NOÇÕES SOBRE CORRELAÇÃO

A correlação tem por objetivo verificar a existência da relação entre duas variáveis quantitativas, avaliando o comportamento conjunto das mesmas. Você já deve ter ouvido falar que a pressão arterial aumenta quando a idade avança, que o desempenho de um atleta melhora com o treinamento e que o número de cáries diminui com uma higiene oral bem-feita.

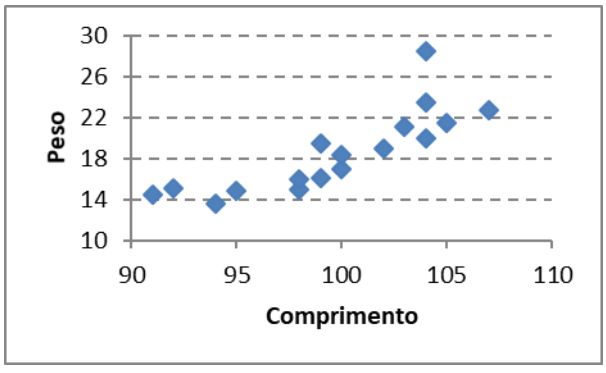
O comportamento conjunto de duas variáveis quantitativas pode ser observado através de um gráfico, denominado *Diagrama de Dispersão* e medido através do *Coeficiente de Correlação Linear de Pearson*.

DIAGRAMA DE DISPERSÃO

Observe os dados de comprimento e peso da tabela.

Tabela - Comprimento e peso

Comprimento	Peso	Comprimento	Peso
104	23,5	98	15
107	22,7	95	14,9
103	21,1	92	15,1
105	21,5	104	22,2
100	17	94	13,6
104	28,5	99	16,1
108	19	98	18
91	14,5	98	16
102	19	104	20
99	19,5	100	18,3



A correlação pode ser positiva ou negativa:

Se as variáveis X e Y crescem no *mesmo sentido*, diz-se que as duas variáveis têm correlação *positiva*.

Se as variáveis X e Y variam em *sentidos contrários*, diz-se que as duas variáveis têm correlação *negativa*.

Logo, observando o Gráfico, chega-se a conclusão de que as duas variáveis têm correlação *positiva*.

COEFICIENTE DE CORRELAÇÃO DE PEARSON

O coeficiente de correlação de Pearson (***r***) é uma medida que varia no intervalo de –1 até +1 que visa quantificar o grau de relacionamento linear entre variáveis quantitativas.

Valores próximos de +1 indicam forte correlação direta entre as variáveis enquanto que valores próximos de –1 indicam forte correlação inversa. Valores em torno de zero indicam ausência de correlação.

$$r = \frac{\sum x \cdot y - \frac{\sum x \cdot \sum y}{n}}{\sqrt{\left[\sum x^2 - \frac{(\sum x)^2}{n}\right] \cdot \left[\sum y^2 - \frac{(\sum y)^2}{n}\right]}}$$

Para entender como se aplica esta fórmula, observe os seguintes exemplos apresentados nas Tabelas II e III e seus respectivos diagramas de dispersão:

Tabela II

X	Y
1	1
2	2
3	4
4	5
5	8

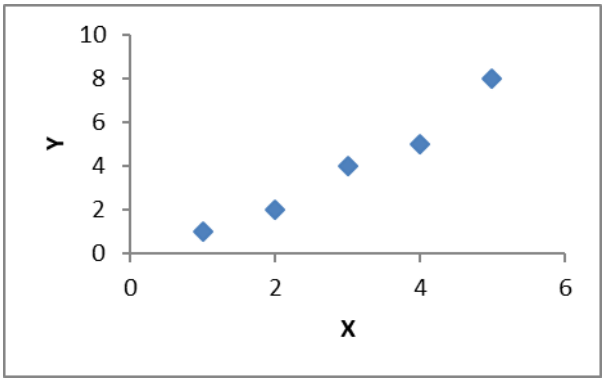
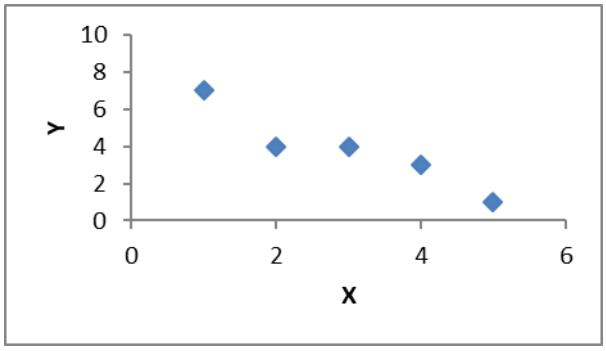


Tabela III

X	Y
1	7
2	4
3	4
4	3
5	1



No primeiro caso (Tabela II), para se obter o coeficiente de correlação realiza-se cálculos intermediários, conforme apresentados nas colunas da Tabela IV.

Tab IV - Cálculos intermediários para obtenção do coeficiente de correlação

x	y	x^2	y^2	x.y
1	1	1	1	1
2	2	4	4	4
3	4	9	16	12
4	5	16	25	20
5	8	25	64	40
Σ = 15	Σ = 20	Σ = 55	Σ = 110	Σ = 77

Com os valores obtidos na Tabela IV, realiza-se então o cálculo de ***r***.

$$r = \frac{77 - \frac{15 \cdot 20}{5}}{\sqrt{\left[55 - \frac{(15)^2}{5}\right] \cdot \left[110 - \frac{(20)^2}{5}\right]}} \qquad r = \frac{77 - 60}{\sqrt{[55 - 45] \cdot [110 - 80]}}$$

$$r = \frac{17}{\sqrt{10 \cdot 30}} \qquad r = \frac{17}{\sqrt{300}} = \frac{17}{17,32} = 0,98$$

No segundo caso (Tabela III), para se obter o coeficiente de correlação realiza-se cálculos intermediários, conforme apresentados nas colunas da Tabela V.

Tab V - Cálculos intermediários para obtenção do coeficiente de correlação

x	y	x^2	y^2	x.y
1	7	1	49	7
2	4	4	16	8
3	4	9	16	12
4	3	16	9	12
5	1	25	1	5
Σ = 15	Σ = 20	Σ = 55	Σ = 91	Σ = 44

Com os valores obtidos na Tabela V, realiza-se então o cálculo de ***r***.

$$r = \frac{44 - \frac{15 \cdot 19}{5}}{\sqrt{\left[55 - \frac{(15)^2}{5}\right] \cdot \left[91 - \frac{(19)^2}{5}\right]}} \qquad r = \frac{44 - 57}{\sqrt{[55 - 45] \cdot [91 - 72,2]}}$$

$$r = \frac{-13}{\sqrt{10 \cdot 18,8}} \qquad r = \frac{-13}{\sqrt{188}} = \frac{-13}{13,71} = -0,95$$

EXERCÍCIOS

1. Sem ver os dados, que tipo de correlação você espera entre:
   a. idade de pessoas adultas e velocidade de corrida;
   b. número de vendedores na loja e volume de vendas diárias;
   c. estatura de um homem e o número de dentes presentes na boca.

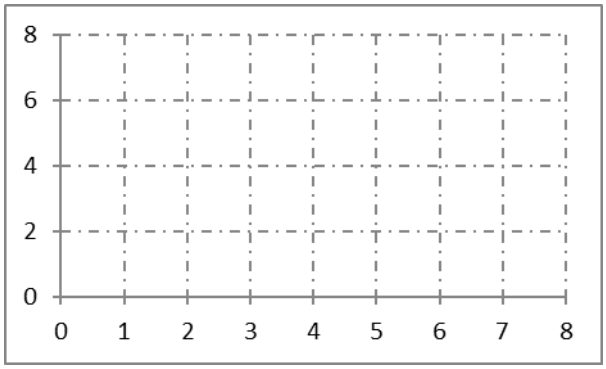
2. Sobre os diagramas de dispersão para duas variáveis é correto afirmar que:
   I. se a nuvem de pontos distribui-se nas proximidades de uma reta, então há correlação linear.
   II. se todos os pontos de diagrama pertencem a uma única reta oblíqua, então há correlação linear perfeita.
   III. se a nuvem de pontos distribui-se nas proximidades de uma reta que represente uma função crescente, então há correlação linear positiva.

   a. Todas as afirmativas são verdadeiras.
   b. Somente a afirmativa II é verdadeira.
   c. Somente a afirmativa III é falsa.
   d. A afirmativa I é falsa.
   e. Todas as afirmativas são falsas.

3. Faça um diagrama de dispersão e calcule o coeficiente de correlação para os dados apresentados na tabela. Discuta o resultado.

Tabela de dados

X	Y
3	2
5	2
4	7
2	7
1	2



4. Se todos os valores de Y forem iguais, qual será o valor de r?

5. A tabela seguinte constitui uma amostra aleatória referente às alturas, em centímetros, de 10 meninas e suas respectivas mães. Calcule o coeficiente de correlação. r = 0,92 / forte positiva

altura das mães	altura das filhas
165	168
169	170
166	168
158	160
172	172
160	164
167	165
155	158
169	172
166	164

6. Calcule os coeficientes de correlação de Pearson para os dados dos dois conjuntos a seguir. Discuta a razão de os valores de r serem tão diferentes, embora os dados sejam tão semelhantes.

Dois conjuntos de pares de valores de duas variáveis

Conjunto A		Conjunto B	
X	Y	X	Y
1	2	1	2
2	4	2	4
3	6	3	6
4	8	4	8
5	10	5	0

7. O volume máximo de oxigênio inalado VO_2 máx tem sido usado como medida da situação cardíaca tanto de indivíduos saudáveis como de pessoas que sofrem de doenças cardíacas. Os dados de VO_2 máx em mililitros por quilograma por minuto para 12 homens saudáveis depois dos exercícios estão na tabela. Desenhe o diagrama de dispersão. Olhando o diagrama de dispersão, você diria que VO_2 máx diminui quando aumenta a atividade? r = –0,93 / forte negativa

Duração do exercício	VO_2 máx
10	82
9,5	73
10,2	68
10,5	74
11	66
11,3	63
11,6	58
12	54
12,1	56
12,5	51
12,8	55
13	44

8. Verificar a correlação existente entre as variáveis a seguir:

Indivíduo	Número de erros (X)	Horas de sono (Y)
1	8	12
2	7	13
3	9	9
4	12	6
5	14	5
Média	10	9
Desvio padrão	2,92	3,54

NOÇÕES DE REGRESSÃO LINEAR SIMPLES

Algumas vezes estamos interessados não apenas saber se existe associação entre duas variáveis quantitativas x e y, mas também uma hipótese a respeito de uma provável relação de causa e efeito entre variáveis. Desejamos saber se y "depende" de x. Neste caso, y é chamado de variável dependente ou variável resposta e x é chamado de variável independente ou explanatória que, na linguagem epidemiológica, é denominada "fator de risco".

A regressão é usada basicamente com duas finalidades: de previsão (prever o valor de y a partir do valor de x) e estimar o quanto x influencia ou modifica y.

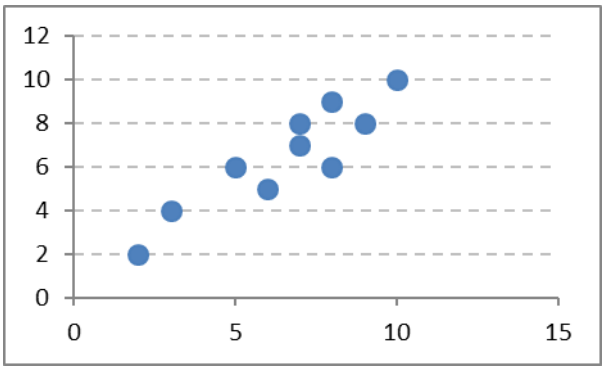
RETA DE REGRESSÃO

A variação de Y em função de X dever ser observada no gráfico. Se os pontos ficam dispersos em torno de uma reta, é razoável traçar uma reta no meio desses pontos. A melhor reta (melhor, no sentido que tem propriedades estatísticas) recebe o nome de *reta de regressão*.

Sejam duas variáveis X e Y, entre as quais exista uma correlação, por exemplo:

X	5	8	7	10	6	7	9	3	8	2
Y	6	9	8	10	5	7	8	4	6	2

Cujo diagrama de dispersão é dado por:



Podemos concluir, pela forma do diagrama, que se trata de uma correlação linear, portanto, pode ser descrita por meio de uma reta de regressão, dada uma função definida por:

$$Y = a \cdot X + b$$

Para ajustar uma reta de regressão, é preciso obter o *coeficiente linear* (***b***) e o *coeficiente*

angular (***a***) da reta, também chamados coeficientes de regressão.

Em estatística, o coeficiente angular da reta é obtido por meio da fórmula:

$$a = \frac{n \cdot \sum x \cdot y - \sum x \cdot \sum y}{n \cdot \sum x^2 - (\sum x)^2}$$

e o coeficiente linear é obtido por meio da fórmula:

$$b = \bar{y} - a \cdot \bar{x}$$

em que:

n é o número de observações;

$\bar{x}$ é a média dos valores de x;

$\bar{y}$ é a média dos valores de y.

Vamos, então, calcular os valores dos parâmetros ***a*** e ***b*** com a ajuda das fórmulas. Formemos a tabela de valores:

x	y	x.y	x^2
5	6	30	25
8	9	72	64
7	8	56	49
10	10	100	100
6	5	30	36
7	7	49	49
9	8	72	81
3	4	12	9
8	6	48	64
2	2	4	4
Σ = 65	Σ = 65	Σ = 473	Σ = 481

Temos assim:

$$a = \frac{10 \cdot 473 - 65 \cdot 65}{10 \cdot 481 - (65)^2} = \frac{4730 - 4225}{4810 - 4225} = 0{,}86$$

Calculando o valor das médias:

$$\bar{x} = \frac{65}{10} = 6{,}5 \qquad \bar{y} = \frac{65}{10} = 6{,}5$$

Calculando o coeficiente linear ***b***:

$$b = 6{,}5 - 0{,}86 \cdot 6{,}5 = 6{,}5 - 5{,}6108 = 0{,}89$$

Logo, com o coeficiente angular (***a***) e o coeficiente linear (***b***):

$$\hat{Y} = a \cdot x + b$$

$$\hat{Y} = 0{,}86 \cdot x + 0{,}89$$

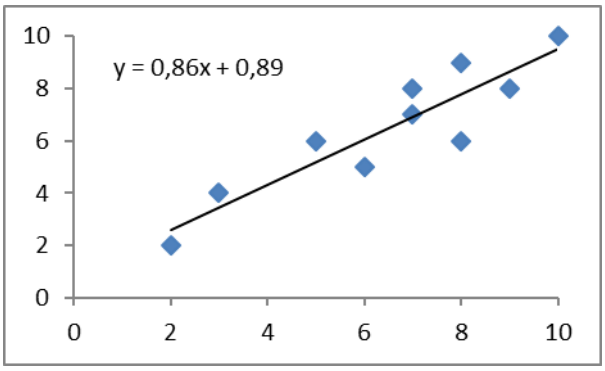
Nota: como estamos fazendo uso de uma amostra, o resultado, na realidade, é uma estimativa da verdadeira equação de regressão. Sendo assim, escrevemos: $\hat{Y} = a \cdot x + b$, onde $\hat{Y}$ é o Y estimado.

Para traçarmos a reta no gráfico, basta determinar dois de seus pontos:

$$X = 0 \rightarrow \hat{Y} = 0{,}89$$

$$X = 5 \rightarrow \hat{Y} = 5{,}19$$

Assim temos,



EXERCÍCIOS

1. A tabela abaixo apresenta o número de exames realizados anualmente por um laboratório:

Ano	Qtd.(mil)
2000	34
2001	36
2002	36
2003	38
2004	41
2005	42
2006	43
2007	44
2008	46

Calcule:

a. a reta ajustada; y = 1,50.x – 32,50
b. a produção estimada para 2009. 47,50

2. A tabela seguinte constitui uma amostra aleatória referente às alturas, em centímetros, de 10 meninas e suas respectivas mães:

altura das mães	altura das filhas
165	168
169	170
166	168
158	160
172	172
160	164
167	165
155	158
169	172
166	164

Calcule:

a. a reta ajustada; y = 0,816.x + 31,66
b. preveja a altura de uma menina cuja mãe tenha 150 cm de altura. 154,06 cm

LISTA – Correlação e Regressão linear

Nome:.....

Curso:.....

1. A partir da tabela:

X	1	2	3	4	5	6
Y	70	50	40	30	20	10

a. calcule o coeficiente de correlação; r = – 0,989/ forte negativa
b. determine a reta ajustada; y = – 11,42.x + 76,66
c. estime o valor de y para x = 0. y = 76,66

2. Certa empresa, estudando a variação da demanda de seu produto em relação à variação de preço de venda, obteve a tabela:

Preço (X)	Demanda (Y)
38	350
42	325
50	297
56	270
59	256
63	246
70	238
80	223
95	215
110	208

a. determine o coeficiente de correlação; r = –0,90/ forte negativa
b. estabeleça a equação da reta ajustada; y = – 1,87.x + 386,8
c. estime y para x = 60 e x = 120. 274,6 e 162,4

3. A tabela seguinte apresenta os graus finais de Estatística e de Bioquímica obtidos por 10 estudantes, selecionados ao acaso entre um grande grupo de estudantes.
   a. determine o coeficiente de correlação; r = 0,87/ forte positiva
   b. determinar a reta ajustada aos dados, adotado X como variável independente. y = 0,661.x + 29,12
   c. se um estudante obteve grau 75 em Estatística, qual é seu grau esperado em Bioquímica? y = 78,69
   d. se um estudante obteve grau 75 em Bioquímica, qual é seu grau esperado em Estatística? x = 69,4

Estatística (X)	Bioquímica (Y)
75	82
80	78
93	86
65	72
87	91
71	80
98	95
68	72
84	89
77	74

4. A tabela a seguir apresenta o número de unidades produzidas (P) por 10 operadores de uma fábrica e o número de unidades produzidas com defeitos (D).

Operador (*i*)	Produção (P_i)	Defeituosa (D_i)
1	94	4
2	98	5
3	106	6
4	114	7
5	107	6
6	93	5
7	98	6
8	88	4
9	103	7
10	95	5

O coeficiente de correlação entre P e D é:

a. – 0,855
b. – 0,731
c. 0,0000
d. 0,855
e. 0,731

5. Para se verificar se existe relação linear entre umidade relativa (UR) do ar de secagem de sementes e a germinação das mesmas, um pesquisador realizou um experimento com 4 valores diferentes para a %UR do ar, obtendo-se os seguintes dados (dados hipotéticos):

Umidade relativa	Germinação
20	94
30	96
40	95
50	97

a. Verificar se existe correlação da UR do ar de secagem na % de germinação (r). r = 0,8/ forte positiva
b. Verificar se existe efeito da UR do ar de secagem na % de germinação (reta ajustada para o modelo). y = 0,08.x + 92,7
c. Estimar qual seria a % de germinação esperada quando UR=45%? 96,3

REFERÊNCIAS BIBLIOGRÁFICAS

STEVENSON, Willian J. Estatística aplicada à administração. São Paulo: Haper, 1981.

FREUND, John. Estatística aplicada à economia, administração e contabilidade. Porto Alegre: Bookman, 2000.

CRESPO, Antônio Arnot. Estatística Fácil. 19 ed. São Paulo: Ed. Saraiva, 2009.

VIEIRA, Sônia. Introdução à bioestatística. 4 ed. Rio de Janeiro: Elsevier, 2008.

NOVAES, D. V.; COUTINHO, C. Q. S. Estatística para educação profissional. São Paulo: Atlas, 2009.

MORETTIN, Luiz Gonzaga. Estatística básica – Probabilidade. São Paulo: Makron Books, 2005.

SPIEGEL, Murray R. Estatística. São Paulo: Makron Books, 1993.

OLIVEIRA, Francisco. E. M. Estatística e Probabilidade. São Paulo: Atlas, 1995.

BUSSAD, Wilton de O. Estatística Básica. 5 ed. São Paulo: Ed. Saraiva, 2004.

BITTENCOURT, Hélio Radke. Apostila Bioestatística. PUC-RS.