



UNIDADE II

Estatística e Probabilidade

Prof. Me. Antônio Palmeira

Estatística Indutiva

- Caracterizada pelo uso de técnicas que possibilitam avaliar características de uma população por meio do estudo de uma amostra dela.
- Ela nasce da ideia de que nem sempre uma amostra grande é uma amostra boa.

Exemplo:

- Encontrar a idade média dos 600 alunos da Escola ABC, que tem estudantes de 6 a 17 anos matriculados no Ensino Fundamental (10 turmas) e no Ensino Médio (6 turmas).
 - Se pegarmos uma amostra grande, com 400 alunos, mas todos do Ensino Fundamental, e calcularmos a média das idades desses alunos, NÃO teremos uma boa aproximação da idade média de todos os 600 estudantes da Escola ABC.
 - Observação: Amostra boa é amostra que traz consigo todas as características presentes na população e na proporção em que ocorrem na população.

Bases da Estatística Indutiva

- Ao trabalharmos com estatística indutiva, de modo geral, chamamos de X a variável aleatória que representa a característica que queremos estudar em dada população.
- Dessa população, retiramos uma amostra de tamanho n , representada por $(X_1, X_2, \dots, X_i, \dots, X_n)$.
- A partir disso precisamos definir: parâmetro, estimador e estimativa.

Parâmetro

- É a quantidade da característica da população que estamos estudando.
- Na maioria das vezes, não conhecemos tal valor.
- Por isso utilizamos uma estimativa para fazer inferências.

Exemplo:

- μ (média) é o parâmetro cujo valor fornece o peso médio das pessoas entre 15 e 65 anos que moram em uma cidade fictícia, chamada de Novo Brasil, que tem cerca de 5 mil habitantes;
 - σ^2 (variância) é o parâmetro cujo valor fornece a variância do peso médio das pessoas entre 15 e 65 anos que moram na cidade Novo Brasil.
 - Veja que, nos exemplos apresentados, a população é formada por todas as pessoas entre 15 e 65 anos que moram na cidade Novo Brasil.

Estimador

- Representa o resultado da amostra usado para estimar determinado parâmetro populacional.
- É uma variável aleatória que depende dos componentes $X_1, X_2, \dots, X_i, \dots, X_n$ da amostra.

Exemplo

- \bar{x} é o símbolo do estimador usado para estimar o parâmetro peso médio das pessoas entre 15 e 65 anos que moram na cidade Novo Brasil.
- Imagine que, para constituirmos esse estimador, usamos uma amostra aleatória formada por 12 pessoas entre 15 e 65 anos que moram na cidade Novo Brasil.
- Essa amostra, de tamanho $n = 12$, é representada por $(X_1, X_2, \dots, X_{12})$.

Suponha que o estimador \bar{x} seja a média das observações X_1, X_2, \dots, X_{12} . Assim, nesse caso, temos:

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_{12}}{12}$$

Estimador

Um bom estimador deve ser:

- não viciado (seu valor esperado é o valor do parâmetro em foco);
- consistente (quanto mais aumentamos o tamanho da amostra, mais seu valor converge para o “valor” do parâmetro em foco e mais sua variância vai para 0).

	Parâmetro	Estimador	Estimativa
Média	μ	$\bar{X} = \frac{X_1 + X_2 + \dots + X_i + \dots + X_n}{n}$	\bar{X}_{obs}
Variância	σ^2	$S^2 = \frac{1}{n-1} \left(\sum_{i=1}^n X_i^2 - n\bar{X}^2 \right)$	S^2_{obs}

Estimativa

- Uma estimativa é um valor “específico” de um estimador quando usamos valores “específicos” de determinada amostra.
- Exemplo: Imagine que, para determinada amostra de 12 pessoas entre 15 e 65 anos que moram na cidade Novo Brasil, tenhamos observado os valores a seguir de pesos, em kg.
 $(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9, X_{10}, X_{11}, X_{12}) = (58, 67, 76, 57, 69, 77, 72, 63, 65, 54, 51, 66)$

Admita que a “fórmula” do estimador \bar{X} empregado para estimar o peso médio da população da cidade Novo Brasil seja a média das observações X_1, X_2, \dots, X_{12} , conforme já vimos:

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_{12}}{12}$$

Se aplicarmos os valores da amostra coletada à expressão anterior, obteremos uma estimativa pontual, indicada por \bar{x}_{obs} , para a média populacional μ :

$$\bar{X}_{obs} = \frac{58 + 67 + 76 + 57 + 69 + 77 + 72 + 63 + 65 + 54 + 51 + 66}{12} = \frac{775}{12} = 64,6 \text{ kg}$$

Teorema central do limite (TCL)

Amostra aleatória simples de tamanho “n” de uma população cujos parâmetros são μ e σ^2 . Considere os seus estimadores média amostral \bar{X} e variância amostral S^2 calculadas por:

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_i + \dots + X_n}{n} \quad S^2 = \frac{1}{n-1} \left(\sum_{i=1}^n X_i^2 - n\bar{X}^2 \right)$$

- Observe que diferentes amostras geram diferentes médias amostrais. Por exemplo A, B, C e D, geram, respectivamente, médias amostrais $\bar{X}_A, \bar{X}_B, \bar{X}_C$ e \bar{X}_D .
- Se n é suficientemente grande, a distribuição das médias amostrais comporta-se como uma distribuição normal, que tem como média a média populacional (μ) e como variância a variância populacional (σ^2) dividida pela raiz quadrada do tamanho da amostra (\sqrt{n}).

Assim, o TCL garante que, em amostras aleatórias simples grandes, a distribuição da média amostral é a seguinte:

$$\bar{X} \sim N \left(\mu; \frac{\sigma^2}{n} \right)$$

Importância do TCL

- O TCL é extremamente importante para a estatística indutiva, porque assegura que a média amostral de uma amostra aleatória simples é um estimador não viciado para a média populacional.
- Isso significa que, se extraíssemos muitas amostras aleatórias simples de uma mesma população e calculássemos a média das médias amostrais, ela seria muito próxima da média populacional verdadeira.
- Esse teorema aponta que a média amostral é um estimador bastante eficiente, pois a variância da média amostral é inversamente proporcional ao tamanho da amostra.
 - Por exemplo, como a variância da média amostral é igual a σ^2/n , caso tenhamos uma amostra do peso de 1000 crianças, a variância da média amostral do peso será 1000 vezes menor do que a variância amostral do peso das crianças.

Estimadores pontuais x Estimadores intervalares

- Os estimadores apresentados até agora são chamados de estimadores pontuais, pois estimam, por meio de “números fixos”, os valores (parâmetros) da média populacional μ e da variância populacional σ^2 .

	Parâmetro	Estimador	Estimativa
Média	μ	$\bar{X} = \frac{X_1 + X_2 + \dots + X_i + \dots + X_n}{n}$	\bar{X}_{obs}
Variância	σ^2	$S^2 = \frac{1}{n-1} \left(\sum_{i=1}^n X_i^2 - n\bar{X}^2 \right)$	S^2_{obs}

Fonte: Adaptado do livro-texto.

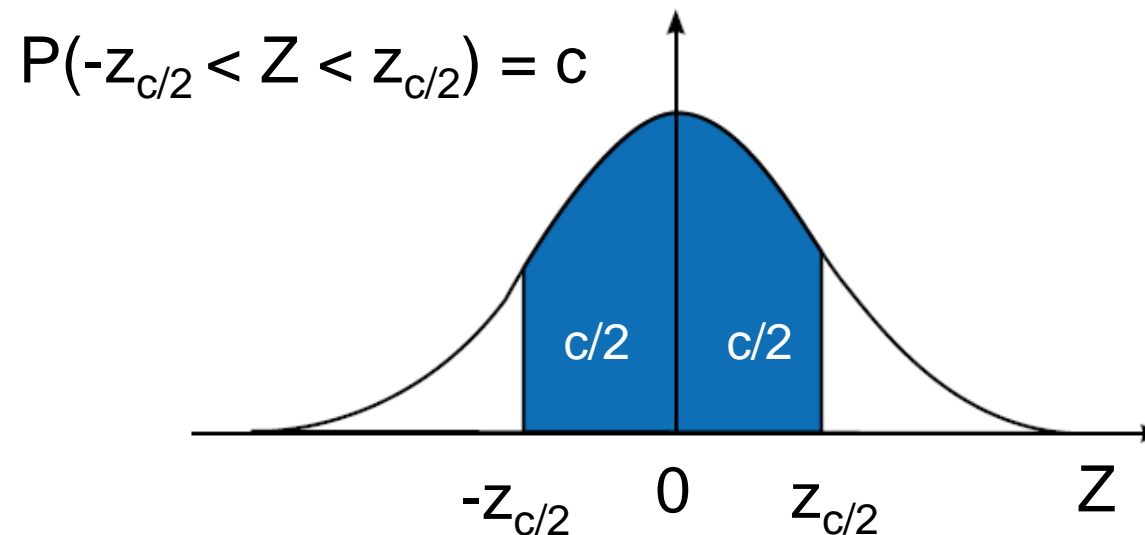
Muitas vezes, é interessante usar estimadores intervalares, em vez de estimadores pontuais, dizendo que:

- a média populacional μ situa-se, com determinado coeficiente de confiança c , entre $\mu - a$ e $\mu + a$, ou seja, no intervalo de confiança $IC = [\mu - a; \mu + a]$;
- a variância populacional σ^2 situa-se, com determinado coeficiente de confiança c , entre $\sigma^2 - b$ e $\sigma^2 + b$, ou seja, no intervalo de confiança $IC = [\sigma^2 - b; \sigma^2 + b]$.

Intervalo de confiança para a média com variância populacional conhecida

- Vamos pensar, inicialmente, no intervalo de confiança para a média μ de uma população que segue modelo normal e cuja variância σ^2 seja conhecida.
- Imagine que, dessa população, retiremos uma amostra de tamanho n representada pelas variáveis aleatórias independentes (X_1, X_2, \dots, X_n) , sendo \bar{X} a média amostral.
- De acordo com o TCL, a média amostral também segue distribuição normal de probabilidades com média μ e variância σ^2/n . Logo:
$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim N(0;1)$$

Fixado determinado coeficiente de confiança c , com $0 < c < 1$, podemos encontrar $Z_{c/2}$, de modo que:

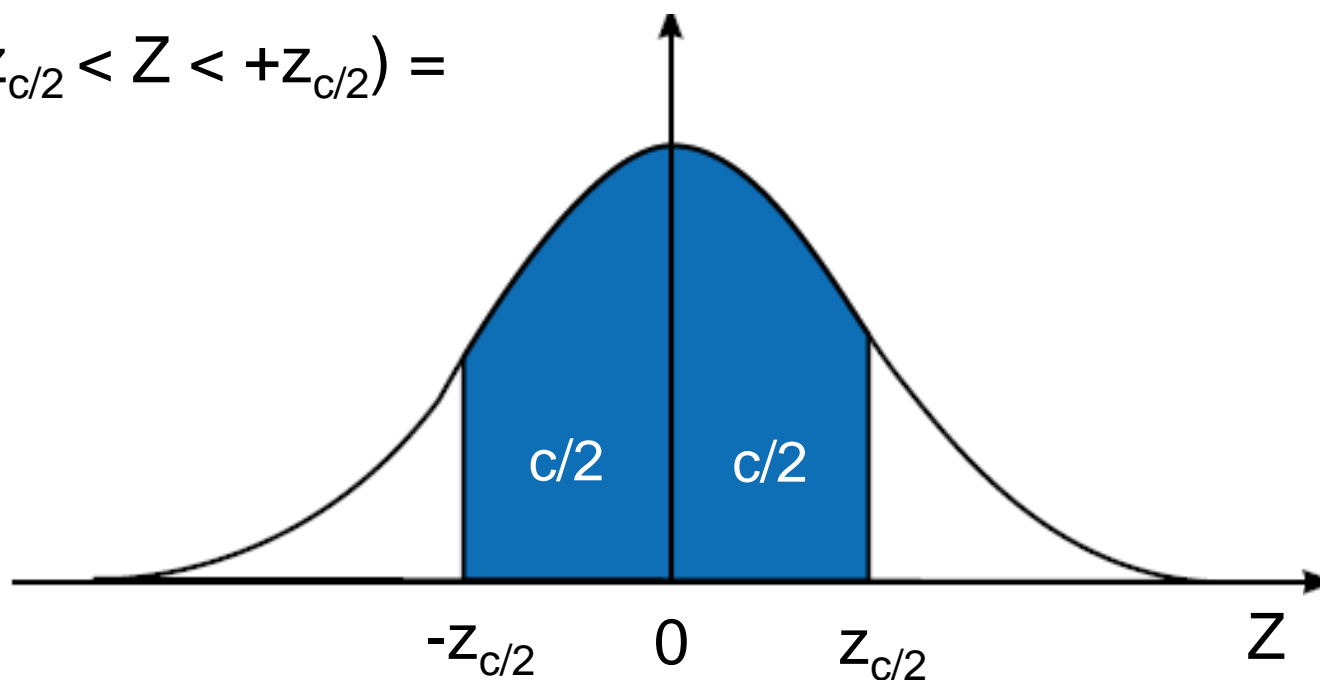


Intervalo de confiança para a média com variância populacional conhecida

Nesse caso, o intervalo de confiança para a média μ , com coeficiente de confiança c , indicado por $IC(\mu; c)$, para dado valor de média amostral observada \bar{X}_{obs} , é calculado por:

$$IC(\mu; c) = \left[\bar{X}_{obs} - z_{c/2} \cdot \frac{\sigma}{\sqrt{n}}; \bar{X}_{obs} + z_{c/2} \cdot \frac{\sigma}{\sqrt{n}} \right]$$

$$P(-z_{c/2} < Z < +z_{c/2}) = c$$



Interatividade

A quantidade da característica da população que estudamos em um processo utilizando estatística indutiva é chamada de:

- a) Parâmetro.
- b) Correlação.
- c) Estimativa.
- d) Estimador.
- e) Média.

Resposta

A quantidade da característica da população que estudamos em um processo utilizando estatística indutiva é chamada de:

- a) **Parâmetro.**
- b) Correlação.
- c) Estimativa.
- d) Estimador.
- e) Média.

Exemplo 1 de estimativa intervalar

- Imagine que a distribuição das alturas das pessoas com mais de 18 anos que moram na cidade fictícia Novo Mundo obedeça a um modelo normal com média μ desconhecida e com variância σ^2 igual a $1,06 \text{ m}^2$.
- Foi feita uma amostra aleatória de 55 dessas pessoas, o que forneceu média amostral observada \bar{X}_{obs} igual a $1,71 \text{ m}$.

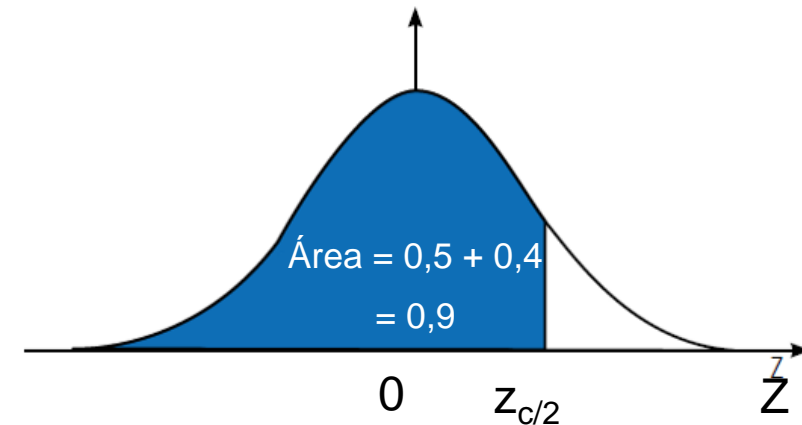
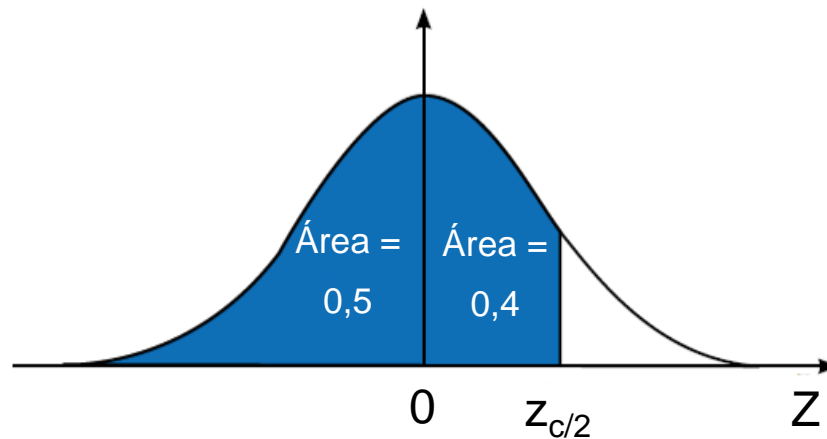
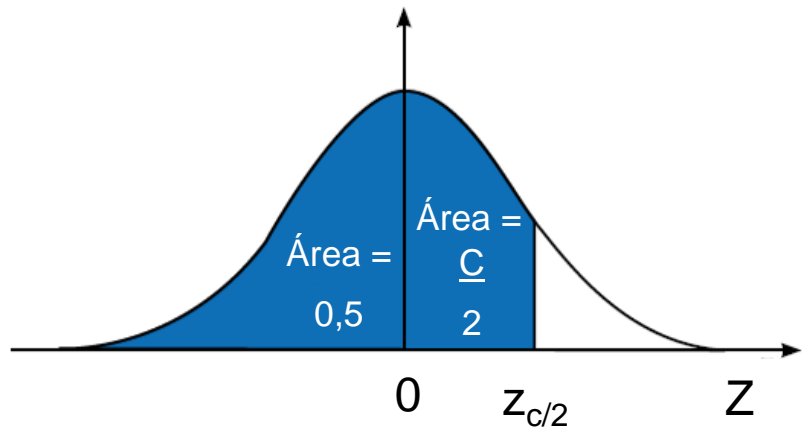
Para essa situação, qual é a estimativa intervalar da média populacional μ com coeficiente de confiança de 80%?

Resumo dos dados fornecidos pelo exemplo 1

- Modelo de distribuição de probabilidades das alturas: normal.
- Média populacional das alturas: parâmetro μ desconhecido.
- Variância populacional das alturas: parâmetro $\sigma^2 = 1,06 \text{ m}^2$.
- Desvio padrão populacional das alturas: parâmetro $\sigma = \sqrt{\sigma^2} = 1,03 \text{ m}$.
- Média amostral das alturas: estimador \bar{X} .
- Tamanho da amostra: $n = 55$.
- Média amostral das alturas observada na amostra: estimativa $\bar{X}_{\text{obs}} = 1,71 \text{ m}$.
- Coeficiente de confiança da estimativa intervalar: $c = 0,80$.

Resolução do Exemplo 1

- Se o coeficiente de confiança “c” vale 0,80, então $c/2 = 0,40$.
- Precisamos achar $Z_{c/2}$, tal que tenhamos as configurações ilustradas a seguir.



- Observando a tabela normal reduzida, o valor 0,9, encontramos 0,8997, e ele corresponde a $Z_{c/2} = 1,28$.

Z	0,08
1,2	0,8997 \approx 0,9

Resolução do Exemplo 1

- Agora, podemos calcular o intervalo de confiança para a média populacional das alturas μ , com coeficiente de confiança $c = 0,8$ (80%), indicado por $IC(\mu;c)$, para o valor de média amostral observada $\bar{X}_{obs} = 1,71$ m, com $Z_{c/2} = 1,28$, $n = 55$ e $\sigma = 1,03$ m.

$$IC(\mu;c) = \left[\bar{X}_{obs} - Z_{c/2} \cdot \frac{\sigma}{\sqrt{n}}; \bar{X}_{obs} + Z_{c/2} \cdot \frac{\sigma}{\sqrt{n}} \right]$$

$$IC(\mu;0,8) = \left[1,71 - 1,28 \cdot \frac{1,03}{\sqrt{55}}; 1,71 + 1,28 \cdot \frac{1,03}{\sqrt{55}} \right]$$

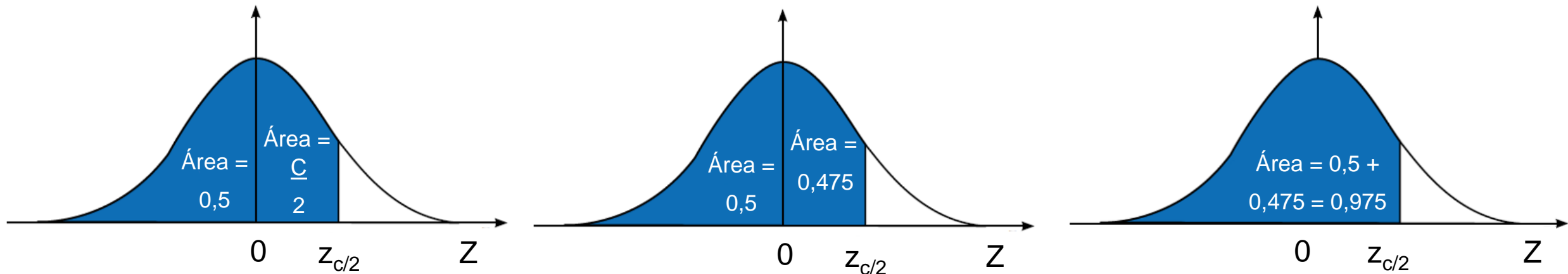
$$IC(\mu;0,8) = [1,71 - 0,18; 1,71 + 0,18] = [1,53; 1,89]$$

- Veja que, com confiança de 80%, “acreditamos” que a média populacional das alturas μ das pessoas com mais de 18 anos que moram na cidade fictícia Novo Mundo esteja entre 1,53 m e 1,89 m.

Exemplo 2 de estimativa intervalar

- Considerando a mesma situação e aumentando a confiança para 95%, o que acontecerá com o intervalo de confiança para a média populacional?

Como c vale 0,95, $c/2$ vale 0,475. Encontramos o $Z_{c/2}$ a partir das funções a seguir:



- Observando a tabela normal reduzida, o valor 0,975 corresponde a $Z_{c/2} = 1,96$.

Z	0,06
1,9	0,975

Resolução do Exemplo 2

- Agora, podemos calcular o intervalo de confiança para a média populacional das alturas μ , com coeficiente de confiança $c = 0,95$ (95%), indicado por $IC(\mu;c)$, para o valor de média amostral observada $\bar{X}_{obs} = 1,71$ m, com $Z_{c/2} = 1,96$, $n = 55$ e $\sigma = 1,03$ m.

$$IC(\mu;c) = \left[\bar{X}_{obs} - z_{c/2} \cdot \frac{\sigma}{\sqrt{n}}; \bar{X}_{obs} + z_{c/2} \cdot \frac{\sigma}{\sqrt{n}} \right] \quad IC(\mu;0,95) = \left[1,71 - 1,96 \cdot \frac{1,03}{\sqrt{55}}; 1,71 + 1,96 \cdot \frac{1,03}{\sqrt{55}} \right]$$

$$IC(\mu;0,95) = [1,71 - 0,27; 1,71 + 0,27] = [1,44; 1,98]$$

- Veja que, com confiança de 95%, “acreditamos” que a média populacional das alturas μ das pessoas com mais de 18 anos que moram na cidade fictícia Novo Mundo esteja entre 1,44 m e 1,98 m.
- Observação: Quanto mais aumentamos a confiança, mais aumentamos a amplitude do intervalo de confiança.

Introdução ao teste de hipóteses

- Um teste de hipóteses tem o objetivo de, com base nas características de uma amostra representativa de uma população, concluir informações sobre a população como um todo, atividade conhecida como inferência estatística.

Exemplo:

- O tempo de retífica de peça em uma máquina é uma variável aleatória contínua e segue uma distribuição normal de probabilidades com média igual a 120s e desvio padrão igual a 5s.
- Essa máquina será substituída por outra, mais nova, cujo tempo de retífica segue a mesma distribuição.
- É verdadeira a suspeita de que o tempo médio de retífica da peça diminua?
 - A resposta pode ser dada por meio de um teste de hipóteses.
 - Nesse teste, tomamos uma hipótese como referência: trata-se da hipótese nula, indicada por H_0 .
 - Prosseguimos fazendo uma comparação entre a hipótese nula e uma hipótese alternativa, indicada por H_a .

Resposta para o exemplo de teste de hipótese (Apresentando as hipóteses)

- Para a situação em estudo, vamos chamar de X a variável aleatória contínua que representa o tempo, em segundos, que a máquina leva para retificar a peça.
- Sabemos que X segue uma distribuição normal de média $\mu = 120\text{s}$ e desvio padrão $\sigma = 5\text{s}$, ou seja, de variância $\sigma^2 = 5^2 = 25\text{ s}^2$, o que é indicado por $X \sim N(120;25)$.

Queremos testar as hipóteses a seguir.

- Hipótese nula (H_0): o tempo médio de retífica permanece igual a 120s com a máquina nova. $H_0: \mu = 120\text{s}$.
- Hipótese alternativa (H_a): o tempo médio de retífica é menor do que 120s com a máquina nova. $H_a: \mu < 120\text{s}$.
 - Não sabemos exatamente o que irá acontecer, visto que a hipótese nula H_0 e a hipótese alternativa H_a são conjecturas (suposições) que fazemos sobre um parâmetro populacional que não conhecemos.

Resposta para o exemplo de teste de hipótese (Explicando os tipos de erros)

Erros quando rejeitamos H_0 e quando aceitamos H_0 :

- Erro tipo I (falso positivo) – ocorre quando rejeitamos H_0 , sendo H_0 , na realidade, verdadeira.
- Erro tipo II (falso negativo) – ocorre quando não rejeitamos H_0 , sendo H_0 , na realidade, falsa.

As decisões em que não cometemos erros são as seguintes.

- Rejeitamos H_0 , e H_0 é falsa.
- Não rejeitamos H_0 , e H_0 é verdadeira.

		Realidade	
		Ho é verdadeira	Ho é falsa
Decisão	Rejeito H_0	Erro tipo I	Sem erro
	Não rejeito H_0	Sem erro	Erro tipo II

Resposta para o exemplo de teste de hipótese (Explicando as probabilidades de erros)

Vamos chamar de “ α ” a probabilidade de ocorrência de erro tipo I e de “ β ” a probabilidade de ocorrência de erro tipo II. Assim temos:

- A probabilidade α é chamada de nível de significância do teste e está relacionada ao controle do erro tipo I.
- $\alpha = P(\text{erro tipo I}) = P(\text{rejeitar } H_0/H_0 \text{ é verdadeira})$
- A probabilidade β é chamada de poder do teste e está relacionada ao controle do erro tipo II.
- $\beta = P(\text{erro tipo II}) = P(\text{não rejeitar } H_0/H_0 \text{ é falsa})$

Observação: a soma das probabilidades α e β não resulta em 1 (ou 100%). Mas, quanto mais diminuirmos α , mais aumentamos β .

Resposta para o exemplo de teste de hipótese (Voltando ao exemplo)

- Voltemos à situação da nova máquina retificadora.
- Imagine que façamos uma amostra de 30 tempos de retificação.
- Para dado nível de significância α do teste de hipóteses, descrevemos o valor crítico x_c , conforme segue.

$$\alpha = P(\text{erro tipo I}) = P(\text{rejeitar } H_0/H_0 \text{ é verdadeira}) = P(\bar{X} < 120/\mu = 120)$$

Imagine que adotemos nível de significância de 5% ($\alpha = 0,05$). Assim, ficamos com:

$$0,05 = P(\bar{X} < 120/\mu = 120)$$

Vimos, pelo TCL, que, se a variável X segue uma distribuição normal de média μ e variância σ^2 , ou seja, $X \sim N(\mu; \sigma^2)$, então a média amostral também segue distribuição normal de probabilidades com média μ e variância σ^2/n , ou seja, $\bar{X} \sim N(\mu; \sigma^2/n)$. Logo:

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim N(0; 1)$$

Interatividade

Em um teste de hipótese, aquela conhecida como hipótese nula é conhecida como:

- a) H_{nulo} .
- b) H_0 .
- c) H_a .
- d) H .
- e) H_1 .

Resposta

Em um teste de hipótese, aquela conhecida como hipótese nula é conhecida como:

- a) H_{nulo} .
- b) H_0 .**
- c) H_a .
- d) H .
- e) H_1 .

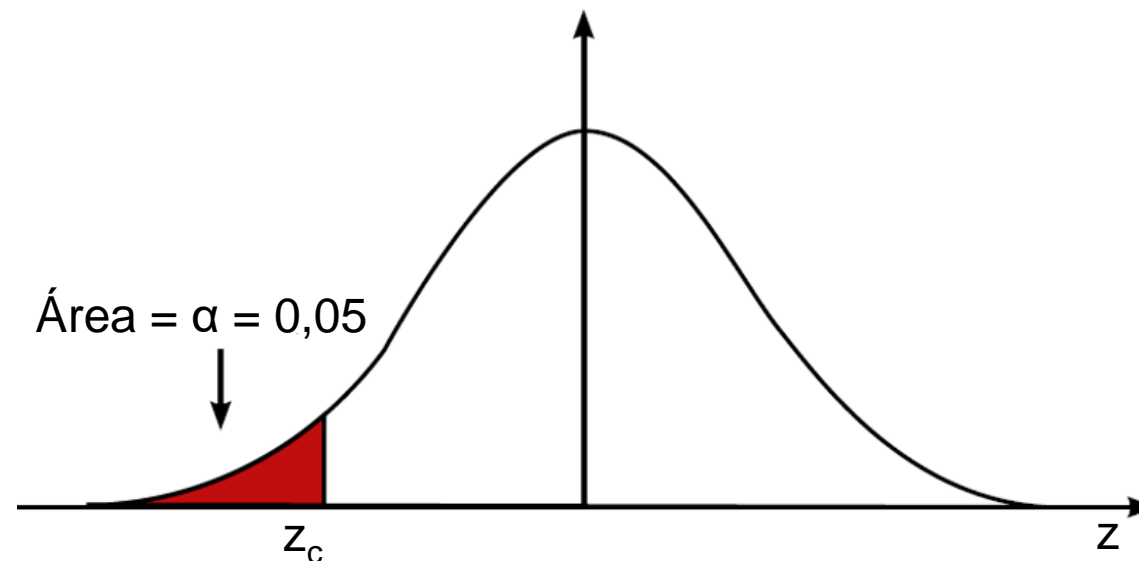
Resposta para o exemplo de teste de hipótese (Continuando o exemplo)

Imagine que façamos uma amostra de 30 tempos de retificação da máquina nova, ou seja, obtemos uma amostra de tamanho $n = 30$. Assim, para o exemplo em estudo, temos:

$$0,05 = P(\bar{X} < 120 / \mu = 120) = P\left(\frac{\bar{X} - \mu}{\sigma / \sqrt{n}} < \frac{x_c - 120}{5 / \sqrt{30}}\right)$$

Chamamos $\frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$ de Z e $\frac{x_c - 120}{5 / \sqrt{30}}$ de z_c e chegamos a: $0,05 = P(Z < z_c)$

- Queremos achar Z_c de modo que tenhamos o que se ilustra na gaussiana a seguir.



Fonte: Adaptado do livro-texto.

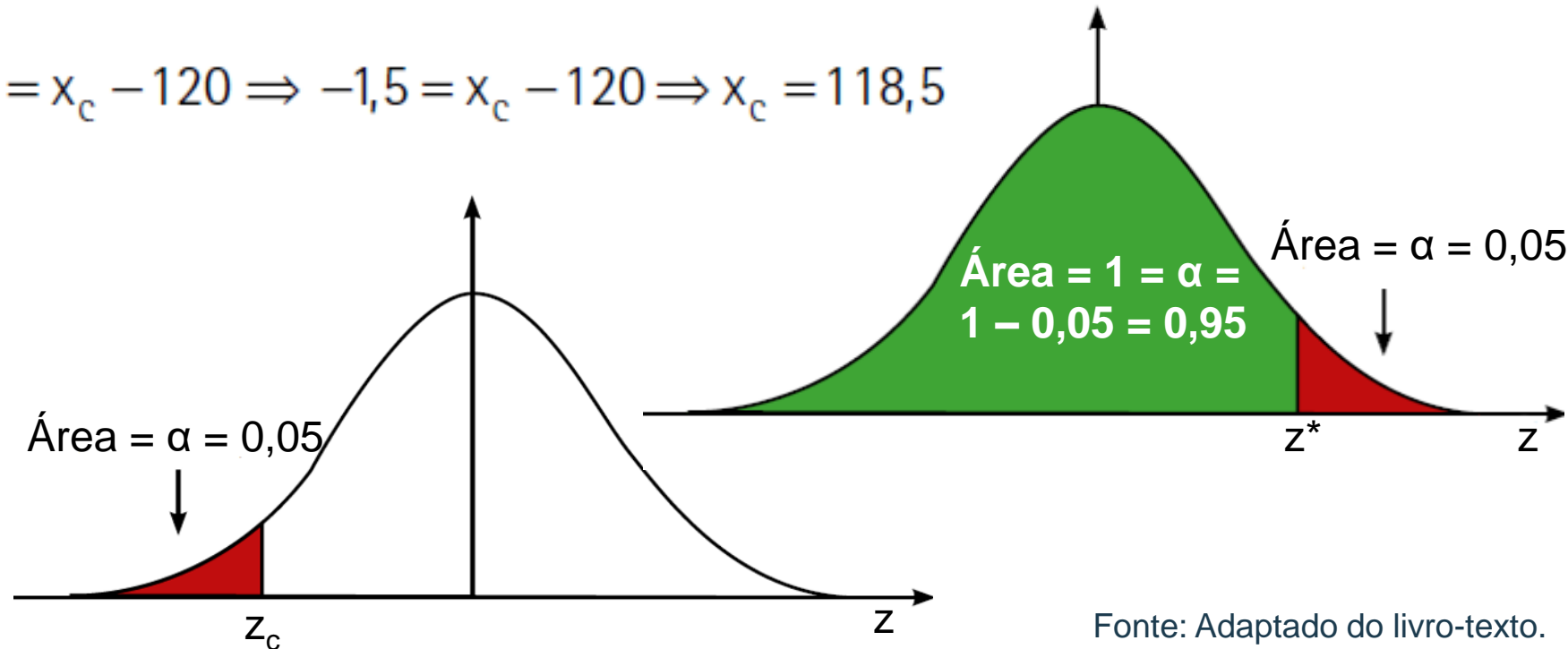
Resposta para o exemplo de teste de hipótese (Continuando o exemplo)

- Precisamos encontrar, “dentro” da tabela normal reduzida, o valor 0,95, em que o valor mais próximo de é 0,9505, que corresponde a $z^* = 1,65$.
- Como $z^* = 1,65$ e $z^* = -Z_c$, então $Z_c = -1,65$.

$$Z_c = \frac{x_c - 120}{5 / \sqrt{30}}$$

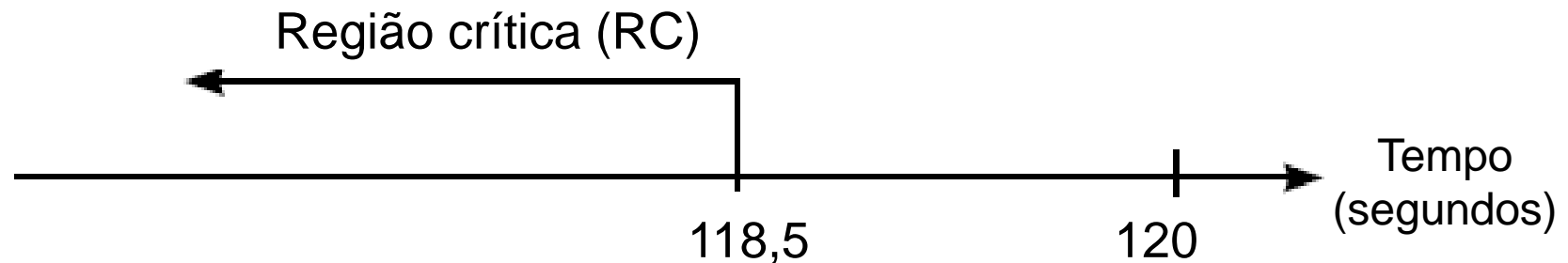
Z	0,05
1,6	0,9505 \approx 0,95

$$-1,65 = \frac{x_c - 120}{5 / \sqrt{30}} \Rightarrow -1,65 \cdot \frac{5}{\sqrt{30}} = x_c - 120 \Rightarrow -1,5 = x_c - 120 \Rightarrow x_c = 118,5$$



Resposta para o exemplo de teste de hipótese (Continuando o exemplo)

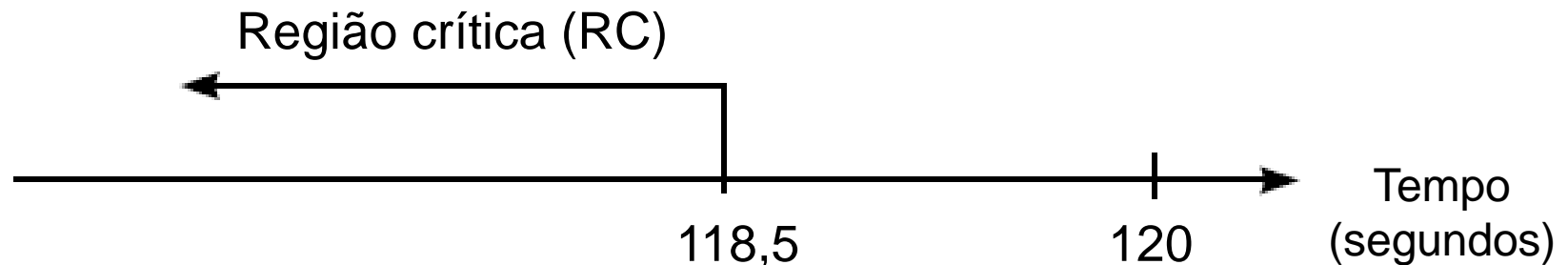
- Podemos dizer que testar uma hipótese estatística é estabelecer uma regra que possibilite, com base na informação de uma amostra, decidir pela rejeição ou pela não rejeição da hipótese nula H_0 .
- No caso em análise, com uma amostra de tamanho $n = 30$, essa regra, expressa pela região crítica (RC) ao nível de significância de 5% ($\alpha = 0,05$), ilustrada a seguir, é dada por $\bar{x} < 118,5$ s.



Fonte: Adaptado do livro-texto.

Resposta para o exemplo de teste de hipótese (Conclusão)

- Com uma amostra de tamanho $n = 30$, o conjunto de valores de tempo, representado pela variável X , que levam à não aceitação da hipótese nula H_0 é dado por $x < 118,5s$.
- Ou seja, se a média amostral dos tempos resultar em valor menor do que $118,5s$, ao nível de significância de 5%, não aceitaremos que o tempo médio de retificação da nova máquina seja igual a 120 segundos (nesse caso, acataremos a hipótese H_a , segundo a qual o tempo médio de retificação da nova máquina é menor do que 120s).
- No entanto, se em uma amostra de 30 tempos de retificação da máquina nova tivermos a média amostral de 119 s, por exemplo, não vamos concluir, ao nível de significância de 5%, que o tempo médio de retificação da nova máquina é menor do que 120s. Fazendo com que não rejeitemos a hipótese nula H_0 .



Etapas em um teste de hipóteses para média com variância populacional conhecida

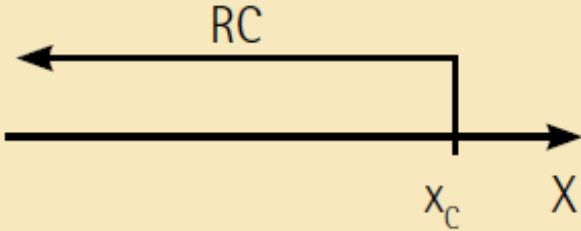
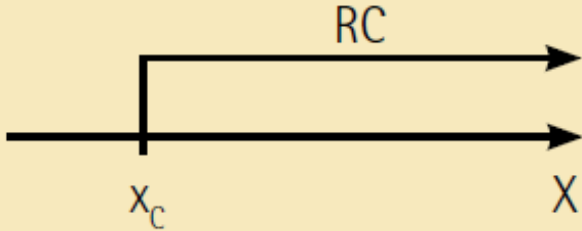
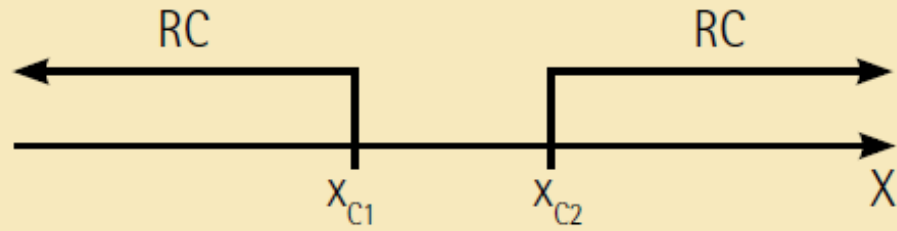
- **Etapa 1:** Estabelecer as hipóteses H_0 e H_a em relação ao valor de média amostral de referência μ_0 , em que podemos ter o caso 1 (unilateral à esquerda), o caso 2 (unilateral à direita) ou o caso 3 (bilateral).

Caso 1	Caso 2	Caso 3
$H_0: \mu = \mu_0$ $H_a: \mu < \mu_0$	$H_0: \mu = \mu_0$ $H_a: \mu > \mu_0$	$H_0: \mu = \mu_0$ $H_a: \mu \neq \mu_0$

Fonte: Adaptado do livro-texto.

Etapas em um teste de hipóteses para média com variância populacional conhecida

- **Etapa 2:** Definir uma regra de decisão com base em H_a , em que X_c representa o valor crítico e RC representa a região crítica (zona de rejeição de H_0).

Regra 1	Regra 2	Regra 3
$H_a: \mu < \mu_0$ Rejeitar H_0 se $\bar{X} \leq x_c$	$H_a: \mu > \mu_0$ Rejeitar H_0 se $\bar{X} \geq x_c$	$H_a: \mu \neq \mu_0$ Rejeitar H_0 se $\bar{X} \leq x_{c1}$ ou se $\bar{X} \geq x_{c2}$
		

Fonte: Adaptado do livro-texto.

Etapas em um teste de hipóteses para média com variância populacional conhecida

- **Etapa 3:** Identificar o estimador e o tipo de distribuição de probabilidades que essa variável aleatória segue.
- Por exemplo, se a variável aleatória populacional X segue uma normal de média μ e variância σ^2 , ou seja, $X \sim N(\mu; \sigma^2)$, então a variável aleatória amostral \bar{X} , relativa a uma amostra de tamanho n , segue uma normal de média μ e variância σ^2/n , ou seja, $\bar{X} \sim N(\mu; \sigma^2/n)$.
 - **Etapa 4.** Fixar o nível de significância α (ou seja, a probabilidade de rejeitar H_0 , sendo H_0 verdadeira) e determinar a RC.
 - **Etapa 5.** Concluir o teste verificando se o valor de média observado na amostra, indicado por \bar{X}_{obs} , pertence ou não pertence à RC.

Testes qui-quadrado

Teste de aderência

- Visam a testar se dado modelo probabilístico é adequado a determinado conjunto de dados.
- Nesses testes, verificamos se a distribuição das frequências absolutas de fato observadas de uma variável é significativamente diferente da distribuição das frequências absolutas esperadas para essa variável.

Teste de independência

- Visam a testar se há independência entre duas variáveis A e B .

Teste de independência

- Inicialmente, organizamos as frequências “O” observadas em uma tabela de dupla entrada, com r linhas e s colunas, como a mostrada a seguir.

A/B	B_1	B_2	...	B_s	Total
A_1	O_{11}	O_{12}	...	O_{1s}	
A_2	O_{21}	O_{22}	...	O_{2s}	
...	
A_r	O_{r1}	O_{r2}	...	O_{rs}	
Total					n

Fonte: Adaptado do livro-texto.

Então, estabelecemos as hipóteses H_0 e H_a indicadas a seguir:

- Hipótese nula (H_0): as variáveis A e B são independentes.
- Hipótese alternativa (H_a): as variáveis A e B não são independentes.

Teste de independência

- Sendo E_{ij} a frequência esperada para a medida ij , fazemos a quantificação das diferenças entre as frequências observadas e suas respectivas frequências esperadas por meio da estatística a seguir, indicada por Q^2 .

$$Q^2 = \sum_{i=1}^r \sum_{j=1}^s \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

- Se H_0 é verdadeira, então a variável aleatória Q^2 segue aproximadamente uma distribuição χ^2 (letra grega qui elevada ao quadrado) com q graus de liberdade (χ^2_q).
- Isso é válido para número total de observações n “grande” ($n \geq 30$) e para no mínimo 5 frequências absolutas esperadas em cada categoria.
 - Quando aplicamos o cálculo de Q^2 a um conjunto específico de observações, obtemos o valor Q^2_{obs} .

$$Q^2_{\text{obs}} = \sum_{i=1}^r \sum_{j=1}^s \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \approx \chi_q^2$$

Teste de independência

- Na expressão, q é o número de graus de liberdade, sendo $q = (r - 1) \cdot (s - 1)$.
- Identificamos $P = P(\chi^2_q \geq Q^2_{\text{obs}})$, em que, como acabamos de dizer, Q^2_{obs} é o valor calculado para Q^2 com base nos dados observados.
- Finalmente, se, para determinado nível α fixado, temos $P \leq \alpha$, então rejeitamos H_0 .

Exemplo de Teste de independência

- Imagine que a fábrica Chocolate Delicioso produza 4 tipos de chocolate: Choc A, Choc B, Choc C e Choc D.
- O gestor dessa fábrica quer saber se a preferência por tipo de chocolate é ou não é independente do local de moradia do consumidor (zona sul, zona norte ou zona central). Para isso, fez uma pesquisa com 1320 consumidores e obteve as observações apresentadas a seguir.

O que se pode concluir dessas observações ao nível de significância de 5%?

Local de moradia	Tipo de chocolate preferido				
	Choc A	Choc B	Choc C	Choc D	Total
Zona sul	11	9	5	28	53
Zona norte	91	165	126	75	457
Zona central	202	257	221	130	810
Total	304	431	352	233	1320

Interatividade

A região expressa pela regra que possibilite, com base na informação de uma amostra, decidir pela rejeição ou pela não rejeição da hipótese nula H_0 , é chamada de:

- a) Região nula.
- b) Região hipotética.
- c) Região crítica.
- d) Região normal.
- e) Região transiente.

Resposta

A região expressa pela regra que possibilite, com base na informação de uma amostra, decidir pela rejeição ou pela não rejeição da hipótese nula H_0 , é chamada de:

- a) Região nula.
- b) Região hipotética.
- c) **Região crítica.**
- d) Região normal.
- e) Região transiente.

Exemplo de Teste de independência (Estabelecendo hipóteses e apresentando o número de observações)

As hipóteses a serem testadas são:

- Hipótese nula (H_0): a preferência por determinado tipo de chocolate não depende do local em que o consumidor mora.
- Hipótese alternativa (H_a): a preferência por determinado tipo de chocolate depende do local em que o consumidor mora.

Encontramos o total “n” de 1320 observações dividido em:

- 4,02% dos 1320 consumidores entrevistados moram na zona sul, pois $(53/1320) \cdot 100\% = 4,02\%$.
- 34,62% dos 1320 consumidores entrevistados moram na zona norte, pois $(457/1320) \cdot 100\% = 34,62\%$.
- 61,36% dos 1320 consumidores entrevistados moram na zona central, pois $(810/1320) \cdot 100\% = 61,36\%$.

Exemplo de Teste de independência (Construção da tabela de valores observados e esperados)

- Se há independência entre o local em que o consumidor mora e o tipo de chocolate que ele prefere, temos as quantidades esperadas mostradas na tabela que será apresentada.
- Como se trata de quantidades teóricas, para fins de cálculos, consideraremos valores não inteiros, mesmo sabendo que o número de consumidores é inteiro.

Exemplo de cálculo:

- Quantidade esperada de consumidores que moram na zona sul e preferem Choc A: 12,22.
$$\frac{(\% \text{ de moradores da zona sul}) \cdot (\text{Número de consumidores que preferem Choc A})}{100\%}$$
$$= \frac{4,02.304}{100\%} = 12,22$$

Exemplo de Teste de independência

(Construção da tabela de valores observados e esperados)

- Na tabela a seguir, temos, nas colunas azuis, as quantidades observadas e, nas colunas verdes, as quantidades esperadas de consumidores.

Local de moradia	Tipo de chocolate								
	Choc A		Choc B		Choc C		Choc D		Total
Zona sul	11	12,22	9	17,33	5	14,15	28	9,37	53
Zona norte	91	105,24	165	149,21	126	121,86	75	80,66	457
Zona central	202	186,53	257	264,46	221	215,99	130	142,97	810
Total	304		431		352		233		1320

Fonte: Adaptado do livro-texto.

Exemplo de Teste de independência

- Na tabela a seguir, temos o destaque das indicações das quantidades observadas (em azul) e das quantidades esperadas (em verde) de consumidores.

Choc A		Choc B		Choc C		Choc D	
$O_{11} = 11$	$E_{11} = 12,22$	$O_{12} = 9$	$E_{12} = 17,33$	$O_{13} = 5$	$E_{13} = 14,15$	$O_{14} = 28$	$E_{14} = 9,37$
$O_{21} = 91$	$E_{21} = 105,24$	$O_{22} = 165$	$E_{22} = 149,21$	$O_{23} = 126$	$E_{23} = 121,86$	$O_{24} = 75$	$E_{24} = 80,66$
$O_{31} = 202$	$E_{31} = 186,53$	$O_{32} = 257$	$E_{32} = 264,46$	$O_{33} = 221$	$E_{33} = 215,99$	$O_{34} = 130$	$E_{34} = 142,97$

- Fazemos a quantificação das diferenças entre as frequências observadas e suas respectivas frequências esperadas:

$$Q^2_{\text{obs}} = \sum_{i=1}^r \sum_{j=1}^s \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = 54,01$$

Como temos $r = 3$ categorias de locais de moradia (zona sul, zona norte e zona central) e $s = 4$ categorias de tipos de chocolate (Choc A, Choc B, Choc C e Choc D), o número de graus de liberdade q vale 6, pois:

$$q = (r - 1) \cdot (s - 1) = (3 - 1) \cdot (4 - 1) = 2 \cdot 3 = 6$$

Fonte: Livro-texto.

Exemplo de Teste de independência

- Assim, temos $\chi_6^2 = 54,01$ e procuramos, em uma tabela de distribuição qui-quadrado, na linha de graus de liberdade (g.l.) igual a 6 ($q = 6$).

g.l.	0,995	0,990	0,975	0,950	0,900	0,500	0,100	0,050	0,025	0,010	0,005
6	0,676	0,872	1,237	1,635	2,204	5,348	10,645	12,592	14,449	16,812	18,548

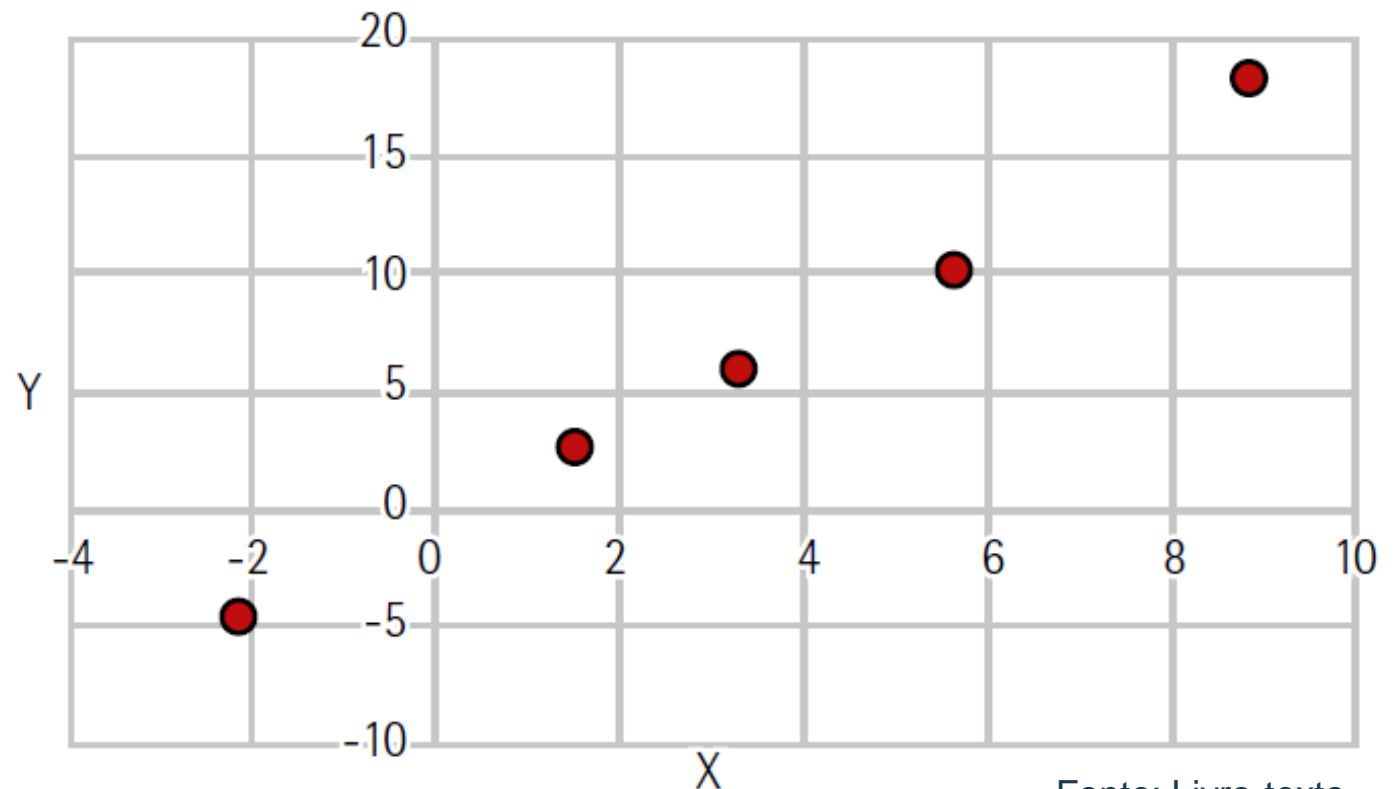
Fonte: Livro-texto.

- O máximo valor na linha de graus de liberdade (g.l.) igual a 6 ($q = 6$) é 18,548, que resulta em $P(\chi_6^2 \geq 18,548) = 0,005$.
 - Logo, temos a certeza de que $P(X_6^2 \geq 54,01) < 0,005$.
 - Assim, concluímos que, ao nível de significância de 5% ($\alpha = 0,05$), rejeitamos H_0 , visto que $P(\chi_6^2 \geq 54,01) < \alpha$.
 - Ou seja, ao nível de significância de 5%, chegamos à conclusão de que a preferência por determinado tipo de chocolate depende do local em que o consumidor mora.

Coeficiente de correlação

Será que essas duas variáveis estão relacionadas ou correlacionadas entre si de maneira aproximadamente linear?

X	-2,1	1,5	3,3	5,6	8,8
Y	-4,4	2,8	6,1	10,1	18,3



Fonte: Livro-texto.

Coeficiente de correlação

- Temos um coeficiente, chamado de coeficiente de correlação e indicado por R, que quantifica o grau de associação entre duas variáveis. Esse coeficiente é calculado pela expressão a seguir, em que n é o número de pares (X;Y).

$$R = \frac{\sum_{i=1}^n x_i \cdot y_i - n \cdot \bar{x}_{obs} \cdot \bar{y}_{obs}}{\sqrt{\left[\sum_{i=1}^n x_i^2 - n \cdot \bar{x}_{obs}^2 \right] \cdot \left[\sum_{j=1}^n y_j^2 - n \cdot \bar{y}_{obs}^2 \right]}}$$

Calculando o Coeficiente de correlação

Calculando o coeficiente de correlação:

$$\bar{X}_{obs} = \frac{-2,1 + 1,5 + 3,3 + 5,6 + 8,8}{5} = \frac{17,1}{5} = 3,42$$

$$\bar{Y}_{obs} = \frac{-4,4 + 2,8 + 6,1 + 10,1 + 18,3}{5} = \frac{32,9}{5} = 6,58$$

$$\sum_{i=1}^n x_i \cdot y_i = x_1 \cdot y_1 + x_2 \cdot y_2 + x_3 \cdot y_3 + x_4 \cdot y_4 + x_5 \cdot y_5$$

$$\sum_{i=1}^n x_i \cdot y_i = (-2,1) \cdot (-4,4) + 1,5 \cdot 2,8 + 3,3 \cdot 6,1 + 5,6 \cdot 10,1 + 8,8 \cdot 18,3$$

$$\sum_{i=1}^n x_i \cdot y_i = 9,24 + 4,2 + 20,13 + 56,56 + 161,04 = 251,17$$

$$n \cdot \bar{x}_{obs} \cdot \bar{y}_{obs} = 5 \cdot 3,42 \cdot 6,58 = 112,518$$

$$\sum_{i=1}^n x_i^2 = (-2,1)^2 + 1,5^2 + 3,3^2 + 5,6^2 + 8,8^2$$

$$\sum_{i=1}^n x_i^2 = 4,41 + 2,25 + 10,89 + 31,36 + 77,44 = 126,35$$

$$\sum_{i=1}^n y_i^2 = (-4,4)^2 + 2,8^2 + 6,1^2 + 10,1^2 + 18,3^2$$

$$\sum_{i=1}^n y_i^2 = 19,36 + 7,84 + 37,21 + 102,01 + 334,89 = 501,31$$

Calculando o Coeficiente de correlação

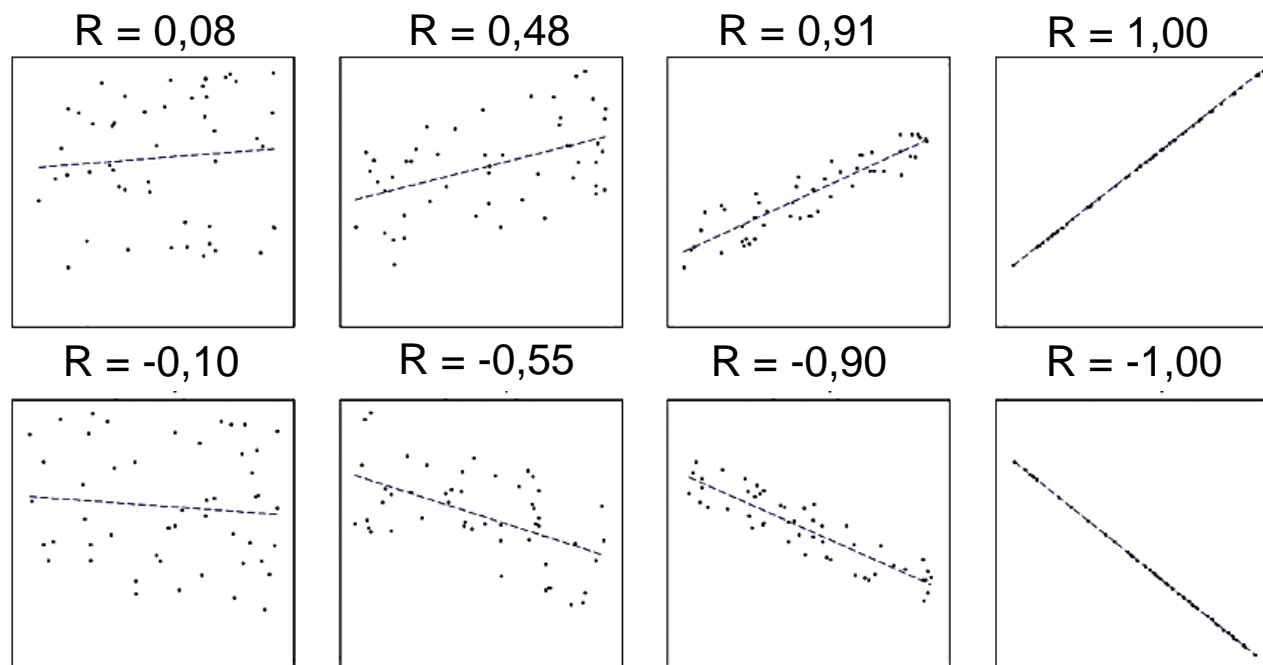
- O coeficiente R igual a 0,997 indica que as variáveis X e Y em estudo têm forte correlação positiva: se X aumenta, Y aumenta, e esse aumento é praticamente linear.
- O coeficiente de correlação R também é chamado de coeficiente de Pearson.

$$R = \frac{\sum_{i=1}^n x_i \cdot y_i - n \cdot \bar{x}_{obs} \cdot \bar{y}_{obs}}{\sqrt{\left[\sum_{i=1}^n x_i^2 - n \cdot \bar{x}_{obs}^2 \right] \cdot \left[\sum_{j=1}^n y_j^2 - n \cdot \bar{y}_{obs}^2 \right]}}$$
$$R = \frac{251,17 - 112,518}{\sqrt{[126,35 - 5 \cdot (3,42)^2] \cdot [501,31 - 5 \cdot (6,58)^2]}}$$
$$R = \frac{138,652}{\sqrt{[126,35 - 58,482] \cdot [501,31 - 216,482]}} = \frac{138,652}{\sqrt{19330,71}}$$

$$R = 0,997$$

Variação do coeficiente de correlação

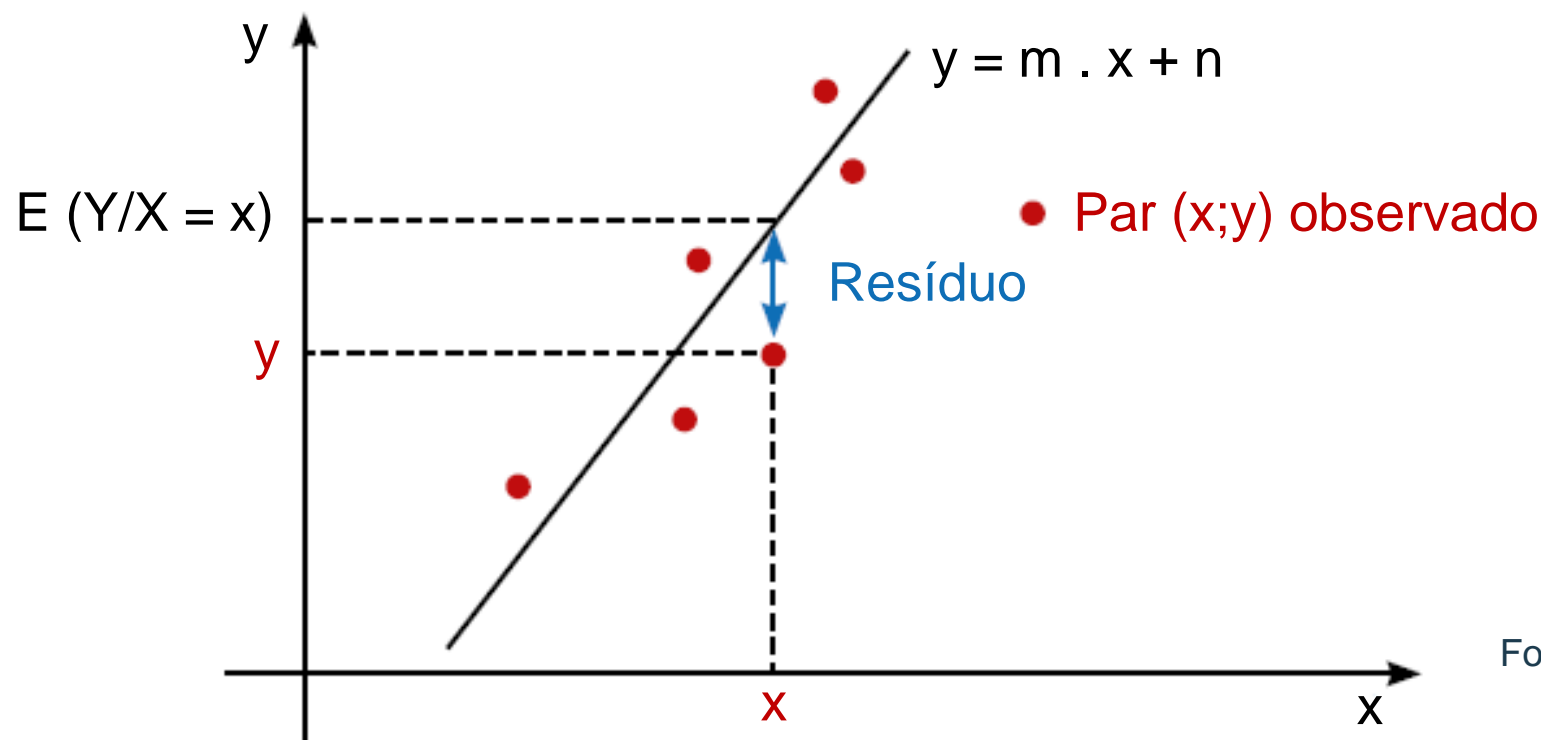
- se $R > 0$, quando uma variável aumenta, a outra variável aumenta;
- se $R < 0$, quando uma variável aumenta, a outra variável diminui;
- se $R = 0$, as variáveis não apresentam associação linear;
- se $R = 1$, as variáveis apresentam associação linear positiva tão forte que os pontos do gráfico de dispersão são pontos de uma reta crescente;
- se $R = -1$, as variáveis apresentam associação linear negativa tão forte que os pontos do gráfico de dispersão são pontos de uma reta decrescente.



Fonte: Adaptado
do livro-texto.

Regressão linear

- Muitas vezes, queremos saber qual é a reta que se ajusta da melhor maneira aos pontos de um gráfico de dispersão.
- Nesse caso, usamos o chamado método dos mínimos quadrados para estimar os coeficientes m e n de uma função do 1º grau $y = m \cdot x + n$ (cujo gráfico é uma reta) que minimizam a soma dos quadrados dos resíduos vindos da diferença entre os valores y efetivamente observados e os seus valores esperados $E(Y/X = x)$. Nesse sentido, observe a figura a seguir.



Regressão linear

- É possível demonstrar que os coeficientes m e n são calculados segundo as equações a seguir.

$$m = \frac{\sum_{i=1}^n x_i \cdot y_i - n \cdot \bar{x}_{obs} \cdot \bar{y}_{obs}}{\sum_{i=1}^n x_i^2 - n \cdot \bar{x}_{obs}^2} \quad n = \bar{y}_{obs} - m \cdot \bar{x}_{obs}$$

Fonte: Adaptado do livro-texto.

- Vamos aplicar essas fórmulas para acharmos a melhor reta, de equação $y = m \cdot x + n$, para representar a relação entre os pares $(x;y)$

X	-3	-1	0	2	5
Y	-10,5	-3	0,5	5	17

Fonte: Adaptado do livro-texto.

- $(x_1; y_1) = (-3; -10,5)$
- $(x_2; y_2) = (-1; -3)$
- $(x_3; y_3) = (0; 0,5)$
- $(x_4; y_4) = (2; 5)$
- $(x_5; y_5) = (5; 17)$

Cálculo do coeficiente

- Como o cálculo do coeficiente m é extenso, é interessante que o façamos por partes.

$$\bar{x}_{obs} = \frac{-3 + (-1) + 0 + 2 + 5}{5} = \frac{3}{5} = 0,6$$

$$\bar{y}_{obs} = \frac{-10,5 + (-3) + 0,5 + 5 + 17}{5} = \frac{9}{5} = 1,8$$

$$\sum_{i=1}^n x_i \cdot y_i = x_1 \cdot y_1 + x_2 \cdot y_2 + x_3 \cdot y_3 + x_4 \cdot y_4 + x_5 \cdot y_5$$

$$\sum_{i=1}^n x_i \cdot y_i = (-3) \cdot (-10,5) + (-1) \cdot (-3) + 0 \cdot 0,5 + 2 \cdot 5 + 5 \cdot 17$$

$$\sum_{i=1}^n x_i \cdot y_i = 31,5 + 3 + 0 + 10 + 85 = 129,5$$

$$n \cdot \bar{x}_{obs} \cdot \bar{y}_{obs} = 5 \cdot 0,6 \cdot 1,8 = 5,4$$

$$\sum_{i=1}^n x_i^2 = (-3)^2 + (-1)^2 + 0^2 + 2^2 + 5^2$$

$$\sum_{i=1}^n x_i^2 = 9 + 1 + 0 + 4 + 25 = 39$$

Logo:

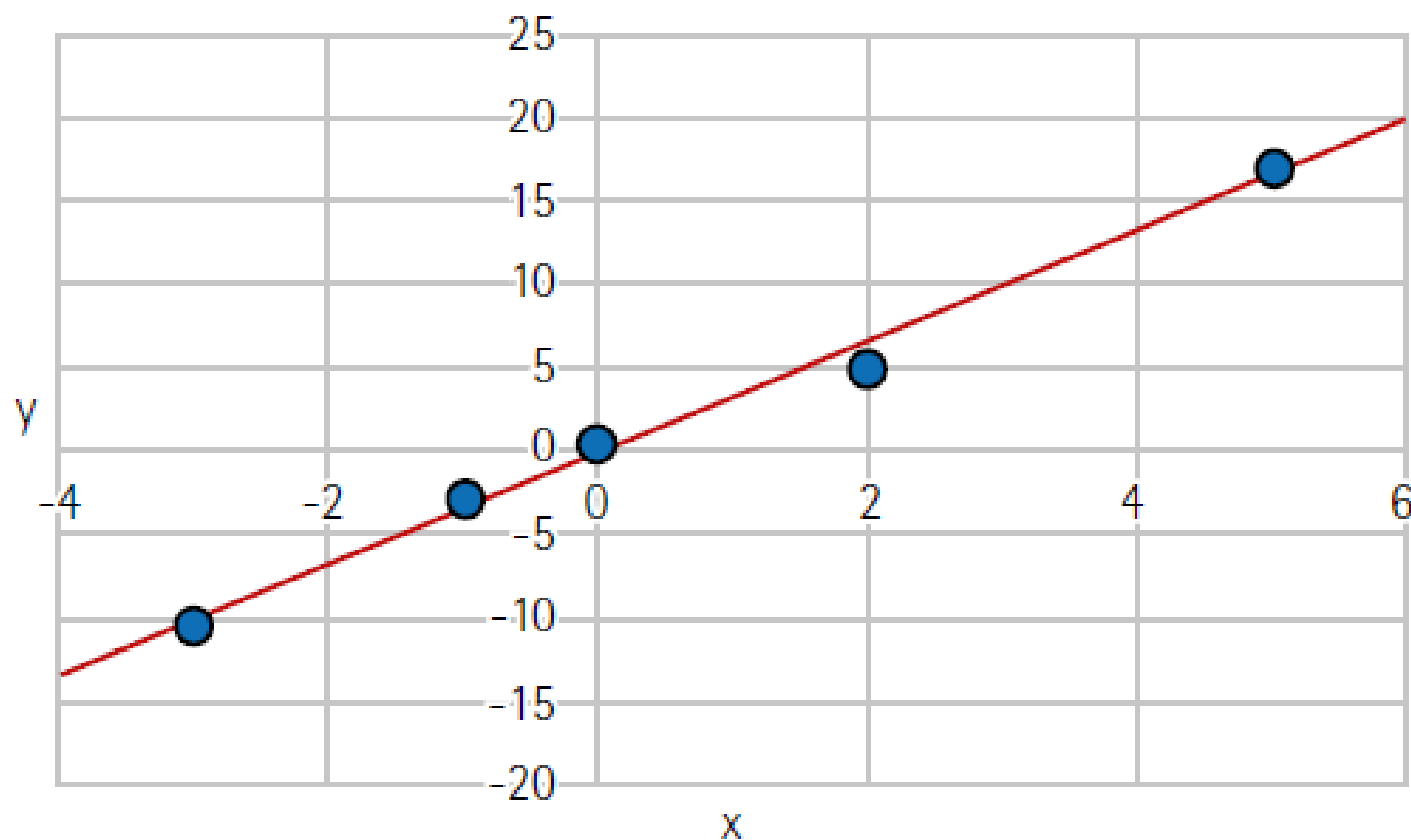
$$m = \frac{\sum_{i=1}^n x_i \cdot y_i - n \cdot \bar{x}_{obs} \cdot \bar{y}_{obs}}{\sum_{i=1}^n x_i^2 - n \cdot \bar{x}_{obs}^2} = \frac{129,5 - 5,4}{39 - 5 \cdot 0,6 \cdot 0,6} = \frac{124,1}{37,2} = 3,336$$

Cálculo do coeficiente

- No gráfico, temos os 5 pares (x;y) observados e a reta $y = 3,336x - 0,202$.

$$n = \bar{y}_{obs} - m \cdot \bar{x}_{obs} = 1,8 - 3,336 \cdot 0,6 = 1,8 - 2,002 = -0,202$$

$$y = m \cdot x + n \Rightarrow y = 3,336x - 0,202$$



Interatividade

Qual é o coeficiente que informa se duas variáveis estão relacionadas entre si de maneira aproximadamente linear?

- a) Coeficiente de independência.
- b) Coeficiente de pesquisa.
- c) Coeficiente de correlação.
- d) Coeficiente médio.
- e) Coeficiente de dispersão.

Resposta

Qual é o coeficiente que informa se duas variáveis estão relacionadas entre si de maneira aproximadamente linear?

- a) Coeficiente de independência.
- b) Coeficiente de pesquisa.
- c) Coeficiente de correlação.
- d) Coeficiente médio.
- e) Coeficiente de dispersão.

ATÉ A PRÓXIMA!