<u>CORRELAÇÃO</u> – “Aqui me tens de regresso”

O assunto Correlação fez parte, acompanhado de Regressão, do programa de Auditor Fiscal, até 1998, desaparecendo a partir do concurso do ano 2000 para agora retornar sozinho.

Mesmo sabendo que Regressão não faz mais parte do programa, devemos, pelo menos, entender a diferença entre Correlação e Regressão, que são assuntos próximos, mas não são a mesma coisa.

- A CORRELAÇÃO mede a força, a intensidade ou grau de relacionamento entre duas ou mais variáveis.
- A REGRESSÃO fornece uma equação que descreve esse relacionamento em termos matemáticos.

Para um modelo linear simples podemos encontrar essa equação (chamada de equação da reta ajustante e igual a: $\hat{Y} = \hat{\alpha} + \hat{\beta}X$) pelo Método dos Mínimos Quadrados e através dessa equação poder estimar valores não observados da variável Y. Os valores de β (coeficiente angular da reta) e de α (intercepto) serão calculados com base nas observações de X e de Y.

É claro que só poderemos estabelecer um modelo de regressão entre duas (regressão simples) ou mais (regressão múltipla) variáveis se estas forem <u>dependentes</u>, ou seja, o valor de Y depende do valor de X e nesse caso existe correlação entre X e Y (correlação diferente de zero), pois as variáveis se relacionarão.

Se as variáveis forem <u>independentes</u>, (não há relacionamento entre elas), não existirá correlação (correlação nula ou ausência de correlação) e nesse caso não poderemos estabelecer um modelo de regressão.

A Correlação entre duas variáveis pode ser:

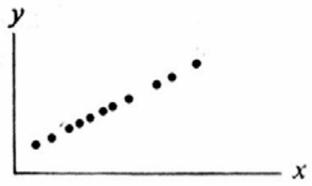
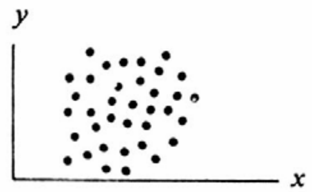
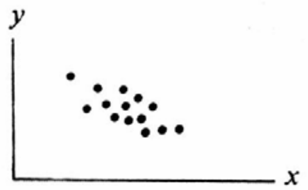
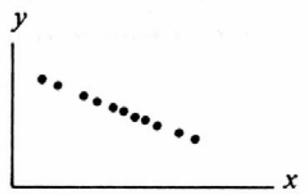
1) **<u>POSITIVA</u>** (Correlação Direta) → Quando, para valores altos de uma variável corresponderão valores altos para outra e para valores baixos de uma, associaremos também valores baixos para outra. Nesse caso, o valor do Coeficiente de Correlação entre X e Y (ρ_{xy}) estará entre 0 (exclusive) e 1 (inclusive), ou seja, $0 < \rho_{xy} \le 1$. Quando ρ_{xy} = 1, dizemos que a Correlação entre X e Y é Direta e Perfeita. Assim, para cada aumento na variável X, a variável Y aumentará na mesma proporção. Exemplos de correlação direta: salário e investimento em poupança (em geral quanto maior for o salário maior será o valor poupado); idade e pressão arterial (em geral, as pessoas mais idosas têm maior pressão arterial); nota em Matemática e nota em Estatística (geralmente as pessoas com maior dificuldade em Matemática, notas baixas, terão maior dificuldade em Estatística, apresentando também, notas baixas);

2) **<u>NEGATIVA</u>** (Correlação Inversa) → Quando as variáveis têm sentidos opostos, ou seja, à medida que X aumenta, o valor de Y diminui. Nesse caso, o valor do Coeficiente de Correlação entre X e Y (ρ_{xy}) estará entre – 1 (inclusive) e 0 (exclusive), ou seja, $-1 \le \rho_{xy} < 0$.

Quando $\rho_{xy} = -1$, dizemos que a Correlação entre X e Y é perfeitamente inversa e assim, a cada aumento da variável X, a variável Y diminuirá na mesma proporção. Um exemplo de Correlação Inversa: considerando automóveis de mesmo ano, marca e modelo, quanto maior for a quilometragem do veículo, menor será o preço de venda.

3) **<u>NULA</u>** (ausência de correlação) → Quando não é possível estabelecer uma relação entre as variáveis X e Y, e nesse caso $\rho_{xy} = 0$. É o que ocorre quando as VARIÁVEIS forem INDEPENDENTES.

Analisando os três casos possíveis com relação à Correlação, vemos que o seu coeficiente (ρ_{xy}) varia apenas no intervalo de – 1 até 1, ou seja, ρ_{xy} = [–1; +1]. A Correlação será forte (positiva ou negativamente) quando estiver próxima de 1 ou de –1, e fraca quando estiver próxima de zero. Abaixo, temos os diagramas de dispersão para os três tipos de correlação descritos anteriormente:

Valor de *r*	Descrição do relacionamento linear	Diagrama de dispersão
+1,00	Relacionamento positivo, perfeito	
Cerca de +0,70	Relacionamento positivo, moderado	
0,00	Ausência de relacionamento	
Cerca de –0,70	Relacionamento negativo, moderado	
–1,00	Relacionamento negativo, perfeito	

Extraídos do livro Estatística Aplicada à Administração – William J. Stevenson – Editora Harbra.

Uma fórmula para encontrar o valor do Coeficiente de Correlação Linear Simples é dada por:

$\rho_{xy} = \dfrac{n\sum XY - \sum X \cdot \sum Y}{\sqrt{n\sum X^2 - \left(\sum X\right)^2} \cdot \sqrt{n\sum Y^2 - \left(\sum Y\right)^2}}$, onde n é o número de pares de observações.

Mas essa fórmula pode ser resumida por: $\boxed{\rho_{xy} = \dfrac{COV(x,y)}{\sigma_x \cdot \sigma_y}}$ ou seja:

<u>É a Covariância entre X e Y dividida pelo produto dos desvios padrões de X e de Y</u>.

Observe que, na 1ª fórmula dada, no denominador temos:

$\sqrt{n\sum X^2 - \left(\sum X\right)^2}$ que nada mais é do que $n \cdot \sigma_x$; e

$\sqrt{n\sum Y^2 - \left(\sum Y\right)^2}$ que nada mais é do que $n \cdot \sigma_y$;

Observe agora que o numerador da 1ª fórmula é:

$n\sum XY - \sum X \cdot \sum Y$ que nada mais é do que $n^2 \cdot COV(x,y)$.

Então, substituindo $n\sum XY - \sum X \cdot \sum Y$ <u>por</u> $n^2 \cdot COV(x,y)$, $\sqrt{n\sum X^2 - \left(\sum X\right)^2}$ <u>por</u> $n \cdot \sigma_x$ e $\sqrt{n\sum Y^2 - \left(\sum Y\right)^2}$ <u>por</u> $n \cdot \sigma_y$ teremos: $\rho_{xy} = \dfrac{n^2 \cdot COV(x,y)}{n \cdot \sigma_x \cdot n \cdot \sigma_y}$.

Dividindo o numerador e o denominador por n^2, fica: $\rho_{xy} = \dfrac{COV(x,y)}{\sigma_x \cdot \sigma_y}$, que é bem mais fácil de guardar do que a 1ª fórmula. Então, é importante saber que a Covariância entre X e Y nada mais é do que a <u>esperança conjunta de X e Y</u> menos o <u>produto das esperanças individuais</u>, ou seja: $\boxed{COV(x,y) = E[XY] - E[X] \cdot E[Y]}$.

<u>Provando que as transformações estão corretas</u>:

$E[XY] = \dfrac{\sum XY}{n}$; $E[X] = \dfrac{\sum X}{n}$ e $E[Y] = \dfrac{\sum Y}{n}$.

Assim, $COV(x,y) = \dfrac{\sum XY}{n} - \dfrac{\sum X}{n} \cdot \dfrac{\sum Y}{n} \Rightarrow COV(x,y) = \dfrac{\sum XY}{n} - \dfrac{\sum X \cdot \sum Y}{n^2}$.

Multiplicando por n^2 ambos os membros da igualdade:

$n^2 \cdot COV(x,y) = n^2 \cdot \left(\dfrac{\sum XY}{n} - \dfrac{\sum X \cdot \sum Y}{n^2}\right) \Rightarrow \boxed{n^2 \cdot COV(x,y) = n\sum XY - \sum X \cdot \sum Y}$

A variância de X é dada por: $\sigma_x^2 = \dfrac{1}{n}\left\{\sum X^2 - \dfrac{\left(\sum X\right)^2}{n}\right\}$.

Multiplicando por n^2 ambos os membros da igualdade fica: $n^2 \cdot \sigma_x^2 = n^2 \cdot \dfrac{1}{n}\left\{\sum X^2 - \dfrac{\left(\sum X\right)^2}{n}\right\}$

Logo: $n^2 \cdot \sigma_x^2 = n \cdot \left\{\sum X^2 - \dfrac{\left(\sum X\right)^2}{n}\right\} \Rightarrow n^2 \cdot \sigma_x^2 = n\sum X^2 - \left(\sum X\right)^2$

Portanto, o desvio padrão de X será: $\boxed{n \cdot \sigma_x = \sqrt{n\sum X^2 - \left(\sum X\right)^2}}$

O mesmo procedimento será válido para provar que: $\boxed{n \cdot \sigma_y = \sqrt{n\sum Y^2 - \left(\sum Y\right)^2}}$

Qual das duas fórmulas utilizar?

$$\rho_{xy} = \frac{n\sum XY - \sum X \cdot \sum Y}{\sqrt{n\sum X^2 - \left(\sum X\right)^2} \cdot \sqrt{n\sum Y^2 - \left(\sum Y\right)^2}} \ \underline{\text{ou}} \ \rho_{xy} = \frac{COV(x,y)}{\sigma_x \cdot \sigma_y} \ ?$$

Como já comentado anteriormente, ambas são equivalentes, mas a segunda é bem mais fácil de guardar do que a primeira. Para decidir qual das duas usar, dependerá dos dados fornecidos na questão que for proposta, mas em qualquer uma das duas fórmulas o ideal é começar calculando o numerador, pois se as variáveis forem independentes, a covariância será igual a zero e podemos responder que a correlação também será zero, sem necessidade de cálculo das variâncias de X e de Y, o que já evitará trabalho e perda de tempo.

Veja na segunda fórmula, $\rho_{xy} = \dfrac{COV(x,y)}{\sigma_x \cdot \sigma_y}$, que o denominador será sempre positivo, pois cada desvio padrão é a raiz quadrada positiva da variância (não existe desvio padrão negativo). Ora, o produto de dois valores positivos sempre será positivo. O que irá determinar se a Correlação será positiva, negativa ou nula, será a Covariância, pois se:

1) COV(x,y) > 0 $\Rightarrow \ \rho_{xy} = \dfrac{+}{+} \ \Rightarrow \boxed{\rho_{xy} > 0}$

2) COV(x,y) = 0 $\Rightarrow \ \rho_{xy} = \dfrac{0}{+} \ \Rightarrow \boxed{\rho_{xy} = 0}$

3) COV(x,y) < 0 $\Rightarrow \ \rho_{xy} = \dfrac{-}{+} \ \Rightarrow \boxed{\rho_{xy} < 0}$

E quando a covariância será positiva, negativa ou nula?

Dependerá do valor da esperança conjunta e dos valores das esperanças individuais, pois sabemos que: $COV(x,y) = E[XY] - E[X] \cdot E[Y]$

Logo, se:

$E[XY] > E[X] \cdot E[Y] \Rightarrow COV(x,y) > 0$

$E[XY] = E[X] \cdot E[Y] \Rightarrow COV(x,y) = 0$

$E[XY] < E[X] \cdot E[Y] \Rightarrow COV(x,y) < 0$

Embora o nosso foco principal seja Correlação, é importante saber que, quando $\rho_{xy} \neq 0$, podemos estabelecer a reta de regressão, dada por: $\hat{Y} = \hat{\alpha} + \hat{\beta}X$. Quando $\rho_{xy} > 0$, o valor do coeficiente angular da reta ($\hat{\beta}$) também será positivo e a reta terá inclinação para cima. Quando $\rho_{xy} < 0$, o valor do coeficiente angular da reta ($\hat{\beta}$) também será negativo e a reta terá inclinação para baixo (veja os diagramas de dispersão mostrados anteriormente).

Vejamos um exemplo numérico para melhor entendimento:

Abaixo temos uma tabela com as notas obtidas por 10 alunos em Matemática (variável X) e Estatística (variável Y):

ALUNO	X	Y
A	6	7
B	5	6
C	9	10
D	10	9
E	3	2
F	4	3
G	8	9
H	7	5
I	6	6
J	2	3

Sem fazer cálculo algum, podemos entender que deverá haver relação entre essas variáveis, ou seja, o aluno que tem dificuldade em Matemática deverá ter dificuldade também em Estatística, que é uma disciplina que depende essencialmente da Matemática. Portanto para valores altos de X teremos associados valores altos em Y e para valores baixos em X valores baixos em Y (correlação direta).

Vejamos então se a covariância é diferente de zero, lembrando que $COV(x,y) = E[XY] - E[X] \cdot E[Y]$. Lembrando ainda que: $E[XY] = \frac{\sum XY}{n}$; $E[X] = \frac{\sum X}{n}$ e $E[Y] = \frac{\sum Y}{n}$.

Vamos então calcular, na tabela dada, esses somatórios:

ALUNO	X	Y	XY
A	6	7	42
B	5	6	30
C	9	10	90
D	10	9	90
E	3	2	6
F	4	3	12
G	8	9	72
H	7	5	35
I	6	6	36
J	2	3	6
TOTAL	60	60	419

Assim, $E[XY] = \frac{419}{10} = 41{,}9$; $E[X] = \frac{60}{10} = 6$ e $E[Y] = \frac{60}{10} = 6$

Portanto, $COV(x,y) = 41{,}9 - 6 \cdot 6 \Rightarrow COV(x,y) = 41{,}9 - 36 \Rightarrow COV(x,y) = 5{,}9$.

Observamos que, como já havíamos previsto, há uma correlação direta entre X e Y, pois COV(x,y) > 0 e assim sendo, $\boxed{\rho_{xy} > 0}$.

Vamos então avaliar o grau, a força ou intensidade desse relacionamento calculando o Coeficiente de Correlação entre X e Y. Se tivéssemos obtido COV(x,y) = 0 nem calcularíamos as variâncias, pois se o numerador fosse igual a zero, ρ_{xy} seria igual a zero.

Precisamos agora, calcular as variâncias de X e de Y para obter os seus desvios padrões. Completando a tabela anterior com os quadrados de X e de Y, temos:

ALUNO	X	Y	XY	X^2	Y^2
A	6	7	42	36	49
B	5	6	30	25	36
C	9	10	90	81	100
D	10	9	90	100	81
E	3	2	6	9	4
F	4	3	12	16	9
G	8	9	72	64	81
H	7	5	35	49	25
I	6	6	36	36	36
J	2	3	6	4	9
TOTAL	60	60	419	420	430

$$V[X] = \frac{1}{n}\left\{\sum X^2 - \frac{(\sum X)^2}{n}\right\} \Rightarrow V[X] = \frac{1}{10}\left\{420 - \frac{(60)^2}{10}\right\} \Rightarrow V[X] = \frac{1}{10}\{420 - 360\} \Rightarrow V[X] = \frac{60}{10} = \mathbf{6}.$$

$$V[Y] = \frac{1}{n}\left\{\sum Y^2 - \frac{(\sum Y)^2}{n}\right\} \Rightarrow V[Y] = \frac{1}{10}\left\{430 - \frac{(60)^2}{10}\right\} \Rightarrow V[Y] = \frac{1}{10}\{430 - 360\} \Rightarrow V[Y] = \frac{70}{10} = \mathbf{7}.$$

Podemos agora calcular o Coeficiente de Correlação, utilizando a fórmula mais simples $\rho_{xy} = \frac{COV(x,y)}{\sigma_x \cdot \sigma_y} \Rightarrow \rho_{xy} = \frac{5,9}{\sqrt{6}\cdot\sqrt{7}} \Rightarrow \rho_{xy} = \frac{5,9}{\sqrt{42}} \Rightarrow \rho_{xy} = 0,91039$, o que indica haver uma <u>forte correlação</u> entre estas variáveis.

Considerando que o assunto Regressão não faz parte do programa de Auditor Fiscal, mas apenas Correlação, não vamos nos aprofundar no assunto, mas é importante saber que, como já dito no início, a equação de regressão ($\hat{Y} = \hat{\alpha} + \hat{\beta}X$) permitirá fazermos estimativas para valores que constam e, principalmente, para valores que não constam do conjunto de observações. É importante notar também que, para calcular o valor de $\hat{\alpha}$ (intercepto), precisaremos calcular primeiro o valor de $\hat{\beta}$. Para calcular o $\hat{\beta}$ da equação, temos a fórmula:

$\boxed{\hat{\beta} = \frac{n\sum XY - \sum X \cdot \sum Y}{n\sum X^2 - (\sum X)^2}}$. Mas, se já calculamos o valor do Coeficiente de Correlação e os valores dos desvios de X e de Y, fica bem mais fácil usar uma fórmula equivalente:

$\boxed{\hat{\beta} = \rho_{xy} \cdot \frac{\sigma_y}{\sigma_x}}$. No exemplo dado, $\hat{\beta}$ será igual a: $\hat{\beta} = 0,91 \cdot \frac{\sqrt{7}}{\sqrt{6}} \Rightarrow \boxed{\hat{\beta} \cong 0,98}$

Conhecendo o valor de $\hat{\beta}$, podemos encontrar o valor de $\hat{\alpha}$ fazendo (pelas propriedades da média): $\overline{Y} = \hat{\alpha} + \hat{\beta}\overline{X} \Rightarrow \boxed{\hat{\alpha} = \overline{Y} - \hat{\beta}\overline{X}}$. Ou usar a fórmula equivalente: $\hat{\alpha} = \frac{\sum Y - \beta \sum X}{n}$.

No nosso exemplo, $\overline{Y} = 6$, $\overline{X} = 6$ e $\hat{\beta} = 0,98$. Logo: $\hat{\alpha} = 6 - 0,98 \cdot 6 \Rightarrow \boxed{\hat{\alpha} = 0,12}$

Assim, a equação da reta de regressão será dada por: $\boxed{\hat{Y} = 0,12 + 0,98X}$.

Com esta equação obteremos as seguintes estimativas para Y em função do valor de X:

ALUNO	X foi igual a:	A estimativa para Y será:
A	6	$\hat{Y} = 0{,}12 + (0{,}98 \cdot 6) \Rightarrow \hat{Y} =$ **6,00**
B	5	$\hat{Y} = 0{,}12 + (0{,}98 \cdot 5) \Rightarrow \hat{Y} =$ **5,02**
C	9	$\hat{Y} = 0{,}12 + (0{,}98 \cdot 9) \Rightarrow \hat{Y} =$ **8,94**
D	10	$\hat{Y} = 0{,}12 + (0{,}98 \cdot 10) \Rightarrow \hat{Y} =$ **9,92**
E	3	$\hat{Y} = 0{,}12 + (0{,}98 \cdot 3) \Rightarrow \hat{Y} =$ **3,06**
F	4	$\hat{Y} = 0{,}12 + (0{,}98 \cdot 4) \Rightarrow \hat{Y} =$ **4,04**
G	8	$\hat{Y} = 0{,}12 + (0{,}98 \cdot 8) \Rightarrow \hat{Y} =$ **7,96**
H	7	$\hat{Y} = 0{,}12 + (0{,}98 \cdot 7) \Rightarrow \hat{Y} =$ **6,98**
I	6	$\hat{Y} = 0{,}12 + (0{,}98 \cdot 6) \Rightarrow \hat{Y} =$ **6,00**
J	2	$\hat{Y} = 0{,}12 + (0{,}98 \cdot 2) \Rightarrow \hat{Y} =$ **2,08**

Um dos pressupostos básicos da regressão é que o valor esperado dos resíduos (diferença entre o valor real e o valor estimado) seja igual a <u>zero</u>. Vamos verificar que isto ocorre no nosso exemplo, pois:

Aluno	X	Y (real)	$\hat{Y}$ (estimado)	Resíduos: $Y - \hat{Y}$
A	6	7	6	1,00
B	5	6	5,02	0,98
C	9	10	8,94	1,06
D	10	9	9,92	–0,92
E	3	2	3,06	–1,06
F	4	3	4,04	–1,04
G	8	9	7,96	1,04
H	7	5	6,98	–1,98
I	6	6	6	0,00
J	2	3	2,08	0,92
			SOMA DOS RESÍDUOS ➔	**0,00**

Usando a equação que encontramos para a reta, podemos fazer diversas estimativas, como por exemplo, podemos estimar que:

a) Um aluno que tirou 1 em Matemática obteria:

$\hat{Y} = 0{,}12 + (0{,}98 \cdot 1)$ = 1,10 em Estatística;

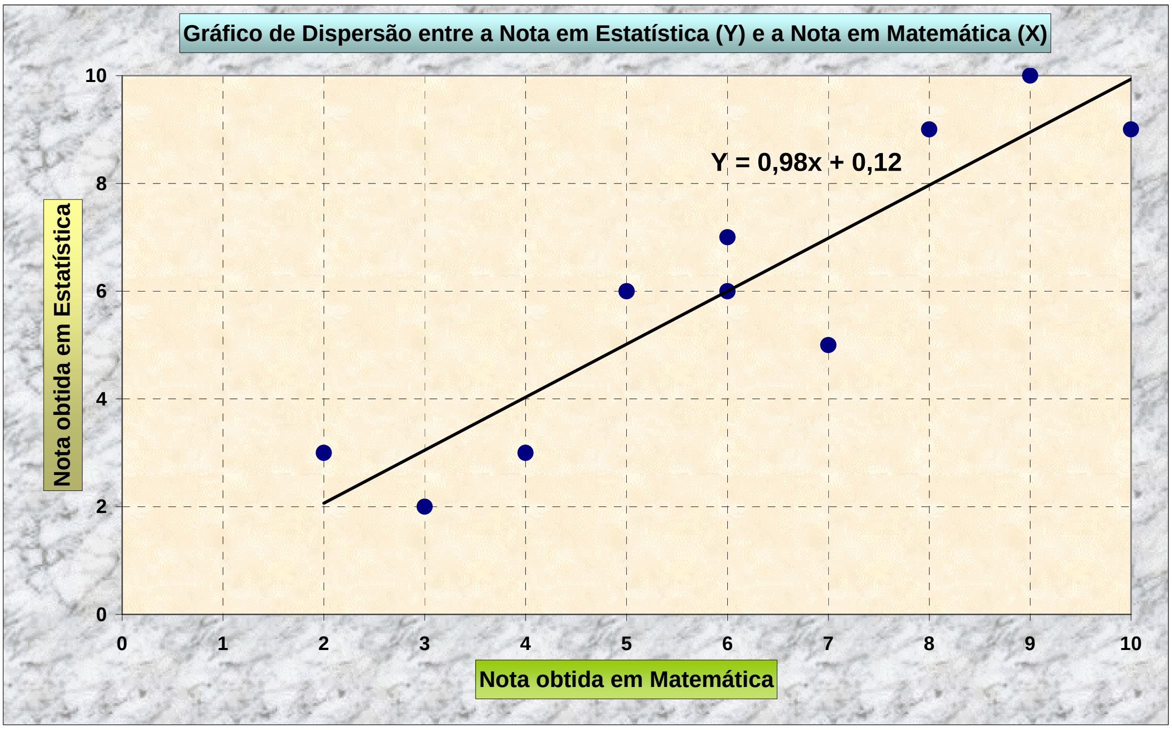
b) Um aluno que tirou 3,5 em Matemática obteria:

$\hat{Y} = 0{,}12 + (0{,}98 \cdot 3{,}5)$ = 3,55 em Estatística;

c) Um aluno que tirou 8,5 em Matemática obteria:

$\hat{Y} = 0{,}12 + (0{,}98 \cdot 8{,}5)$ = 8,45 em Estatística;

A seguir, vemos o gráfico de dispersão para os valores observados de X e de Y e os valores estimados pela reta de regressão:



Neste exemplo, fica difícil buscar valores a estimar porque a variável nota, em geral, fica limitada entre zero e dez.

Vamos utilizar outro exemplo, apenas para fazer estimativas, sendo já fornecidos o Coeficiente de Correlação e a equação de regressão.

A tabela abaixo indica as idades (X) e as pressões arteriais (Y) de 12 mulheres:

IDADE (X)	56	42	72	36	63	47	55	49	38	42	68	60
PRESSÃO ARTERIAL (Y)	147	125	160	118	149	128	150	145	115	140	152	155

Coeficiente de Correlação: $\rho_{xy} = 0{,}8961$. Isto mostra um forte relacionamento entre as variáveis X e Y.

Reta de regressão: $\hat{Y} = 80{,}78 + 1{,}138X$

Com base na equação de regressão, quais as estimativas de pressão arterial para:

a) Uma mulher de 45 anos?

b) Uma mulher de 30 anos?

c) Uma mulher de 70 anos?

RESPOSTAS:

a) 132 (131,988);

b) 115 (114,9179);

c) 160 (160,4381).

ALGUMAS QUESTÕES - EXEMPLOS INTERESSANTES SOBRE O ASSUNTO:

Questão 1 [BACEN-98] Duas variáveis X e Y têm coeficiente de correlação linear igual a 0,9. Obtendo-se a reta de regressão linear simples de Y sobre X, pode-se dizer que seu coeficiente angular:

(a) Será menor que 0,9
(b) Será maior que 0,9
(c) Poderá ser negativo
(d) Poderá ser nulo
(e) Será positivo

RESPOSTA: LETRA E. Esta é extremamente fácil, mas é só para começar. É claro que, se o coeficiente de correlação linear for positivo o coeficiente angular também será positivo, pois a reta de regressão terá inclinação para cima.

Questão 2 [IBGE-99] Se X é uma variável aleatória e Y = 5 – 2X, então o coeficiente de correlação linear entre X e Y é igual a:

(a) 2,5
(b) 1,0
(c) 0
(d) –0,4
(e) –1,0

RESPOSTA: LETRA E. Também é uma questão fácil. Basta ver que, se substituirmos X, na equação dada, pelos valores 0, 1, 2, 3 por exemplo, vamos obter valores de Y iguais a, respectivamente, 5, 3, 1, –1, ou seja, esses pontos formarão uma reta inclinada para baixo e a cada aumento de 1 unidade em X, teremos uma redução de 2 unidades em Y, o que nos mostra que há uma relação perfeitamente inversa. Logo o Coeficiente de Correlação só pode ser –1,0.

Questão 3 [SUSEP-94] Se as variáveis aleatórias **X** e **Y** são tais que **Y** = **2X**, o coeficiente de correlação linear γ entre **X** e **Y** é tal que:

(a) $\gamma = 1$
(b) $\gamma = 0$
(c) $\gamma = -1$
(d) $0 < \gamma < 1$
(e) $-1 < \gamma < 0$

RESPOSTA: LETRA A. Nem é preciso dizer que o raciocínio é idêntico ao da questão anterior, com a diferença que, nesta, a cada aumento de 1 unidade em X, teremos um aumento de 2 unidades em Y, o que nos mostrará uma relação perfeitamente direta entre X e Y. Logo o Coeficiente de Correlação só pode ser igual a 1.

Questão 4 [SUSEP-98] Considere X e Y duas variáveis aleatórias com variâncias de 4 e 1, respectivamente, e coeficiente de correlação igual a 1/4. A variância de Z = (X + Y) é:

(a) 5
(b) 6
(c) 7
(d) 41/8
(e) 21/4

RESPOSTA: LETRA B. Ainda é uma questão fácil, mas não tão óbvia como as anteriores. Veja que, se as variâncias de X e Y são, respectivamente, 4 e 1, os desvios padrões serão respectivamente 2 e 1. Agora, peguemos aquela fórmula do Coeficiente de Correlação em função da Covariância dividida pelo produto dos desvios: $\boxed{\rho_{xy} = \frac{COV(x,y)}{\sigma_x \cdot \sigma_y}}$. O valor do Coeficiente de Correlação (rô), foi dado no enunciado e é igual a 1/4. Substituindo na fórmula, encontraremos o valor da Covariância, pois: $\frac{1}{4} = \frac{COV(x,y)}{2 \cdot 1} \Rightarrow COV(x,y) = \frac{1}{2}$. Agora veja (à página 41 do meu livro "Estatística Básica para Concursos") que a Variância da soma de duas variáveis aleatórias é dada por: V[X **+** Y] = V[X] **+** V[Y] **+** 2cov(x,y). A variável Z é definida como sendo a soma das variáveis X e Y, logo, V[Z] = V[X + Y] = 4 + 1 + $2 \cdot \frac{1}{2}$ = 4 + 1 + 1 = **6**.

Questão 5 [AFC-94] A tabela abaixo 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 (Pi)	Defeituosa (Di)
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

Da tabela foram obtidos os seguintes valores:

$$\sum_{i=1}^{10} P_i = 996 \qquad \sum_{i=1}^{10} P_i^2 = 99.752 \qquad \sum_{i=1}^{10} \left(P_i - \bar{P}\right)^2 = 550,4 \qquad \sum_{i=1}^{10} \left(P_i - \bar{P}\right) \cdot \left(D_i - \bar{D}\right) = 65$$

$$\sum_{i=1}^{10} D_i = 55 \qquad \sum_{i=1}^{10} D_i^2 = 313 \qquad \sum_{i=1}^{10} \left(D_i - \bar{D}\right)^2 = 10,5 \qquad \sum_{i=1}^{10} P_i D_i = 5.543$$

O coeficiente de correlação linear entre P e D é:

(a) –0,855
(b) –0,731
(c) 0,000
(d) 0,731
(e) 0,855

<u>RESPOSTA</u>: LETRA E. Nesta questão, como são fornecidos os valores de todos os somatórios possíveis, fica melhor de usar a primeira fórmula dada para encontrar o valor do Coeficiente de Correlação entre P e D: $\rho_{PD} = \dfrac{n\sum PD - \sum P \cdot \sum D}{\sqrt{n\sum P^2 - \left(\sum P\right)^2} \cdot \sqrt{n\sum D^2 - \left(\sum D\right)^2}}$ e encontrar 0,85503.

Questão 6 [IBGE-2002] **X** e **Y** são duas variáveis aleatórias com variâncias 144 e 64 respectivamente. Assinale o item que **NÃO** indica um valor possível para a covariância entre X e Y.

(a) 87,5
(b) 18,7
(c) 0
(d) –0,3
(e) –100

<u>RESPOSTA</u>: LETRA E. Uma ótima questão, muito inteligente e bem bolada.

$$\left.\begin{array}{l} V[X] = 144 \Rightarrow \sigma_X = 12; \\ V[Y] = 64 \Rightarrow \sigma_y = 8; \end{array}\right\} \text{Temos então que: } \sigma_X \cdot \sigma_y = 96$$

Sabemos que o Coeficiente de Correlação (ρ_{xy}) varia apenas no intervalo [–1; 1] e que $\rho_{xy} = \dfrac{COV(x,y)}{\sigma_x \cdot \sigma_y}$. Logo, temos 2 hipóteses: $\rho_{xy} = -1$ ou $\rho_{xy} = 1$.

$$\left.\begin{array}{l} \text{Se } \rho_{xy} = -1 \Rightarrow COV(x,y) = \rho_{xy} \cdot \sigma_X \cdot \sigma_y \Rightarrow COV(x,y) = -96. \\ \text{Se } \rho_{xy} = 1 \Rightarrow COV(x,y) = \rho_{xy} \cdot \sigma_X \cdot \sigma_y \Rightarrow COV(x,y) = 96. \end{array}\right\} \Rightarrow COV(x,y) = \left[-96;96\right]$$

O único valor fora desse intervalo é –100.

Questão 7 [IBGE-2002] Os dados a seguir apresentam os investimentos (em milhares de reais) e os lucros (em milhares de reais) no ano seguinte realizados por cinco empresas escolhidas aleatoriamente:

Empresa	***Investimento***	***Lucro***
1	10	1,5
2	15	2,0
3	5	0,5
4	12	1,5
5	18	2,5

O coeficiente de correlação linear amostral destes dados é, aproximadamente, igual a:

(a) –0,74
(b) –0,26
(c) 0,48
(d) 0,72
(e) 0,98

<u>RESPOSTA</u>: LETRA E. Use uma das duas fórmulas para calcular o Coeficiente de Correlação entre o Investimento (X) e o Lucro (Y), e encontre aproximadamente 0,9875.

Questão 8 O coeficiente de correlação entre duas variáveis, X e Y, é r = 0,60. Se $S_x = 1,50$, $S_y = 2$, $\overline{X} = 10$ e $\overline{Y} = 20$, determinar as equações das retas de regressão de:

a) Y para X;
b) X para Y.

RESPOSTAS:

a) Y para X: $\hat{\beta} = r \cdot \frac{\sigma_y}{\sigma_x} \Rightarrow \hat{\beta} = 0,60 \cdot \frac{2}{1,50} \Rightarrow \hat{\beta} = \frac{1,20}{1,50} \Rightarrow \hat{\beta} = 0,80$

$\hat{\alpha} = \overline{Y} - \hat{\beta}\overline{X} \Rightarrow \hat{\alpha} = 20 - 0,8 \cdot 10 \Rightarrow \hat{\alpha} = 20 - 8 \Rightarrow \hat{\alpha} = 12$. Logo, $\boxed{\hat{Y} = 12 + 0,8X}$

b) X para Y: $\hat{\beta} = r \cdot \frac{\sigma_x}{\sigma_y} \Rightarrow \hat{\beta} = 0,60 \cdot \frac{1,50}{2} \Rightarrow \hat{\beta} = \frac{0,90}{2} \Rightarrow \hat{\beta} = 0,45$

$\hat{\alpha} = \overline{X} - \hat{\beta}\overline{Y} \Rightarrow \hat{\alpha} = 10 - 0,45 \cdot 20 \Rightarrow \hat{\alpha} = 10 - 9 \Rightarrow \hat{\alpha} = 1$. Logo, $\boxed{\hat{X} = 1 + 0,45Y}$

Questão 9 Com base na questão anterior, calcular:

a) O erro padrão da estimativa de Y para X, $S_{y \cdot x}$;
b) O erro padrão da estimativa de X para Y, $S_{x \cdot y}$.

RESPOSTAS:

a) $S_{y \cdot x} = S_y \cdot \sqrt{1 - r^2} \Rightarrow S_{y \cdot x} = 2 \cdot \sqrt{1 - 0,60^2} \Rightarrow S_{y \cdot x} = 2 \cdot \sqrt{0,64} \Rightarrow \boxed{S_{y \cdot x} = 1,60}$

b) $S_{x \cdot y} = S_x \cdot \sqrt{1 - r^2} \Rightarrow S_{x \cdot y} = 1,5 \cdot \sqrt{1 - 0,60^2} \Rightarrow S_{x \cdot y} = 1,5 \cdot \sqrt{0,64} \Rightarrow \boxed{S_{x \cdot y} = 1,20}$

Questão 10 Se $S_{y \cdot x} = 3$ e $S_y = 5$, determinar *r*.

RESPOSTA:

$S_{y \cdot x} = S_y \cdot \sqrt{1 - r^2} \Rightarrow 3 = 5 \cdot \sqrt{1 - r^2} \Rightarrow 0,60 = \sqrt{1 - r^2} \Rightarrow 1 - r^2 = 0,36 \Rightarrow r^2 = 0,64 \Rightarrow \boxed{r = 0,80}$

Questão 11 Se o coeficiente de correlação entre X e Y é 0,50, que percentagem da variação total permanece não-explicada pela equação de regressão?

RESPOSTA: O Coeficiente de Determinação, que mede o grau de explicação da variável Y pela variável X, se Y for dependente de X, é dado por: $\mathbf{r^2}$.

A variação não-explicada será dada por: $1 - r^2$.

Logo, se r = 0,50, então $1 - r^2 = 1 - 0,5^2 = 1 - 0,25 = 0,75 =$ **75%.**

DESEJO BONS ESTUDOS E EXCELENTE PROVA DE ESTATÍSTICA A TODOS!

PROFESSOR PEDRO BELLO