja conhecida.

BIBLIOGRAFIA

Aitken, Alexandre C. — "Statistical Mathematics", Oliver & Boyd, Edinburgh, 1945.

Arkin, Herbert e Colton, Raymond R. — "Gráficos", Serviço Gráfico do I.B.G.E., Rio de Janeiro, 1946.

Baños, Olegario Fernandez — "Tratado de Estatística", Nuevas Gráficas, Madrid, 1945.

Carvalho Bulhões — "Estatística, Método e Aplicação", Rio de Janeiro, 1933.

Castro, Lauro Sodré Viveiros de — "Pontos de Estatística", Rio de Janeiro, 1949.

— "Exercícios de Estatística", Rio de Janeiro, 1948.

Chalier, C. V. L. — "Elementos de Estatistica Matemática", Buenos Aires, 1936.

Fisher, R. A. — "Statistical methods for research workes", Oliver & Boyd, Edinburgh, 1944.

Garret, H. E. — "Statistical in Psychology and Education", Longman's, Green & Co., New York, 1940.

Guiford, J. P. — "Fundamental Statistics in Psychology and Education", McGraw-Hill, New York, 1942.

— "Psychology methods", McGraw-Hill, New York, 1936.

Hoz, Vitor Garcia — "Formulário y tablas de Estatística aplicada a la Pedagogía", Graficas Versal, Madrid, 1944.

Kafuri, Jorge — "Lições de Estatística Matemática", Rio de Janeiro, 1934.

Kingston, Jorge — "A teoria da indução estatística", Serviço Gráfico do I.B.G.E., Rio de Janeiro, 1945.

Lindquist, E. F. — "Statistical analysis in educational research", Houghton Mifflin Co., Boston, 1940.

Martins, Otávio, A. Lins — "Sôbre o coeficiente de correlação bisserial", Revista Brasileira de Estatística — Ano VII, n.º 28, 1946.

Rodrigues, Milton da Silva — "Elementos de Estatística geral", Companhia Editôra Nacional, São Paulo, 1941.

Vessereau, André — "La Statistique", Coleção "Que sais-je ?" — Presses Universitaires de France, Paris, 1947.

Yule, G. Udny e Kendall, M. C. — "Introdução à teoria da Estatística" — Serviço gráfico do I.B.G.E., Rio de Janeiro, 1948.

Weinberg, D. — "Méthodes d'unification des mesures en biométrie et biotypologie", Herman et Cie., Paris, 1937.