

Unidade II

5 ESTATÍSTICA INDUTIVA: PARTE 1

5.1 Noções gerais sobre estatística indutiva

Na estatística indutiva, usamos técnicas que nos possibilitam avaliar características de uma população por meio do estudo de uma amostra dela.

Parece natural pensarmos que uma amostra grande, ou seja, com muitos elementos, é uma amostra boa. Mas esse pensamento não é necessariamente verdadeiro.

Vejamos um exemplo. Imagine que queiramos saber a idade média dos 600 alunos das 16 turmas 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, teremos uma boa aproximação da idade média de todos os 600 estudantes da Escola ABC?

Obviamente, não, pois nossa amostra não deu chance para que as idades dos alunos do Ensino Médio aparecessem. Assim, podemos concluir que amostra boa é amostra que traz consigo todas as características presentes na população e na proporção em que ocorrem na população.

E se escolhêssemos, por sorteio, uma amostra de 160 alunos da Escola ABC, sendo 10 alunos de cada turma, e calculássemos a média das idades desses alunos?

Certamente, a média das idades dos alunos dessa amostra seria bem mais representativa da idade média da população (600 estudantes da Escola ABC) do que a média das idades da primeira amostra.

Agora, suponha que fizéssemos 5 sorteios com amostras de 160 alunos da Escola ABC, sendo 10 alunos de cada turma, e calculássemos a média das idades dos alunos em cada uma das 5 amostras. Teríamos o mesmo valor de média nessas 5 amostras?

Provavelmente, não, pois a natureza aleatória presente no processo amostral feito por sorteio não assegura que repetições de amostragens conduzam ao mesmo resultado.



Observação

Evidentemente, se usarmos toda a população na amostra, não teremos um processo aleatório, mas o oposto disso. Ou seja, nesse caso, teremos um processo certo ou determinado.

De modo geral, vamos chamar 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 $(X_1, X_2, \dots, X_i, \dots, X_n)$.

Para prosseguirmos nossos estudos em estatística indutiva, precisamos definir alguns termos, a saber:

- parâmetro;
- estimador;
- estimativa.

5.1.1 Parâmetro

Um parâmetro é a quantidade da característica da população que estamos estudando. Na maioria das vezes, não conhecemos tal valor. Como veremos, usamos uma estimativa para fazer inferências.

Por exemplo:

- μ 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 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.

5.1.2 Estimador

Um estimador (ou estimador pontual) representa o resultado da amostra que é usado para estimar determinado parâmetro populacional. Veja que o estimador é uma variável aleatória que depende dos componentes $X_1, X_2, \dots, X_i, \dots, X_n$ da amostra.

Por 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:

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

Note que o estimador \bar{X} foi apresentado como uma função de X_1, X_2, \dots, X_{12} , ou seja, o estimador não é um resultado numérico. Veja que o estimador \bar{X} é uma função das variáveis aleatórias X_i . Portanto, o estimador \bar{X} também é uma variável aleatória.

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).



Observação

Dizemos, também, que um estimador não viciado é um estimador não viesado (sem viés).

Na tabela a seguir, temos um resumo dos principais parâmetros (média populacional μ e variância populacional σ^2) e seus estimadores (respectivamente, média amostral \bar{X} e variância amostral S^2) não viciados e consistentes.

Tabela 56 – Parâmetros e estimadores

	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}



Lembrete

Dissemos que os estimadores (como o estimador \bar{X} para a média populacional μ e o estimador S^2 para a variância populacional σ^2 , mostrados na tabela 56) são variáveis aleatórias. Logo, os estimadores também seguem distribuições de probabilidade.

5.1.3 Estimativa

Uma estimativa é um valor "específico" de um estimador quando usamos valores "específicos" de determinada amostra.

Por 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 :

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

Assim, para os valores da amostra coletada, $\bar{X}_{\text{obs}} = 64,6 \text{ kg}$ é a estimativa usada para estimar o parâmetro peso médio μ das pessoas entre 15 e 65 anos que moram na cidade Novo Brasil.

Fica claro que, se fizermos nova amostragem, usando o mesmo estimador \bar{X} , provavelmente obteremos diferente valor de estimativa \bar{X}_{obs} .

Vamos usar a fórmula a seguir, retirada da tabela 56, para fazermos o cálculo da estimativa da variância S^2 .

$$S^2 = \frac{1}{n-1} \left(\sum_{i=1}^n X_i^2 - n\bar{X}^2 \right)$$

Se aplicarmos os valores da amostra coletada à expressão anterior, obteremos uma estimativa, indicada por S_{obs}^2 para a variância populacional σ^2 :

$$S_{\text{obs}}^2 = \frac{1}{12-1} \left(\sum_{i=1}^{12} X_i^2 - n\bar{X}_{\text{obs}}^2 \right)$$

$$S_{\text{obs}}^2 = \frac{1}{11} \left(\sum_{i=1}^{12} X_i^2 - 12 \cdot \bar{X}_{\text{obs}}^2 \right)$$

$$S_{\text{obs}}^2 = \frac{1}{11} \left((58^2 + 67^2 + 76^2 + 57^2 + 69^2 + 77^2 + 72^2 + 63^2 + 65^2 + 54^2 + 51^2 + 66^2) - 12(64,6)^2 \right)$$

$$S_{\text{obs}}^2 = \frac{1}{11} (50819 - 50077,92) = \frac{741,08}{11} = 67,37 \text{ kg}^2$$

Assim, para os valores da amostra coletada, $S_{\text{obs}}^2 = 67,37 \text{ kg}^2$ é a estimativa usada para estimar o parâmetro variância σ^2 do peso das pessoas entre 15 e 65 anos que moram na cidade Novo Brasil.

5.2 Teorema central do limite (TCL)

Imagine que retiremos uma amostra aleatória simples de tamanho n de uma população cujos parâmetros são média μ e variância σ^2 . Veja que, em princípio, não sabemos os valores de μ e de σ^2 .

Vamos representar essa amostra por n variáveis aleatórias independentes (X_1, X_2, \dots, X_n) , sendo a média amostral \bar{X} e a variância amostral S^2 calculadas, conforme mostrado na tabela 56, por:

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_i + \dots + X_n}{n}$$

$$S^2 = \frac{1}{n-1} \left(\sum_{i=1}^n X_i^2 - n\bar{X}^2 \right)$$

Observe que diferentes amostras geram diferentes médias amostrais. Ou seja, as amostras A, B, C e D, por exemplo, geram, respectivamente, médias amostrais \bar{X}_A , \bar{X}_B , \bar{X}_C e \bar{X}_D .

Segundo o teorema central do limite (TCL), para n suficientemente grande, a distribuição das médias amostrais devidamente padronizadas 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:

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

Note que, independentemente de como seja a distribuição da variável aleatória relativa a determinada população, a distribuição das médias amostrais de amostras aleatórias simples grandes segue modelo normal.



Observação

O que é uma amostra suficientemente grande depende muito da aplicação e do grau de precisão desejado. Alguns autores sugerem que a aproximação da média (ou seja, da proporção p) da distribuição binomial pela aproximação normal, por exemplo, é boa quando $np > 5$ e $n(1 - p) > 5$. Assim, se a proporção de acertos em uma binomial é 0,5 (ou seja, 50%), uma amostra de tamanho 11 já permite uma aproximação razoável pela distribuição normal. Se a proporção de acertos é 10%, uma amostra de tamanho 51 possibilita uma aproximação adequada pela distribuição normal. Há matemáticos que sugerem regras empíricas alternativas para estabelecer esse tamanho da amostra, como $np(1 - p) \geq 3$. Para outras distribuições, a convergência da média amostral para a distribuição normal pode exigir amostras muito maiores, de milhares ou mesmo dezenas de milhares de observações.

Vale destacar que, qualquer que seja o tamanho n da amostra, a distribuição amostral \bar{X} originária de uma população cuja variável X segue modelo normal também segue modelo normal.

O TCL é extremamente importante para a estatística indutiva. Ele 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.

Além disso, 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.



Saiba mais

Para saber mais sobre o teorema central do limite (TCL), assista ao vídeo indicado.

TEOREMA do limite central. 2018. 1 vídeo (14 min). Publicado pelo canal Professor Marcos Moreira. Disponível em: <https://bit.ly/3qZp0QV>. Acesso em: 17 nov. 2021.

6 ESTATÍSTICA INDUTIVA: PARTE 2

6.1 Intervalo de confiança (IC)

Os estimadores mostrados na tabela 56 são chamados de estimadores pontuais, pois estimam, por meio de "números fixos", os valores da média populacional μ e da variância populacional σ^2 . Por exemplo, para o caso estudado no item "Estimativa", vimos que:

- a estimativa $\bar{X}_{obs} = 64,6$ kg, vinda do estimador média amostral \bar{X} , é um número usado para estimar o parâmetro peso médio μ das pessoas entre 15 e 65 anos que moram na cidade Novo Brasil;
- a estimativa $S^2_{obs} = 67,37$ kg², vinda do estimador variância amostral S^2 , é um número usado para estimar o parâmetro variância σ^2 do peso das pessoas entre 15 e 65 anos que moram na cidade Novo Brasil.

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 $IC = [\sigma^2 - b; \sigma^2 + b]$.

Veja que associamos o intervalo de confiança ao coeficiente de confiança. Por exemplo, podemos dizer que, com 80% de confiança, determinada medida encontra-se entre 21,3 cm e 25,2 cm, ou seja,

IC = [21,3 cm;25,2 cm]. Se quisermos aumentar a confiança, nosso intervalo terá de ser aumentado, como veremos adiante.

Essas ideias ficarão mais claras se você ler o texto a seguir, que escrevi em 2018.

O que significam os números de uma pesquisa eleitoral?

Nos últimos meses, com as pesquisas feitas em função do período eleitoral em nosso país, temos visto e ouvido os termos "margem de erro", "grau de confiança", "tamanho da amostra" e muitos outros.

Qual é o sentido dessas expressões? E mais: o que significam os números relacionados a elas?

Para responder a essas questões, podemos analisar uma situação específica, como a exposta a seguir.

Segundo pesquisa divulgada pelo instituto Datafolha em 10 de outubro de 2018 sobre o segundo turno da eleição presidencial no Brasil, o candidato Jair Bolsonaro tinha 58% dos votos válidos e o candidato Fernando Haddad tinha 42% dos votos válidos.

O Datafolha também informou que:

- o levantamento de dados foi realizado em 10 de outubro de 2018;
- foram entrevistados 3.235 eleitores em 227 municípios;
- 6% dos entrevistados não sabiam em quem votar;
- 8% dos entrevistados votavam em branco ou anulavam o voto;
- a margem de erro foi de 2 pontos percentuais para cima ou para baixo;
- o nível de confiança da pesquisa foi de 95%.

Vamos analisar essa pesquisa.

Os resultados de votos válidos "valeram" para o dia em que a pesquisa foi feita e não são uma previsão do que vai realmente acontecer nas urnas. Vemos, inclusive, que os 6% de indecisos podem votar tanto em um candidato quanto no outro, ou podem anular seus votos.

A margem de erro de 2% indica que, no momento da realização da entrevista, Bolsonaro poderia ter entre 56% (58% menos 2%) e 60% (58% mais 2%) e Haddad poderia ter entre 40% (42% menos 2%) e 44% (42% mais 2%).

No entanto, como o nível de confiança da pesquisa foi de 95%, a chance, na ocasião, de um candidato ter entre 56% e 60% e do outro ter entre 40% e 44% foi de 95%. Ou seja, mesmo com a margem de erro, não há 100% de certeza da verdadeira intenção dos eleitores em 10 de outubro de 2018, mas há elevada probabilidade de os resultados da pesquisa coincidirem com essa intenção.

Há ainda que se considerar que o Brasil tem cerca de 5.570 municípios e que, segundo o Tribunal Superior Eleitoral (TSE), no primeiro turno das eleições, ocorrido em 7 de outubro de 2018, houve o comparecimento de 117.364.560 eleitores, com 107.050.673 de votos válidos.

O leitor pode pensar: uma pesquisa feita com 3.235 eleitores em 227 municípios pode ser válida para estimar o que pensam mais de 100 milhões de eleitores em mais de 5.500 municípios?

A resposta é sim. Vejamos um exemplo que trata de um caso bem mais simples do que o caso que estamos analisando, mas útil para entendermos o problema.

Imagine que você compre uma garrafa com 750 mL de um vinho de altíssimo padrão. Você precisa tomar todo esse volume para atestar que o vinho é de excelência? Não. A ingestão de um cálice com 30 mL de vinho, ou até menos, é suficiente, pois esse volume é uma amostra que representa todo o conteúdo da garrafa.

De modo geral, quase sempre "o todo" (população) que queremos estudar é inacessível, pois é muito grande, como no caso de mais de 100 milhões de eleitores em mais de 5.500 municípios, ou é desconhecido. Assim, a ideia é coletar uma amostra para fazer uma inferência sobre a população que queremos estudar.

Na pesquisa eleitoral que usamos como exemplo, a população é o eleitorado brasileiro com 16 anos ou mais. Uma amostra representativa dessa população deve ser formada por um conjunto de pessoas com as mesmas características de idade, gênero e distribuição regional da população, traduzindo fielmente o conjunto de todo o eleitorado. Ou seja, toda a diversidade da população deve "aparecer" na amostra na mesma proporção em que ocorre na população.

Concluimos que uma amostra de eleitores não deve ser necessariamente grande para representar o conjunto "completo" de eleitores: o importante é que o método de amostragem garanta a representatividade da amostra. Se esse método não for eficiente, uma amostra "muito grande", com elevada quantidade de entrevistados, pode não ser "boa".

No caso de pesquisas eleitorais como as do Datafolha, trabalha-se com amostra estratificada. Inicialmente, os 5.570 municípios brasileiros são classificados em três estratos: capital, região metropolitana e interior. Para cada estrato, são feitas, com base em critérios estatísticos robustos, que incluem a proporcionalidade, a seleção aleatória do município que fará parte da amostra, a seleção aleatória dos pontos de abordagem do município e a

seleção aleatória do entrevistado com base na distribuição de gênero e de faixa etária do eleitorado brasileiro.

Enfim, números relacionados à "margem de erro", ao "grau de confiança" e ao "tamanho da amostra" em pesquisas eleitorais não são simplesmente valores que fazem uma previsão de reais resultados: eles refletem a realidade da data da pesquisa e estão vinculados a probabilidades. São inferências.

Fonte: Doi (2018).

6.1.1 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 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:

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

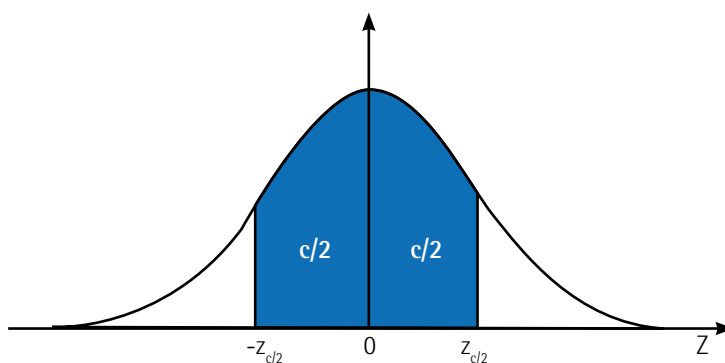


Figura 55 – A área indicada na gaussiana fornece $P(-z_{c/2} < Z < z_{c/2}) = \frac{c}{2} + \frac{c}{2} = c$

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:

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

Vejamos um exemplo. 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%?

Vamos fazer um resumo dos dados fornecidos no exemplo.

- **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 $\sigma = \sqrt{\sigma^2} = \sqrt{1,06} = 1,03 \text{ m}$.
- **Média amostral das alturas:** estimador \bar{X} .
- **Tamanho da amostra:** $n = 55$.
- **Média amostral das alturas observada na amostra:** estimativa $\bar{X}_{obs} = 1,71 \text{ m}$.
- **Coeficiente de confiança da estimativa intervalar:** $c = 0,80$.

Como c vale $0,80$, $c/2$ vale $0,40$, pois $c/2 = 0,80/2 = 0,40$. Precisamos achar $z_{c/2}$ tal que tenhamos as configurações ilustradas a seguir.

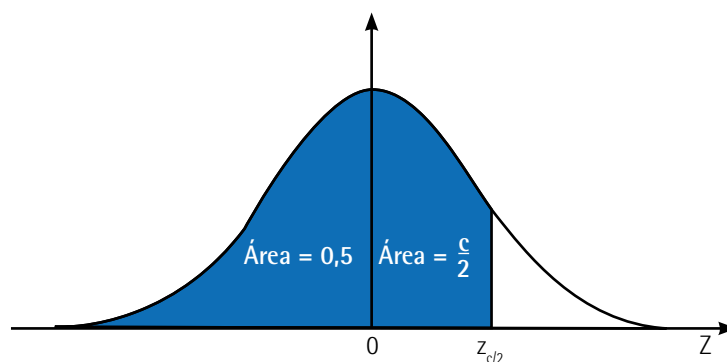


Figura 56

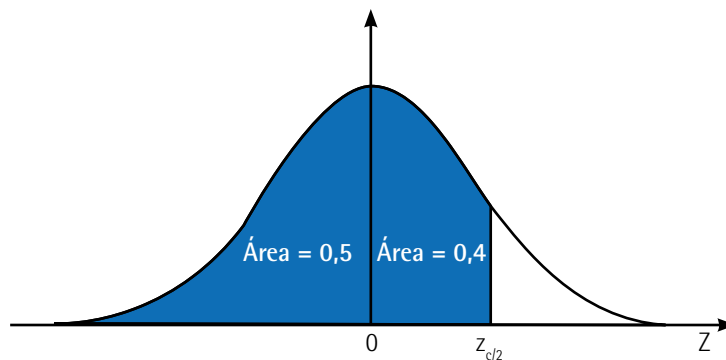


Figura 57

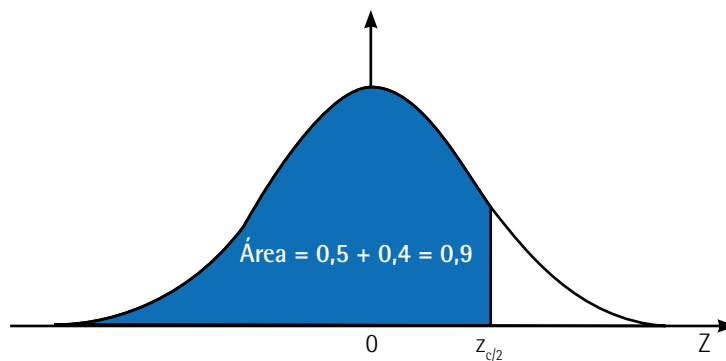


Figura 58

Precisamos encontrar, "dentro" da tabela normal reduzida, o valor 0,9. Vemos que, na tabela 39, o valor mais próximo de 0,9 é 0,8997, e ele corresponde a $z_{c/2} = 1,28$ (1,2 na horizontal e 0,08 na vertical).

Tabela 57

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

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.

E se, para a mesma situação, quisermos aumentar nossa confiança para 95%? Nesse caso, o que acontecerá com o intervalo de confiança para a média populacional? Certamente, esse intervalo será mais amplo. Façamos os cálculos.

Como c vale 0,95, $c/2$ vale 0,475, pois $c/2 = 0,90/2 = 0,475$. Precisamos achar $z_{c/2}$ tal que tenhamos as configurações ilustradas a seguir.

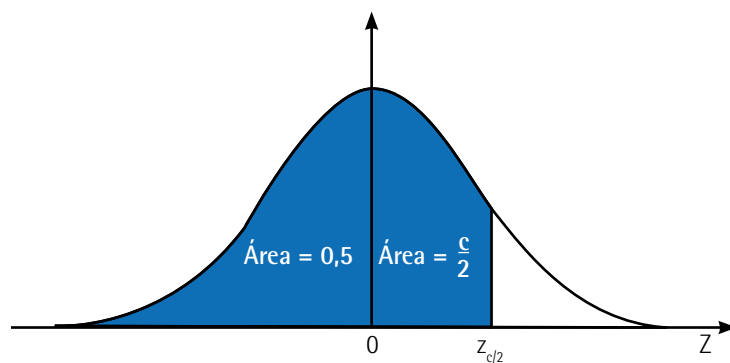


Figura 59

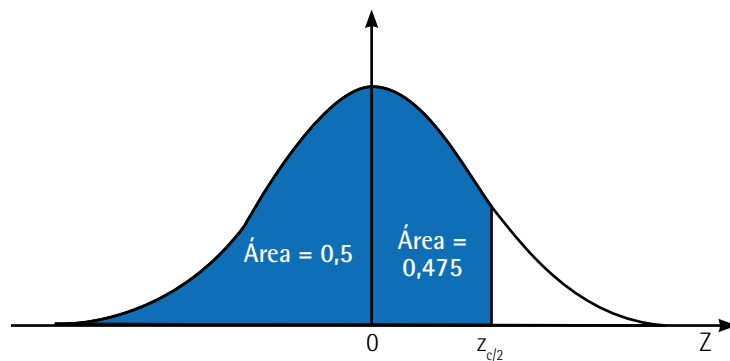


Figura 60

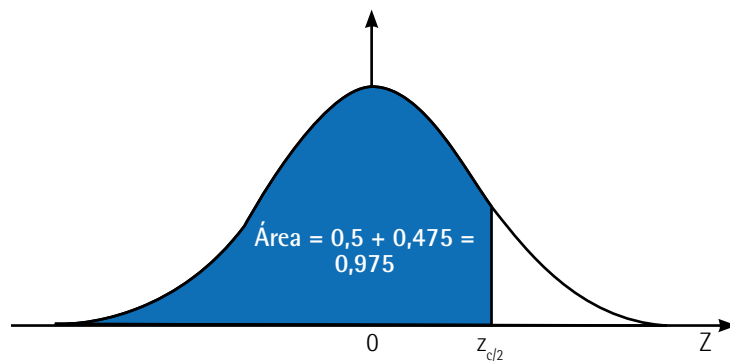


Figura 61

Precisamos encontrar, "dentro" da tabela normal reduzida, o valor 0,975. Vemos que, na tabela 39, o valor de 0,975 corresponde a $z_{c/2} = 1,96$ (1,9 na horizontal e 0,06 na vertical).

Tabela 58

Z	0,06
1,9	0,975

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.

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

$$IC(\mu; 0,95) = \left[1,71 - 1,96 \cdot \frac{1,03}{\sqrt{55}}; 1,71 + 1,96 \cdot \frac{1,03}{\sqrt{55}} \right]$$

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

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. Podemos dizer que o aumento da confiança foi "pago" com um intervalo amplo de alturas, que começa com pessoas de estaturas menores do que 1,45 m e chega a pessoas com quase 2 m.



Lembrete

Quanto mais aumentamos a confiança, mais aumentamos a amplitude do intervalo de confiança.

7 ESTATÍSTICA INDUTIVA: PARTE 3

7.1 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. Para entendermos o que chamamos de teste de hipóteses, vamos pensar na situação a seguir.

Imagine que o tempo que uma máquina leva para retificar certo modelo de peça seja uma variável aleatória contínua, que segue uma distribuição normal de probabilidades com média igual a 120 s e desvio padrão igual a 5 s, em que s representa segundos. Essa máquina será substituída por outra, mais nova, cujo tempo de retífica segue a mesma distribuição. Suspeita-se que, com isso, o tempo médio de retífica da peça diminua. Será que tal suspeita é verdadeira?

A resposta a essa questão 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 .

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 normal de média $\mu = 120$ s e desvio padrão $\sigma = 5$ s, ou seja, de variância $\sigma^2 = 5^2 = 25$ s², 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 120 s com a máquina nova.
- **Hipótese alternativa (H_a):** o tempo médio de retífica é menor do que 120 s com a máquina nova.

Podemos escrever essas hipóteses de modo mais sintético, como mostrado a seguir.

- **H_0 :** $\mu = 120$ s
- **H_a :** $\mu < 120$ 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. Assim, estamos sujeitos a erros quando rejeitamos H_0 e quando aceitamos H_0 . Nesse contexto, os erros recebem nomes especiais:

- erro tipo I;
- erro tipo II.

O erro tipo I ocorre quando rejeitamos H_0 , sendo H_0 , na realidade, verdadeira (V). Esse é o tipo de erro que, nas ciências ligadas à saúde, gera o diagnóstico chamado de **falso positivo**. Veja que, na situação em estudo, cometemos erro tipo I quando rejeitamos que o tempo médio de retífica da nova máquina é igual a 120 s, sendo que, na realidade, esse tempo é de 120 s.

O erro tipo II ocorre quando não rejeitamos H_0 , sendo H_0 , na realidade, falsa (F). Esse tipo de erro gera o diagnóstico **falso negativo**. Veja que, na situação em estudo, cometemos erro tipo II quando

aceitamos que o tempo médio de retífica da nova máquina é igual a 120 s, sendo que, na realidade, esse tempo é menor do que 120 s.

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.

No quadro a seguir, temos um resumo do que acabamos de ponderar.

Quadro 1 – Resumo dos casos de um teste de hipóteses

		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



Observação

Em estatística, não falamos em aceitar a hipótese nula. Pense, por exemplo, em um processo criminal no qual a hipótese nula é que o réu é inocente, e a hipótese alternativa é que o réu é culpado. Se o réu for absolvido por falta de provas, não rejeitamos a hipótese nula, de que ele é inocente. Isso, no entanto, não quer dizer que o réu é inocente, apenas que não há evidências suficientes que corroborem a hipótese alternativa de que ele seja culpado. Note, portanto, que não rejeitar a hipótese nula não é a mesma coisa que aceitá-la.

Vamos chamar de α a probabilidade de ocorrência de erro tipo I e de β a probabilidade de ocorrência de erro tipo II. Logo, temos o que segue.

- $\alpha = P(\text{erro tipo I}) = P(\text{rejeitar } H_0 / H_0 \text{ é verdadeira})$
- $\beta = P(\text{erro tipo II}) = P(\text{não rejeitar } H_0 / H_0 \text{ é falsa})$

A probabilidade α é chamada de nível de significância do teste e está relacionada ao controle do erro tipo I. A probabilidade β é chamada de poder do teste e está relacionada ao controle do erro tipo II.



Lembrete

A probabilidade α é lida como "a probabilidade de rejeitarmos a hipótese H_0 dado que essa hipótese é verdadeira". A probabilidade β é lida como "a probabilidade de não rejeitarmos a hipótese H_0 dado que essa hipótese é falsa".

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

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 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:

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

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:

$$0,05 = P(Z < z_c)$$

Visualmente, queremos achar z_c de modo que tenhamos o que se ilustra na gaussiana a seguir.

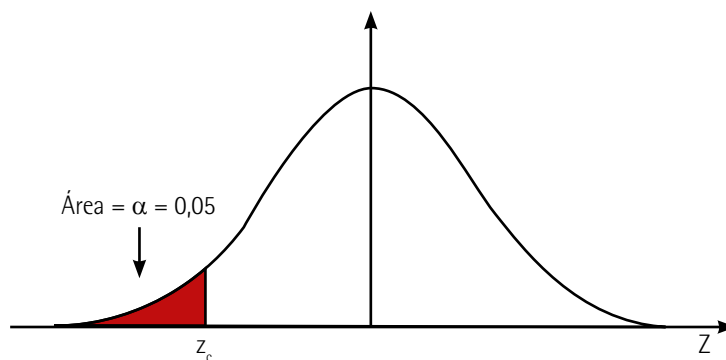


Figura 62

Para usarmos a tabela 39, é interessante que observemos a ilustração a seguir, em que $z^* = -z_c$.

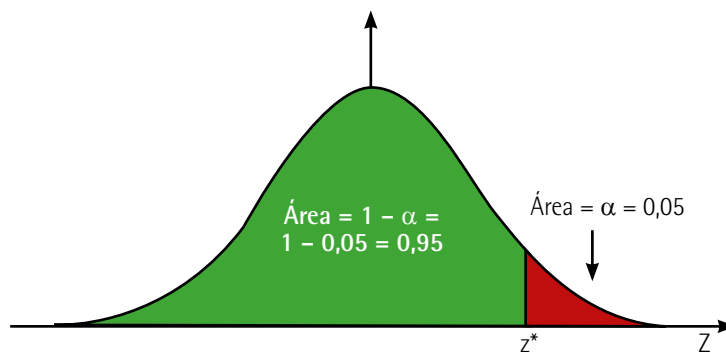


Figura 63

Precisamos encontrar, "dentro" da tabela normal reduzida, o valor 0,95. Vemos que, na tabela 39, o valor mais próximo de 0,95 é 0,9505, que corresponde a $z^* = 1,65$ (1,6 na horizontal e 0,05 na vertical).

Tabela 59

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

Como $z^* = 1,65$ e $z^* = -z_c$, então $z_c = -1,65$.

Fizemos $z_c = \frac{x_c - 120}{5/\sqrt{30}}$. Logo:

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

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 $x < 118,5$ s.

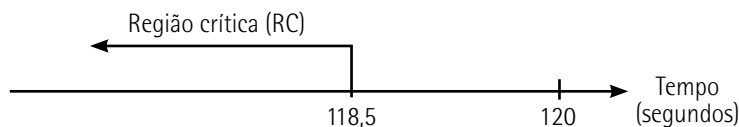


Figura 64

No exemplo, de acordo com a regra de decisão adotada, 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 $x < 118,5$ s (com $x > 0$). Ou seja, se a média amostral dos tempos resultar em valor menor do que 118,5 s, 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 120 s).

Veja que uma amostra de 30 tempos de retificação da máquina nova com média amostral de 119 s, por exemplo, não permite que concluamos, ao nível de significância de 5%, que o tempo médio de retificação da nova máquina é menor do que 120 s.

O que acontece com a RC do exemplo em estudo se mantivermos o nível de significância de 5%, mas usarmos uma amostra com 15 elementos ($n = 15$)?

Note que estamos diminuindo o tamanho da amostra (n passa a ser 15) e permanecendo com $\alpha = 0,05$. Logo:

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

$$0,05 = P(\bar{X} < 120 / \mu = 120) = P\left(\frac{\bar{X} - \mu}{\sigma / \sqrt{n}} < \frac{x_c - 120}{5 / \sqrt{15}}\right)$$

Chamamos $\frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$ de Z e $\frac{x_c - 120}{5 / \sqrt{15}}$ de z_c e chegamos a:

$$0,05 = P(Z < z_c)$$

Já vimos que a equação anterior é verdadeira para $z_c = -1,65$.

Fizemos $z_c = \frac{x_c - 120}{5 / \sqrt{15}}$. Logo:

$$-1,65 = \frac{x_c - 120}{5 / \sqrt{15}} \Rightarrow -1,65 \cdot \frac{5}{\sqrt{15}} = x_c - 120 \Rightarrow -2,1 = x_c - 120 \Rightarrow x_c = 117,9$$

No caso em análise, com uma amostra de tamanho $n = 15$, a RC ao nível de significância de 5% ($\alpha = 0,05$), ilustrada a seguir, é dada por $x < 117,9$ segundos.



Figura 65

Para $\alpha = 0,05$ e para tamanho de amostra $n = 15$, o conjunto de valores de tempo, representado pela variável X , que levam à não aceitação da hipótese nula H_0 é dado por $x < 117,9$ s (com $x > 0$). Ou seja, se a média amostral dos tempos resultar em valor menor do que 117,9 s, ao nível de significância de 5%, não aceitaremos que o tempo médio de retificação da nova máquina seja igual a 120 s (nesse caso, acataremos a hipótese H_a , segundo a qual o tempo médio de retificação da nova máquina é menor do que 120 s).

Veja que, para $\alpha = 0,05$, no caso da amostra de tamanho 30, a média amostral de 118,2 s faz com que não aceitemos H_0 (rejeitamos tempo médio de 120 s). Já para o caso da amostra de tamanho 15, para $\alpha = 0,05$, a média amostral de 118,2 s faz com que não rejeitemos H_0 (acatamos tempo médio de 120 s).

O que acontece com a RC do exemplo em estudo se usarmos uma amostra com 15 elementos ($n = 15$), mas, agora, ao nível de significância de 1% ($\alpha = 0,01$)?

Vamos estudar esse caso.

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

$$0,01 = P(\bar{X} < 120 / \mu = 120)$$

$$0,01 = P(\bar{X} < 120 / \mu = 120) = P\left(\frac{\bar{X} - \mu}{\sigma / \sqrt{n}} < \frac{x_c - 120}{5 / \sqrt{15}}\right)$$

Chamamos $\frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$ de Z e $\frac{x_c - 120}{5 / \sqrt{15}}$ de z_c e chegamos a:

$$0,01 = P(Z < z_c)$$

Visualmente, queremos achar z_c de modo que tenhamos o que se ilustra na gaussiana a seguir.

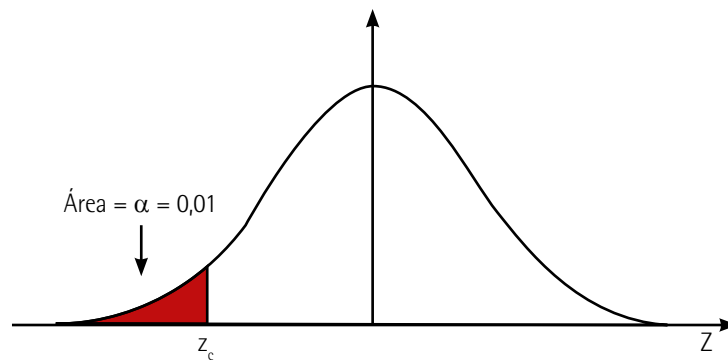


Figura 66

Para usarmos a tabela 39, é interessante que observemos a ilustração a seguir, em que $z^* = -z_c$.

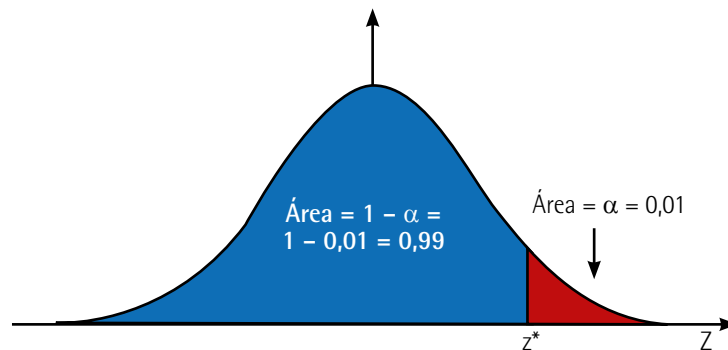


Figura 67

Precisamos encontrar, "dentro" da tabela normal reduzida, o valor 0,99. Vemos que, na tabela 39, o valor mais próximo de 0,99 é 0,9901, que corresponde a $z^* = 2,33$ (2,3 na horizontal e 0,03 na vertical).

Tabela 60

Z	0,03
2,3	0,9901 \approx 0,99

Como $z^* = 2,33$ e $z^* = -z_c$, então $z_c = -2,33$.

Fizemos $z_c = \frac{x_c - 120}{5/\sqrt{15}}$. Logo:

$$-2,33 = \frac{x_c - 120}{5/\sqrt{15}} \Rightarrow -2,33 \cdot \frac{5}{\sqrt{15}} = x_c - 120 \Rightarrow -3,00 = x_c - 120 \Rightarrow x_c = 117,0$$

No caso em análise, com uma amostra de tamanho $n = 15$, a RC ao nível de significância de 1% ($\alpha = 0,01$), ilustrada a seguir, é dada por $x < 117,0$ segundos.

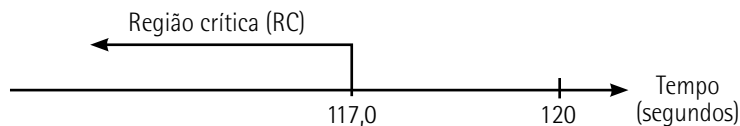


Figura 68

Para $\alpha = 0,01$ e para tamanho de amostra $n = 15$, o conjunto de valores de tempo, representado pela variável X , que levam à rejeição da hipótese nula H_0 é dado por $x < 117,0$ s (com $x > 0$). Ou seja, se a média amostral dos tempos resultar em valor menor do que 117,0 s, ao nível de significância de 1%, não aceitaremos que o tempo médio de retificação da nova máquina seja igual a 120 s (nesse caso, acataremos a hipótese H_a , segundo a qual o tempo médio de retificação da nova máquina é menor do que 120 s).

Veja que, para $\alpha = 0,01$ e amostra de tamanho 15, a média amostral de 117,5 s faz com que não aceitemos H_0 (rejeitamos tempo médio de 120 s). Já para o caso da amostra de tamanho 15, mas com $\alpha = 0,05$, a média amostral de 117,5 s faz com que não rejeitemos H_0 (acatamos tempo médio de 120 s).



Observação

Um teste de hipóteses sempre se refere a um parâmetro populacional. Não faz sentido fazer um teste de hipóteses sobre estimativas amostrais, pois o valor de uma estimativa amostral é conhecido quando temos os valores da amostra.

Não faz sentido, portanto, fazer um teste de hipóteses como o exemplificado a seguir, relativo à média amostral.

- $H_0: \bar{X} = 15$
- $H_a: \bar{X} > 15$

Isso porque, se alguém quiser saber a média amostral \bar{X} , basta que a calcule. O teste de hipóteses correto seria sobre a média populacional.

- $H_0: \mu = 15$
- $H_a: \mu > 15$

De modo geral, as hipóteses alternativas podem ser:

- hipóteses alternativas bilaterais, em que detectamos desvios em qualquer sentido;
- hipóteses alternativas unilaterais, em que detectamos desvios em um único sentido.

O exemplo que fizemos, relativo ao tempo de retificação de uma máquina, tratou de uma H_a unilateral, visto que somente detectamos desvios à esquerda de dada referência de parâmetro μ (no caso, 120 s), isto é, apenas estamos interessados em saber se a média populacional é menor do que 120 s.

A seguir, temos um resumo de como fazer um teste de hipóteses.

7.1.1 Etapas para a realização de um teste de hipóteses para média com variância populacional conhecida

Na sequência, apresentamos um resumo das etapas envolvidas na realização de 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).

Tabela 61

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

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).

Tabela 62

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 $\bar{X} \geq x_{c2}$

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, $\bar{X} \sim N(\mu; \sigma^2/n)$.

Etapas 4. Fixar o nível de significância α e determinar a RC.

Etapas 5. Concluir o teste verificando se o valor de média observado na amostra, indicado por \bar{x}_{obs} , pertence ou não pertence à RC.

Vamos aplicar essas etapas a uma situação específica.

Suponha que a concentração X de glicose no sangue seja uma variável aleatória que siga uma distribuição normal de probabilidades de desvio padrão igual a 7 mg/dL, em que miligramas por decilitro (mg/dL) é a unidade de medida. Para essa situação, teste a hipótese nula de que a média populacional μ seja igual a 93 mg/dL contra a hipótese alternativa de que essa média não seja igual a 93 mg/dL com base em uma amostra de 45 pessoas, na qual se observou média de 88 mg/dL. Use nível de significância de 5%.

A seguir, temos uma síntese dos dados fornecidos.

- **X:** concentração de glicose no sangue em mg/dL
- **X:** variável aleatória que segue distribuição normal de probabilidades de desvio padrão $\sigma = 7$ mg/dL
- **Variância de X:** $\sigma^2 = (7 \text{ mg/dL})^2$ ou $\sigma^2 = 49 \text{ mg}^2/\text{dL}^2$
- **Ho:** $\mu = 93 \text{ mg/dL}$
- **Ha:** $\mu \neq 93 \text{ mg/dL}$
- **Tamanho da amostra:** $n = 45$
- **Média amostral observada:** $\bar{x}_{obs} = 88 \text{ mg/dL}$
- $\alpha = 0,05$

Temos o que segue.

- $X \sim N(\mu; \sigma^2) \rightarrow X \sim N(\mu; 49)$
- $\bar{X} \sim N\left(\mu; \frac{\sigma^2}{n}\right) \rightarrow \bar{X} \sim N\left(\mu; \frac{49}{45}\right) \rightarrow \bar{X} \sim N(\mu; 1,09)$

Vamos aplicar as etapas descritas anteriormente para a situação em análise.

Etapla 1. Estabelecer as hipóteses H_0 e H_a para o exemplo, em que temos um teste bilateral.

- **H_0 :** $\mu = 93 \text{ mg/dL}$
- **H_a :** $\mu \neq 93 \text{ mg/dL}$

Etapla 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).

- **H_a :** $\mu \neq 93 \text{ mg/dL}$

Rejeitar H_0 se $\bar{X} \leq x_{c1}$ ou se $\bar{X} \geq x_{c2}$

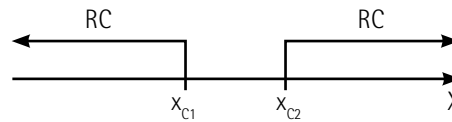


Figura 69

Etapla 3. Identificar o estimador e o tipo de distribuição de probabilidades que essa variável aleatória segue.

Vimos que a variável aleatória populacional X segue uma normal de média μ e variância $\sigma^2 = 49$, ou seja, $X \sim N(\mu; 49)$. A variável aleatória amostral \bar{X} , relativa a uma amostra de tamanho $n = 45$, segue uma normal de média μ e variância $\sigma^2/n = 49/45 = 1,09$, ou seja, $\bar{X} \sim N(\mu; 1,09)$.

Etapla 4. Fixar o nível de significância α e determinar a RC.

Foi dito que $\alpha = 0,05$. Logo:

$$\alpha = P(\text{erro tipo I}) = P(\text{rejeitar } H_0 / H_0 \text{ é verdadeira}) = P(\bar{X} \neq 93 / \mu = 93)$$

$$0,05 = P(\bar{X} \neq 93 / \mu = 93) = P\left(\frac{\bar{X} - \mu}{\sigma / \sqrt{n}} < \frac{x_{c1} - 93}{7 / \sqrt{45}} \text{ ou } \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} > \frac{x_{c2} - 93}{7 / \sqrt{45}}\right)$$

Chamamos $\frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$ de Z , $\frac{x_{c1} - 93}{7 / \sqrt{45}}$ de z_{c1} e $\frac{x_{c2} - 93}{7 / \sqrt{45}}$ de z_{c2} e chegamos a:

$$0,05 = P(Z < z_{c1} \text{ ou } Z > z_{c2}) = P(Z < z_{c1}) + P(Z > z_{c2})$$

Visualmente, queremos achar z_{c1} e z_{c2} de modo que tenhamos o que se ilustra na gaussiana a seguir.

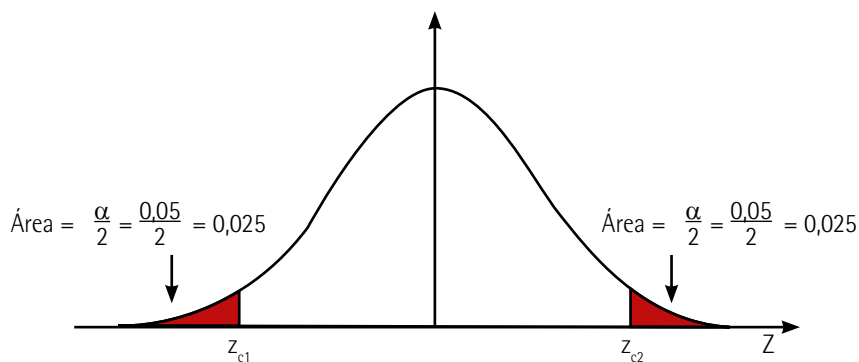


Figura 70

Para usarmos a tabela 39, é interessante observar a ilustração a seguir.

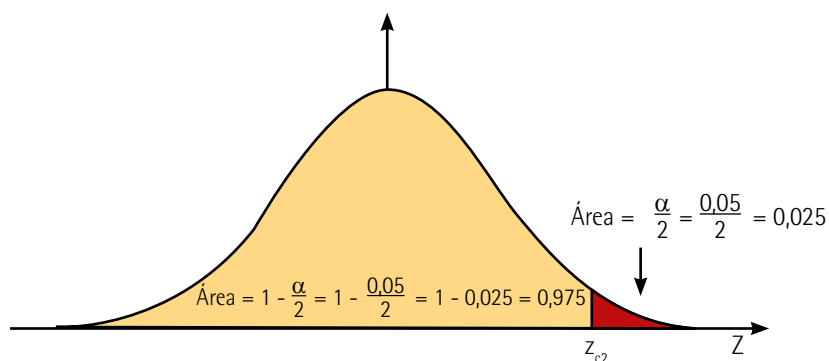


Figura 71

Precisamos encontrar, "dentro" da tabela normal reduzida, o valor 0,975. Vemos que, na tabela 39, o valor 0,975, que corresponde a $z_{c2} = 1,96$ (1,9 na horizontal e 0,06 na vertical).

Tabela 63

Z	0,06
1,9	0,975

Como $z_{c2} = 1,96$, então $z_{c1} = -1,96$.

Fizemos $z_{c1} = \frac{x_{c1} - 93}{7/\sqrt{45}}$ e $z_{c2} = \frac{x_{c2} - 93}{7/\sqrt{45}}$. Logo:

$$-1,96 = \frac{x_{c1} - 93}{7/\sqrt{45}} \Rightarrow -1,96 \cdot \frac{7}{\sqrt{45}} = x_{c1} - 93 \Rightarrow -2,05 = x_{c1} - 93 \Rightarrow x_{c1} = 90,95$$

$$1,96 = \frac{x_{c2} - 93}{7 / \sqrt{45}} \Rightarrow 1,96 \cdot \frac{7}{\sqrt{45}} = x_{c1} - 93 \Rightarrow 2,05 = x_{c1} - 93 \Rightarrow x_{c1} = 95,05$$

No caso em análise, com uma amostra de tamanho $n = 45$, a RC ao nível de significância de 5% ($\alpha = 0,05$), ilustrada a seguir, é dada por $x < 90,95$ e $x > 95,05$.

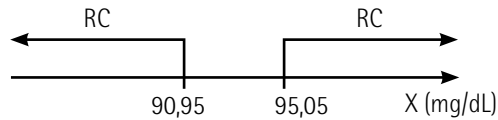


Figura 72

No exemplo, de acordo com a regra de decisão adotada, com uma amostra de tamanho $n = 45$, o conjunto de valores de tempo, representado pela variável X , que levam à não aceitação da hipótese nula H_0 é dado por $x < 90,95$ mg/dL e por $x > 95,05$ mg/dL.

Etapla 5. Concluir o teste verificando se o valor da média observado na amostra, $\bar{x}_{\text{obs}} = 88$ mg/dL, pertence ou não pertence à RC.

Como indicado na ilustração a seguir, o valor $\bar{x}_{\text{obs}} = 88$ mg/dL pertence à região crítica.

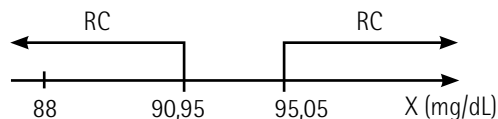


Figura 73

Concluimos que, para $\alpha = 0,05$ e para tamanho de amostra $n = 45$, o conjunto de valores observados que levam à não aceitação da hipótese nula H_0 é dado por $x < 90,95$ mg/dL e $x > 95,05$ mg/dL (com $x > 0$). Ou seja, como a média amostral é 88 mg/dL, ao nível de significância de 5%, não aceitamos que a média populacional seja igual a 93 mg/dL (nesse caso, acatamos a hipótese H_a , segundo a qual a média populacional é diferente de 93 mg/dL).

7.2 Testes qui-quadrado

Dois dos principais testes chamados de testes qui-quadrado são:

- teste de aderência;
- teste de independência.

7.2.1 Teste de aderência

Podemos dizer que os testes 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. Nesses testes, O indicará o valor observado, e E indicará o respectivo valor esperado.

Inicialmente, organizamos os dados observados em k categorias, com $k \geq 2$ (duas categorias no mínimo) e registramos as frequências observadas O em cada categoria, como apresentado a seguir.

Tabela 64

Categoria	Frequência observada
1	O_1
2	O_2
3	O_3
...	...
i	O_i
...	...
k	O_k

Então, estabelecemos uma hipótese H_0 para os dados observados. Calculamos as frequências esperadas E no caso de H_0 ser verdadeira e fazemos a tabela a seguir.

Tabela 65

Categoria	Frequência observada	Frequência esperada
1	O_1	E_1
2	O_2	E_2
3	O_3	E_3
...
i	O_i	E_i
...
k	O_k	E_k

Fazemos a quantificação das diferenças entre as frequências observadas e suas respectivas frequências esperadas por meio da estatística a seguir, indicada por Q^2 .

$$Q^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}$$



Observação

Veja que o cálculo de Q^2 é feito com diferenças $(O_i - E_i)$ elevadas ao quadrado a fim de que não haja "cancelamentos" entre termos positivos e termos negativos. Além disso, cada $(O_i - E_i)^2$ é dividido por E_i , a fim de que, como veremos, possamos usar a tabela de distribuição de probabilidades chamada de qui-quadrado.

Supondo que H_0 seja verdadeira, fazemos:

$$Q^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i} \approx \chi_q^2$$

Na expressão, q é o número de graus de liberdade, calculado como o número de categorias menos 1.

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_q^2). 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_{obs}^2 .

Identificamos $P = P(\chi_q^2 \geq Q_{obs}^2)$, em que, como acabamos de dizer, Q_{obs}^2 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 .

Vejamos um exemplo. Imagine que o gestor de saúde de um município queira saber se a quantidade de internações nos hospitais por ele geridos varia de acordo com o dia da semana. Para isso, em determinada semana, fez as observações apresentadas a seguir.

Tabela 66

Dia da semana	Quantidade de internações
Segunda-feira	18
Terça-feira	12
Quarta-feira	14
Quinta-feira	13
Sexta-feira	31
Sábado	23
Domingo	36

Temos, portanto, uma amostra com 7 observações da variável quantidade de interações por dia. O que se pode concluir dessas observações ao nível de significância de 5%?

Vamos fazer um teste qui-quadrado de aderência e verificar seu resultado.

Primeiramente, estabelecemos as hipóteses a serem testadas.

- **Hipótese nula (H_0):** a quantidade de interações não varia com o dia da semana.
- **Hipótese alternativa (H_a):** há pelo menos 1 dia da semana em que a quantidade de interações é diferente das quantidades dos outros dias.

Na tabela, temos o total n de 147 interações, pois:

$$n = 18 + 12 + 14 + 13 + 31 + 23 + 36 = 147$$

Se a quantidade de interações não variar com o dia da semana, deveremos ter, em cada um dos 7 dias da semana, $1/7$ das 147 interações, o que resultará em 21 interações diárias. Ou seja, se H_0 for verdadeira, o valor esperado para cada frequência será igual a 21. Assim, obtemos a tabela a seguir.

Tabela 67

Dia da semana	Categoria	Quantidade de interações observadas	Quantidade de interações esperadas
Segunda-feira	1	$O_1 = 18$	$E_1 = 21$
Terça-feira	2	$O_2 = 12$	$E_2 = 21$
Quarta-feira	3	$O_3 = 14$	$E_3 = 21$
Quinta-feira	4	$O_4 = 13$	$E_4 = 21$
Sexta-feira	5	$O_5 = 31$	$E_5 = 21$
Sábado	6	$O_6 = 23$	$E_6 = 21$
Domingo	7	$O_7 = 36$	$E_7 = 21$

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

$$Q^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}$$

$$Q^2_{\text{obs}} = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} + \frac{(O_3 - E_3)^2}{E_3} + \dots + \frac{(O_7 - E_7)^2}{E_7}$$

$$Q^2_{\text{obs}} = \frac{(18-21)^2}{21} + \frac{(12-21)^2}{21} + \frac{(14-21)^2}{21} + \frac{(13-21)^2}{21} + \frac{(31-21)^2}{21} + \frac{(23-21)^2}{21} + \frac{(36-21)^2}{21}$$

$$Q^2_{\text{obs}} = \frac{(-3)^2}{21} + \frac{(-9)^2}{21} + \frac{(-7)^2}{21} + \frac{(-8)^2}{21} + \frac{(10)^2}{21} + \frac{(2)^2}{21} + \frac{(15)^2}{21}$$

$$Q^2_{\text{obs}} = \frac{9}{21} + \frac{81}{21} + \frac{49}{21} + \frac{64}{21} + \frac{100}{21} + \frac{4}{21} + \frac{225}{21} = \frac{532}{21}$$

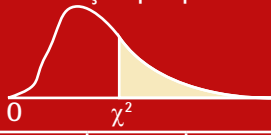
$$Q^2_{\text{obs}} = 25,3$$

Como temos 7 categorias, o número de graus de liberdade é 6 ($q = 6$), visto que ele é dado pelo número de categorias menos 1.

Assim, temos $Q^2_{\text{obs}} = 25,3$. Procuramos, em uma tabela de distribuição qui-quadrado, como a tabela 68, na linha de graus de liberdade (g.l.) igual a 6 ($q = 6$), 25,3.

Tabela 68 – Distribuição qui-quadrado

Distribuição qui-quadrado						A tabela fornece os valores c, tais que $P(\chi^2 > c) = p$, onde n é o número de graus de liberdade e p é a probabilidade de sucesso					
g.l.	0,995	0,990	0,975	0,950	0,900	0,500	0,100	0,050	0,025	0,010	0,005
1	0,000	0,000	0,001	0,004	0,016	0,455	2,706	3,841	5,024	6,635	7,879
2	0,010	0,020	0,051	0,103	0,211	1,386	4,605	5,991	7,378	9,210	10,597
3	0,072	0,115	0,216	0,352	0,584	2,366	6,251	7,815	9,348	11,345	12,838
4	0,207	0,297	0,484	0,711	1,064	3,357	7,779	9,488	11,143	13,277	14,860
5	0,412	0,554	0,831	1,145	1,610	4,351	9,236	11,070	12,833	15,086	16,750
6	0,676	0,872	1,237	1,635	2,204	5,348	10,645	12,592	14,449	16,812	18,548
7	0,989	1,239	1,690	2,167	2,833	6,346	12,017	14,067	16,013	18,475	20,278
8	1,344	1,646	2,180	2,733	3,490	7,344	13,362	15,507	17,535	20,090	21,955
9	1,735	2,088	2,700	3,325	4,168	8,343	14,684	16,919	19,023	21,666	23,589
10	2,156	2,558	3,247	3,940	4,865	9,342	15,987	18,307	20,483	23,209	25,188
11	2,603	3,053	3,816	4,575	5,578	10,341	17,275	19,675	21,920	24,725	26,757
12	3,074	3,571	4,404	5,226	6,304	11,340	18,549	21,026	23,337	26,217	28,300
13	3,565	4,107	5,009	5,892	7,042	12,340	19,812	22,362	24,736	27,688	29,819
14	4,075	4,660	5,629	6,571	7,790	13,339	21,064	23,685	26,119	29,141	31,319
15	4,601	5,229	6,262	7,261	8,547	14,339	22,307	24,996	27,488	30,578	32,801
16	5,142	5,812	6,908	7,962	9,312	15,338	23,542	26,296	28,845	32,000	34,267
17	5,697	6,408	7,564	8,672	10,085	16,338	24,769	27,587	30,191	33,409	35,718

Distribuição qui-quadrado							A tabela fornece os valores c, tais que $P(\chi^2 > c) = p$, onde n é o número de graus de liberdade e p é a probabilidade de sucesso				
											
g.l.	0,995	0,990	0,975	0,950	0,900	0,500	0,100	0,050	0,025	0,010	0,005
18	6,265	7,015	8,231	9,390	10,865	17,338	25,989	28,869	31,526	34,805	37,156
19	6,844	7,633	8,907	10,117	11,651	18,338	27,204	30,144	32,852	36,191	38,582
20	7,434	8,260	9,591	10,851	12,443	19,337	28,412	31,410	34,170	37,566	39,997
21	8,034	8,897	10,283	11,591	13,240	20,337	29,615	32,671	35,479	38,932	41,401
22	8,643	9,542	10,982	12,338	14,041	21,337	30,813	33,924	36,781	40,289	42,796
23	9,260	10,196	11,689	13,091	14,848	22,337	32,007	35,172	38,076	41,638	44,181
24	9,886	10,856	12,401	13,848	15,659	23,337	33,196	36,415	39,364	42,980	45,559
25	10,520	11,524	13,120	14,611	16,473	24,337	34,382	37,652	40,646	44,314	46,928
26	11,160	12,198	13,844	15,379	17,292	25,336	35,563	38,885	41,923	45,642	48,290
27	11,808	12,879	14,573	16,151	18,114	26,336	36,741	40,113	43,195	46,963	49,645
28	12,461	13,565	15,308	16,928	18,939	27,336	37,916	41,337	44,461	48,278	50,993
29	13,121	14,256	16,047	17,708	19,768	28,336	39,087	42,557	45,722	49,588	52,336
30	13,787	14,953	16,791	18,493	20,599	29,336	40,256	43,773	46,979	50,892	53,672
35	17,192	18,509	20,569	22,465	24,797	34,336	46,059	49,802	53,203	57,342	60,275
40	20,707	22,164	24,433	26,509	29,051	39,335	51,805	55,758	59,342	63,691	66,766
45	24,311	25,901	28,366	30,612	33,350	44,335	57,505	61,656	65,410	69,957	73,166
50	27,991	29,707	32,357	34,764	37,689	49,335	63,167	67,505	71,420	76,154	79,490
55	31,735	33,570	36,398	38,958	42,060	54,335	68,796	73,311	77,380	82,292	85,749
60	35,534	37,485	40,482	43,188	46,459	59,335	74,397	79,082	83,298	88,379	91,952
70	43,275	45,442	48,758	51,739	55,329	69,334	85,527	90,531	95,023	100,425	104,215
80	51,172	53,540	57,153	60,391	64,278	79,334	96,578	101,879	106,629	112,329	116,321
90	59,196	61,754	65,647	69,126	73,291	89,334	107,565	113,145	118,136	124,116	128,299
100	67,328	70,065	74,222	77,929	82,358	99,334	118,498	124,342	129,561	135,807	140,169
110	75,550	78,458	82,867	86,792	91,471	109,334	129,385	135,480	140,917	147,414	151,948
120	83,852	86,923	91,573	95,705	100,624	119,334	140,233	146,567	152,211	158,950	163,648

Disponível em: <https://bit.ly/30XtiZ9>. Acesso em: 17 nov. 2021.

Vemos, na tabela 68, que o máximo valor na linha de graus de liberdade (g.l.) igual a 6 ($q = 6$) é 18,548, que resulta em $P(\chi^2_6 \geq 18,548) = 0,05$. Logo, temos a certeza de que $P(\chi^2_6 \geq 25,3) = 0,005$.

Assim, concluímos que, ao nível de significância de 5% ($\alpha = 0,05$), não rejeitamos H_0 , visto que $P(\chi^2_6 \geq 25,3) < \alpha$. Ou seja, ao nível de significância de 5%, chegamos à conclusão de que a quantidade de interações não é a mesma todos os dias da semana.

7.2.2 Teste de independência

Podemos dizer que os testes de independência visam a testar se há independência entre duas variáveis A e B.

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

Tabela 69

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

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.

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_q^2). 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_{obs}^2 .

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

Na expressão, q é o número de graus de liberdade, sendo $q = (r - 1) \cdot (s - 1)$.

Identificamos $P = P(\chi_q^2 \geq Q_{obs}^2)$, em que, como acabamos de dizer, Q_{obs}^2 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 .

Vejamos um exemplo. 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.

Tabela 70

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

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

Vamos fazer um teste qui-quadrado de independência e verificar seu resultado.

Primeiramente, vamos estabelecer as hipóteses a serem testadas.

- **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.

Na tabela, temos o total n de 1320 observações. Vemos que, independentemente do tipo de chocolate preferido, temos o que segue.

- 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 $(810/1320) \cdot 100\% = 61,36\%$.

Se há independência entre o local em que o consumidor mora e o tipo de chocolate que ele prefere, temos as quantidades esperadas mostradas a seguir. Veja que, como se trata de quantidades teóricas, para fins de cálculos, consideraremos valores não inteiros, mesmo sabendo que o número de consumidores é inteiro.

- Quantidade esperada de consumidores que moram na zona sul e preferem Choc A: 12,22.

$$\frac{(\% \text{ de moradores da zona sul}) \cdot (\text{Número de consumidores que preferem Choc A})}{100\%} =$$
$$= \frac{4,02\% \cdot 304}{100\%} = 12,22$$

- Quantidade esperada de consumidores que moram na zona norte e preferem Choc A: 105,24.

$$\frac{(\% \text{ de moradores da zona norte}) \cdot (\text{Número de consumidores que preferem Choc A})}{100\%} =$$
$$= \frac{34,62\% \cdot 304}{100\%} = 105,24$$

- Quantidade esperada de consumidores que moram na zona central e preferem Choc A: 186,53.

$$\frac{(\% \text{ de moradores da zona central}) \cdot (\text{Número de consumidores que preferem Choc A})}{100\%} =$$
$$= \frac{61,36\% \cdot 304}{100\%} = 186,53$$

- Quantidade esperada de consumidores que moram na zona sul e preferem Choc B: 17,33.

$$\frac{(\% \text{ de moradores da zona sul}) \cdot (\text{Número de consumidores que preferem Choc B})}{100\%} =$$
$$= \frac{4,02\% \cdot 431}{100\%} = 17,33$$

- Quantidade esperada de consumidores que moram na zona norte e preferem Choc B: 149,21.

$$\frac{(\% \text{ de moradores da zona norte}) \cdot (\text{Número de consumidores que preferem Choc B})}{100\%} =$$
$$= \frac{34,62\% \cdot 431}{100\%} = 149,21$$

- Quantidade esperada de consumidores que moram na zona central e preferem Choc B: 264,46.

$$\frac{(\% \text{ de moradores da zona central}) \cdot (\text{Número de consumidores que preferem Choc B})}{100\%} =$$
$$= \frac{61,36\% \cdot 431}{100\%} = 264,46$$

- Quantidade esperada de consumidores que moram na zona sul e preferem Choc C: 14,15.

$$\frac{(\% \text{ de moradores da zona sul}) \cdot (\text{Número de consumidores que preferem Choc C})}{100\%} =$$
$$= \frac{4,02\% \cdot 352}{100\%} = 14,15$$

- Quantidade esperada de consumidores que moram na zona norte e preferem Choc C: 121,86.

$$\frac{(\% \text{ de moradores da zona norte}) \cdot (\text{Número de consumidores que preferem Choc C})}{100\%} =$$
$$= \frac{34,62\% \cdot 352}{100\%} = 121,86$$

- Quantidade esperada de consumidores que moram na zona central e preferem Choc C: 215,99.

$$\frac{(\% \text{ de moradores da zona central}) \cdot (\text{Número de consumidores que preferem Choc C})}{100\%} =$$
$$= \frac{61,36\% \cdot 352}{100\%} = 215,99$$

- Quantidade esperada de consumidores que moram na zona sul e preferem Choc D: 9,37.

$$\frac{(\% \text{ de moradores da zona sul}) \cdot (\text{Número de consumidores que preferem Choc D})}{100\%} =$$
$$= \frac{4,02\% \cdot 233}{100\%} = 9,37$$

- Quantidade esperada de consumidores que moram na zona norte e preferem Choc D: 80,66.

$$\frac{(\% \text{ de moradores da zona norte}) \cdot (\text{Número de consumidores que preferem Choc D})}{100\%} =$$
$$= \frac{34,62\% \cdot 233}{100\%} = 80,66$$

- Quantidade esperada de consumidores que moram na zona central e preferem Choc D: 142,97.

$$\frac{(\% \text{ de moradores da zona central}) \cdot (\text{Número de consumidores que preferem Choc D})}{100\%} =$$
$$= \frac{61,36\% \cdot 233}{100\%} = 142,97$$

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

Tabela 71

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

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

Tabela 72

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:

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

$$Q^2_{\text{obs}} = \frac{(11-12,22)^2}{12,22} + \frac{(9-17,33)^2}{17,33} + \frac{(5-14,15)^2}{14,15} + \frac{(28-9,37)^2}{9,37} +$$

$$+ \frac{(91-105,24)^2}{105,24} + \frac{(165-149,21)^2}{149,21} + \frac{(126-121,86)^2}{121,86} + \frac{(75-80,66)^2}{80,66} +$$

$$+ \frac{(202-186,53)^2}{186,53} + \frac{(257-264,46)^2}{264,46} + \frac{(221-215,99)^2}{215,99} + \frac{(130-142,97)^2}{142,97}$$

$$Q^2_{\text{obs}} = 0,12 + 4,00 + 5,96 + 37,04 + 1,93 + 1,67 + 0,14 + 0,40 + 1,28 + 0,21 + 0,12 + 1,18$$

$$Q^2_{\text{obs}} = 54,01$$

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:

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

Assim, temos $\chi^2_6 = 54,01$. Procuramos, em uma tabela de distribuição qui-quadrado, como a tabela 68, na linha de graus de liberdade (g.l.) igual a 6 ($q = 6$).

Vemos, na tabela 68, que o máximo valor na linha de graus de liberdade (g.l.) igual a 6 ($q = 6$) é 18,548, que resulta em $P(\chi^2_6 \geq 18,548) = 0,005$. Logo, temos a certeza de que $P(\chi^2_6 \geq 54,01) < 0,005$.

Assim, concluímos que, ao nível de significância de 5% ($\alpha = 0,05$), rejeitamos H_0 , visto que $P(\chi^2_6 \geq 54,01) < \alpha$. 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.

7.2.3 Aplicação que usa testes de aderência e de independência

População, amostra e testes

Imagine que tenhamos aplicado um questionário a uma amostra de 132 alunos matriculados no 2º ano do Ensino Médio de 5 escolas localizadas em 5 diferentes regiões de determinada cidade, a saber: uma na zona sul, uma na zona norte, uma na zona central, uma na zona leste e uma na zona oeste. A amostra foi construída de maneira a ser representativa da população de alunos do 2º ano do Ensino Médio da cidade.

Por meio do teste de aderência, queremos verificar se a distribuição das idades observadas pode ou não ser modelada pelo modelo normal de distribuição de probabilidades.

Por meio de testes de independência, queremos verificar se as seguintes variáveis são ou não são independentes.

- "Gostar de matemática" e "Relacionar a matemática com situações do dia a dia"
- "Gostar de matemática" e "Perceber relações entre a matemática e a apresentação de informações em jornais, em revistas, na TV e em sites por meio de tabelas e gráficos"
- "Gostar de matemática" e "Achar que a matemática ajuda no entendimento de outras disciplinas"
- "Gostar de matemática" e "Achar que o professor desenvolve atividades que possibilitam relacionar a matemática com outras disciplinas ou com situações do dia a dia"
- "Gostar de matemática" e "Idade"

A população em estudo corresponde aos alunos que cursam o 2º ano do Ensino Médio em determinada cidade. A inferência dos parâmetros populacionais é feita pela análise de uma amostra composta por 132 alunos que responderam a um questionário constituído por 9 perguntas, mostradas a seguir.

Questionário e outras informações

1. Qual é a sua idade?

2. Você vê relações entre a matemática e situações do seu dia a dia?

A) Sim, em várias situações.

B) Apenas em algumas situações.

C) Não, em praticamente nenhuma situação.

3. Você percebe relações entre a matemática e a apresentação de informações em jornais, em revistas, na TV e em sites por meio de tabelas e gráficos?

A) Sim, percebo claramente essas relações em diversas situações.

B) Apenas em algumas situações.

C) Não percebo essas relações em praticamente nenhuma situação.

4. Você acha que a matemática ajuda no entendimento de outras disciplinas?

A) Sim.

B) Em algumas situações.

C) Não.

5. Você gosta de matemática?

A) Sim.

B) Mais ou menos.

C) Não.

6. O seu professor desenvolve atividades que possibilitam relacionar a matemática com outras disciplinas ou com situações do dia a dia?

A) Sim.

B) Não.

7. Quanto tempo por semana, em horas, você se dedica ao estudo de matemática?

8. Você trabalha?

A) Sim.

B) Não.

9. Você acha que a disciplina matemática deveria ter exigido mais de você?

A) Sim, deveria ter exigido mais.

B) Exigiu na medida certa.

C) Não, deveria ter exigido menos.

Seguem os dados sobre os colégios cujos alunos do 2º ano do Ensino Médio participaram da resposta ao questionário.

- **Colégio situado na zona sul.** Tamanho da amostra: 26 alunos.
- **Colégio situado na zona norte.** Tamanho da amostra: 26 alunos.
- **Colégio situado na zona central.** Tamanho da amostra: 22 alunos.
- **Colégio situado na zona leste.** Tamanho da amostra: 33 alunos.
- **Colégio situado na zona oeste.** Tamanho da amostra: 25 alunos.

O tamanho total da amostra final, igual a 132, composta pela soma das amostras de cada zona, foi superior a $n = 30$, o que permite fazer a estimativa de parâmetros populacionais.

Resultados do questionário

Nas tabelas e nas figuras a seguir, encontramos informações que descrevem os resultados obtidos pela aplicação do questionário.

Tabela 73 – Questão: qual é a sua idade?

Idade (anos)	Frequência absoluta	Frequência relativa	Percentual
14	1	0,008	0,76
15	16	0,121	12,12
16	77	0,583	58,33
17	30	0,227	22,73
18	6	0,045	4,55
19	1	0,008	0,76
22	1	0,008	0,76
	132	1	100

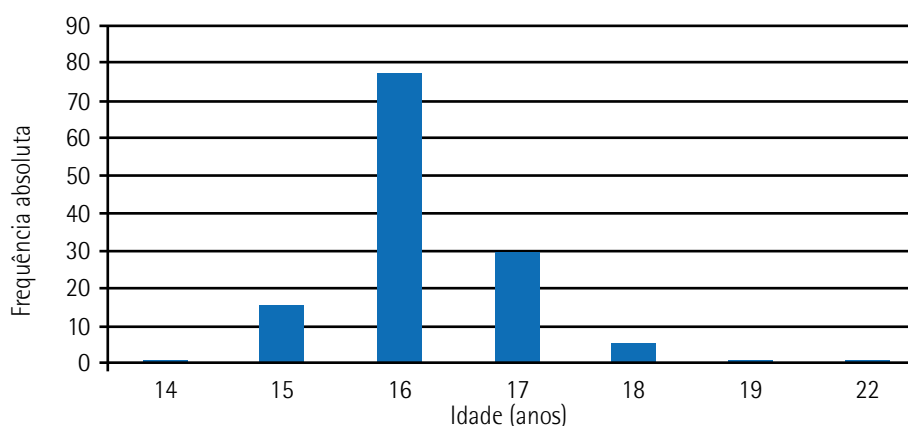


Figura 74 – Frequências absolutas: idades

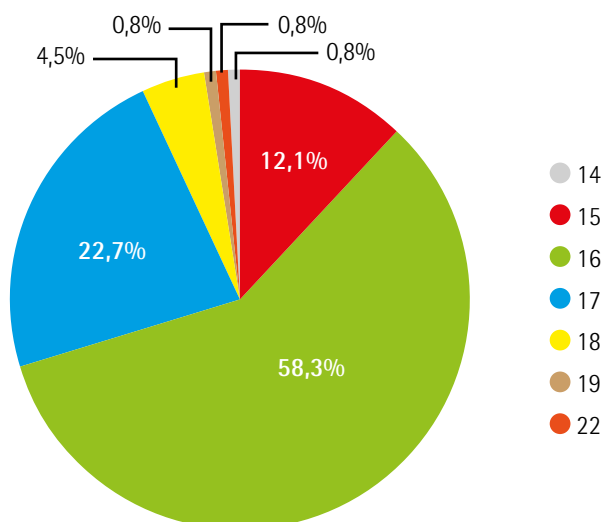


Figura 75 – Percentuais: idades

Tabela 74 – Questão: você vê relações entre a matemática e situações do seu dia a dia?

Você vê relações entre a matemática e situações do seu dia a dia?	Frequência absoluta	Frequência relativa	Percentual
Sim, em várias situações	88	0,667	66,7%
Apenas em algumas situações	42	0,318	31,8%
Não, em praticamente nenhuma situação	2	0,015	1,5%
	132	1	100%

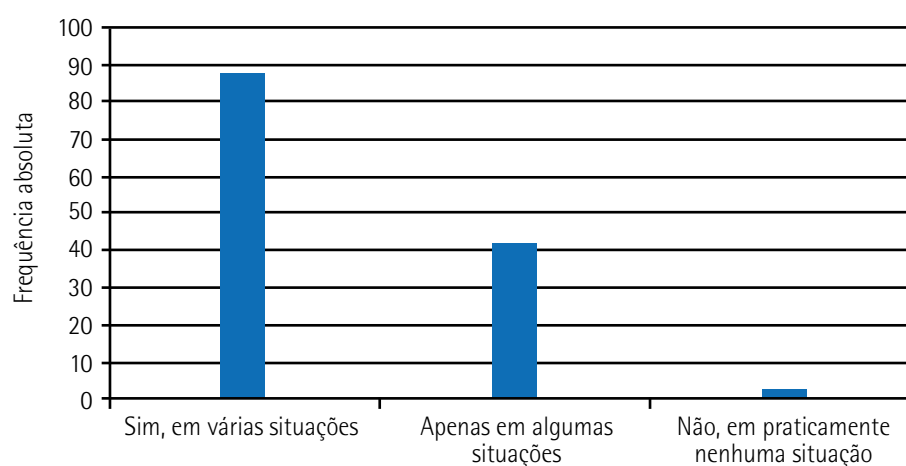


Figura 76 – Frequências absolutas: relações entre a matemática e situações do dia a dia

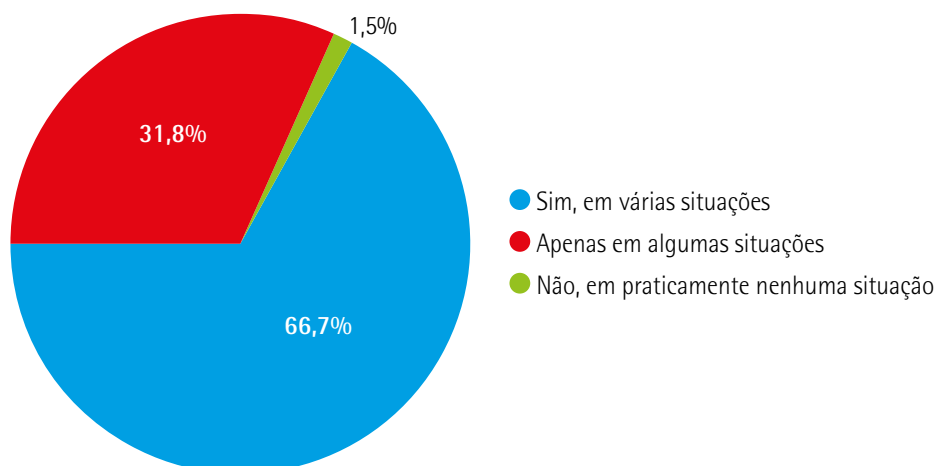


Figura 77 – Percentuais: relações entre a matemática e situações do dia a dia

Tabela 75 – Questão: você percebe relações entre a matemática e a apresentação de informações em jornais, em revistas, na TV e em sites por meio de tabelas e gráficos?

Você percebe relações entre a matemática e a apresentação de informações em jornais, em revistas, na TV e em sites por meio de tabelas e gráficos?	Frequência absoluta	Frequência relativa	Percentual
Sim, percebo claramente essas relações em diversas situações	76	0,576	57,6%
Apenas em algumas situações	54	0,409	40,9%
Não percebo essas relações em praticamente nenhuma situação	2	0,015	1,5%
	132	1	100%

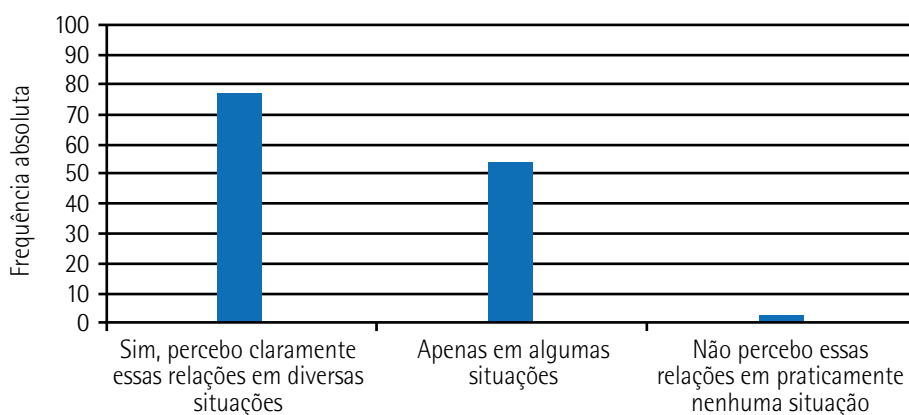


Figura 78 – Frequências absolutas: relações entre a matemática e a apresentação de informações em jornais, em revistas, na TV e em sites por meio de tabelas e gráficos

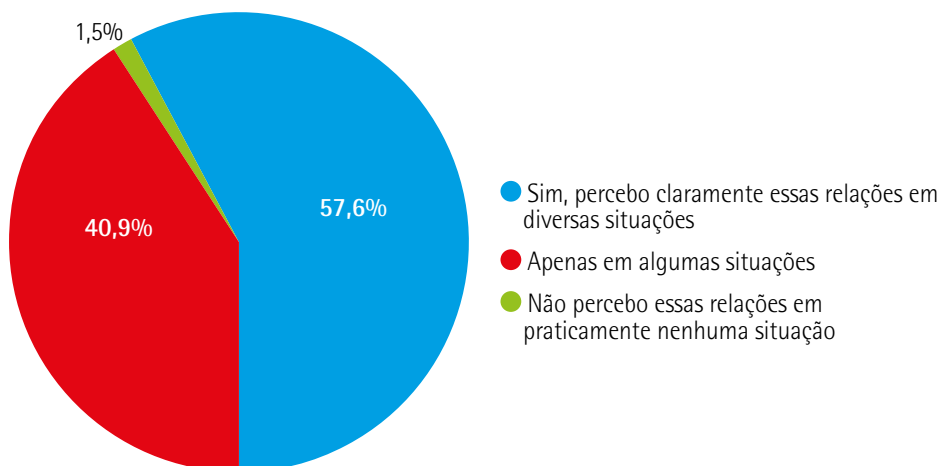


Figura 79 – Percentuais: relações entre a matemática e a apresentação de informações em jornais, em revistas, na TV e em sites por meio de tabelas e gráficos

Tabela 76 – Questão: você acha que a matemática ajuda no entendimento de outras disciplinas?

Você acha que a matemática ajuda no entendimento de outras disciplinas?	Frequência absoluta	Frequência relativa	Percentual
Sim	80	0,606	60,6%
Em algumas situações	50	0,379	37,9%
Não	2	0,015	1,5%
	132	1	100%

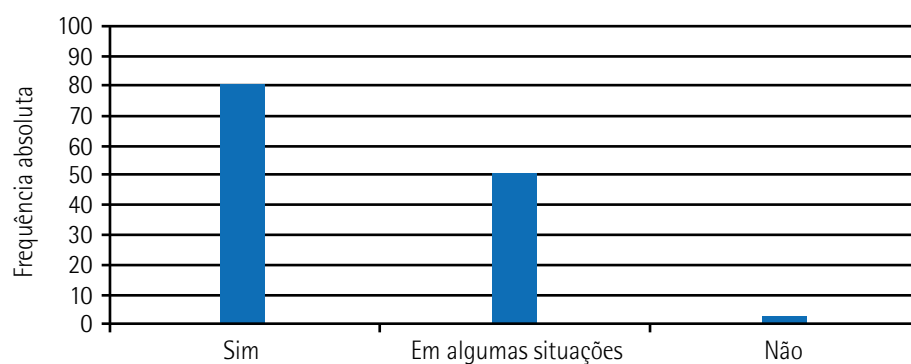


Figura 80 – Frequências absolutas: a matemática ajuda no entendimento de outras disciplinas

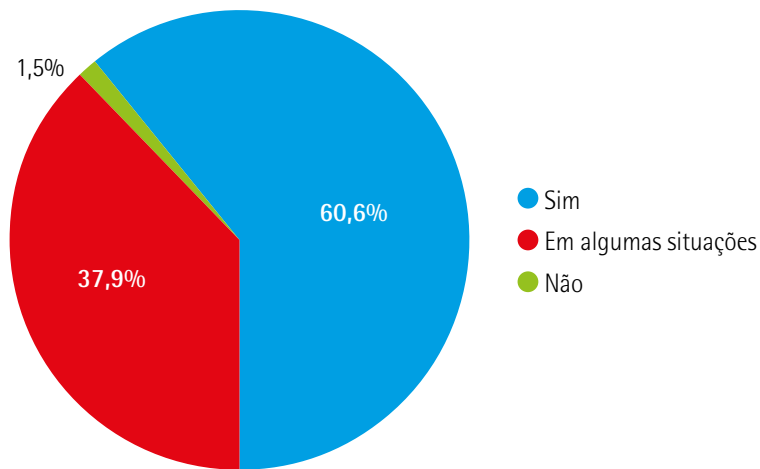


Figura 81 – Percentuais: a matemática ajuda no entendimento de outras disciplinas

Tabela 77 – Questão: você gosta de matemática?

Você gosta de matemática?	Frequência absoluta	Frequência relativa	Percentual
Sim	46	0,348	34,8%
Mais ou menos	67	0,508	50,8%
Não	19	0,144	14,4%
	132	1	100%

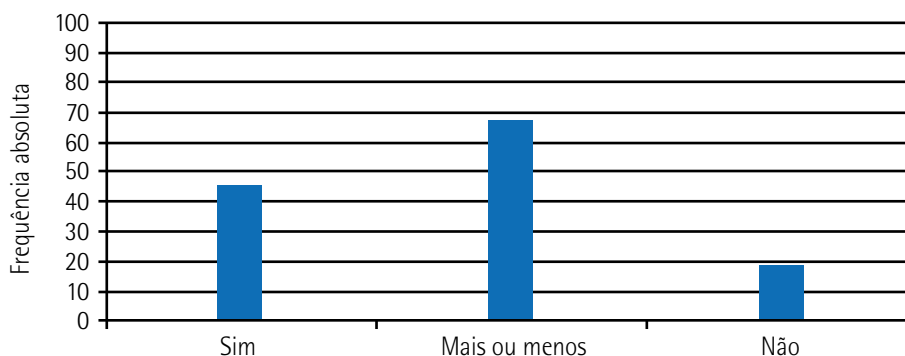


Figura 82 – Frequências absolutas: gostar de matemática

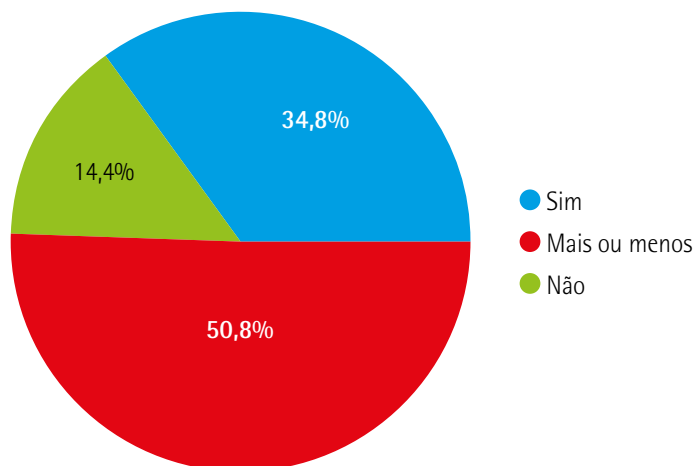


Figura 83 – Percentuais: gostar de matemática

Tabela 78 – Questão: o seu professor desenvolve atividades que possibilitam relacionar a matemática com outras disciplinas ou com situações do dia a dia?

O seu professor desenvolve atividades que possibilitam relacionar a matemática com outras disciplinas ou com situações do dia a dia?	Frequência absoluta	Frequência relativa	Percentual
Sim	82	0,621	62,1%
Não	50	0,379	37,9%
	132	1	100%

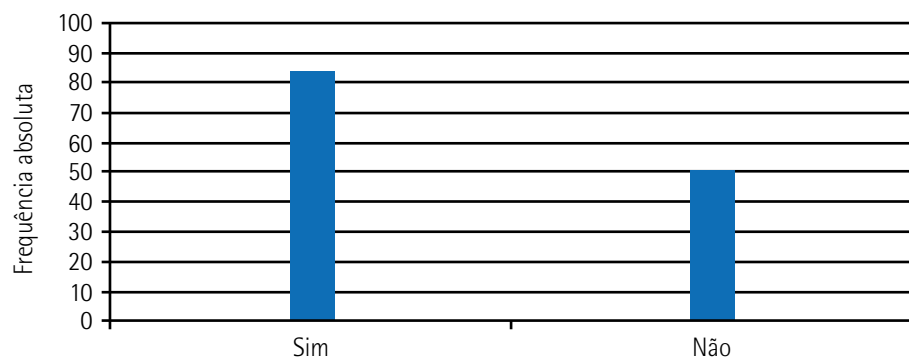


Figura 84 – Frequências absolutas: o professor desenvolve atividades que possibilitam relacionar a matemática com outras disciplinas ou com situações do dia a dia

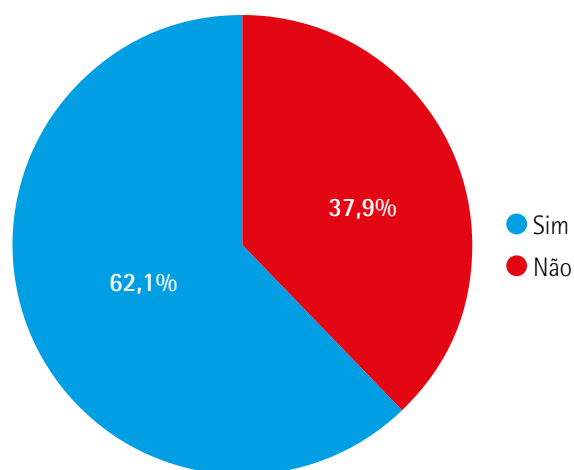


Figura 85 – Percentuais: o professor desenvolve atividades que possibilitam relacionar a matemática com outras disciplinas ou com situações do dia a dia

Tabela 79 – Questão: quanto tempo por semana, em horas, você se dedica ao estudo de matemática?

Horas semanais de estudo de matemática	Frequência absoluta	Frequência relativa	Percentual
0	42	0,318	31,8%
0,5	4	0,030	3,0%
0,75	2	0,015	1,5%
0,83	1	0,008	0,8%
1	36	0,273	27,3%
2	13	0,098	9,8%
3	17	0,129	12,9%
4	4	0,030	3,0%
5	9	0,068	6,8%
6	1	0,008	0,8%
9	1	0,008	0,8%

Horas semanais de estudo de matemática	Frequência absoluta	Frequência relativa	Percentual
10	1	0,008	0,8%
30	1	0,008	0,8%
	132	1	100%

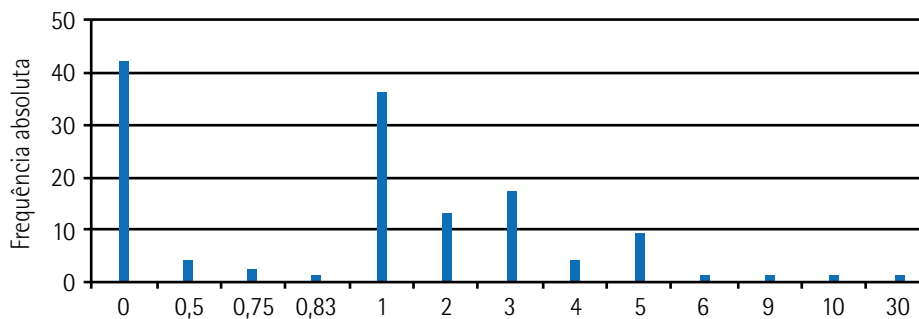


Figura 86 – Frequências absolutas: horas semanais de estudo de matemática

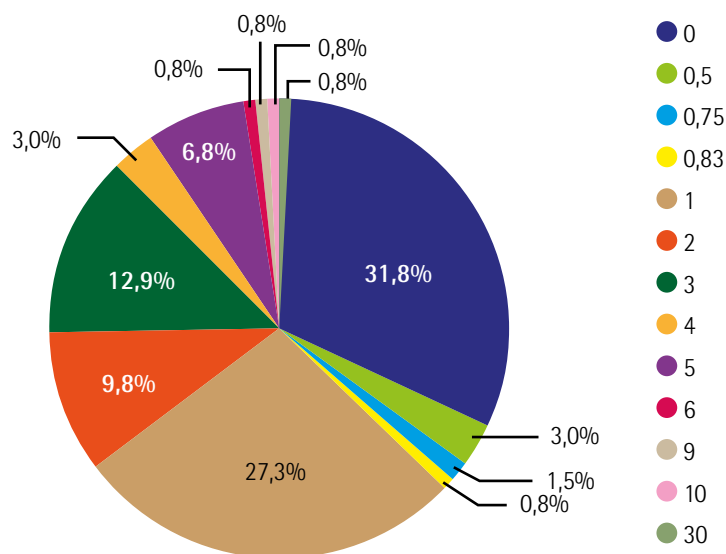


Figura 87 – Percentuais: horas semanais de estudo de matemática

Tabela 80 – Questão: você trabalha?

Você trabalha?	Frequência absoluta	Frequência relativa	Percentual
Sim	55	0,417	41,7%
Não	77	0,583	58,3%
	132	1	100%

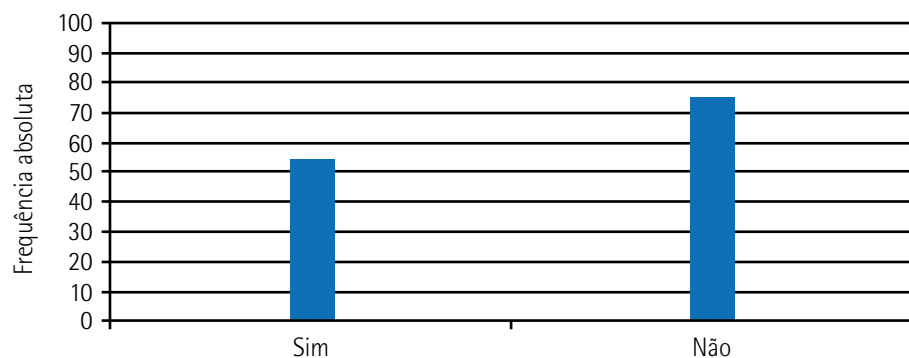


Figura 88 – Frequências absolutas: trabalha ou não trabalha

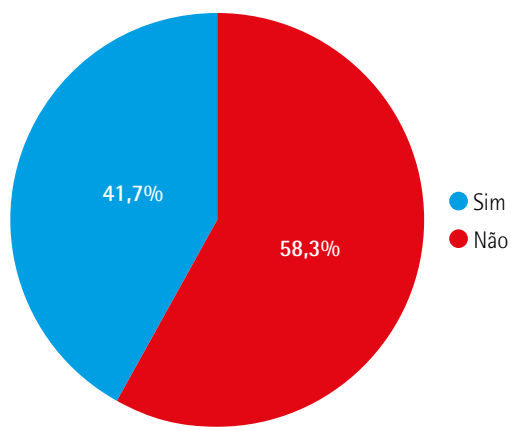


Figura 89 – Percentuais: trabalha ou não trabalha

Tabela 81 – Questão: você acha que a disciplina matemática deveria ter exigido mais de você?

Você acha que a disciplina matemática deveria ter exigido mais de você?	Frequência absoluta	Frequência relativa	Percentual
Sim, deveria ter exigido mais	64	0,485	48,5%
Exigiu na medida certa	64	0,485	48,5%
Não, deveria ter exigido menos	4	0,030	3,0%
	132	1	100%

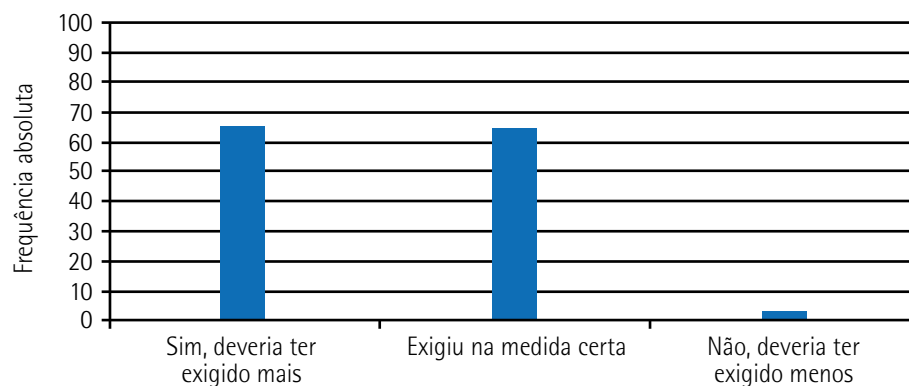


Figura 90 – Frequências absolutas: exigência da disciplina matemática

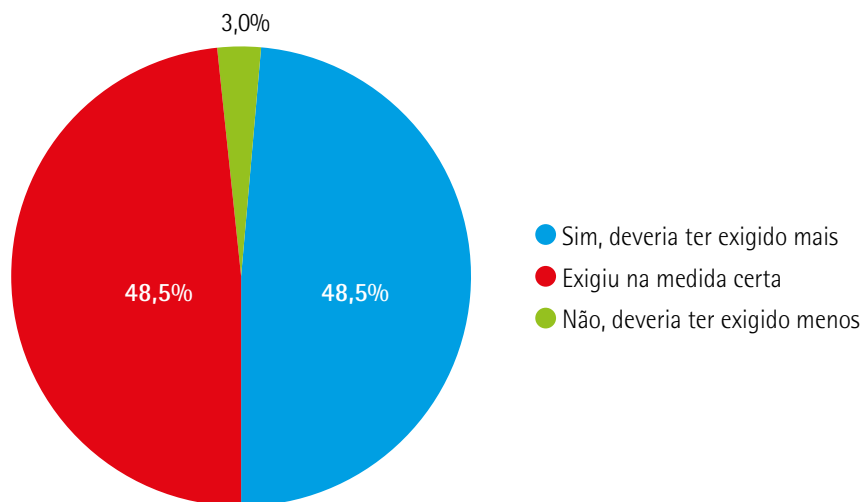


Figura 91 – Percentuais: exigência da disciplina matemática

Na figura 92, encontramos o gráfico de frequências absolutas de idades observadas por região pesquisada. Na tabela 82, temos as medidas de posição (média, mediana e moda) e as medidas de dispersão para a variável idade por zona de localização da escola.

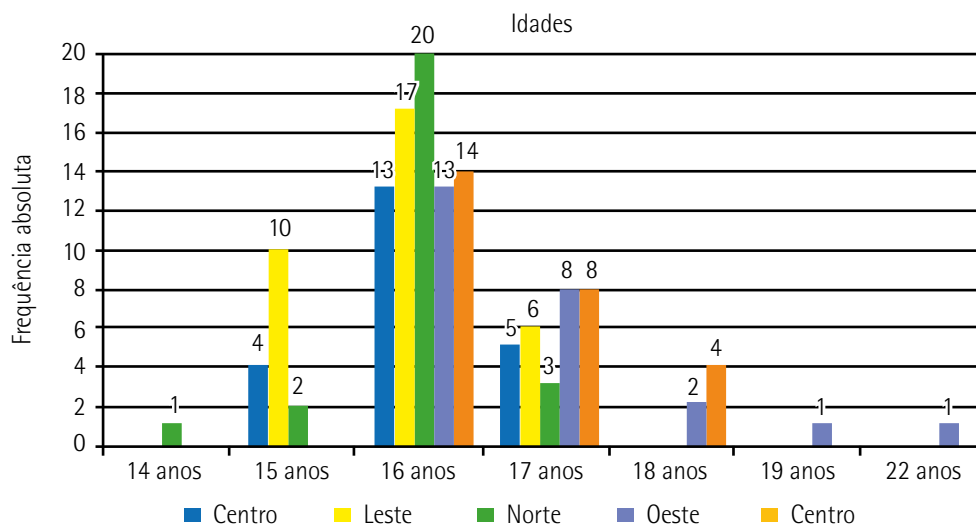


Figura 92 – Frequências absolutas: idades por região

Tabela 82 – Medidas de posição e de dispersão para a variável idade (por região)

	Variável idade (em anos)					
	Centro	Leste	Norte	Oeste	Sul	Geral
Média	16,05	15,88	15,96	16,84	16,62	16,25
Mediana	16,00	16,00	16,00	16,00	16,00	16,00
Moda	16,00	16,00	16,00	16,00	16,00	16,00

	Variável idade (em anos)					
	Centro	Leste	Norte	Oeste	Sul	Geral
Variância	0,43	0,48	0,36	1,81	0,57	0,85
Desvio padrão	0,65	0,70	0,60	1,34	0,75	0,92

A média etária mais elevada é a dos alunos da escola da região oeste. Nessa escola, também observamos o maior desvio padrão. Embora a maioria dos seus estudantes do 2º ano do Ensino Médio tenha entre 16 e 18 anos, há um aluno com 19 anos e um aluno com 22 anos.

Vemos que 66,7% dos estudantes da amostra identificam relações entre a matemática e várias situações cotidianas. Apenas 1,5% dos alunos não observam essas relações em praticamente nenhuma situação.

Na figura a seguir, encontra-se o gráfico de frequências absolutas de respostas à pergunta "Você vê relações entre a matemática e situações do seu dia a dia?" por região pesquisada. Verifica-se que, independentemente da localização da escola, a maioria das respostas é "Sim, em várias situações".

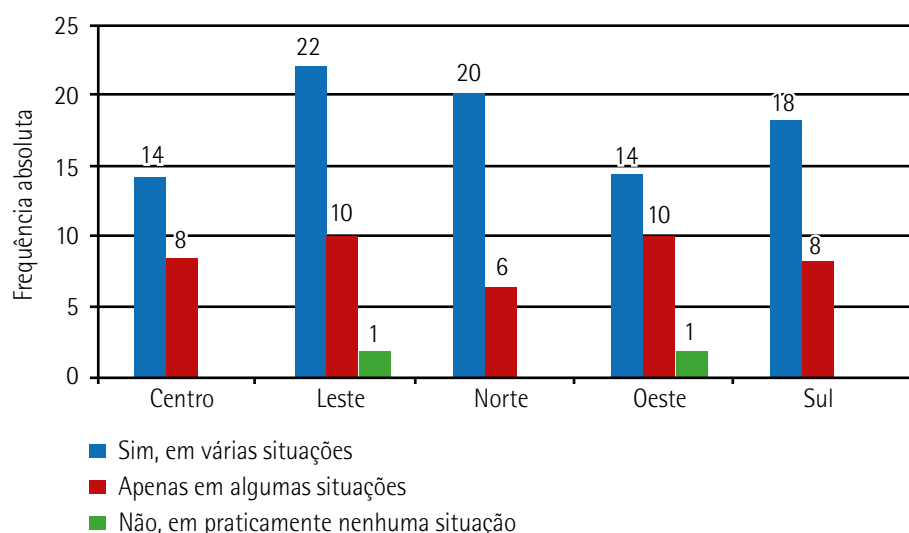


Figura 93 – Frequências absolutas: relações entre a matemática e situações do dia a dia, por região

Para mais da metade (57,6%) dos alunos da amostra, há relações entre a matemática e a apresentação de informações em jornais, em revistas, na TV e em sites por meio de tabelas e gráficos. Somente 1,5% dos alunos não observam essas relações.

Na figura a seguir, está o gráfico de frequências absolutas de respostas à pergunta "Você percebe relações entre a matemática e a apresentação de informações em jornais, em revistas, na TV e em sites por meio de tabelas e gráficos?" por região pesquisada. Verificamos que, independentemente da localização da escola, a maioria das respostas é "Sim, percebo claramente essas relações em várias situações".

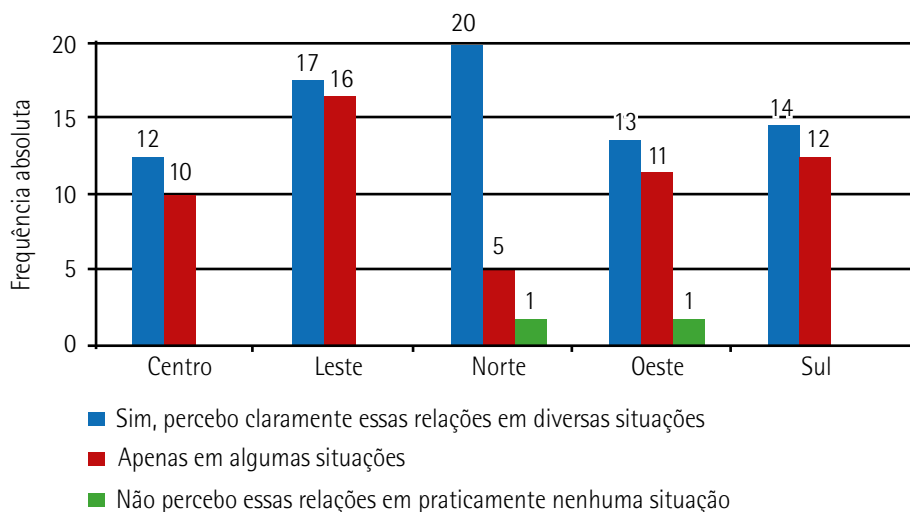


Figura 94 – Frequências absolutas: relações entre a matemática e a apresentação de informações em jornais, em revistas, na TV e em sites por meio de tabelas e gráficos, por região

Aproximadamente 60,6% dos alunos da amostra acham que a matemática auxilia no entendimento de outras disciplinas.

Na figura a seguir, vemos o gráfico de frequências absolutas de respostas à pergunta "Você acha que a matemática ajuda no entendimento de outras disciplinas?" por região pesquisada. Verificamos que, independentemente da localização da escola, nenhuma ou a minoria das respostas é "não".

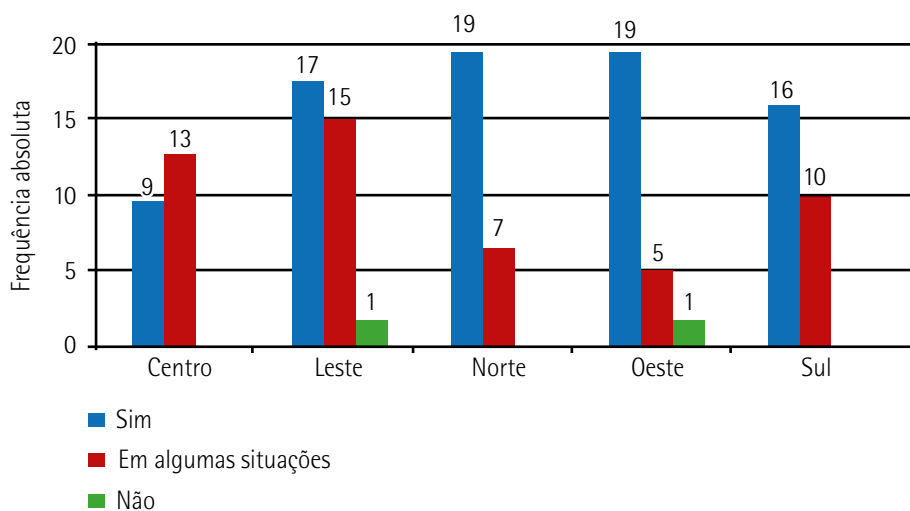


Figura 95 – Frequências absolutas: a matemática ajuda no entendimento de outras disciplinas, por região

A frequência mais elevada de respostas à questão "Você gosta de matemática?" é a opção "Mais ou menos". Isso corresponde às respostas de 50,8% dos alunos da amostra de 132 estudantes.

Na figura a seguir, encontra-se o gráfico de frequências absolutas de respostas à pergunta "Você gosta de matemática?" por região pesquisada. Verificamos que, independentemente da localização da escola, a minoria das respostas é "Não".

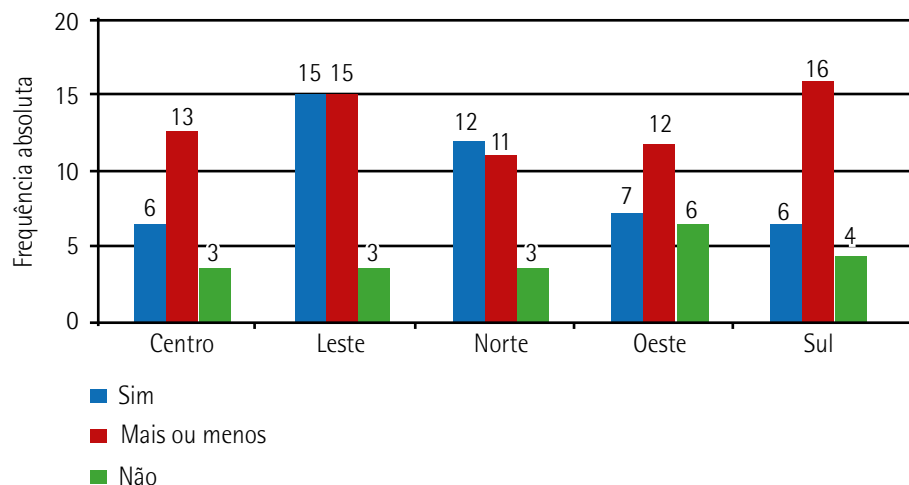


Figura 96 – Frequências absolutas: gostar de matemática, por região

Dos alunos da amostra, 63,6% acreditam que o professor desenvolve atividades que possibilitam relacionar a matemática com outras disciplinas ou com situações cotidianas.

Na figura a seguir, temos o gráfico de frequências absolutas de respostas à pergunta "O seu professor desenvolve atividades que possibilitam relacionar a matemática com outras disciplinas ou com situações do dia a dia?" por região pesquisada. Verifica-se que, independentemente da localização da escola, a maioria das respostas é "Sim", à exceção da região norte, em que a resposta "Não" é mais frequente.

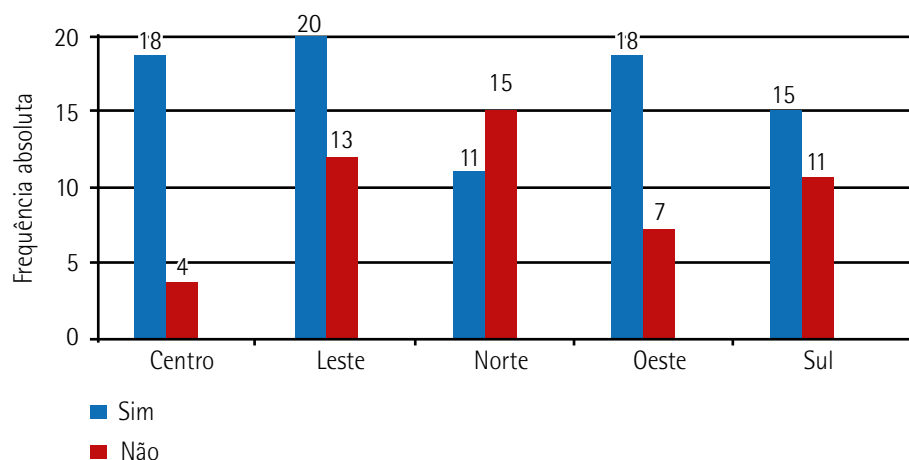


Figura 97 – Frequências absolutas: o professor desenvolve atividades que possibilitam relacionar a matemática com outras disciplinas ou com situações do dia a dia, por região

Percentual considerável dos alunos da amostra (31,8%) afirma que não dedica tempo algum (zero horas) ao estudo de matemática. Metade dos estudantes usa de 1 a 3 horas semanais para o estudo de matemática.

Na figura a seguir, temos as frequências absolutas do tempo médio semanal (em horas) de estudo de matemática por região pesquisada, e na tabela 83 estão mostradas as medidas de posição (média, mediana e moda) e as medidas de dispersão para essa variável por zona de localização da escola.

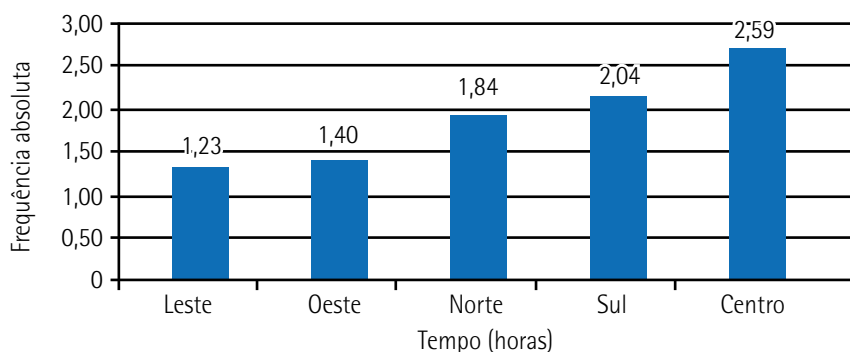


Figura 98 – Frequências absolutas: tempo médio semanal (em horas) de estudo de matemática, por região

Tabela 83 – Medidas de posição e de dispersão para o tempo de estudo (por região)

	Variável tempo de estudo (em horas semanais)					
	Centro	Leste	Norte	Oeste	Sul	Geral
Média	2,59	1,23	1,84	1,40	2,04	1,77
Mediana	3,00	1,00	1,00	1,00	-	1,00
Moda	1,00	-	1,00	-	-	-
Variância	2,54	2,89	5,62	2,58	34,36	9,45
Desvio padrão	1,65	1,70	2,37	1,61	5,86	3,07

O maior tempo médio semanal de estudo de matemática é o da escola da zona central (2,59 horas) e o menor é o da escola da zona leste (1,23 hora).

Boa parte dos alunos da amostra (41,7%) já trabalha. Na figura a seguir, estão apresentadas as frequências absolutas das respostas à pergunta "Você trabalha?" por região pesquisada.

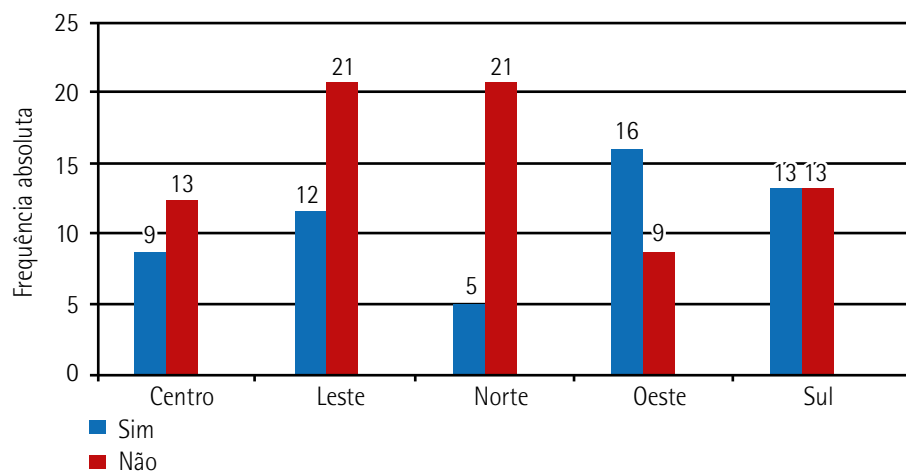


Figura 99 – Frequências absolutas: trabalha ou não trabalha, por região

Observamos que 48,5% dos alunos opinaram que a disciplina matemática deveria ter exigido mais deles e houve mesmo percentual de respostas afirmando que a disciplina matemática exigiu na medida certa. Apenas 3% dos alunos acham que a disciplina matemática deveria ter exigido menos deles.

Na figura a seguir, estão apresentadas as frequências absolutas das respostas à pergunta "Você acha que a disciplina matemática deveria ter exigido mais de você?" por região pesquisada.

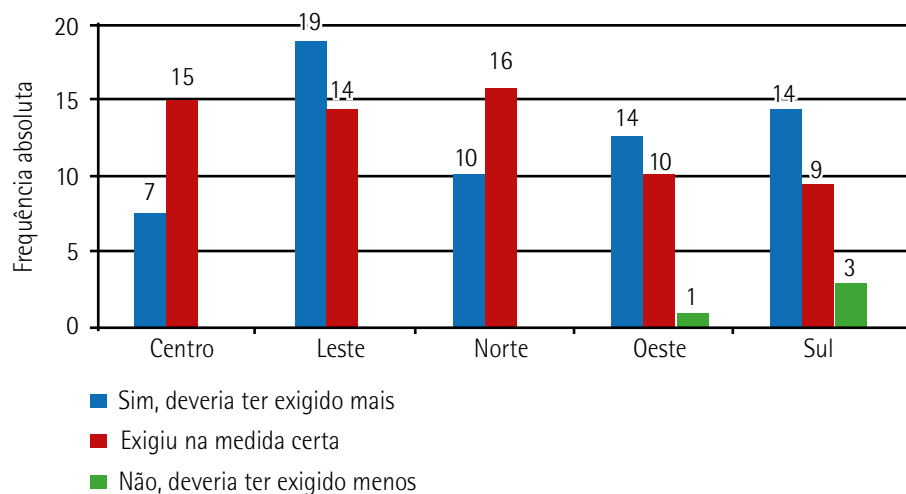


Figura 100 – Frequências absolutas: exigência da disciplina matemática, por região

Teste de aderência

A seguir, é apresentado o teste de aderência para os resultados da variável quantitativa relativa à questão das idades dos alunos.

Podemos conjecturar que a variável idade dos alunos, representada por I , seguiria, de maneira aproximada, o modelo normal de distribuição de probabilidades.

Com base nessa conjectura, as hipóteses a serem testadas são as descritas a seguir.

- **Ho (hipótese nula):** I segue o modelo normal de distribuição de probabilidades.
- **Ha (hipótese alternativa):** I não o segue modelo normal de distribuição de probabilidades.

A média amostral \bar{I} das idades das 132 observações ($n = 132$), calculada por $\bar{I} = \frac{\sum_{i=1}^n I_i}{n}$, é utilizada como estimador consistente e não viciado da média populacional. Como se trata de amostra de tamanho grande, a variância amostral S^2 , onde $S^2 = \frac{1}{n-1} \sum_{i=1}^n (I_i - \bar{I})^2$, pode ser usada como estimador da variância populacional σ^2 .

Os resultados obtidos são os citados a seguir.

- **Média amostral (estimativa para a média populacional):** $\bar{I} = 16,25$ anos.
- **Variância amostral (estimativa para a variância populacional):** $S^2 = 0,85$ anos².

Assim, testaremos a seguinte aproximação: $I \sim N(16,25; 0,85)$.

Para medirmos a diferença entre os valores esperados e os valores observados, usamos a variável aleatória Q^2 . Se houver um número suficientemente grande de observações, a distribuição de Q^2 comporta-se, aproximadamente, como um modelo qui-quadrado com $q = (k - 1)$ graus de liberdade, em que k é o número de categorias.

O valor observado de Q^2 , denominado Q^2_{obs} , é calculado por:

$$Q^2_{obs} = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}$$

Na expressão, O_i indica o i -ésimo valor observado e E_i indica o respectivo i -ésimo valor esperado.

Para a avaliação das hipóteses em estudo, agrupamos as idades nas categorias indicadas na sequência.

- **Categoria C1:** entre 13 anos (inclusive) e 16 anos (exclusive).
- **Categoria C2:** entre 16 anos (inclusive) e 17 anos (exclusive).
- **Categoria C3:** entre 17 anos (inclusive) e 18 anos (exclusive).
- **Categoria C4:** entre 18 anos (inclusive) e 23 anos (exclusive).

Vejamos, a seguir, os cálculos das frequências esperadas E_i no caso de a hipótese nula H_0 ser válida.

- $E_1 = 132 \cdot P(13 \leq I < 16 / H_0 \text{ é verdadeira})$

$$E_1 = 132 \cdot P((13 - 16,25)/\sqrt{0,85} \leq z < (16 - 16,25)/\sqrt{0,85})$$

$$E_1 = 132 \cdot P(-3,53 \leq z < -0,27) = 132 \cdot (P(0 < z < 3,53) - P(0 < z < 0,27))$$

$$E_1 = 132 \cdot (0,4998 - 0,1064) = 51,9$$

- $E_2 = 132 \cdot P(16 \leq I < 17 / H_0 \text{ é verdadeira})$

$$E_2 = 132 \cdot P((16 - 16,25)/\sqrt{0,85} \leq z < (17 - 16,25)/\sqrt{0,85})$$

$$E_2 = 132 \cdot P(-0,27 \leq z < 0,81) = 132 \cdot (P(-0,27 \leq z < 0) + P(0 < z < 0,81))$$

$$E_2 = 132 \cdot (P(0 \leq z < 0,27) + P(0 < z < 0,81)) = 132 \cdot (0,1064 + 0,2910) = 52,5$$

- $E_3 = 132 \cdot P(17 \leq I < 18 / H_0 \text{ é verdadeira})$

$$E_3 = 132 \cdot P((17 - 16,25)/\sqrt{0,85} \leq z < (18 - 16,25)/\sqrt{0,85})$$

$$E_3 = 132 \cdot P(0,81 \leq z < 1,90) = 132 \cdot (P(0 \leq z < 1,90) - P(0 < z < 0,81))$$

$$E_3 = 132 \cdot (0,4713 - 0,2910) = 23,8$$

- $E_4 = 132 \cdot P(18 \leq I < 23 / H_0 \text{ é verdadeira})$

$$E_4 = 132 \cdot P((18 - 16,25)/\sqrt{0,85} \leq z < (23 - 16,25)/\sqrt{0,85})$$

$$E_4 = 132 \cdot P(1,90 \leq z < 7,38) = 132 \cdot (P(0 \leq z < 7,38) - P(0 < z < 1,90))$$

$$E_4 = 132 \cdot (0,5 - 0,4713) = 3,8$$

Na tabela a seguir, estão apresentadas as frequências observadas e esperadas para cada categoria de idade (em anos).

Tabela 84 – Idades: frequências observadas e esperadas

Categoria	Idades (anos)	Frequências observadas	Frequências esperadas
C1	Entre 13 (inclusive) e 16 (exclusive)	17	51,9
C2	Entre 16 (inclusive) e 17 (exclusive)	77	52,5

Categoria	Idades (anos)	Frequências observadas	Frequências esperadas
C3	Entre 17 (inclusive) e 18 (exclusive)	30	23,8
C4	Entre 18 (inclusive) e 23 (exclusive)	8	3,8

O valor observado de Q^2 , ou seja, Q^2_{obs} , para os dados valores apresentados na tabela, resulta em 41,1, pois:

$$Q^2_{\text{obs}} = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i} = \frac{(17 - 51,9)^2}{51,9} + \frac{(77 - 52,5)^2}{52,5} + \frac{(30 - 23,8)^2}{23,8} + \frac{(8 - 3,8)^2}{3,8} = 41,1$$

Como há um número grande de observações ($n = 132$), vimos que temos, de forma aproximada, um modelo qui-quadrado com $(k - 1) = (4 - 1) = 3$ graus de liberdade.

Da tabela da distribuição qui-quadrado com 3 graus de liberdade e assumindo o nível de significância de 5%, temos o valor crítico de 7,815. Desse modo, a região crítica (RC) corresponde a valores maiores do que 7,815.

O valor calculado de Q^2 para a situação em análise vale 41,1, ou seja, ele pertence à RC e, consequentemente, rejeitamos a hipótese H_0 ao nível de significância de 5%. Em outras palavras, ao nível de significância de 5%, decidimos que a variável I não segue um modelo normal de distribuição de probabilidades com média de 16,25 anos e variância de 0,85 anos².

A figura a seguir mostra as densidades de frequências observadas e esperadas da variável idade.

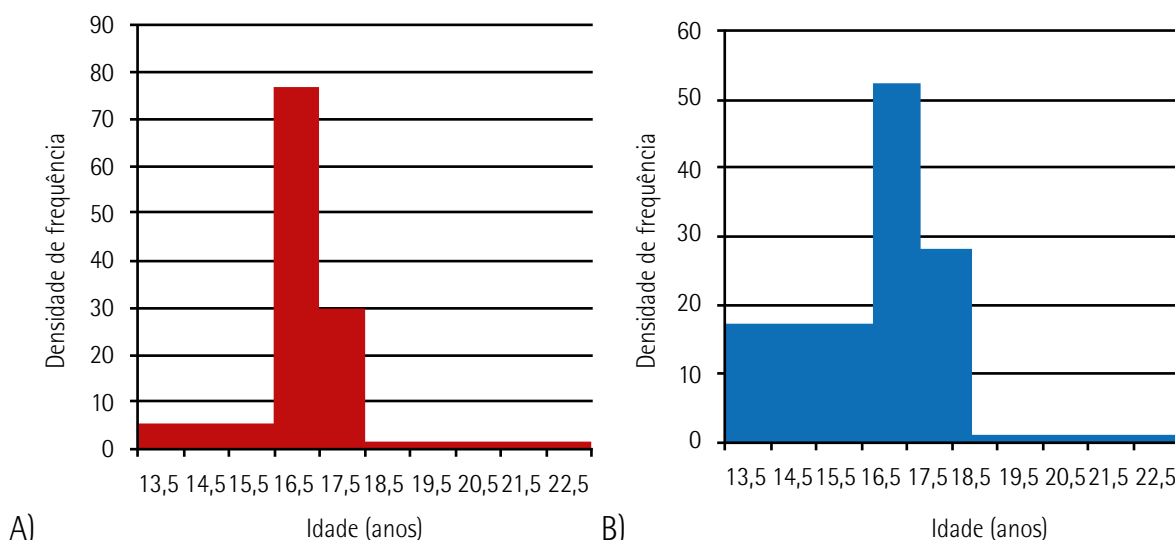


Figura 101 – Densidades de frequências observadas (A) e esperadas (B) da variável idade

Testes de independência

A seguir, são mostrados testes de independência para a avaliação da ocorrência ou da não ocorrência de dependência entre diversas variáveis pesquisadas no questionário.

1) "Gostar de matemática" e "Relacionar a matemática com situações do dia a dia"

Desejamos testar se existe ou não existe dependência entre "Gostar de matemática" e "Relacionar a matemática com situações do dia a dia". Para tanto, consideremos as variáveis a seguir.

- **X:** relacionar a matemática com situações do dia a dia.
- **Y:** gostar de matemática.

Na apresentação quantitativa das respostas observadas para X ("Relacionar a matemática com situações do dia a dia") são consideradas as categorias "Sim, em várias situações" e "Apenas em algumas situações" ou "Não, em praticamente nenhuma situação". No caso de Y ("Gostar de matemática"), são consideradas as categorias "Sim" ou "Mais ou menos" e "Não". Os resultados observados nos questionários encontram-se na tabela a seguir.

Tabela 85 – Dupla entrada: valores observados: "Relacionar a matemática com situações do dia a dia" e "Gostar de matemática"

Y \ X	Sim, em várias situações	Apenas em algumas situações Não, em praticamente nenhuma situação	Total
Sim Mais ou menos	77	36	113
Não	11	8	19
Total	88	44	132

Testaremos as hipóteses descritas na sequência.

- **Ho (hipótese nula):** as variáveis X e Y são independentes.
- **Ha (hipótese alternativa):** as variáveis X e Y não são independentes.

Na tabela a seguir, estão apresentados os valores esperados para as variáveis em estudo sob a hipótese Ho de independência entre elas.

Tabela 86 – Dupla entrada: valores esperados: “Relacionar a matemática com situações do dia a dia” e “Gostar de matemática”

Y \ X	Sim, em várias situações	Apenas em algumas situações Não, em praticamente nenhuma situação	Total
Sim Mais ou menos	75,3	37,7	113
Não	12,7	6,3	19
Total	88	44	132

Para medirmos a diferença entre os valores esperados e observados, usaremos a variável aleatória Q^2 . Para um número suficientemente grande de observações, a distribuição de Q^2 comporta-se, aproximadamente, como um modelo qui-quadrado com $(r - 1) \cdot (s - 1)$ graus de liberdade, em que r é o número de linhas e s é o número de colunas da tabela de dupla entrada.

No caso, temos que $(r - 1) \cdot (s - 1) = (2 - 1) \cdot (2 - 1) = 1$ grau de liberdade.

Considerando α igual a 5%, a tabela da distribuição qui-quadrado com 1 grau de liberdade fornece 3,841. Desse modo, a região crítica (RC) corresponde a valores maiores do que 3,841.

Sob a validade da hipótese nula H_0 , o valor observado de Q^2 , denominado Q^2_{obs} , é calculado por:

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

Na expressão, O indica o valor observado e E indica o respectivo valor esperado.

Para os valores das tabelas, o valor de Q^2_{obs} resulta em 0,801, pois:

$$Q^2_{obs} = \sum_{i=1}^r \sum_{j=1}^s \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \frac{(77 - 75,3)^2}{75,3} + \frac{(36 - 37,7)^2}{37,7} + \frac{(11 - 12,7)^2}{12,7} + \frac{(8 - 6,3)^2}{6,3} = 0,801$$

Como Q^2_{obs} igual a 0,801 não pertence à RC, concluímos pela não rejeição da hipótese nula, ou seja, as duas variáveis são independentes segundo um nível de significância de 5%.

2) “Gostar de matemática” e “Perceber relações entre a matemática e a apresentação de informações em jornais, em revistas, na TV e em sites por meio de tabelas e gráficos”

Desejamos testar se existe ou não existe dependência entre “Gostar de matemática” e “Perceber relações entre a matemática e a apresentação de informações em jornais, em revistas, na TV e em sites por meio de tabelas e gráficos”. Para tanto, consideremos as variáveis a seguir.

- **X:** relações entre a matemática e a apresentação de informações em jornais, em revistas, na TV e em sites por meio de tabelas e gráficos.
- **Y:** gostar de matemática.

Na apresentação quantitativa das respostas observadas para X ("Perceber relações entre a matemática e a apresentação de informações em jornais, em revistas, na TV e em sites por meio de tabelas e gráficos"), são consideradas as categorias "Perceber claramente as relações" e "Perceber relações em algumas situações" ou "Não perceber as relações". No caso de Y ("Gostar de matemática"), são consideradas as categorias "Sim" ou "Mais ou menos" e "Não". Os resultados observados nos questionários encontram-se na tabela a seguir.

Tabela 87 – Dupla entrada: valores observados: "Perceber relações entre a matemática e a apresentação de informações em jornais, em revistas, na TV e em sites por meio de tabelas e gráficos" e "Gostar de matemática"

Y \ X	Perceber claramente as relações	Perceber relações em algumas situações Não perceber as relações	Total
Sim Mais ou menos	67	46	113
Não	9	10	19
Total	76	56	132

Testaremos as hipóteses descritas na sequência.

- **Ho (hipótese nula):** as variáveis X e Y são independentes.
- **Ha (hipótese alternativa):** as variáveis X e Y não são independentes.

Na tabela a seguir, estão apresentados os valores esperados para as variáveis em estudo sob a hipótese Ho de independência entre elas.

Tabela 88 – Dupla entrada: valores esperados: "Perceber relações entre a matemática e a apresentação de informações em jornais, em revistas, na TV e em sites por meio de tabelas e gráficos" e "Gostar de matemática"

Y \ X	Perceber claramente as relações	Perceber relações em algumas situações Não perceber as relações	Total
Sim Mais ou menos	65,1	47,9	113
Não	10,9	8,1	19
Total	76	56	132

Analogamente ao que já fizemos, para medirmos a diferença entre os valores esperados e observados, usaremos a variável aleatória Q^2 aproximada por um modelo qui-quadrado com 1 grau de liberdade. Considerando o nível de significância α igual a 5%, a tabela da distribuição qui-quadrado com 1 grau de liberdade fornece 3,841. Desse modo, a região crítica (RC) corresponde a valores maiores do que 3,841.

Para os valores das tabelas, o valor de Q^2_{obs} resulta em 0,908, pois:

$$Q^2_{obs} = \sum_{i=1}^r \sum_{j=1}^s \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \frac{(67 - 65,1)^2}{65,1} + \frac{(46 - 47,9)^2}{47,9} + \frac{(9 - 10,9)^2}{10,9} + \frac{(10 - 8,1)^2}{8,1} = 0,908$$

Como Q^2_{obs} igual a 0,908 não pertence à RC, concluímos pela não rejeição da hipótese nula, ou seja, as duas variáveis são independentes segundo um nível de significância de 5%.

3) "Gostar de matemática" e "Achar que a matemática ajuda no entendimento de outras disciplinas"

Desejamos testar se existe ou não existe dependência entre "Gostar de matemática" e "Achar que a matemática ajuda no entendimento de outras disciplinas". Para tanto, consideremos as variáveis a seguir.

- **X:** achar que a matemática ajuda no entendimento de outras disciplinas.
- **Y:** gostar de matemática.

Na apresentação quantitativa das respostas observadas para X ("Achar que a matemática ajuda no entendimento de outras disciplinas") são consideradas as categorias "Sim" e "Em algumas situações" ou "Não". No caso de Y ("Gostar de matemática"), são consideradas as categorias "Sim" ou "Mais ou menos" e "Não". Os resultados observados nos questionários encontram-se na tabela a seguir.

Tabela 89 – Dupla entrada: valores observados: "Achar que a matemática ajuda no entendimento de outras disciplinas" e "Gostar de matemática"

Y \ X	Sim	Em algumas situações Não	Total
Sim Mais ou menos	72	41	113
Não	8	11	19
Total	80	52	132

Testaremos as hipóteses descritas na sequência.

- **Ho (hipótese nula):** as variáveis X e Y são independentes.
- **Ha (hipótese alternativa):** as variáveis X e Y não são independentes.

Na tabela a seguir, estão apresentados os valores esperados para as variáveis em estudo sob a hipótese H_0 de independência entre elas.

Tabela 90 – Dupla entrada: valores esperados: "Achar que a matemática ajuda no entendimento de outras disciplinas" e "Gostar de matemática"

Y \ X	Sim	Em algumas situações Não	Total
Sim Mais ou menos	68,5	44,5	113
Não	11,5	7,5	19
Total	80	52	132

Analogamente ao que já fizemos, para medirmos a diferença entre os valores esperados e observados, usaremos a variável aleatória Q^2 aproximada por um modelo qui-quadrado com 1 grau de liberdade. Considerando o nível de significância α igual a 5%, a tabela da distribuição qui-quadrado com 1 grau de liberdade fornece 3,841. Desse modo, a região crítica (RC) corresponde a valores maiores do que 3,841.

Para os valores das tabelas, o valor de Q^2_{obs} resulta em 3,15, pois:

$$Q^2_{obs} = \sum_{i=1}^r \sum_{j=1}^s \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \frac{(72 - 68,5)^2}{68,5} + \frac{(41 - 44,5)^2}{44,5} + \frac{(8 - 11,5)^2}{11,5} + \frac{(11 - 7,5)^2}{7,5} = 3,15$$

Como Q^2_{obs} igual a 3,15 não pertence à RC, concluímos pela aceitação da hipótese nula, ou seja, as duas variáveis são independentes segundo um nível de significância de 5%.

4) "Gostar de matemática" e "Achar que o professor desenvolve atividades que possibilitam relacionar a matemática com outras disciplinas ou com situações do dia a dia"

Desejamos testar se existe ou não existe dependência entre "Gostar de matemática" e "Achar que o professor desenvolve atividades que possibilitam relacionar a matemática com outras disciplinas ou com situações do dia a dia". Para tanto, consideremos as variáveis a seguir.

- **X:** achar que o professor desenvolve atividades que possibilitam relacionar a matemática com outras disciplinas ou com situações do dia a dia.
- **Y:** gostar de matemática.

Na apresentação quantitativa das respostas observadas para X ("Achar que o professor desenvolve atividades que possibilitam relacionar a matemática com outras disciplinas ou com situações do dia a dia") foram consideradas as categorias "Sim" e "Não". No caso de Y ("Gostar de matemática"), foram consideradas as categorias "Sim" ou "Mais ou menos" e "Não". Os resultados observados nos questionários encontram-se na tabela a seguir.

Tabela 91 – Dupla entrada: valores observados: “Achar que o professor desenvolve atividades que possibilitam relacionar a matemática com outras disciplinas ou com situações do dia a dia” e “Gostar de matemática”

Y \ X	Sim	Não	Total
Sim Mais ou menos	69	44	113
Não	13	6	19
Total	82	50	132

Testaremos as hipóteses descritas na sequência.

- **Ho (hipótese nula):** as variáveis X e Y são independentes.
- **Ha (hipótese alternativa):** as variáveis X e Y não são independentes.

Na tabela a seguir, estão apresentados os valores esperados para as variáveis em estudo sob a hipótese Ho de independência das variáveis.

Tabela 92 – Dupla entrada: valores esperados: “Achar que o professor desenvolve atividades que possibilitam relacionar a matemática com outras disciplinas ou com situações do dia a dia” e “Gostar de matemática”

Y \ X	Sim	Não	Total
Sim Mais ou menos	70,2	42,8	113
Não	11,8	7,2	19
Total	82	50	132

Analogamente ao que já fizemos, para medirmos a diferença entre os valores esperados e observados, usaremos a variável aleatória Q^2 aproximada por um modelo qui-quadrado com 1 grau de liberdade. Considerando o nível de significância α igual a 5%, a tabela da distribuição qui-quadrado com 1 grau de liberdade fornece 3,841. Desse modo, a região crítica (RC) corresponde a valores maiores do que 3,841.

Para os valores das tabelas, o valor de Q^2_{obs} resulta em 0,376, pois:

$$Q^2_{\text{obs}} = \sum_{i=1}^r \sum_{j=1}^s \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \frac{(69 - 70,2)^2}{70,2} + \frac{(44 - 42,8)^2}{42,8} + \frac{(13 - 11,8)^2}{11,8} + \frac{(6 - 7,2)^2}{7,2} = 0,376$$

Como Q^2_{obs} igual a 0,376 não pertence à RC, concluímos pela não rejeição da hipótese nula, ou seja, as duas variáveis são independentes segundo um nível de significância de 5%.

5) "Gostar de matemática" e "Idade"

Desejamos testar se existe ou não existe dependência entre "Gostar de matemática" e "Idade". Para tanto, consideremos as variáveis a seguir.

- **X:** idade (em anos).
- **Y:** gostar de matemática.

Na apresentação quantitativa das respostas observadas para X ("Idade"), são consideradas as categorias "14 a 16 anos" e "17 a 22 anos". No caso de Y ("Gostar de matemática"), são consideradas as categorias "Sim" ou "Mais ou menos" e "Não". Os resultados observados nos questionários encontram-se na tabela a seguir.

Tabela 93 – Dupla entrada: valores observados: "Idade" e "Gostar de matemática"

Y \ X	14 a 16 anos	17 a 22 anos	Total
Sim	83	30	113
Mais ou menos			
Não	11	8	19
Total	94	38	132

Testaremos as hipóteses descritas na sequência.

- **Ho (hipótese nula):** as variáveis X e Y são independentes.
- **Ha (hipótese alternativa):** as variáveis X e Y não são independentes.

Na tabela a seguir, estão apresentados os valores esperados para as variáveis em estudo sob a hipótese Ho de independência entre elas.

Tabela 94 – Dupla entrada: valores esperados: "Idade" e "Gostar de matemática"

Y \ X	14 a 16 anos	17 a 22 anos	Total
Sim	80,4	32,5	113
Mais ou menos			
Não	13,6	5,5	19
Total	94	38	132

Analogamente ao que já fizemos, para medirmos a diferença entre os valores esperados e observados, usaremos a variável aleatória Q^2 aproximada por um modelo qui-quadrado com 1 grau de liberdade. Considerando o nível de significância α igual a 5%, a tabela da distribuição qui-quadrado com 1 grau de liberdade fornece 3,841. Desse modo, a região crítica (RC) corresponde a valores maiores do que 3,841.

Para os valores das tabelas, o valor de Q^2_{obs} resulta em 1,910, pois:

$$Q^2_{\text{obs}} = \sum_{i=1}^r \sum_{j=1}^s \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \frac{(83 - 80,4)^2}{80,4} + \frac{(30 - 32,5)^2}{32,5} + \frac{(11 - 13,6)^2}{13,6} + \frac{(8 - 5,5)^2}{5,5} = 1,910$$

Como Q^2_{obs} igual a 1,910 não pertence à RC, concluímos pela não rejeição da hipótese nula, ou seja, as duas variáveis são independentes segundo um nível de significância de 5%.

Conclusões vindas dos testes

No teste de aderência, verificamos que a distribuição das idades observadas, ao nível de significância de 5%, não pode ser modelada pelo modelo normal. Esse resultado já era esperado em função dos elevados afastamentos entre os valores de fato observados e os valores esperados. A média das médias etárias dos alunos que participaram da pesquisa é aproximadamente 16 anos, e a mediana das 132 observações é igual a 16 anos (58,3% dos estudantes têm essa idade, e cerca de 93% deles têm entre 15 e 17 anos). Esse resultado faz sentido, uma vez que estudantes de um mesmo ano do Ensino Médio tendem a ter idades muito parecidas, o que não condiz com uma distribuição normal.

No primeiro teste de independência, verificamos que, ao nível de significância de 5%, as variáveis "Gostar de matemática" e "Relacionar a matemática com situações do dia a dia" são independentes.

No segundo teste de independência, verificamos que, ao nível de significância de 5%, as variáveis "Gostar de matemática" e "Perceber relações entre a matemática e a apresentação de informações em jornais, em revistas, na TV e em sites por meio de tabelas e gráficos" são independentes.

No terceiro teste de independência, verificamos que, ao nível de significância de 5%, as variáveis "Gostar de matemática" e "Achar que a matemática ajuda no entendimento de outras disciplinas" são independentes. Isso significa que não há associação entre as variáveis.

No quarto teste de independência, verificamos que, ao nível de significância de 5%, as variáveis "Gostar de matemática" e "Achar que o professor desenvolve atividades que possibilitam relacionar a matemática com outras disciplinas ou com situações do dia a dia" são independentes, ou seja, observamos que não há associação entre essas variáveis.

No quinto teste de independência, verificamos que, ao nível de significância de 5%, as variáveis "Gostar de matemática" e "Idade" são independentes. Logo, a idade do aluno não interfere no fato de ele gostar ou não gostar de matemática.

8 ESTATÍSTICA INDUTIVA: PARTE 4

8.1 Correlação e regressão linear

8.1.1 Coeficiente de correlação

Observe a tabela a seguir, em que temos valores de duas variáveis, X e Y.

Tabela 95 – Valores de duas variáveis, X e Y

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

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

Podemos começar a pensar na resposta a essa pergunta pela observação de um gráfico em que plotamos os 5 pares ordenados mostrados na tabela, detalhados a seguir.

- $(x_1; y_1) = (-2,1; -4,4)$
- $(x_2; y_2) = (1,5; 2,8)$
- $(x_3; y_3) = (3,3; 6,1)$
- $(x_4; y_4) = (5,6; 10,1)$
- $(x_5; y_5) = (8,8; 18,3)$

Vejamos, na figura a seguir, uma representação gráfica desses pontos no plano cartesiano, chamada de **gráfico de dispersão**.

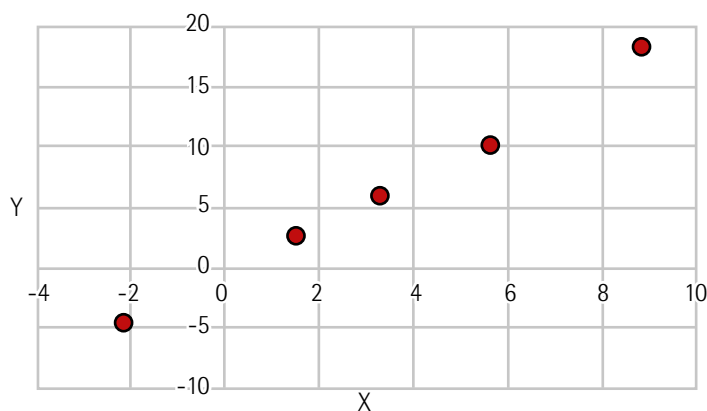


Figura 102 – Valores de duas variáveis, X e Y

Visualmente, parece que as variáveis X e Y estão linearmente relacionadas, pois os 5 pares de pontos parecem estar alinhados.

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).

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

Vamos aplicar essa fórmula aos dados da tabela 95. Como se trata de um cálculo extenso, é interessante que o façamos por partes.

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

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

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

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

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

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

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

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

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

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

Logo, ficamos com:

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

$$R = \frac{251,17 - 112,518}{\sqrt{\left[126,35 - 5 \cdot (3,42)^2 \right] \cdot \left[501,31 - 5 \cdot (6,58)^2 \right]}}$$

$$R = \frac{138,652}{\sqrt{\left[126,35 - 58,482 \right] \cdot \left[501,31 - 216,482 \right]}} = \frac{138,652}{\sqrt{19330,71}}$$

$$R = 0,997$$

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.



Observação

O coeficiente de correlação R também é chamado de **coeficiente de Pearson**.

O coeficiente de correlação R varia de -1 a 1 , sendo que:

- 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.

Na figura a seguir, podemos ver exemplos de gráficos de dispersão para diferentes valores de R .

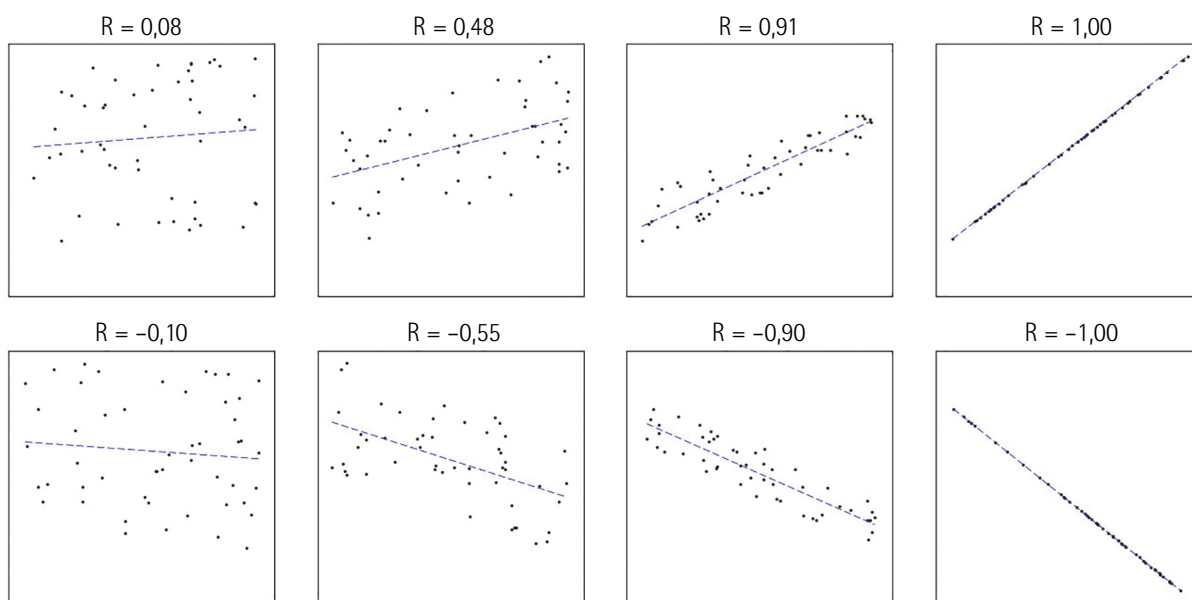


Figura 103 – Exemplos de gráficos de dispersão



Observação

Se elevarmos R ao quadrado, obteremos R^2 , que é o coeficiente de determinação.

8.1.2 Regressão linear

Muitas vezes, queremos saber qual é a reta que se ajusta da melhor maneira aos pontos de um gráfico de dispersão, como o gráfico da figura 104. 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.

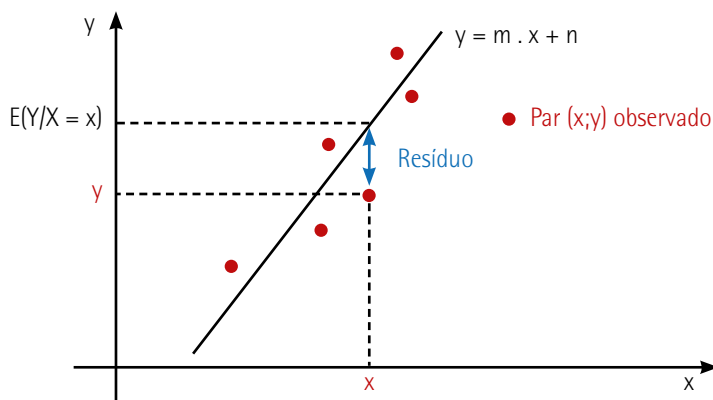


Figura 104 – Reta $y = m \cdot x + n$ cujos coeficientes m e n minimizam a soma dos quadrados dos resíduos vindos da diferença entre os valores y efetivamente observados e os seus valores esperados $E(Y/X = x)$

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

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

$$n = \bar{y}_{obs} - m \cdot \bar{x}_{obs}$$

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) da tabela 96.

Tabela 96 – Valores de duas variáveis, X e Y

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

Na tabela 96, temos os 5 pares (x,y) observados a seguir.

- $(x_1; y_1) = (-3; -10,5)$
- $(x_2; y_2) = (-1; -3)$

- $(x_3; y_3) = (0; 0,5)$
- $(x_4; y_4) = (2; 5)$
- $(x_5; y_5) = (5; 17)$

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

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

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

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

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

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

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

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

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

Logo:

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

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

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

Na figura 105, temos os 5 pares (x;y) observados e a reta $y = 3,336x - 0,202$.

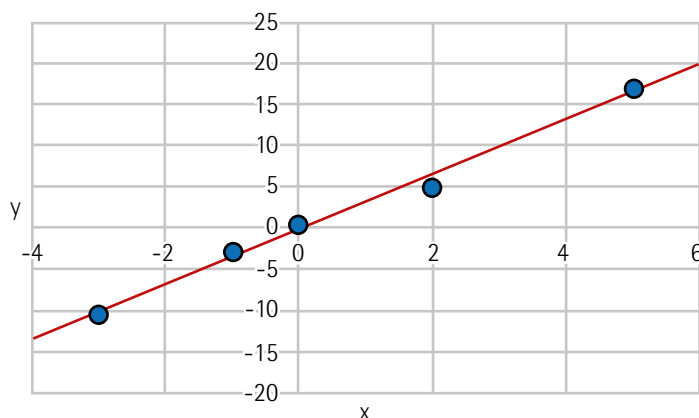


Figura 105 – Reta $y = 3,336x - 0,202$

Para mais uma exemplificação, na figura 106, temos os pares (x;y) observados da tabela 97 e a reta de equação $y = -5,917x - 0,945$, que representa a relação entre as variáveis em foco. O traçado da reta e o cálculo da sua equação foram feitos com ferramental do software Microsoft Excel.

Tabela 97 – Valores de duas variáveis, X e Y

X	-7,9	-4,8	-1,2	2,1	5,7	8,8
Y	38,1	25,2	7	-11,3	-31,7	-47

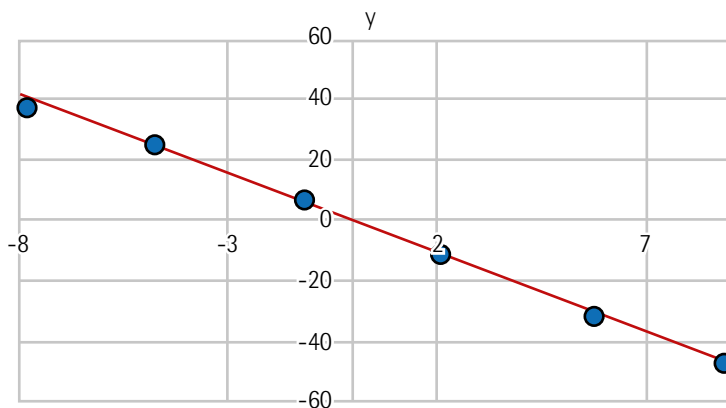


Figura 106 – Reta $y = -5,917x - 0,945$



Resumo

Iniciamos a unidade II estudando a estatística indutiva. Dissemos que, nessa parte da estatística, utilizamos técnicas que permitem que avaliemos características de uma população por meio do estudo de uma amostra. Destacamos que uma amostra ideal é aquela que apresenta toda a variabilidade da característica populacional em estudo e na proporção em que ocorre na população. Dissemos que, se usarmos toda a população na amostra, não teremos um processo aleatório, mas o oposto disso (nesse caso, teremos um processo certo ou determinado).

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 $(X_1, X_2, \dots, X_i, \dots, X_n)$. Com base nisso, definimos parâmetro, estimador e estimativa.

Um parâmetro é a quantidade da característica da população que estamos estudando. Na maioria das vezes, não conhecemos tal valor. Usamos, como veremos, uma estimativa para fazermos inferência. Por exemplo, μ é o parâmetro cujo valor fornece o peso médio das pessoas de uma cidade, e σ^2 é o parâmetro cujo valor fornece a variância do peso médio dessas pessoas.

Um estimador é o símbolo do resultado da amostra que é usado para estimar determinado parâmetro populacional. Veja que o estimador é uma variável aleatória que depende dos componentes $X_1, X_2, \dots, X_i, \dots, X_n$ da amostra. Por exemplo, \bar{X} é o símbolo do estimador usado para estimar o parâmetro peso médio das pessoas de uma cidade.

Vimos que um bom estimador deve ser não viciado (seu valor esperado é o valor do parâmetro em foco) e 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).

Observamos que a média amostral \bar{X} e a variância populacional S^2 são estimadores não viciados e consistentes para, respectivamente, a média populacional μ e a variância populacional σ^2 , conforme resumido a seguir.

Tabela 98

	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}

Dissemos que os estimadores (como o estimador \bar{X} para a média populacional μ e o estimador S^2 para a variância populacional σ^2 , mostrados na tabela anterior) são variáveis aleatórias. Logo, os estimadores seguem distribuições de probabilidade.

Uma estimativa é um valor "específico" de um estimador quando usamos valores "específicos" de determinada amostra. Se aplicarmos os valores de uma amostra coletada para o cálculo da média amostral observada, chegaremos ao que chamamos de \bar{X}_{obs} . Se aplicarmos os valores de uma amostra coletada para o cálculo da variância amostral observada, chegaremos ao que chamamos de S^2_{obs} .

Imagine que retiremos uma amostra aleatória simples de tamanho n de uma população cujos parâmetros são média μ e variância σ^2 . Segundo o teorema central do limite (TCL), para n suficientemente grande, a distribuição das médias amostrais devidamente padronizadas comporta-se como uma distribuição normal de média 0 e variância 1.

Adicionalmente, qualquer que seja a distribuição da variável aleatória relativa a determinada população, a distribuição das médias amostrais de amostras "grandes" segue modelo normal.

Prosseguimos com o estudo de intervalos de confiança, em que usamos estimadores intervalares, 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 $IC = [\sigma^2 - b; \sigma^2 + b]$.

Logo, associamos o intervalo de confiança ao coeficiente de confiança.

Passamos para os testes de hipóteses, aqueles em que não sabemos exatamente o que irá ocorrer, 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. Assim, estamos sujeitos a erros quando rejeitamos H_0 e quando aceitamos H_0 . Nesse contexto, os erros recebem nomes especiais.

- **Erro tipo I:** ocorre quando rejeitamos H_0 , e na realidade H_0 é verdadeira (V).
- **Erro tipo II:** ocorre quando não rejeitamos H_0 , e na realidade H_0 é falsa (F).

As decisões em que não cometemos erros são as seguintes.

- Rejeitamos H_0 , e H_0 é falsa.
- Não rejeitamos (aceitamos) H_0 , e H_0 é verdadeira.

No quadro a seguir, temos um resumo do que acabamos de ponderar.

Quadro 2 – Resumo dos casos de um teste de hipóteses

		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

Chamamos de α a probabilidade de ocorrência de erro tipo I e de β a probabilidade de ocorrência de erro tipo II. Logo, temos o que segue.

- $\alpha = P(\text{erro tipo I}) = P(\text{rejeitar } H_0 / H_0 \text{ é verdadeira})$
- $\beta = P(\text{erro tipo II}) = P(\text{não rejeitar } H_0 / H_0 \text{ é falsa})$

A probabilidade α é chamada de nível de significância do teste e está relacionada ao controle do erro tipo I. A probabilidade β é chamada de poder do teste e está relacionada ao controle do erro tipo II.

Mostramos as etapas necessárias para a realização de um teste de hipóteses, indicando a região crítica (RC).

Continuamos com os dois dos principais testes chamados de testes qui-quadrado.

- **Testes de aderência:** visam a testar se dado modelo probabilístico é adequado a determinado conjunto de dados.
- **Testes de independência:** visam a testar se há independência entre duas variáveis.

Finalmente, abordamos os temas correlação e regressão linear.

Vimos que o coeficiente de correlação, indicado por R , quantifica o grau de associação entre duas variáveis por meio de valores observados. O coeficiente de correlação R varia de -1 a 1 , sendo que:

- 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.

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



Exercícios

Questão 1. Imagine que a distribuição dos pesos das pessoas com mais de 18 anos que moram na cidade fictícia Vila Feliz obedeça a um modelo normal com média μ desconhecida e com variância σ^2 igual a $15,7 \text{ kg}^2$. Foi feita uma amostra aleatória de 40 dessas pessoas, o que forneceu média amostral observada \bar{X}_{obs} igual a 73 kg. Para essa situação, assinale a alternativa que apresenta corretamente a estimativa intervalar da média populacional μ com coeficiente de confiança de 90%.

- A) [61,8 kg; 79,3 kg]
- B) [60,0 kg; 80,0 kg]
- C) [65,0 kg; 75,0 kg]
- D) [68,8 kg; 75,2 kg]
- E) [71,9 kg; 74,1 kg]

Resposta correta: alternativa E.

Análise da questão

Vamos fazer um resumo dos dados fornecidos na questão.

- **Modelo de distribuição de probabilidades dos pesos:** normal.
- **Média populacional dos pesos:** parâmetro μ desconhecido.
- **Variância populacional dos pesos:** parâmetro $\sigma^2 = 15,7 \text{ kg}^2$.
- **Desvio padrão populacional dos pesos:** parâmetro $\sigma = \sqrt{\sigma^2} = \sqrt{15,7} = 3,96 \text{ kg}$.
- **Média amostral dos pesos:** estimador \bar{X} .
- **Tamanho da amostra:** $n = 40$.
- **Média amostral dos pesos observada na amostra:** estimativa $\bar{X}_{\text{obs}} = 73 \text{ kg}$.
- **Coeficiente de confiança da estimativa intervalar:** $c = 0,90$.

Como c vale 0,90, $c/2$ vale 0,45, pois $c/2 = 0,90/2 = 0,45$. Precisamos achar $z_{c/2}$ tal que tenhamos as configurações ilustradas a seguir.

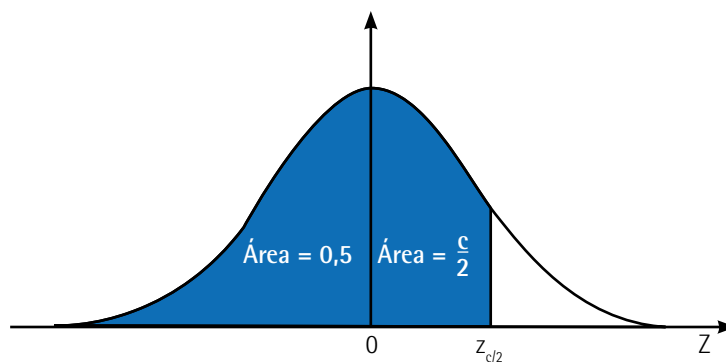


Figura 107

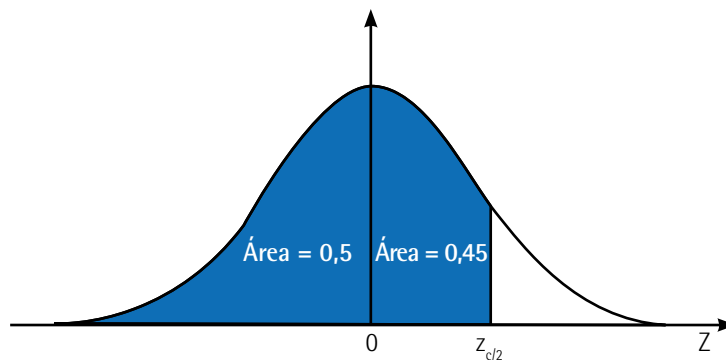


Figura 108

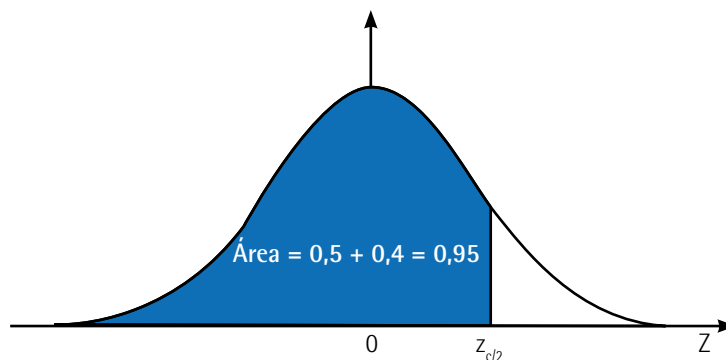


Figura 109

Precisamos encontrar, "dentro" da tabela normal reduzida, o valor 0,95. Vemos que, na tabela 39, o valor mais próximo de 0,95 é 0,9505, e ele corresponde a $z_{c/2} = 1,65$ (1,6 na horizontal e 0,05 na vertical).

Tabela 99

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

Agora, podemos calcular o intervalo de confiança para a média populacional dos pesos μ , com coeficiente de confiança $c = 0,9$ (90%), indicado por $IC(\mu; c)$, para o valor de média amostral observada $\bar{X}_{obs} = 73$ kg, com $z_{c/2} = 1,65$, $n = 40$ e $\sigma = 3,96$ kg.

$$IC(\mu; 0,90) = \left[\bar{X}_{obs} - z_{c/2} \cdot \frac{\sigma}{\sqrt{n}}; \bar{X}_{obs} + z_{c/2} \cdot \frac{\sigma}{\sqrt{n}} \right] = \left[73 - 1,65 \cdot \frac{3,96}{\sqrt{40}}; 73 + 1,65 \cdot \frac{3,96}{\sqrt{40}} \right]$$

$$IC(\mu; 0,90) = [73 - 1,03; 73 + 1,03] = [71,97; 74,03]$$

Com confiança de 90%, "acreditamos" que a média populacional dos pesos μ das pessoas com mais de 18 anos que moram na cidade fictícia Vila Feliz esteja entre 71,9 kg e 74,1 kg.

Questão 2. Imagine que Juca, gerente da loja Enxoval Perfeito, queira saber se a quantidade mensal de toalhas vendidas varia de acordo com a cor da toalha. Para isso, em determinado mês, Juca fez as observações apresentadas a seguir.

Tabela 100

Cor da toalha	Quantidade de peças vendidas
Branca	33
Azul	25
Verde	12
Lilás	8
Rosa	13
Cinza	2

Com base no exposto e nos seus conhecimentos, assinale a alternativa que mostra a conclusão correta para essas observações ao nível de significância de 5%.

- A) Ao nível de significância de 5%, chegamos à conclusão de que a quantidade de toalhas vendidas depende da cor da toalha, e o valor Q^2_{obs} é 42,2.
- B) Ao nível de significância de 5%, chegamos à conclusão de que a quantidade de toalhas vendidas não depende da cor da toalha, e o valor Q^2_{obs} é 42,2.
- C) Ao nível de significância de 5%, chegamos à conclusão de que a quantidade de toalhas vendidas depende da cor da toalha, e o valor Q^2_{obs} é 15,5.
- D) Ao nível de significância de 5%, chegamos à conclusão de que a quantidade de toalhas vendidas não depende da cor da toalha, e o valor Q^2_{obs} é 15,5.
- E) Ao nível de significância de 5%, chegamos à conclusão de que a quantidade de toalhas vendidas não depende da cor da toalha, e o valor Q^2_{obs} é 16,75.

Resposta correta: alternativa A.

Análise da questão

Vamos fazer um teste qui-quadrado de aderência e verificar seu resultado.

Primeiramente, estabelecemos as hipóteses a serem testadas.

- **Hipótese nula (Ho):** a quantidade de toalhas vendidas não varia com a cor da toalha.
- **Hipótese alternativa (Ha):** há pelo menos uma cor de toalha que é mais ou menos vendida do que as outras cores.

Na tabela, temos o total n de 93 peças vendidas, pois:

$$n = 33 + 25 + 12 + 8 + 13 + 2 = 93$$

Se a quantidade de toalhas vendidas não variar com a cor da toalha, deveremos ter, em cada uma das 6 cores listadas na tabela do enunciado, $1/6$ do total de 93 peças vendidas no mês, o que resulta em 15,5 unidades vendidas por cor. Ou seja, se H_0 for verdadeira, o valor esperado para cada frequência será igual a 15,5. Assim, obtemos a tabela a seguir.

Tabela 101

Cor da toalha	Categoria	Quantidade observada de peças vendidas	Quantidade esperada de peças vendidas
Branca	1	$O_1 = 33$	$E_1 = 15,5$
Azul	2	$O_2 = 25$	$E_2 = 15,5$
Verde	3	$O_3 = 12$	$E_3 = 15,5$
Lilás	4	$O_4 = 8$	$E_4 = 15,5$
Rosa	5	$O_5 = 13$	$E_5 = 15,5$
Cinza	6	$O_6 = 2$	$E_6 = 15,5$

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

$$Q_2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}$$

$$Q_{obs}^2 = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} + \frac{(O_3 - E_3)^2}{E_3} + \dots + \frac{(O_6 - E_6)^2}{E_6}$$

$$Q^2_{\text{obs}} = \frac{(33-15,5)^2}{15,5} + \frac{(25-15,5)^2}{15,5} + \frac{(12-15,5)^2}{15,5} + \frac{(8-15,5)^2}{15,5} + \frac{(13-15,5)^2}{15,5} + \frac{(2-15,5)^2}{15,5}$$

$$Q^2_{\text{obs}} = \frac{(17,5)^2}{15,5} + \frac{(9,5)^2}{15,5} + \frac{(-3,5)^2}{15,5} + \frac{(-7,5)^2}{15,5} + \frac{(-2,5)^2}{15,5} + \frac{(-13,5)^2}{15,5}$$

$$Q^2_{\text{obs}} = \frac{306,25}{15,5} + \frac{90,25}{15,5} + \frac{12,25}{15,5} + \frac{56,25}{15,5} + \frac{6,25}{15,5} + \frac{182,25}{15,5}$$

$$Q^2_{\text{obs}} = \frac{653,5}{15,5} = 42,2$$

Como temos 6 categorias, o número de graus de liberdade é 5 ($q = 5$), visto que ele é dado pelo número de categorias menos 1.

Procuramos, na tabela de distribuição qui-quadrado, na linha de graus de liberdade (g.l.) igual a 5 ($q = 5$), o valor 42,2.

Ocorre, como podemos ver na tabela a seguir, que o máximo valor na linha de graus de liberdade (g.l.) igual a 5 ($q = 5$) é 16,750, que resulta em $P(\chi^2_5 \geq 16,750) = 0,005$. Logo, temos a certeza de que $P(\chi^2_5 \geq 42,2) < 0,005$.

Tabela 102

g.l.	0,995	0,990	0,975	0,950	0,900	0,500	0,100	0,050	0,025	0,010	0,005
1	0,000	0,000	0,001	0,004	0,016	0,455	2,706	3,841	5,024	6,635	7,879
2	0,010	0,020	0,051	0,103	0,211	1,386	4,605	5,991	7,378	9,210	10,597
3	0,072	0,115	0,216	0,352	0,584	2,366	6,251	7,815	9,348	11,345	12,838
4	0,207	0,297	0,484	0,711	1,064	3,357	7,779	9,488	11,143	13,277	14,860
5	0,412	0,554	0,831	1,145	1,610	4,351	9,236	11,070	12,833	15,086	16,750
6	0,676	0,872	1,237	1,635	2,204	5,348	10,645	12,592	14,449	16,812	18,548

Adaptada de: <https://bit.ly/30XtiZ9>. Acesso em: 17 nov. 2021.

Assim, concluímos que, ao nível de significância de 5%, a quantidade de toalhas vendidas depende da cor da toalha.

REFERÊNCIAS

Audiovisuais

PROBABILIDADE: como calcular? 2019. 1 vídeo (6 min). Publicado pelo canal Ferreto Matemática. Disponível em: <https://bit.ly/3Duyouq>. Acesso em: 17 nov. 2021.

TEOREMA do limite central. 2018. 1 vídeo (14 min). Publicado pelo canal Professor Marcos Moreira. Disponível em: <https://bit.ly/3qZpQQV>. Acesso em: 17 nov. 2021.

Textuais

ALMEIDA, C. M. V. B.; DOI, C. M. *Explicando matemática*. Rio de Janeiro: Ciência Moderna, 2018.

ANTUNES, C. *Jogos para a estimulação das múltiplas inteligências*. Petrópolis: Vozes, 2002.

BERGER, R. I.; CASELLA, G. *Inferência estatística*. São Paulo: Cengage Learning, 2010.

BUSSAB, W. O.; MORETTIN, P. A. *Estatística básica*. 5. ed. São Paulo: Saraiva, 2007.

COSTA NETO, P. L. *Estatística*. 2. ed. São Paulo: Edgard Blucher, 2002.

DOI, C. M. O que significam os números de uma pesquisa eleitoral? *Jornal da Manhã*, 15 out. 2018. Disponível em: <https://bit.ly/3DrnWUC>. Acesso em: 17 nov. 2021.

KARL Friedrich Gauss. *Biblioteca Matemática*, [s.d.]. Disponível em: <https://bit.ly/30FHFRK>. Acesso em: 17 nov. 2021.

MAGALHÃES, M. N.; LIMA, A. C. P. *Noções de probabilidade e estatística*. 6. ed. São Paulo: Edusp, 2008.

NOTAÇÃO de somatório. *Khan Academy*, [s.d.]. Disponível em: <https://bit.ly/3CuCzFd>. Acesso em: 17 nov. 2021.

SPIEGEL, M. R. *Estatística*. São Paulo: McGraw-Hill do Brasil, 1974.



Handwriting practice lines consisting of 30 horizontal blue lines. Each line is preceded by a small blue vertical margin line on the left side.



A series of horizontal lines for writing, consisting of 30 evenly spaced lines across the page.



Handwriting practice lines consisting of 30 horizontal blue lines. Each line is preceded by a small blue dot, serving as a starting point for letter formation. The lines are evenly spaced and extend across the width of the page.



Handwriting practice lines consisting of 30 horizontal blue lines. Each line is preceded by a small blue dot, serving as a guide for letter height and placement.



Interativa

Informações:
www.sepi.unip.br ou 0800 010 9000