PROBABILIDADE E ESTATÍSTICA

Covariância e Correlação

Livro Eletrônico

GRAN CURSOS ONLINE

SUMÁRIO

Covariância e Correlação 3
1. Covariância 3
1.1. Expressão para a Covariância 6
1.2. Propriedades da Covariância 7
2. Correlação 8
3. Interpretações da Correlação 11
4. Propriedades da Correlação 17
5. Variância da Soma 18
Resumo 20
Questões Comentadas em Aula 21
Questões de Concursos 23
Gabarito 33

COVARIÂNCIA E CORRELAÇÃO

O desvio padrão e a variância são duas medidas de variabilidade que se aplicam **a uma única variável aleatória.**

Elas são muito úteis para medir **a oscilação** dessas variáveis **em torno da média.**

A covariância e a correlação são ferramentas matemáticas para avaliar se duas variáveis oscilam juntas em torno de suas respectivas médias ou se suas oscilações são independentes.

Em primeiro lugar, é importante lembrar que covariância e correlação somente se aplicam a variáveis quantitativas. Não é possível calcular essas métricas para variáveis qualitativas.

1. Covariância

Vamos nos lembrar de que a covariância tem por objetivo avaliar **os desvios conjuntos** de duas variáveis aleatórias.

Uma forma de fazer isso é tomar a média dos produtos dos desvios das duas variáveis.

$$Cov(X,Y) = E[(X-\mu_X)(Y-\mu_Y)]$$

Uma forma interessante de representar a covariância é $\mathbf{S_{XY}}$. Essa representação guarda um paralelo com a representação da variância $\mathbf{S_{XX}}$.

$$S_{XY} = E[(X-\mu_X).(Y-\mu_Y)] = \frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x}) \ . \ (y_i-\bar{y})$$

desvios de X
desvios de Y
desvios de X
desvios de Y
desvios de X
desvios de Y

$$S_{XX} = E[(X-\mu_X)^2] = \frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})^2$$

desvios de X
desvios de X
desvios de X ao quadrado
desvios de X ao quadrado

Observe que podemos chegar a uma expressão matemática para a covariância:

$$Cov(X,Y) = S_{XY} = E[(X-\mu_X)(Y-\mu_Y)] = \frac{\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y})}{N}$$

Porém, da mesma maneira que o que vimos no cálculo do desvio padrão, não é necessário decorar essa longa expressão matemática. Fique tranquilo. Basta você seguir um procedimento.

Para aprendermos esse passo a passo, vamos considerar um par de variáveis aleatórias:

N.	Salário (X)	Idade (Y)
1	4,00	26
2	4,56	32
3	5,25	36
4	5,73	20
5	6,26	40

- O primeiro passo é calcular as médias de ambas as variáveis aleatórias:

$$\mu_X = \frac{4{,}00 + 4{,}56 + 5{,}25 + 5{,}73 + 6{,}26}{5} = \frac{25{,}8}{5} = 5{,}16$$

$$\mu_Y = \frac{26 + 32 + 36 + 20 + 40}{5} = \frac{154}{5} = 30{,}8$$

O segundo passo é calcular os desvios de ambas as variáveis em torno da média:

N.	Salário (X)	Desvios ($X - \mu_X$)	Idade (Y)	Desvios ($Y - \mu_Y$)
1	4,00	4,00 – 5,16 = –1,16	26	26 – 30,8 = –4,8
2	4,56	4,56 – 5,16 = –0,60	32	32 – 30,8 = 1,2
3	5,25	5,25 – 5,16 = 0,09	36	36 – 30,8 = 5,2
4	5,73	5,73 – 5,16 = 0,57	20	20 – 30,8 = –10,8
5	6,26	6,26 – 5,16 = 1,10	40	40 – 30,8 = 9,2

- O terceiro passo é calcular os produtos dos desvios de ambas as variáveis em torno da média:

N.	Salário (X)	(X – μX)	Idade (Y)	(Y – μY)	Produto
1	4,00	–1,16	26	–4,8	(–1,16).(–4,8) = 5,568
2	4,56	–0,60	32	1,2	(–0,60).(1,2) = –0,72
3	5,25	0,09	36	5,2	(0,09).(5,2) = 0,468
4	5,73	0,57	20	–10,8	(0,57).(–10,8) = –6,156
5	6,26	1,10	40	9,2	(1,10).(9,2) = 10,12

- O quarto e último passo consiste em tomar a média aritmética das observações. Vale lembrar que, assim como o desvio padrão, no caso de uma amostra, devemos usar o fator de ajuste N – 1 no denominador.

$$S_{XY} = \frac{5{,}568 - 0{,}72 + 0{,}468 - 6{,}156 + 10{,}12}{5 - 1} = \frac{9{,}28}{4} = 2{,}32$$

Direto do Concurso

001. (CESPE/TRT-5ª REGIÃO/ANALISTA JUDICIÁRIO – ESTATÍSTICA/2008) Um estudo acerca de cursos de qualificação profissional envolveu a participação de 100 trabalhadores. A amostra foi classificada em função da rotatividade (número de empregos em até 30 dias após a realização do curso) e da opinião do trabalhador a respeito do curso (satisfação = 0, se o trabalhador entrevistado estava insatisfeito, ou satisfação = 1, se o trabalhador estava satisfeito com o curso realizado).

Os resultados desse estudo são apresentados na tabela a seguir.

		rotatividade		
		0	1	total
satisfação	0	10	10	20
	1	60	20	80
	total	70	30	100

Considerando essas informações, julgue os itens subsequentes.

A covariância entre a rotatividade e a satisfação é inferior a zero.

Vamos seguir o passo a passo do cálculo da covariância.

- Vamos calcular as médias das duas variáveis aleatórias:

$$\mu_{Satisfação} = \frac{(10+10).0 + (60+20).1}{10+10+60+20} = \frac{20.0 + 80.1}{100} = \frac{80}{100} = 0{,}80$$

$$\mu_{Rotatividade} = \frac{(10+60).0 + (10+20).1}{10+60+10+20} = \frac{70.0 + 30.1}{100} = \frac{30}{100} = 0{,}30$$

- Vamos calcular os desvios de cada valor em relação à média:

Frequência	Sat	(X – μ)	Rot	(Y – μ)
10	0	0 – 0,80 = – 0,80	0	0 – 0,30 = – 0,30
10	0	0 – 0,80 = – 0,80	1	1 – 0,30 = 0,70

Frequência	Sat	(X – μ)	Rot	(Y – μ)
60	1	1 – 0,80 = 0,20	0	0 – 0,30 = – 0,30
20	1	1 – 0,80 = 0,20	1	1 – 0,30 = 0,70

- Calculemos o produto dos desvios:

Frequência	Sat	(X – μ)	Rot	(Y – μ)	Produto
10	0	–0,80	0	–0,30	(–0,80). (–0,30) = 0,24
10	0	–0,80	1	0,70	(–0,80).(0,70) = –0,56
60	1	0,20	0	–0,30	(0,20). (–0,30) = –0,06
20	1	0,20	1	0,70	(0,20).(0,70) = 0,14

-
- Calculemos a covariância como a média ponderada dos produtos dos desvios:

$$Cov = \frac{10.0{,}24 + 10.(-0{,}56) + 60.(-0{,}06) + 20.(0{,}14)}{100 - 1}$$

$$Cov = \frac{2{,}4 - 5{,}6 - 3{,}6 + 2{,}8}{99} = \frac{-4}{99} < 0$$

De fato, a covariância é negativa entre as duas variáveis.
Certo.

1.1. Expressão para a Covariância

Uma expressão muito útil para a covariância, particularmente naquelas questões que fornecem somatórios, pode ser deduzida a partir das propriedades da média.

Já vimos que:

$$S_{XY} = E[(X - \mu_X)(Y - \mu_Y)]$$

Podemos fazer a multiplicação das variáveis aleatórias dentro dos parênteses.

$$S_{XY} = E[XY - \mu_X Y - \mu_Y X + \mu_X \mu_Y]$$

Agora podemos usar as propriedades do valor esperado, sempre lembrando que as médias são números reais.

$$S_{XY} = E[XY] - \mu_X E[Y] - \mu_Y E[X] + \mu_X \mu_Y$$

Vamos observar que E[X] = μ_X e E[Y] = μ_Y e, com isso, chegamos à seguinte expressão:

$$\therefore S_{XY} = E[XY] - \mu_X\mu_Y - \mu_Y\mu_X + \mu_X\mu_Y = E[XY] - \mu_X\mu_Y$$

Vamos esquematizar essa importante expressão:

S_{XY}	=	$E[XY]$	−	$\mu_X\mu_Y$
Covariância	=	**Esperança do Produto**	−	**Produto das Esperanças**

Direto do Concurso

002. (FGV/COMPESA/ANALISTA DE GESTÃO/2016) Seja X e Y, duas variáveis aleatórias. Uma forma de mensurar a covariância entre ambas é por meio da seguinte expressão:

a) E[X²] – E[Y²].

b) Var[X] – Var[Y].

c) E[X – E(X)]E[Y-E(Y)].

d) E[XY] – E[X]E[Y].

e) E[X|Y] – E[X]E[Y].

Questão que cobrou a lembrança direta de uma importante fórmula. Não podemos nos esquecer dela.

A covariância é igual à esperança do produto menos o produto das esperanças.

S_{XY}	=	$E[XY]$	−	$\mu_X\mu_Y$
Covariância	=	**Esperança do Produto**	−	**Produto das Esperanças**

Letra d.

1.2. Propriedades da Covariância

Quando as variáveis aleatórias são multiplicadas por uma constante, a covariância entre elas ficará multiplicada por essas mesmas constantes:

$$Cov\,(3X, 4Y) = 3.4.Cov(X, Y) = 12.Cov\,(X, Y)$$

O mesmo vale para uma constante negativa:

$$Cov\,(3X, -4Y) = 3.(-4).Cov(X, Y) = -12.Cov\,(X, Y)$$

Note, porém, que não existe nenhuma propriedade para a covariância quando se trata da soma de duas variáveis aleatórias:

$$Cov\,(X, X+Y) = ?$$

Outra propriedade interessante é a covariância entre variáveis independentes. Por definição, duas variáveis X e Y são independentes quando:

$$E[XY] = E[X].E[Y] = \mu_X \mu_Y$$

Podemos calcular a covariância entre X e Y utilizando essa relação. Vimos que a covariância entre duas variáveis aleatórias é igual à esperança do produto menos o produto das esperanças.

$$Cov(X,Y) = E[XY] - E[X].E[Y]$$

Porém, se X e Y são independentes, temos que a esperança do produto é igual ao produto das esperanças:

$$Cov(X,Y) = E[XY] - E[X].E[Y] = E[X].E[Y] - E[X].E[Y] = 0$$

Portanto, **a covariância entre duas variáveis independentes é sempre igual a zero.**

Guarde essa informação, porque é muito comum os enunciados de questões se referirem a variáveis independentes. Nesse caso, você pode, de cara, utilizar duas propriedades:

- O produto das esperanças é igual à esperança dos produtos;
- A covariância entre elas é nula. E, por consequência, como veremos, mais adiante, a correlação também será.

2. Correlação

A correlação é talvez a medida de dispersão mais importante. Normalmente, é calculada pelo chamado **coeficiente de correlação de Pearson**, que é obtido pela razão da covariância com os desvios padrões de ambas as variáveis:

$$\rho_{XY} = \frac{S_{XY}}{\sigma_X \sigma_Y}$$

O coeficiente de Pearson também é chamado de **coeficiente de correlação linear**, porque avalia a existência de relação linear entre duas variáveis.

A correlação sempre terá módulo menor ou igual a 1. Podemos escrever:

$$-1 \leq \rho \leq 1$$

A correlação positiva indica que as duas variáveis X e Y crescem juntas. Por outro lado, a correlação negativa indica que, quando a variável X cresce, a variável Y diminui.

Por outro lado, a correlação nula é uma situação conhecida como **variáveis descorrelacionadas.** Falaremos mais sobre isso mais adiante.

Vejamos exemplos dos três casos em gráficos de dispersão:

Figura 1: Correlação entre duas Variáveis.

A correlação não tem nada a ver com o coeficiente de inclinação da reta de tendência entre as duas variáveis aleatórias. A correlação trata apenas de quão próximas essas observações estão da reta.

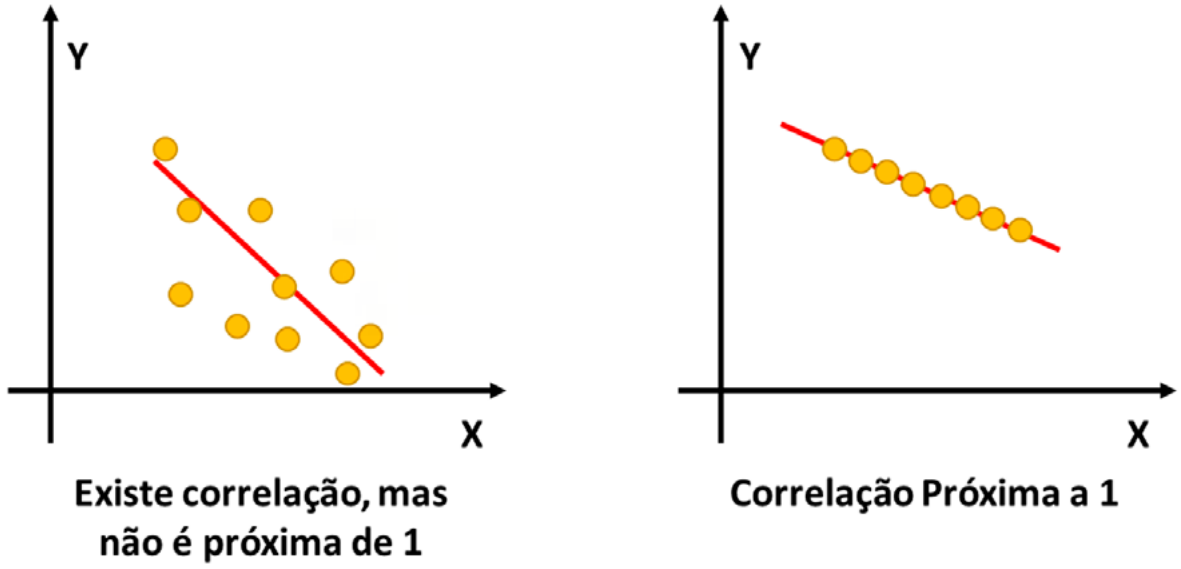


Figura 2: Exemplos de Correlação.

Nos casos mostrados na Figura 2, o conjunto de dados à esquerda tem uma reta de tendência bem mais inclinada. Porém, o ajuste dos dados a essa reta é menor que o da direita. E é isso que importa para a correlação. Perceba que ambas possuem correlação negativa.

Existe um certo conjunto de práticas na Estatística a respeito do que seria uma correlação fraca, moderada ou forte. Vejamos:

$\|\rho\|$	Interpretação
0	Variáveis independentes
0 a 0,20	Correlação muito fraca
0,20 a 0,40	Correlação fraca
0,40 a 0,70	Correlação moderada
0,70 a 0,90	Correlação forte
0,90 a 1,00	Correlação muito forte

Tabela 8: Interpretações do Coeficiente de Correlação.

É importante destacar um caso extremo: **quando as variáveis são independentes, a correlação entre elas é nula .**

Pode-se provar essa propriedade pela definição de correlação. A correlação é igual à razão entre a covariância e o produto dos desvios padrões.

$$\rho = \frac{Cov}{\sigma_X \sigma_Y}$$

No caso de variáveis independentes, a covariância é nula. Portanto, a correlação também será.

$$\rho = \frac{Cov}{\sigma_X \sigma_Y} = 0$$

Agora, vamos nos aprofundar um pouco mais no assunto de correlação.

Direto do Concurso

003. (CESPE/TCE-PR/2016) Se satisfação no trabalho e saúde no trabalho forem indicadores com variâncias populacionais iguais a 8 e 2, respectivamente, e se a covariância populacional

entre esses indicadores for igual a 3, então a correlação populacional entre satisfação no trabalho e saúde no trabalho será superior a 0,70.

Para o cálculo da correlação, precisaremos dos desvios-padrão das variáveis, que são iguais à raiz quadrada das variâncias populacionais. Lembrando-se disso, a correlação é dada por:

$$\rho = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} = \frac{3}{\sqrt{8}.\sqrt{2}} = \frac{3}{\sqrt{16}} = \frac{3}{4} = 0{,}75 > 0{,}70$$

Certo.

004. (CONSULPLAN/TSE/ANALISTA JUDICIÁRIO – ESTATÍSTICA/2012) Para duas variáveis x e y, são dados:

$$\bar{x} = 50; \bar{y} = 4; \overline{xy} = 320; \sigma(x) = 10\sqrt{10}; \sigma(y) = \sqrt{22{,}5}$$

O coeficiente de correlação entre as variáveis é:

a) -0,2.
b) 0,2.
c) 0,6.
d) 0,8.

A covariância é dada pela diferença entre a esperança do produto e o produto das esperanças.

$$Cov(X,Y) = E[XY] - E[X]E[Y] = 320 - 4.50 = 320 - 200 = 120$$

Já a correlação é igual à covariância dividida pelo produto dos desvios-padrão:

$$\rho = \frac{Cov}{\sigma(x)\sigma(y)} = \frac{120}{10\sqrt{10}.\sqrt{22{,}5}} = \frac{120}{10\sqrt{225}} = \frac{120}{10.15} = \frac{12}{15} = \frac{4}{5} = 0{,}8$$

Letra d.

3. Interpretações da Correlação

O primeiro ponto muito importante é que **correlação não implica causalidade.**

O fato de duas grandezas estarem linearmente correlacionadas não constitui absolutamente nenhuma prova de que existe qualquer relação de causa-efeito entre elas.

Quando duas grandezas são bastante correlacionadas, mas não há qualquer relação de causa-efeito entre elas, diz-se que há uma **correlação espúria.**

Sobre esse assunto, há um famoso livro “Como Mentir com Estatísticas” que mostra que a correlação espúria é algo que é bastante sedutor à grande maioria das pessoas e elas se convencem facilmente sobre uma relação de causalidade por causa da existência de uma correlação.

Vejamos algumas variáveis extremamente bem correlacionadas – várias delas com coeficiente de correlação acima de 99% - extraídas do site Tylervigen (http://www.tylervigen.com/spurious-correlations).

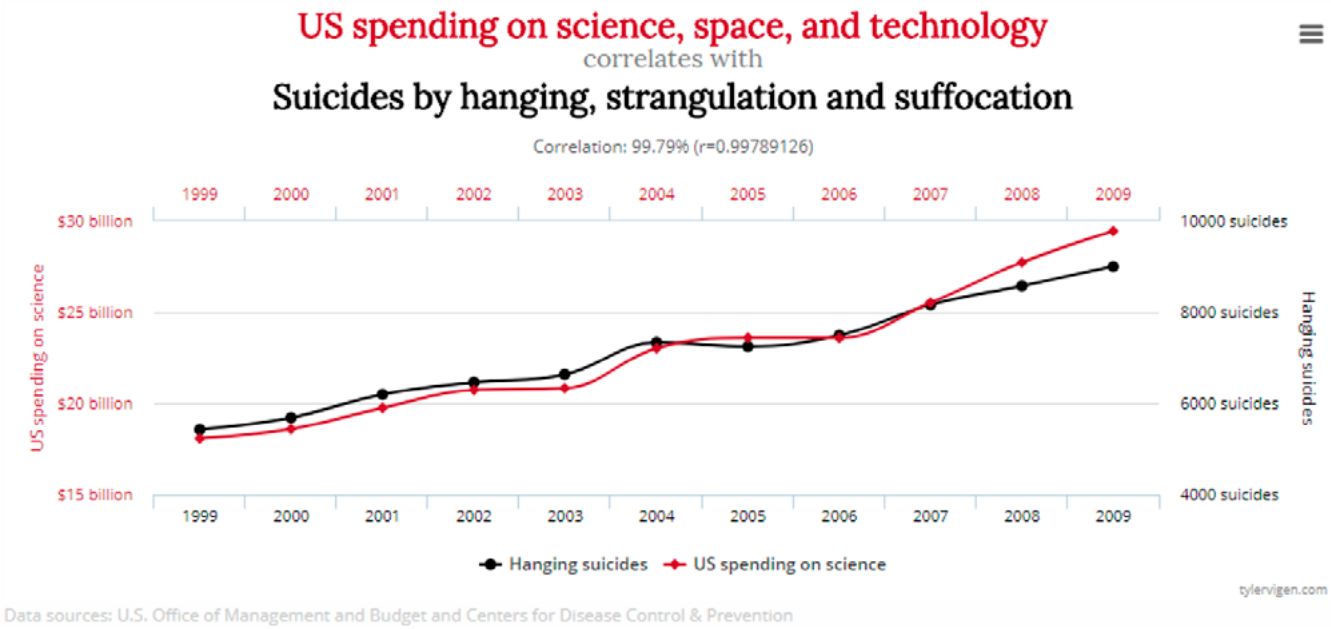


Figura 3: Os gastos do governo americano em ciência, espaço e tecnologia correlacionam quase que perfeitamente com o número de suicídios por enforcamento, estrangulamento e sufocamento.

Perceba que, somente olhando para o gráfico, diríamos que as duas variáveis possuem uma correlação fortíssima, pois é de 99,79%, mas então poderíamos dizer que quanto mais o governo gasta com ciência, mais as pessoas se suicidam pelas causas indicadas? Não, claro que não. Então, cuidado, pois haver uma correlação forte entre duas variáveis pode ser uma mera coincidência, nada mais do que isso.

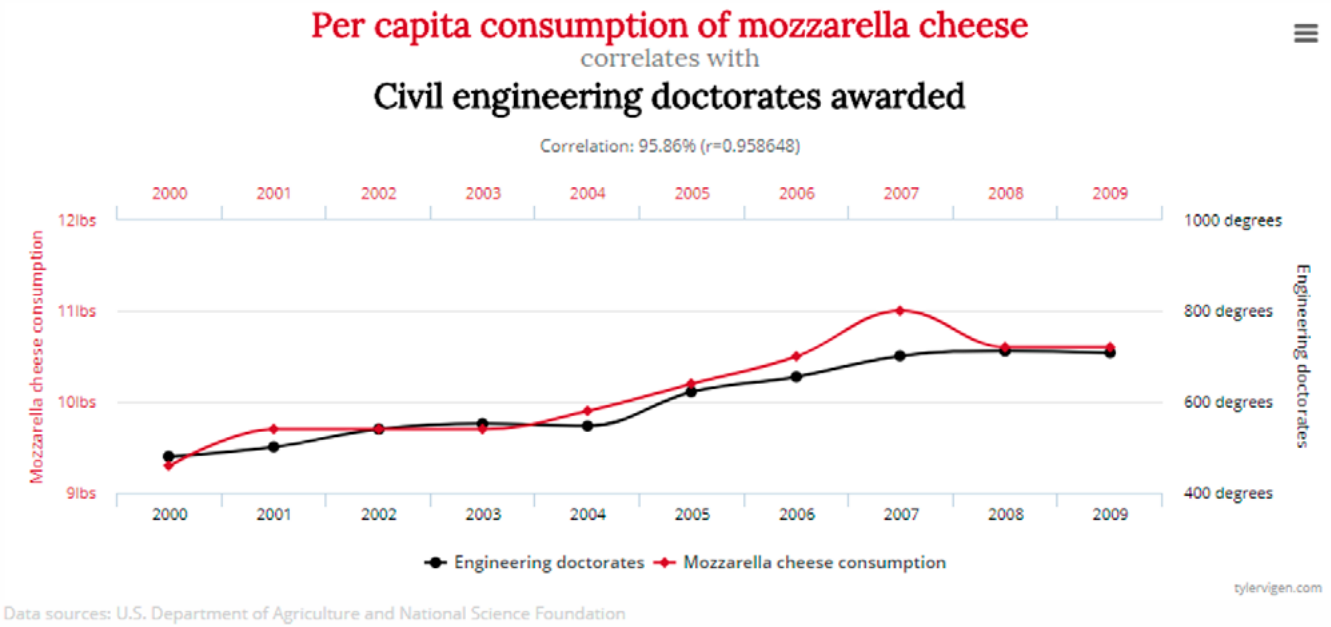


Figura 4: Há uma correlação muito forte entre o consumo per capta de queijo muçarela e o número de doutorados conferidos na área de Engenharia Civil.

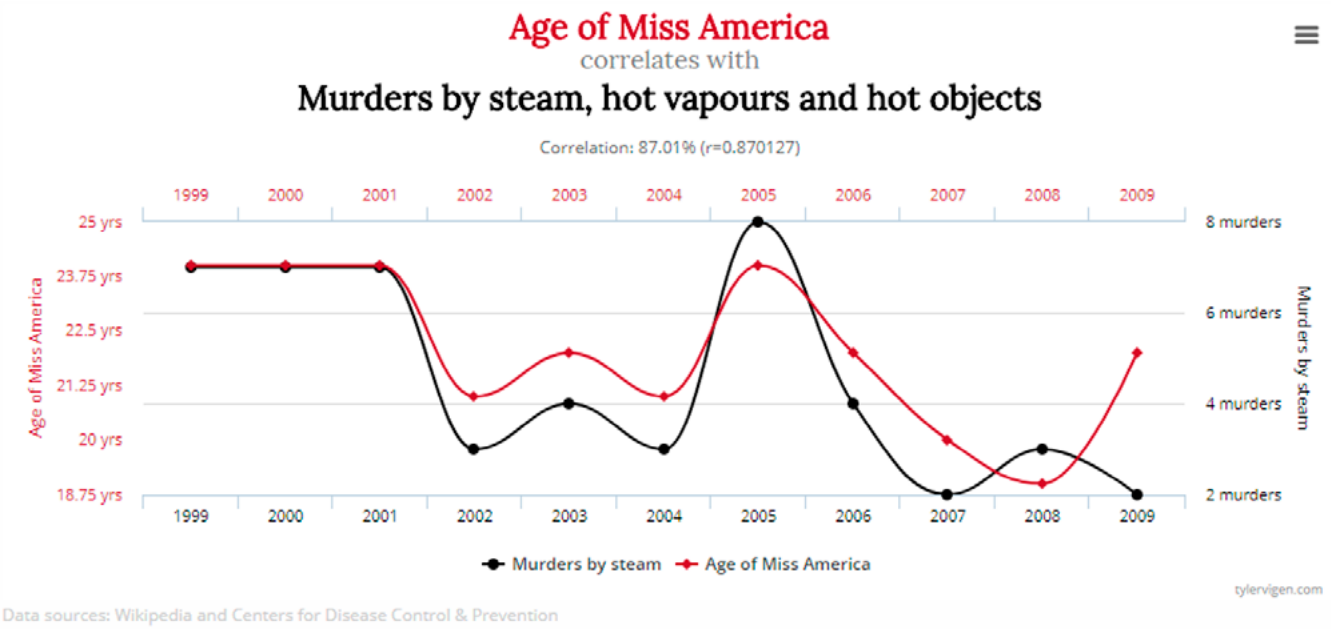


Figura 5: A idade da Miss Estados Unidos correlaciona fortemente com o número de assassinatos por vapor e objetos quentes.

Pense nisso da próxima vez que alguém lhe disser “os números não mentem”.

De fato, os números não mentem. Porém, o ser humano mente. E uma das maiores mentiras com Estatística é dizer que correlação implica causalidade.

Um ponto que muitos alunos podem ter dúvidas é:

Há algum conflito entre os conceitos de que duas variáveis independentes apresentam correlação nula e que correlação implica causalidade?

A resposta é que não. O ponto é que a Estatística normalmente trabalha com amostras. E, muitas vezes, um padrão estatístico pode ser observado com amostras pequenas, mas o mesmo padrão não se repete para a população inteira. Isso pode acontecer por uma mera coincidência.

Outro problema que pode acontecer é a **terceira causa.** Muitas vezes, duas variáveis aleatórias X e Y estão fortemente correlacionadas entre elas, mas porque a sua variação se deveu a uma terceira causa Z comum a ambas.

Suponha que você é um pesquisador e observou que o comportamento do número de horas trabalhadas pelas pessoas de uma cidade está bastante correlacionado com o índice de infartos naquela mesma cidade.

Antes de concluir que o aumento da carga horária levou ao aumento do número de infartos, pode ser interessante analisar outras terceiras causas: por exemplo, o consumo de energéticos. A taurina aumenta a disposição para o trabalho, porém, também aumenta a pulsação cardíaca e seu consumo em excesso pode ocasionar infarto.

E importante destacar que a recíproca também é falsa. Ou seja, **a ausência de correlação também não implica ausência de causalidade.**

Vejamos um exemplo interessantíssimo.

X	$Y = X^2$
-2	4
-1	1
0	0
+1	1
+2	4

Vamos calcular a covariância entre essas duas grandezas. Primeiramente, precisamos calcular as médias.

$$\bar{x} = \frac{-2-1+0+1+2}{5} = 0$$

$$\bar{y} = \frac{4+1+0+1+4}{5} = 2$$

Agora, vamos à covariância:

X	$Y = X^2$	$X - \overline{X}$	$Y - \overline{Y}$	$(X - \overline{X})(Y - \overline{Y})$
-2	4	-2	2	-4
-1	1	-1	-1	1
0	0	0	-2	0
+1	1	1	-1	-1
+2	4	2	2	4

$$Cov(X,Y) = \frac{-4+1+0-1+4}{5} = 0$$

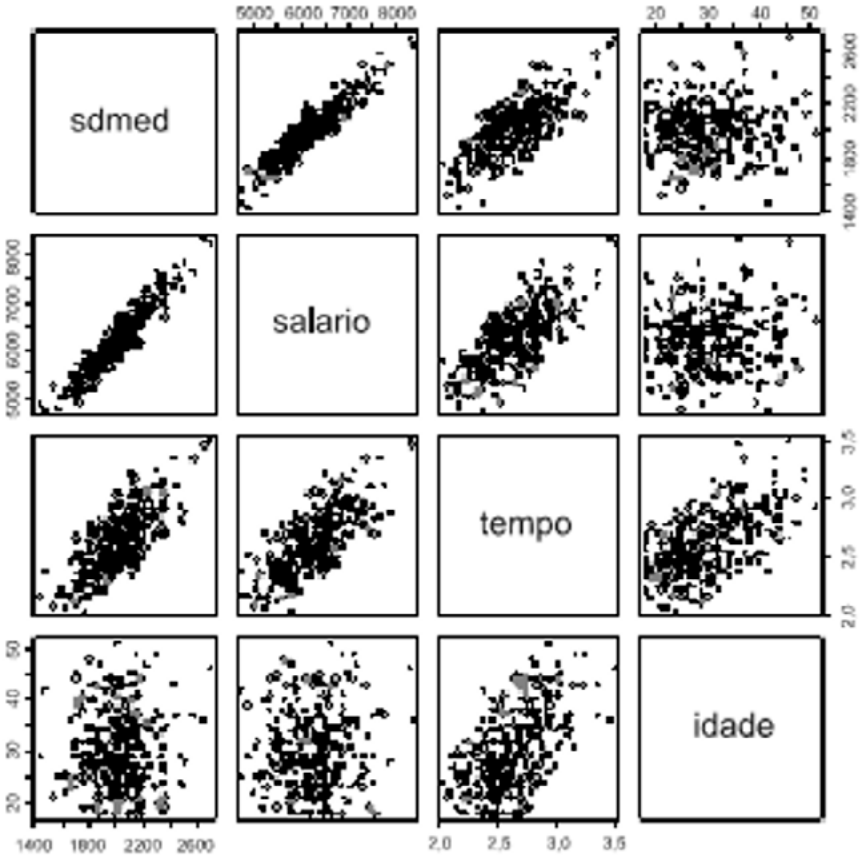
Sendo assim, a covariância é nula entre X e X². Porém, podemos ver que há uma relação de causalidade entre elas, tendo em vista que X² é obtida diretamente a partir de X.

Sendo assim, a ausência de correlação significa apenas **ausência de relação linear**, mas é possível haver outro tipo de relação entre as duas variáveis.

Uma consequência interessante do que acabamos ver é que **a média e o desvio-padrão são descorrelacionadas** quando a distribuição for simétrica. Isso acontece porque o desvio-padrão depende do quadrado da média e já vimos que X e X² não são correlacionadas.

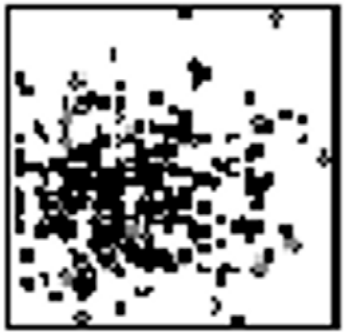
Direto do Concurso

005. (CESPE/BANCO DA AMAZÔNIA/2010) Com o objetivo de estudar as relações entre características de uma carteira de clientes (salário em R$, saldo médio da conta corrente em R$, tempo de conta aberta no banco em anos e idade do correntista), um analista conduziu uma análise multivariada (análise de componentes principais e análise de agrupamento) e obteve os resultados abaixo, gerados por um *software* de análise estatística.

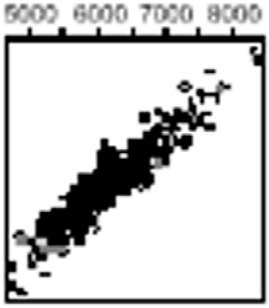


A partir das informações e das figuras apresentadas, julgue o item a seguir.
O salário e a idade dos clientes da amostra possuem uma forte correlação positiva.

A correlação positiva significa que os dados se dispersam ao longo de uma reta. Porém, olhando a dispersão do salário e da idade dos clientes, não notamos nenhum padrão claro.



Nesse gráfico de dispersão, não notamos nenhuma forte tendência de crescimento em linha reta das duas variáveis. Um exemplo melhor de correlação positiva seria o gráfico do saldo médio na conta pelo salário do cliente.



Errado.

4. Propriedades da Correlação

A correlação não é alterada pelas operações algébricas – soma, adição, multiplicação ou divisão –, exceto a multiplicação por um número negativo. Quando multiplicamos uma variável aleatória por um número negativo, o sinal da correlação é invertido.

Por exemplo, suponha que X e Y sejam duas variáveis com duas variáveis aleatórias, tais que o coeficiente de correlação entre elas seja ρ(X,Y) = 0,6. Com base nisso, quais seriam os coeficientes de correlação:

- ρ(X, 2Y + 1) =?
- ρ(3X, 2Y – 1) =?
- ρ(4X, 1 – 2Y) =?

Como vimos, a correlação não é alterada pelas operações algébricas, exceto a multiplicação por um número negativo. Dessa forma, a correlação nos dois primeiros casos é igual a 0,6.

No terceiro caso, como a variável Y foi multiplicada por –2, a correlação terá seu sinal invertido e passará a ser igual a –0,6. Assim, podemos escrever:

- ρ(X, 2Y + 1) = 0,6;
- ρ(3X, 2Y – 1) = 0,6;
- ρ(4X, 1 – 2Y) = –0,6.

Outro ponto interessante de comentarmos é que a correlação entre uma variável aleatória e ela própria é sempre igual a 1. Dessa forma, podemos dizer que:

- ρ(X, X) = 1;
- ρ(X, 2X + 3) = 1;
- ρ(X, 1 – 4X) = –1, pois a multiplicação por número negativo inverte o sinal da correlação.

Vale ressaltar que não existe nenhuma propriedade para a correlação da soma de duas variáveis. Por exemplo, se sabemos que ρ(X, Y) = 0,6, não temos como determinar a correlação ρ(X, X + Y).

5. Variância da Soma

Sejam X e Y duas variáveis aleatórias. A variância da soma pode ser calculada pela seguinte expressão:

$$Var\,(X+Y) = Var\,(X) + Var\,(Y) + 2.Cov\,(X,Y)$$

Portanto, a variância da soma de duas variáveis aleatórias X e Y é igual à soma das variâncias mais duas vezes a covariância entre elas.

É interessante observar o caso particular de quando X e Y são **independentes.** Nesse caso, a covariância entre elas é nula.

Portanto, somente quando X e Y forem duas variáveis independentes, podemos dizer que a variância da soma é igual à soma das variâncias.

$$Var\,(X+Y) = Var\,(X) + Var\,(Y), \text{se X e Y forem independentes}$$

O melhor jeito de aprender essa fórmula é praticando com questões de prova.

Direto do Concurso

006. (CONSULPLAN/TSE/2012) Uma variável X tem desvio-padrão 6, enquanto uma variável Y desvio-padrão 10. A covariância entre X e Y é −50. Assim, a variância de X + Y [Var(X+Y)] é:

a) − 84.
b) 36.
c) 86.
d) 136.

Basta aplicar diretamente a variância da soma.

$$Var\,(X+Y) = Var\,(X) + Var(Y) + 2.Cov(X,Y)$$

O enunciado forneceu os desvios padrões de X e Y. Porém, podemos converter facilmente em variância, pois a variância é igual ao quadrado do desvio padrão.

$$Var(X) = \sigma_X^2 = 6^2 = 36$$

$$Var(Y) = \sigma_Y^2 = 10^2 = 100$$

Vamos substituir os valores fornecidos no enunciado:

$$Var\,(X+Y) = 6^2 + 10^2 + 2.(-50)$$

$$Var\,(X+Y) = 36 + 100 - 100 = 36$$

Letra b.

RESUMO

Passo a Passo para o Cálculo da Covariância:

- Calcule as médias das duas variáveis aleatórias.
- Calcule os desvios de cada observação em relação à média.
- Faça o produto dos desvios de cada observação.
- Some todos os produtos calculados e divida por N, se for uma população, ou por N − 1, se for uma amostra.

Variância da Soma:

$$Var\ (X+Y) = Var(X) + Var(Y) + 2.Cov\ (X,Y)$$

Propriedades da Covariância:

- Quando as variáveis aleatórias são multiplicadas por constantes, essas constantes multiplicam a covariância:

$$Cov\ (5X, 3Y) = 5.3.Cov(X,Y)$$

- Não existe nenhuma propriedade para a covariância quando envolve a soma de variáveis aleatórias:

$$Cov\ (X, X+Y) = ?$$

Correlação:

- É obtida como a covariância dividida pelo produto dos desvios padrões:

$$Var\ (X+Y) = \frac{Cov\ (X,Y)}{\sigma_X \sigma_Y}$$

- A correlação entre duas variáveis aleatórias independentes é igual a zero.
- Correlação não implica causalidade.
- A correlação não é afetada pelas operações algébricas, exceto pela multiplicação por número negativo, caso em que fica multiplicada por −1.

QUESTÕES COMENTADAS EM AULA

001. (CESPE/TRT-5ª REGIÃO/ANALISTA JUDICIÁRIO – ESTATÍSTICA/2008) Um estudo acerca de cursos de qualificação profissional envolveu a participação de 100 trabalhadores. A amostra foi classificada em função da rotatividade (número de empregos em até 30 dias após a realização do curso) e da opinião do trabalhador a respeito do curso (satisfação = 0, se o trabalhador entrevistado estava insatisfeito, ou satisfação = 1, se o trabalhador estava satisfeito com o curso realizado).
Os resultados desse estudo são apresentados na tabela a seguir.

	rotatividade			
satisfação		0	1	total
	0	10	10	20
	1	60	20	80
	total	70	30	100

Considerando essas informações, julgue os itens subsequentes.
A covariância entre a rotatividade e a satisfação é inferior a zero.

002. (FGV/COMPESA/ANALISTA DE GESTÃO/2016) Seja X e Y, duas variáveis aleatórias. Uma forma de mensurar a covariância entre ambas é por meio da seguinte expressão:
a) $E[X^2] - E[Y^2]$.
b) Var[X] – Var[Y].
c) E[X – E(X)]E[Y-E(Y)].
d) E[XY] – E[X]E[Y].
e) E[X|Y] – E[X]E[Y].

003. (CESPE/TCE-PR/2016) Se satisfação no trabalho e saúde no trabalho forem indicadores com variâncias populacionais iguais a 8 e 2, respectivamente, e se a covariância populacional entre esses indicadores for igual a 3, então a correlação populacional entre satisfação no trabalho e saúde no trabalho será superior a 0,70.

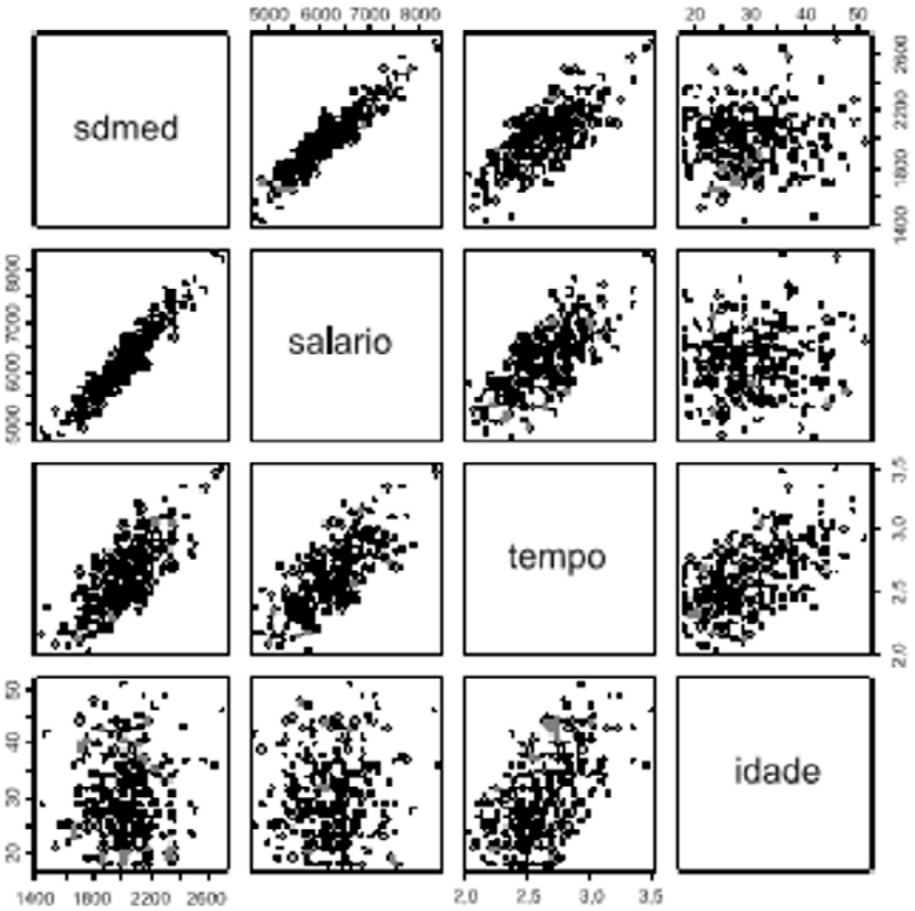
004. (CONSULPLAN/TSE/ANALISTA JUDICIÁRIO – ESTATÍSTICA/2012) Para duas variáveis x e y, são dados:

$$\bar{x} = 50; \bar{y} = 4; \overline{xy} = 320; \sigma(x) = 10\sqrt{10}; \sigma(y) = \sqrt{22{,}5}$$

O coeficiente de correlação entre as variáveis é:
a) -0,2.
b) 0,2.

c) 0,6.
d) 0,8.

005. (CESPE/BANCO DA AMAZÔNIA/2010) Com o objetivo de estudar as relações entre características de uma carteira de clientes (salário em R$, saldo médio da conta corrente em R$, tempo de conta aberta no banco em anos e idade do correntista), um analista conduziu uma análise multivariada (análise de componentes principais e análise de agrupamento) e obteve os resultados abaixo, gerados por um *software* de análise estatística.



A partir das informações e das figuras apresentadas, julgue o item a seguir.
O salário e a idade dos clientes da amostra possuem uma forte correlação positiva.

006. (CONSULPLAN/TSE/2012) Uma variável X tem desvio-padrão 6, enquanto uma variável Y desvio-padrão 10. A covariância entre X e Y é −50. Assim, a variância de X + Y [Var(X+Y)] é:
a) − 84.
b) 36.
c) 86.
d) 136.

QUESTÕES DE CONCURSOS

007. (CESPE/TCE-PR/2016) Se satisfação no trabalho e saúde no trabalho forem indicadores com variâncias populacionais iguais a 8 e 2, respectivamente, e se a covariância populacional entre esses indicadores for igual a 3, então a correlação populacional entre satisfação no trabalho e saúde no trabalho será igual a:

a) 0,8125.
b) 1.
c) 0,1875.
d) 0,30.
e) 0,75.

Vamos utilizar a definição: a correlação é igual à covariância dividida pelo produto dos desvios padrões. Note que o enunciado forneceu as variâncias, que correspondem aos produtos dos desvios padrões.

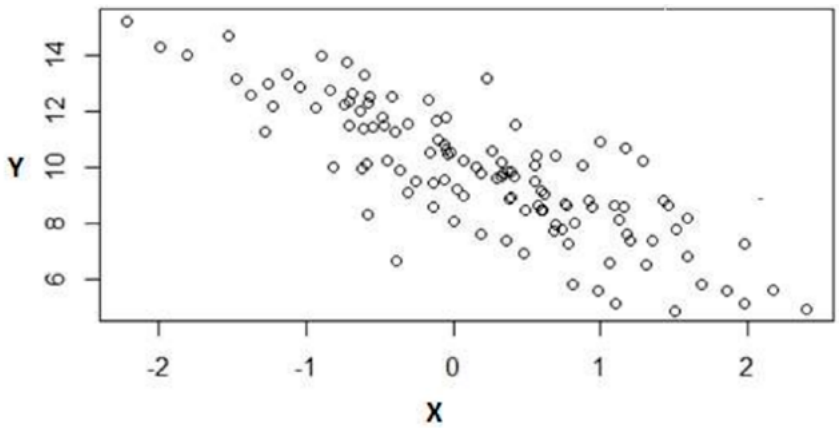
$$\sigma_X^2.\sigma_Y^2 = 8.2 = 16 \therefore \sqrt{\sigma_X^2.\sigma_Y^2} = \sqrt{16} = 4$$

Façamos as contas da correlação:

$$\rho = \frac{Cov}{\sigma_X \sigma_Y} = \frac{3}{4} = 0{,}75$$

Letra e.

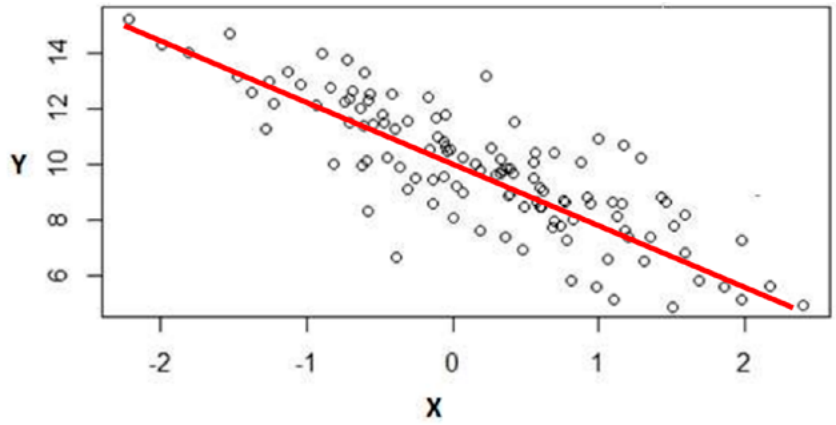
008. (UFU-MG/2019) Considere o seguinte gráfico de dispersão.



Com base na configuração dos pontos no gráfico, assinale a alternativa que apresenta um valor possível para o coeficiente de correlação linear entre X e Y.

a) 0,5.
b) 1.
c) -1.
d) -0,5.

A correlação entre as duas variáveis é negativa, pois notamos uma tendência clara de que a variável Y decresce à medida que X cresce.



Notemos que essa correlação não é perfeita, haja vista que há vários pontos fora da linha reta. Portanto, o valor –0,5 é mais adequado que –1.
Letra d.

009. (CESPE/FUNPRESP/JUD – ANALISTA DE INVESTIMENTOS/2016) Com base na tabela precedente, que apresenta estatísticas referentes a duas variáveis observadas em um estudo previdenciário, julgue o seguinte item.

	variável	
estatística	X	Y
média amostral	25	27
variância amostral	16	9

O valor absoluto da covariância entre X e Y é igual ou inferior a 12.

Sabemos que a correlação é igual à covariância dividida pelo produto dos desvios padrões. Como o valor absoluto da correlação é sempre igual ou inferior a 1, o valor absoluto da covariância será igual ou inferior ao produto dos desvios padrões.

$$|Cov\,(X,Y)| \leq \sqrt{\sigma_X \sigma_Y} = \sqrt{16.9} = \sqrt{144} = 12$$

De fato, o módulo da covariância é realmente igual ou inferior a 12.
Certo.

---

010. (FGV/SEFAZ-RJ/ANALISTA DE CONTROLE INTERNO/2011) A respeito do conceito de covariância, analise as afirmativas a seguir:
I – Se duas variáveis são independentes e ambas apresentam somente valores positivos, a covariância entre as duas é igual a 1.
II – A covariância entre duas variáveis apresenta-se no intervalo entre –1 e 1.
III – A fórmula para o cálculo da covariância entre duas variáveis X e Y é $Cov(x, y) = \overline{XY} - \overline{X}\,\overline{Y}$
Assinale:
a) se apenas a afirmativa II estiver correta.
b) se apenas a afirmativa I estiver correta.
c) se nenhuma afirmativa estiver correta.
d) se apenas a afirmativa III estiver correta.
e) se todas as afirmativas estiverem corretas.

Vamos analisar as afirmações.
I – Se as duas variáveis são independentes, a covariância entre elas é nula. Afirmação incorreta.
II – A covariância pode assumir qualquer valor real. Na verdade, é a correlação que está limitada ao intervalo entre –1 e 1. Afirmação incorreta.
III – É isso mesmo. A covariância é igual à esperança do produto menos o produto das esperanças.
Letra d.

---

011. (FCC/TRT-14ª REGIÃO (RO E AC)/ANALISTA JUDICIÁRIO – ESTATÍSTICA/2011) Seja var(X) variância da variável aleatória X, var(Y) a variância da variável aleatória Y e cov(X, Y) a covariância das variáveis aleatórias X, Y. É correto afirmar que:
a) var(X + Y) < var(X) + var(Y) se cov(X, Y) > 0.
b) var(X + Y) > var(X) + var(Y) se cov(X, Y) > 0.
c) se X e Y são independentes então cov(X, Y) ≠ 0.
d) var(X + c) > var(X) para qualquer c >0.
e) var(cX) = cvar(X) para qualquer c > 0.

Vamos utilizar a propriedade da variância da soma de duas variáveis aleatórias.

$$Var(X+Y) = Var(X) + Var(Y) + 2.Cov(X,Y)$$

Com base nessa expressão, vamos avaliar as afirmações:

a) Se a covariância for positiva, nós podemos escrever que:

$$Var(X+Y) = Var(X) + Var(Y) + 2.Cov(X,Y) > Var(X) + Var(Y)$$

Portanto, se a covariância for positiva, haverá um termo positivo a ser somado à "Var (X) + Var(Y)". Logo, a variância de X + Y será maior que a soma das variâncias. Alternativa incorreta.

b) Como vimos anteriormente, quando a covariância for positiva, a variância da soma é maior que a soma das variâncias. Afirmação correta.

c) Se as duas variáveis são independentes, então a covariância é nula. Afirmação incorreta.

d) A soma de uma constante não afeta o valor da variância. Logo, var(X + c) = var(X). Afirmação incorreta.

e) Uma constante multiplicada deve ser elevada ao quadrado quando retirada da variância. Ou seja, devemos escrever:

$$Var(cX) = c^2.Var(X)$$

Letra b.

012. (INÉDITA/2021) As variáveis aleatórias X e Y têm variâncias iguais e possuem coeficiente de correlação igual a 0,2. O coeficiente de correlação entre as variáveis aleatórias X e 6 – 2Y é:

a) – 0,35.
b) – 0,2.
c) 0,1.
d) 0,56.
e) 0,92.

A correlação não é afetada pelas operações algébricas, exceto pela multiplicação por número negativo. Como foi tomada a correlação com a variável 6 – 2Y, a correlação ficará multiplicada por –1.

Dessa forma, a correlação passará a ser igual a –0,2.

Letra b.

013. (FGV/PREFEITURA DE RECIFE-PE/ANALISTA DE CONTROLE INTERNO – FINANÇAS PÚBLICAS/2014) Uma variável aleatória X tem média igual a 2 e desvio padrão igual a 2. Se Y = 6 – 2X, então a média de Y, a variância de Y e o coeficiente de correlação entre X e Y valem, respectivamente,

a) −2, 4 e 1.
b) −2, 16 e 1.
c) 2, 16 e −1.
d) 10, 2 e −1.
e) 2, 4 e −1.

Questão muito boa para treinarmos as propriedades das variáveis aleatórias.
A média é um operador linear. Isso significa que:
Quando somamos uma constante, essa constante pode ser adicionada do lado de fora da média.
Quando multiplicamos uma variável aleatória por uma constante, essa constante multiplica a sua média também.

$$E[Y] = E[6 - 2X] = 6 - 2.E[X] = 6 - 2.2 = 6 - 4 = 2$$

Para a variância, podemos dizer que:
Quando somamos uma constante, ela não influencia o valor da variância.
Quando multiplicamos a variável aleatória por uma constante, a sua variância fica multiplicada por essa constante ao quadrado.

$$Var[Y] = Var[6 - 2X] = Var(-2X) = (-2)^2.Var(X) = 4.2^2 = 16$$

Por fim, passemos à correlação. Sabemos que a correlação não é afetada pelas operações algébricas, exceto pela multiplicação por número negativo, caso em que ela fica multiplicada por −1.

$$\rho(X,Y) = \rho(X, 6 - 2X) = -\rho(X,X) = -1$$

Letra c.

014. (CESPE/MS/ESTATÍSTICO/2010) Considerando a tabela de valores acima, nas variáveis X e Y, julgue os itens subsequentes.

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

Se Cov (X, Y) é a covariância entre X e Y, V(X) é a variância de X e V(Y) é a variância de Y, então é correto afirmar que o coeficiente de correlação linear é inferior a 0,8.

Vamos seguir o passo a passo para o cálculo da correlação. Como não foi fornecido se a questão trata de uma amostra ou população, vamos considerar uma população, mas vale notar que a correlação não adota fator de correção para a amostra.

1º Passo: calculamos as médias das duas variáveis:

$$\mu_X = \frac{1+2+3+4+5}{5} = \frac{15}{5} = 3$$

$$\mu_Y = \frac{2+3+2+3+4}{5} = \frac{14}{5} = 2{,}8$$

2º Passo: calculamos os desvios de cada observação em relação à média:

X	(X – μ)	**Y**	(Y – μ)
1	1 – 3 = – 2	2	2 – 2,8 = –0,8
2	2 – 3 = – 1	3	3 – 2,8 = 0,2
3	3 – 3 = 0	2	2 – 2,8 = –0,8
4	4 – 3 = 1	3	3 – 2,8 = 0,2
5	5 – 3 = 2	4	4 – 2,8 = 1,2

3º Passo: vamos calcular os produtos das duas variáveis em relação à média. A covariância pode ser obtida como a razão entre essa soma e o número de elementos.

X	(X – μ)	**Y**	(Y – μ)	(X – μ)(Y – μ)
1	1 – 3 = –2	2	2 – 2,8 = –0,8	(–2).(–0,8) = 1,6

X	(X – μ)	Y	(Y – μ)	(X – μ)(Y – μ)
2	2 – 3 = –1	3	3 – 2,8 = 0,2	(–1).(0,2) = –0,2
3	3 – 3 = 0	2	2 – 2,8 = –0,8	(0).(–0,8) = 0
4	4 – 3 = 1	3	3 – 2,8 = 0,2	(1).(0,2) = 0,2
5	5 – 3 = 2	4	4 – 2,8 = 1,2	(2).(1,2) = 2,4

$$S_{XY} = \frac{1,6 - 0,2 + 0 + 0,2 + 2,4}{5} = \frac{4}{5} = 0,8$$

4º Passo: vamos calcular os desvios padrão das duas variáveis. Para isso, precisamos somar o quadrado dos desvios de cada variável em relação à média. Devemos colocar no denominador o número de elementos.

X	(X – μ)	(X – μ)²	Y	(Y – μ)	(Y – μ)²
1	1 – 3 = –2	(–2)² = 4	2	2 – 2,8 = –0,8	(–0,8)² = 0,64
2	2 – 3 = –1	(–1)² = 1	3	3 – 2,8 = 0,2	(0,2)² = 0,04
3	3 – 3 = 0	(0)² = 0	2	2 – 2,8 = –0,8	(–0,8)² = 0,64
4	4 – 3 = 1	(1)² = 1	3	3 – 2,8 = 0,2	(0,2)² = 0,04
5	5 – 3 = 2	(2)² = 4	4	4 – 2,8 = 1,2	(1,2)² = 1,44

$$Var(X) = S_{XX} = \frac{4 + 1 + 0 + 1 + 4}{5} = \frac{10}{5} = 2$$

$$Var(Y) = S_{YY} = \frac{0,64 + 0,04 + 0,64 + 0,04 + 1,44}{5} = \frac{2,8}{5} = 0,7$$

Os desvios padrões são iguais à raiz quadrada da variância:

$$\sigma_X \sigma_Y = \sqrt{2.0,7} = \sqrt{1,4}$$

5º Passo: obtemos a correlação como a razão entre a covariância e o produto dos desvios padrão.

$$\rho = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} = \frac{0,8}{\sqrt{1,4}} < 0,8$$

Observe que nem precisamos calcular a raiz quadrada de 1,4, porque o enunciado perguntou apenas se a correlação seria inferior a 1,4.

Certo.

015. (CESGRANRIO/BNDES/2011/DESAFIO) As variáveis aleatórias X e Y têm variâncias iguais e possuem coeficiente de correlação igual a 0,2. O coeficiente de correlação entre as variáveis aleatórias X e 5X – 2Y é:

a) – 0,35.
b) – 0,2.
c) 0,1.
d) 0,56.
e) 0,92.

Uma questão bem interessante e difícil, por isso a colocamos como desafio. Note que não vimos nenhuma propriedade para a correlação entre duas variáveis aleatórias.
Vamos ter em mente a definição de correlação como a covariância dividida pelo produto dos desvios padrões.

$$\rho = \frac{Cov\,(5X - 2Y, X)}{\sigma(X).\sigma(5X - 2Y)}$$

Porém, de acordo com informações do enunciado, as variâncias de X e Y são iguais.

$$Var\,(X) = Var(Y) = \sigma^2$$

Além disso, foi fornecida também a correlação. Com base nas correlações e nos desvios padrões, podemos calcular a covariância entre X e Y:

$$\rho = \frac{Cov}{\sigma_X \sigma_Y} = \frac{Cov}{\sigma^2} \therefore Cov = \rho\sigma^2 = 0{,}2.\sigma^2$$

Agora, podemos calcular a variância da expressão 5X – 2Y:

$$Var\,(5X - 2Y) = Var(5X) + Var(2Y) - 2.Cov(5X, -2Y)$$

Usando as propriedades da variância, podemos dizer que:
A variância da diferença entre duas variáveis aleatórias é dada pela soma das variâncias menos o dobro da covariância entre elas:

$$Var\,(A - B) = Var\,(A) + Var\,(B) - 2.Cov\,(A, B)$$

Quando uma variável aleatória é multiplicada por uma constante, essa constante pode ser retirada da variância, mas sai elevada ao quadrado.

$$Var\,(5X - 2Y) = Var(5X) + Var(2Y) - 2.Cov(5X, 2Y)$$

$$Var\,(5X - 2Y) = 5^2.Var(X) + 2^2.Var(Y) - 2.(5).(2).Cov(X, Y)$$

Anteriormente, usamos também a propriedade de que a covariância é afetada pelas constantes multiplicativas.

$$Var\,(5X - 2Y) = 25.Var(X) + 4.Var(Y) - 20.Cov(X, Y)$$

$$Var\,(5X - 2Y) = 25.\sigma^2 + 4.\sigma^2 - 20.0{,}2.\sigma^2$$

$$Var\,(5X - 2Y) = 25.\sigma^2 + 4.\sigma^2 - 4.\sigma^2 = 25\sigma^2$$

Por fim, vamos calcular o desvio padrão da variável 5X – 2Y como a raiz quadrada da variância.

$$\therefore \sigma(5X - 2Y) = \sqrt{25\sigma^2} = 5\sigma$$

Agora, vamos à parte mais difícil: calcular a covariância entre a variável aleatória 5X – 2Y e a variável aleatória X. Podemos usar a propriedade de que a covariância é igual à esperança do produto menos o produto das esperanças.

$$Cov\,(5X - 2Y, X) = E[X.(5X - 2Y)] - E[X].E[5X - 2Y]$$

Agora, vamos utilizar o fato de que o operador esperança é linear. Isto é, podemos dizer que: A esperança da soma de duas variáveis aleatórias é igual à soma das esperanças.

$$E[A + B] = E[A] + E[B]$$

Na esperança do produto de uma variável aleatória por uma constante, essa constante multiplicativa pode ser retirada do operador esperança.

$$E[3A] = 3.E[A]$$

Agora, vamos utilizar na expressão:

$$Cov\,(5X - 2Y, X) = E[5X^2 - 2XY] - E[X].(5E[X] - 2E[Y])$$

$$Cov\,(5X - 2Y, X) = E[5X^2] - E[2XY] - E[X].(5E[X] - 2E[Y])$$

$$Cov\,(5X - 2Y, X) = 5\,E[X^2] - 2E[XY] - 5.E[X]^2 + 2E[X]E[Y]$$

Vamos organizar os termos:

$$Cov\,(5X - 2Y, X) = 5.(\,E[X^2] - E[X]^2) - 2.(E[XY] - E[X].E[Y])$$

Notemos que a variância é igual à média dos quadrados menos o quadrado da média e que a covariância é igual à esperança dos produtos menos o produto das esperanças.

$$Var\,(X) = (\,E[X^2] - E[X]^2)$$

$$Cov\,(X, Y) = \;E[XY] - E[X].E[Y]$$

Dessa forma, olhando para essas expressões dentro da expressão da covariância Cov (5X – 2Y, X), enxergamos que:

$$Cov\,(5X - 2Y, X) = 5.Var(X) - 2.Cov(X, Y)$$

$$Cov\,(5X - 2Y, X) = 5.\sigma^2 - 2.0{,}2.\sigma^2 = 5\sigma^2 - 0{,}4\sigma^2$$

$$Cov\,(5X - 2Y, X) = 4{,}6\sigma^2$$

Por fim, vamos utilizar a definição de correlação novamente. A correlação é igual à covariância dividida pelo produto dos desvios padrões.

$$\rho = \frac{Cov\,(5X - 2Y, X)}{\sigma(X).\sigma(5X - 2Y)} = \frac{4{,}6\sigma^2}{\sigma.5\sigma^2} = \frac{4{,}6}{5} = 0{,}92$$

Letra e.

GABARITO

1. C
2. d
3. C
4. d
5. E
6. b
7. e
8. d
9. C
10. d
11. b
12. b
13. c
14. C
15. e

Thiago Cardoso

Engenheiro eletrônico formado pelo ITA com distinção em Matemática, analista-chefe da Múltiplos Investimentos, especialista em mercado de ações. Professor desde os 19 anos e, atualmente, leciona todos os ramos da Matemática para concursos públicos.