



Métodos quantitativos

Métodos quantitativos

Junior Francisco Dias

© 2016 por Editora e Distribuidora Educacional S.A.
Todos os direitos reservados. Nenhuma parte desta publicação poderá ser reproduzida ou transmitida de qualquer modo ou por qualquer outro meio, eletrônico ou mecânico, incluindo fotocópia, gravação ou qualquer outro tipo de sistema de armazenamento e transmissão de informação, sem prévia autorização, por escrito, da Editora e Distribuidora Educacional S.A.

Presidente

Rodrigo Galindo

Vice-Presidente Acadêmico de Graduação

Mário Ghio Júnior

Conselho Acadêmico

Dieter S. S. Paiva

Camila Cardoso Rotella

Emanuel Santana

Alberto S. Santana

Regina Cláudia da Silva Fiorin

Cristiane Lisandra Danna

Danielly Nunes Andrade Noé

Parecerista

Rogério Siqueira Chiacchio

Thiago Barroso Fonte Boa

Editoração

Emanuel Santana

Cristiane Lisandra Danna

André Augusto de Andrade Ramos

Daniel Roggeri Rosa

Adilson Braga Fontes

Diogo Ribeiro Garcia

eGTB Editora

Dados Internacionais de Catalogação na Publicação (CIP)

Dias, Junior Francisco
D541m Métodos quantitativos / Junior Francisco Dias. – Londrina
Editora e Distribuidora Educacional S.A., 2016.
240 p.

ISBN 978-85-8482-354-3

1. Funções. 2. Pesquisa quantitativa. 3. Estatística matemática. 4. Matemática aplicada. I. Título.

CDD 518

2016

Editora e Distribuidora Educacional S.A.
Avenida Paris, 675 – Parque Residencial João Piza
CEP: 86041-100 – Londrina – PR
e-mail: editora.educacional@kroton.com.br
Homepage: <http://www.kroton.com.br/>

Sumário

Unidade 1 Função afim e função quadrática	7
Seção 1.1 - Função	8
Seção 1.2 - Função afim	22
Seção 1.3 - Função quadrática	34
Seção 1.4 - Sinal, mínimo e máximo da função quadrática	45
Unidade 2 Estatística descritiva	55
Seção 2.1 - Amostragem	57
Seção 2.2 - Métodos tabulares e métodos gráficos	75
Seção 2.3 - Medidas de posição	94
Seção 2.4 - Medidas de dispersão	108
Unidade 3 Estatística inferencial (parte I)	123
Seção 3.1 - Noções de probabilidade	124
Seção 3.2 - Distribuição dos estimadores	137
Seção 3.3 - Testes de hipóteses para a média (σ^2 conhecido)	151
Seção 3.4 - Testes de hipóteses para a média (σ^2 desconhecido)	166
Unidade 4 Estatística inferencial (parte II)	181
Seção 4.1 - Correlação entre variáveis quantitativas	183
Seção 4.2 - Teste de significância	196
Seção 4.3 - Regressão linear	209
Seção 4.4 - Estudando resíduos	221

Palavras do autor

Caro aluno, seja bem-vindo!

Nesta unidade curricular estudaremos funções e noções de estatística. Utilizamos esses dois temas o tempo todo, mas nem sempre nos damos conta disso. Observe um exemplo simples com relação à função: no supermercado, ao levarmos os produtos ao caixa, o atendente passa o código de barras pelo leitor e o computador registra o preço do item. Nesse caso, o computador desempenha o papel de uma função, que recebe a informação de um código de barras e, como resposta, registra o preço do produto. Essa é a ideia básica de qualquer função, ou seja, dado certo elemento (que pode ser um objeto, um número, uma pessoa etc.), a função o relaciona a outro, podendo este ser tão diverso quanto o primeiro.

Exemplos como o anterior podem ser adaptados para mostrar a aplicação das funções em qualquer relação de comércio, mas não é somente nesse contexto que as funções são utilizadas. Ao andar de carro você já deve ter reparado a funcionalidade do velocímetro. A ação desse mecanismo também pode ser associada a uma função, pois ele recebe o sinal referente à frequência dos giros da roda do carro, transformando essa informação em registro de velocidade.

Com relação à estatística, também fazemos uso da mesma em nosso cotidiano com muita frequência. Ao fazer um levantamento da quantidade de pessoas que moram em uma região, estamos construindo uma estatística. Quando se realiza uma pesquisa eleitoral, ou ainda ao se comparar preços de itens de supermercado, estamos usando a estatística.

Para que seu estudo ocorra de modo organizado, este material didático foi dividido em 4 unidades de ensino, cada qual subdividida em 4 seções de autoestudo, totalizando 16 seções. A primeira unidade trata das funções afim e quadrática. A unidade 2 aborda a estatística descritiva. As unidades 3 e 4 abordam a estatística inferencial. Desejamos-lhe sucesso nesta empreitada!

Função afim e função quadrática

Convite ao estudo

Olá, aluno! Na Unidade 1 deste livro didático trataremos das funções afim e quadrática. Essas duas classes de funções são muito utilizadas não somente na Matemática, mas também na Física, na Economia, na Engenharia, na Administração etc. Na Física, por exemplo, a trajetória de um projétil pode ser descrita por uma função quadrática; função essa também utilizada na Engenharia para modelar a geometria de algumas estruturas, a exemplo da ponte Juscelino Kubitschek (Figura 1.1), em Brasília, cujos arcos lembram o gráfico dessa função. A afim, por sua vez, é utilizada, por exemplo, na modelagem de alguns problemas nas áreas econômicas e de gestão, em que a utilização de outro tipo de recurso tornaria o problema muito complexo para ser resolvido.

Para tornar o assunto desta unidade mais interessante, veja uma situação em que o emprego de funções pode facilitar a gestão de um negócio.

Imagine que você seja o dono de uma empresa que fabrica bonés. Para melhor analisar os custos e lucros você decidiu estudar esses números utilizando funções e gráficos matemáticos, buscando uma melhor organização e maiores lucros, bem como um planejamento de expansão da empresa.

No decorrer desta unidade você será convidado a desempenhar o papel de dono da empresa e resolver os desafios inerentes à administração dela, mas, para tanto, precisará relacionar diversas grandezas presentes no dia a dia, bem como interpretar números e gráficos.

Seção 1.1

Função

Diálogo aberto

Para gerir melhor sua empresa, você deve analisar os custos, as receitas e o lucro, pois sem lucro a empresa não pode ser mantida.

O custo da produção dos bonés é contabilizado a partir de diversos gastos, como matéria-prima, mão de obra, energia elétrica, entre outros. Com isso, há uma relação direta entre o custo e a quantidade de bonés produzida, ou seja, quanto mais bonés produzidos, maior o custo de produção.

Além do custo, outro item importante na gestão da empresa é a receita, que é o valor recebido com a comercialização dos bonés. Vamos imaginar que o preço de venda dos bonés seja de R\$ 30,00 por unidade. Qual a receita obtida com a venda de 10 unidades? Com um cálculo simples podemos notar que a receita é de R\$ 300,00 ($10 \cdot \text{R}\$ 30,00 = \text{R}\$ 300,00$). Mas, e se quiséssemos escrever isso em uma planilha, de modo que em uma coluna tivéssemos a quantidade vendida e, em outra, a receita correspondente, como podemos agilizar esse cálculo para diversas quantidades comercializadas? Pense um pouco.

Por fim, o lucro é a diferença entre a receita e o custo de produção. Vamos supor que, a partir de balanços financeiros de anos anteriores, chegou-se à conclusão de que, mensalmente, o custo com a produção é composto por um custo fixo de R\$ 9000,00 mais um custo variável de R\$ 20,00 por boné. Nesse caso, com a produção e venda de 750 bonés em um mês, tem-se lucro ou prejuízo? E se forem produzidos e comercializados 1200 bonés?

Para responder a essas e outras perguntas, você deve empregar conceitos de funções. Vamos lá?

Figura 1.1 | Ponte Juscelino Kubitschek, em Brasília



Fonte: <https://commons.wikimedia.org/wiki/File:Ponte_JK_-_Bras%C3%ADlia.jpg>. Acesso em: 19 out. 2015.

Não pode faltar!

Conjuntos

Para compreender a ideia de função, primeiramente é necessário relembrar alguns conceitos, geralmente trabalhados no ensino médio, entre eles, **conjunto**, **elemento** e **pertinência**. Para uma melhor compreensão, observe os seguintes exemplos:

- Conjunto das vogais: $A = \{a, e, i, o, u\}$.
- Conjunto dos planetas do sistema solar: $B = \{\text{Mercúrio, Vênus, Terra, ..., Netuno}\}$.
- Conjunto dos meses do ano: $C = \{\text{janeiro, fevereiro, ..., dezembro}\}$.

$$e \cong 2,71828$$

$$r \cong 3,14159$$



Lembre-se

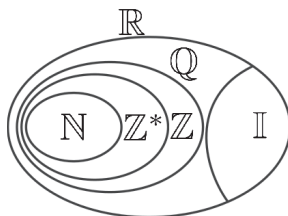
No primeiro exemplo, A é o símbolo utilizado para representar o **conjunto das vogais**; cada vogal é um **elemento do conjunto**. Podemos dizer inclusive que a vogal u **pertence ao conjunto** A , afirmação que pode ser expressa sinteticamente por $u \in A$ (lê-se: u pertence a A). A consoante m não pertence ao conjunto A e escrevemos $m \notin A$ (lê-se: m não pertence a A). Os exemplos mais conhecidos de conjuntos são:

- Números naturais: $N = \{1, 2, 3, 4, 5, 6, \dots, 99, 100, 101, \dots\}$;
- Números inteiros: $Z = \{\dots, -7, -6, \dots, -1, 0, 1, 2, \dots, 5, 6, 7, \dots\}$;
- Números inteiros, sem o zero: $Z^* = \{\dots, -7, -6, \dots, -1, 1, 2, \dots, 5, 6, 7, \dots\}$;
- Números racionais: $Q = \left\{ \frac{a}{b} \mid a \in Z \text{ e } b \in Z^* \right\}$ (lê-se: Q é o conjunto dos números $\frac{a}{b}$ tais que a pertence a Z e b pertence a Z^*);
- Números reais: $R = \{\dots, -50, \dots, -\frac{37}{e}, \dots, \pi, \dots, -2, \dots, -\frac{1}{2}, \dots, 0, \dots, \frac{1}{2}, \dots, 1, \dots, \frac{10}{9}, \dots, 2, \dots, \pi, \dots, 4, \dots, 7\pi, \dots\}$;

• Números irracionais: $I = \{x \mid x \in \mathbb{R} \text{ e } x \notin \mathbb{Q}\}$ (lê-se: I é o conjunto dos números x tais que x pertence a \mathbb{R} e x não pertence a \mathbb{Q}).

Em relação aos conjuntos numéricos, temos as seguintes inclusões (Figura 1.2): (lê-se: \mathbb{N} está contido em \mathbb{Z}^* ; $\mathbb{Z}^* \subset \mathbb{Z}$; $\mathbb{Z} \subset \mathbb{Q}$; $\mathbb{Q} \subset \mathbb{R}$; $\mathbb{I} \subset \mathbb{R}$).

Figura 1.2 | Conjuntos numéricos



Fonte: O autor (2015).

Ainda sobre esses conjuntos numéricos, nenhum elemento de \mathbb{Q} pertence a \mathbb{I} , e nenhum elemento de \mathbb{I} pertence a \mathbb{Q} , ou seja, na **interseção** desses dois conjuntos, não há elementos, e indicamos isso por $\mathbb{Q} \cap \mathbb{I} = \emptyset$, em que \emptyset é o **conjunto vazio**. Por fim, ao **reunir** os dois conjuntos, \mathbb{Q} e \mathbb{I} , obtemos o conjunto dos números reais, ou seja, $\mathbb{Q} \cup \mathbb{I} = \mathbb{R}$; ambos são **subconjuntos** de \mathbb{R} .



Pesquise mais

Para mais detalhes sobre a teoria de conjuntos, acesse o link disponível em: <http://www.uel.br/projetos/matessencial/medio/conjuntos/conjunto.htm>. Acesso em: 20 out. 2015. Elaborado pelo professor Ulysses Sodré, da Universidade Estadual de Londrina, esse site possui alguns dos fundamentos da teoria de conjuntos, notações mais utilizadas e exemplos numéricos com linguagem bastante acessível. Vale a pena conferir!

Produto cartesiano

Outro conceito importante para o entendimento de uma função é o de produto cartesiano.



Assimile

Dados dois conjuntos A e B , o **produto cartesiano** de A por B é o conjunto dos pares ordenados (a,b) tais que $a \in A$ e $b \in B$.

$$A \times B = \{(a,b) \mid a \in A \text{ e } b \in B\}$$



Produto cartesiano de A por B .

Veja um exemplo numérico de produto cartesiano:



Exemplificando

Considerando os conjuntos $A = \{0, 2, 3\}$ e $B = \{-2, 0, 3, 7\}$, escreva o produto cartesiano de A por B.

Resolução:

$$A \times B = \{(a, b) \mid a \in A \text{ e } b \in B\}$$

Para $a = 0$, temos: $(0, -2)$; $(0, 0)$; $(0, 3)$; $(0, 7)$;

Para $a = 2$, temos: $(2, -2)$; $(2, 0)$; $(2, 3)$; $(2, 7)$;

Para $a = 3$, temos: $(3, -2)$; $(3, 0)$; $(3, 3)$; $(3, 7)$.

Logo,

$$A \times B = \{(0, -2), (0, 0), (0, 3), (0, 7), (2, -2), (2, 0), (2, 3), (2, 7), (3, -2), (3, 0), (3, 3), (3, 7)\}.$$

Relação

Outro conceito muito importante para o entendimento de uma função é o de relação.



Assimile

Dados dois conjuntos A e B, uma relação R de A em B é qualquer subconjunto de $A \times B$, ou seja, $R \subset (A \times B)$.



Exemplificando

Considere os conjuntos $A = \{0, 2, 3\}$ e $B = \{-2, 0, 3, 7\}$ e escreva os elementos da relação R descrita pela equação $y = x^2 - 2x$, em que $x \in A$ e $y \in B$.

Resolução:

Para facilitar os cálculos dos elementos de R, vamos utilizar um quadro, como a seguir:

Elementos de A	Elementos de B	Elementos de R
x	$y = x^2 - 2x$	(x, y)
0	$y = x^2 - 2x = 0^2 - 2 \cdot 0 = 0$	$(0, 0)$
2	$y = x^2 - 2x = 2^2 - 2 \cdot 2 = 0$	$(2, 0)$
3	$y = x^2 - 2x = 3^2 - 2 \cdot 3 = 3$	$(3, 3)$

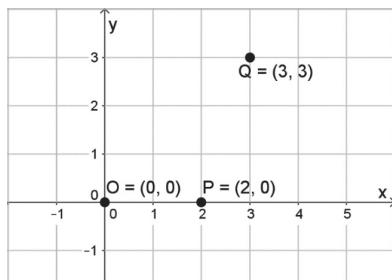
Portanto, $R = \{(0,0), (2,0), (3,3)\}$. Compare os elementos de R com os de $A \# B$ e veja que $R \subset (A \times B)$.

Na relação $R = \{(0,0), (2,0), (3,3)\}$ dizemos que o valor: $0 \in A$ está associado ao valor $0 \in B$; $2 \in A$ está associado ao valor $0 \in B$; $3 \in A$ está associado ao valor $3 \in B$.

Plano cartesiano

Uma relação R pode ser visualizada graficamente em um diagrama denominado **plano cartesiano**. Veja, por exemplo, a representação gráfica da relação $R = \{(0,0), (2,0), (3,3)\}$ no plano cartesiano da Figura 1.3.

Figura 1.3 | Representação gráfica



Fonte: O autor (2015).

Observe que a representação de R corresponde a três pontos no plano. Em relação ao ponto $p = (2,0)$, o par ordenado $(2,0)$ corresponde a suas **coordenadas**. O primeiro valor, 2, é denominado **abscissa** de P e o segundo, 0, a **ordenada**. O valor $x = 2$ corresponde à distância a que o ponto P se encontra do eixo vertical, eixo y (ou **eixo das ordenadas**), e o valor $y = 0$ à distância a que o ponto se encontra do eixo horizontal, eixo x (ou **eixo das abscissas**). O ponto de coordenadas $(0,0)$ é denominado origem.

Em um plano cartesiano, as:

- abscissas são: positivas se estiverem à direita da origem; negativas se estiverem à esquerda da origem;
- ordenadas são: positivas se estiverem acima da origem; negativas se estiverem abaixo da origem.



Veja mais detalhes sobre a construção de um plano cartesiano e a localização de pontos a partir de suas coordenadas no link disponível em: <https://pt.khanacademy.org/math/algebra/introduction-to-algebra/overview_hist_alg/v/descartes-and-cartesian-coordinates>. Acesso em: 22 out. 2015.

Função

A partir dos conceitos aprendidos até agora, podemos definir função.



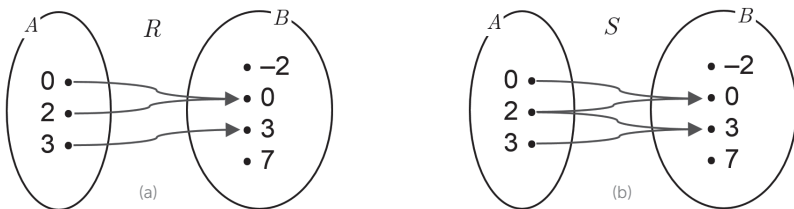
Dados dois conjuntos A e B , uma **função** f de A em B , denotada $f: A \rightarrow B$, é uma relação $f \subset (A \times B)$ tal que para cada $a \in A$ está associado um único $b \in B$.

O conjunto A é o domínio de f (denotado por $D(f)$) e o conjunto B é o contradomínio de f (denotado por $CD(f)$). Convenciona-se utilizar o símbolo x para representar um elemento qualquer de A e y para representar um elemento qualquer de B . Além disso, se x está relacionado a y por meio da função f , escrevemos $y=f(x)$ para simbolizar essa associação, e o par ordenado correspondente será (x,y) ou $(x,f(x))$.

$Im(f) = \{y \in B | y=f(x) \text{ e } x \in A\}$ é denominado conjunto imagem de f . Além disso, se $y=f(x)$, então y é a imagem de x obtida por meio de f .

Para compreender melhor, considere as relações $R = \{(0,0), (2,0), (3,3)\}$ e $S = \{(0,0), (2,0), (3,3), (2,3)\}$ de $A = \{0,2,3\}$ em $B = \{-2,0,3,7\}$. Temos que R é uma função e S não é uma função, pois o valor $2 \in A$ está associado por meio de S a dois elementos de B , a saber, 0 e 3 . Essa constatação pode ser feita mais facilmente por meio de um diagrama de Venn, como os apresentados na Figura 1.4.

Figura 1.4 | Diagrama de Venn: (a) da relação R ; (b) da relação S



Fonte: O autor (2015).

Observe que no caso da relação S há duas setas partindo do número $2 \in A$, uma relacionando-o a 0 e outra relacionando-o a 3, e isso não se encaixa na definição de função.



Exemplificando

Considerando os conjuntos $A = \{-2, -1, 0, 1, 3\}$ e $B = \{0, 1, 2, 4, 3, 9\}$ e a função $f: A \rightarrow B$, de modo que $y = f(x) = x^2$, identifique o domínio, contradomínio e a imagem de f .

Resolução:

Como visto anteriormente, A é o domínio de f e B é o contradomínio, logo:

$D(f) = A = \{-2, -1, 0, 1, 3\}$; $CD(f) = B = \{0, 1, 2, 4, 3, 9\}$;

Para escrevermos o conjunto imagem precisamos determinar os elementos (x, y) pertencentes à relação (vide quadro ao lado). Logo, $Im(f) = \{0, 1, 4, 9\}$.

x	$y = x^2$	(x, y)
-2	$y = (-2)^2 = 4$	$(-2, 4)$
-1	$y = (-1)^2 = 1$	$(-1, 1)$
0	$y = 0^2 = 0$	$(0, 0)$
1	$y = 1^2 = 1$	$(1, 1)$
3	$y = 3^2 = 9$	$(3, 9)$



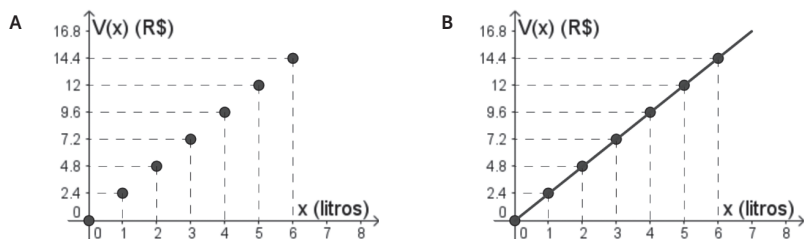
Faça você mesmo

Represente graficamente e elabore um diagrama de Venn para a relação $f: A \rightarrow B$ com $A = \{-2, -1, 0, 1, 3\}$, $B = \{0, 1, 2, 4, 3, 9\}$ e $y = f(x) = x^2$.

Lei de formação e gráfico de uma função

No exemplo anterior, $y = f(x) = x^2$ é o que denominamos **lei de formação** (ou regra de associação) da função $f: A \rightarrow B$. Em alguns problemas conhecemos a lei de formação da função e em outros não. Quando não a conhecemos, em alguns casos, é possível determiná-la a partir de informações do problema. Veja um exemplo: considere que em determinado posto de combustíveis o preço do etanol seja de R\$ 2,40 o litro. Qual é a lei de formação da função que relaciona a quantidade de etanol abastecida (x) e o valor a pagar ($v(x)$)?

Figura 1.5 | Representação gráfica de $v = 2,40 \cdot x$



Fonte: O autor (2015).

A primeira investigação da lei de formação pode ser feita por meio da Tabela 1.1. Observe que, para encontrarmos o valor a ser pago por determinada quantidade de combustível, multiplicamos essa quantidade pelo preço de um litro. Logo, ao adquirirmos x litros de etanol, devemos pagar $2,40 \cdot x$ reais. Portanto, a função $v: A \rightarrow B$, em que A é o conjunto das quantidades de etanol e B é o conjunto dos possíveis preços, possui lei de formação $v(x) = 2,40 \cdot x$.

Os dados apresentados na Tabela 1.1, com o acréscimo de alguns valores, podem ser representados de forma gráfica, como na Figura 1.5 (a). Observe que todos os pontos estão alinhados e, se utilizássemos inúmeros valores intermediários para x ou ainda, se considerássemos $x \in \mathbb{R}$, teríamos uma linha reta, como na Figura 1.5 (b). Para fazer essa constatação de forma mais dinâmica, acesse o link disponível em: <<http://tube.geogebra.org/m/1886475>> acesso em: 23 out. 2015. A linha reta da Figura 1.5 (b) é o que denominamos gráfico da função v . Mais formalmente, o gráfico de uma função $f: A \rightarrow B$ é o conjunto $G(f) = \{ (x, y) \mid x \in A, y \in B \text{ e } y = f(x) \}$.



Exemplificando

Uma empresa de táxi cobra pela corrida um valor fixo de R\$ 4,85 (bandeirada) mais um valor variável de R\$ 2,90 por quilômetro rodado. Construa a lei de formação da função que retorna o preço $f(x)$ para uma distância x percorrida. Além disso, escreva o domínio, a imagem e esboce o gráfico de f . Calcule também o valor a ser pago por uma corrida de 6 km.

Resolução:

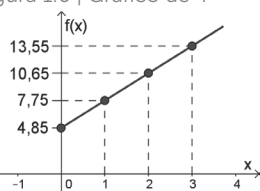
A corrida é composta por um valor fixo de R\$ 4,85 e um valor variável de R\$ 2,90 por quilômetro rodado; matematicamente, essas informações podem ser traduzidas da seguinte forma: $f(x) = 4,85 + 2,90 \cdot x$, em que x é a distância percorrida e $f(x)$ é o preço. Essa é a lei de formação.

A função $f: A \rightarrow B$ é tal que A (domínio) é o conjunto com todos os valores possíveis e adequados ao problema, que pode ser qualquer quantidade maior ou igual a zero, ou seja, $x \geq 0$. Logo, $A = \{x \in \mathbb{R} \mid x \geq 0\}$. A imagem de f é o conjunto $\text{Im}(f) \subset B$ que possui todos os possíveis preços a serem pagos, cujo mínimo é R\$ 4,85; não há valor máximo. Logo, $\text{Im}(f) = \{x \in \mathbb{R} \mid x \geq 4,85\}$.

Para esboçar o gráfico de f , montamos uma tabela com alguns valores de $(x, f(x))$ e esboçamos os pares ordenados em um plano cartesiano (Figura 1.6).

Distância (km)	Preço (R\$)
0	$f(0) = 4,85 + 2,90 \cdot 0 = 4,85$
1	$f(1) = 4,85 + 2,90 \cdot 1 = 7,75$
2	$f(2) = 4,85 + 2,90 \cdot 2 = 10,65$
3	$f(3) = 4,85 + 2,90 \cdot 3 = 13,55$

Figura 1.6 | Gráfico de f



Fonte: O autor (2015).

Por fim, o valor a ser pago por uma corrida de 6 km é $f(6) = 4,85 + 2,90 \cdot 6 = 22,25 \rightarrow$ R\$ 22,25



Pesquise mais

Para esclarecer possíveis dúvidas, leia mais sobre relações, funções e seus gráficos em : <http://www.uel.br/projetos/matessencial/medio/funcoes/funcoes.htm>. Acesso em: 23 out. 2015.

Sem medo de errar!

Vamos retomar o problema proposto no início desta seção. Um dos questionamentos feitos foi: como agilizar os cálculos das receitas para diversas quantidades de bonés comercializados? Como fazer isso em uma planilha, por exemplo?

Lembre-se de que o preço de venda de cada boné é R\$ 30,00.

- Se nenhum boné for vendido, não há receita

$$\left(\frac{0}{\text{Quantidade de bonés}} \cdot \frac{\text{R\$ } 30,00}{\text{Preço por boné}} = \frac{\text{R\$ } 00,00}{\text{Receita}} \right);$$

- Se 1 boné for vendido, a receita é R\$ 30,00

$$\left(\frac{1}{\text{Quantidade de bonés}} \cdot \frac{\text{R\$ } 30,00}{\text{Preço por boné}} = \frac{\text{R\$ } 30,00}{\text{Receita}} \right);$$

- Se 2 bonés forem vendidos, a receita é R\$ 60,00

$$\left(\frac{2}{\text{Quantidade de bonés}} \cdot \frac{\text{R\$ } 30,00}{\text{Preço por boné}} = \frac{\text{R\$ } 60,00}{\text{Receita}} \right);$$

- Se x bonés forem vendidos, a receita é $x \cdot \text{R\$}30,00 = \text{R\$}30,00 \cdot x$. Portanto, a função receita é $R(x) = 30 \cdot x$. Esse cálculo pode ser agilizado em uma planilha, como na Figura 1.7.

Figura 1.7 | Planilha de cálculo da receita de bonés vendidos a R\$ 30,00 por unidade

	A	B
1	Quantidade de bonés	Receita
2	0	=30*A2
3	1	
4	2	
5	3	
6		

(a)

	A	B
1	Quantidade de bonés	Receita
2	0	0
3	1	
4	2	
5	3	
6		

(b)

	A	B
1	Quantidade de bonés	Receita
2	0	0
3	1	30
4	2	60
5	3	90
6		

(c)

Fonte: O autor (2015).

Observe que na Figura 1.7 os valores de x estão inseridos na coluna A; os valores de $y=R(x)$ são calculados na coluna B, sendo cada um

calculado pela função R . A sequência (a), (b) e (c) da Figura 1.7 apenas ilustra como agilizar os cálculos.

Outro questionamento feito foi em relação ao lucro, mas, para isso, precisamos determinar a função custo, traduzindo matematicamente a informação: "o custo com a produção é composto por um custo fixo de R\$ 9000,00 mais um custo variável de R\$ 20,00 por boné". Observe que esse problema é semelhante ao exemplo da corrida de táxi (trabalhado nesta seção). Por analogia, podemos escrever a função custo da seguinte forma: $C(x) = 9000 + 20 \cdot x$, em que x é a quantidade de bonés produzida. Como o lucro/prejuízo é a diferença entre a receita e o custo, podemos analisar o lucro/prejuízo na produção e venda de 750 ou 1200 bonés em um mês:

- 750 bonés: $R(x) = 30 \cdot x \rightarrow R(750) = 30 \cdot 750 = 22500$ (receita); $C(x) = 9000 + 20 \cdot x \rightarrow C(750) = 9000 + 20 \cdot 750 = 24000$ (custo); lucro = receita – custo = $22500 - 24000 = -1500$.

- 1200 bonés: $R(x) = 30 \cdot x \rightarrow R(1200) = 30 \cdot 1200 = 36000$ (receita); $C(x) = 9000 + 20 \cdot x \rightarrow C(1200) = 9000 + 20 \cdot 1200 = 33000$ (custo); lucro = receita – custo = $36000 - 33000 = 3000$.

Portanto, ao produzir e vender 750 bonés, o prejuízo é de R\$ 1500,00; no caso de 1200 bonés, o lucro é de R\$ 3000,00.



Pesquise mais

Veja mais detalhes de como utilizar funções e agilizar cálculos no Excel nos links a seguir:

- Visão geral de fórmulas no Excel. Disponível em: <<https://support.office.com/pt-br/article/Vis%C3%A3o-geral-de-f%C3%B3rmulas-no-Excel-ecfdc708-9162-49e8-b993-c311f47ca173?ui=pt-BR&rs=pt-BR&ad=BR>>. Acesso em: 26 out. 2015.

- Preencher dados automaticamente nas células da planilha. Disponível em: <<https://support.office.com/pt-br/article/Preencher-dados-automaticamente-nas-c%C3%A9lulas-da-planilha-74e31bdd-d993-45da-aa82-35a236c5b5db?omkt=pt-BR&ui=pt-BR&rs=pt-BR&ad=BR>>. Acesso em: 26 out. 2015.

Avançando na prática

Pratique mais

Instrução

Desafiamos você a praticar o que aprendeu transferindo seus conhecimentos para novas situações que pode encontrar no ambiente de trabalho. Realize as atividades e depois as compare com as de seus colegas.

Atualizando preços

1. Competências de Fundamentos de Área

Conhecer os conceitos matemáticos básicos e proporcionar o desenvolvimento do raciocínio lógico e quantitativo.

2. Objetivos de aprendizagem

Aplicar o conceito de função na atualização de preços.

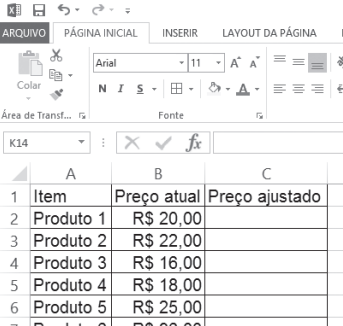
3. Conteúdos relacionados

Função; Lei de formação de uma função.

4. Descrição da SP

Em determinado supermercado será realizada uma remarcação de preços para embutir o aumento da energia elétrica no preço de venda. Após alguns cálculos, foi decidido que cada produto deveria sofrer um aumento de 2% e, para agilizar o trabalho, os novos preços seriam calculados com a ajuda de uma planilha. Veja na Figura 1.8 alguns preços a serem ajustados.

Figura 1.8 | Tabela de preços



The screenshot shows a spreadsheet application interface. At the top, there are menu options: ARQUIVO, PÁGINA INICIAL, INSERIR, LAYOUT DA PÁGINA, and F. Below the menus, there are icons for file operations and text formatting. The main area displays a table with the following data:

	A	B	C
1	Item	Preço atual	Preço ajustado
2	Produto 1	R\$ 20,00	
3	Produto 2	R\$ 22,00	
4	Produto 3	R\$ 16,00	
5	Produto 4	R\$ 18,00	
6	Produto 5	R\$ 25,00	

Fonte: O autor (2015).

Qual função deve ser inserida na célula C2 para que o preço da célula B2 seja reajustado em 2%? Qual o preço ajustado de cada produto?

5. Resolução da SP

Suponha que o preço atual de um produto seja x e que o preço ajustado seja $P(x)$. O preço atual corresponde a 100%; já o preço ajustado (+2%) corresponde a 102%. Logo, por regra de três:

$$\frac{x}{P(x)} = \frac{100}{102} \Rightarrow 102 \cdot x = 100 \cdot P(x) \Rightarrow P(x) = \frac{102 \cdot x}{100} = 1,02x$$

Ao calcular a função $P(x)$ para determinado preço, ela o reajusta em 2%. Adaptando a função para a planilha, temos que, na célula C2, devemos inserir a função $=1,02*B2$. Para os preços apresentados na Figura 1.8, temos:

Item	Preço atual	Preço ajustado
Produto 1	R\$ 20,00	$P(20,00) = 1,02 \cdot 20,00 = 20,40 \rightarrow$ R\$ 20,40
Produto 2	R\$ 22,00	$P(22,00) = 1,02 \cdot 22,00 = 22,44 \rightarrow$ R\$ 22,44
Produto 3	R\$ 16,00	$P(16,00) = 1,02 \cdot 16,00 = 16,32 \rightarrow$ R\$ 16,32
Produto 4	R\$ 18,00	$P(18,00) = 1,02 \cdot 18,00 = 18,36 \rightarrow$ R\$ 18,36
Produto 5	R\$ 25,00	$P(25,00) = 1,02 \cdot 25,00 = 25,50 \rightarrow$ R\$ 25,50



Lembre-se

Uma regra de três pode ser utilizada quando temos duas grandezas proporcionais, sendo que de uma delas conhecemos dois valores e, da outra, um valor. A regra de três é utilizada para determinar o quarto valor. Veja um breve resumo sobre esse assunto em: <http://educacao.globo.com/matematica/assunto/matematica-basica/regra-de-tres.html>. Acesso em: 27 out. 2015.

Faça valer a pena

1. Os conjuntos numéricos são de grande importância para a matemática, principalmente no estudo das funções. Os tipos mais utilizados são: números naturais (N); número inteiros (Z); números inteiros, exceto o zero (Z*); números racionais (Q); números irracionais (I); números reais (R).

Sobre os conjuntos numéricos e seus elementos, é correto afirmar que:

- a) $-1 \in \mathbb{N}$.
- b) $2 \in \mathbb{I}$.
- c) $\sqrt{2} \in \mathbb{R}$.
- d) $0 \in \mathbb{Q}$.
- e) $0 \in \mathbb{Z}^*$.

2. A reunião do conjunto A com o conjunto B é definida como o conjunto $C = \{x \mid x \in A \text{ ou } x \in B\}$ e a simbolizamos por $C = A \cup B$.

Sendo $A = \{1,2,3,4,6\}$ e $B = \{0,2,4,5,8\}$, assinale a alternativa que contém o conjunto $A \cup B$:

- a) $\{0,1,2,3,4,5,6,8\}$.
- b) $\{1,2,3,4,6\}$.
- c) $\{0,2,4,5,8\}$.
- d) $\{0,1,3,4,5,8\}$.
- e) $\{2,4\}$.

3) O produto cartesiano de A por B é o conjunto dos pares ordenados (a,b) tais que $a \in A$ e $b \in B$.

De acordo com o trecho anterior, assinale a alternativa que contém o produto cartesiano de $A = \{1,2,5\}$ por $B = \{3,4,6\}$:

- a) $\{(3,1),(4,1),(6,1),(3,2),(4,2),(6,2),(3,5),(4,5),(6,5)\}$.
- b) $\{(1,3),(1,4),(1,6)\}$.
- c) $\{(1,3),(1,4),(1,6),(2,3),(2,4),(2,6),(5,3),(5,4),(5,6)\}$.
- d) $\{(2,3),(2,4),(2,6)\}$.
- e) $\{(5,3),(5,4),(5,6)\}$.

Seção 1.2

Função afim

Diálogo aberto

Você se lembra de que na seção anterior estudou o lucro e a receita da sua fábrica de bonés? E que para fazer isso foi necessário relembrar alguns conjuntos numéricos, compreender a ideia de produto cartesiano, estudar as relações (que são subconjuntos dos produtos cartesianos) e as funções (que são casos específicos de relações), além de representar esses conjuntos graficamente no plano cartesiano e no diagrama de Venn?

Pois bem, tudo isso abriu caminho para outras possibilidades. Imagine que você precise construir uma apresentação contendo um estudo sobre as finanças da empresa, que será usada para convencer seu sócio a aumentar o investimento na fábrica e expandir o negócio. Um gráfico mostrando os possíveis lucros com o aumento da produção poderia ser interessante e deixá-lo empolgado. Além disso, você poderia incrementar a apresentação com informações detalhadas sobre os lucros (ou prejuízos) e mostrar a ele que você entende do assunto. Quanto mais informação, maior o poder de convencimento, concorda?

Pense um pouco: Será possível determinar uma função que relacione a quantidade produzida e comercializada com o lucro? Será que independentemente da quantidade produzida e comercializada há lucro ou para determinadas quantidades há prejuízo? A partir de que quantidade há lucro? Se aumentarmos a produção em 200 bonés ao mês nos próximos três meses, indo dos atuais 600 para 1200, quanto lucro teremos no trimestre? Essas são algumas das perguntas cujas respostas poderiam estar em sua apresentação. Entretanto, para realizar tudo isso, temos que estudar mais a fundo as funções e, mais especificamente, a função afim e suas propriedades. Vamos lá?

Não pode faltar!

A função afim é um tipo específico de função polinomial e, por este motivo, é também denominada função do 1º grau ou, ainda, função polinomial de grau 1. Mais rigorosamente definimos:



Assimile

Uma função afim é uma função $f: \mathbb{R} \rightarrow \mathbb{R}$ cuja lei de formação é $f(x) = ax + b$, em que $a \in \mathbb{R}$, não nulo, é denominado coeficiente angular e $b \in \mathbb{R}$ é denominado coeficiente linear.

O domínio e contradomínio de uma função afim podem ser intervalos de números reais.



Pesquise mais

Saiba mais sobre intervalos de números reais acessando o site disponível em: <http://www.casadasciencias.org/dmdocuments/intervalo10-11.pdf>. Acesso em: 2 nov. 2015.

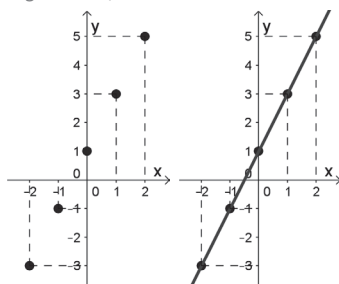
Uma característica interessante da função afim é a forma do seu gráfico, que é uma reta (IEZZI et al., 1977, p. 96-A). Veja um exemplo.



Exemplificando

Dada a função afim $f(x) = 2x + 1$, escreva os pares ordenados (x, y) tais que $x \in A = \{-2, -1, 0, 1, 2\}$ e $y = f(x)$. Em seguida, esboce o gráfico de f .

Figura 1.9 | Gráfico de $f(x) = 2x + 1$



Fonte: O autor (2015).

Resolução: Para escrever os pares ordenados solicitados podemos fazer uso do quadro a seguir:

x	$y = f(x) = 2x + 1$	(x, y)
-2	$y = f(-2) = 2 \cdot (-2) + 1 = -3$	$(-2, -3)$
-1	$y = f(-1) = 2 \cdot (-1) + 1 = -1$	$(-1, -1)$
0	$y = f(0) = 2 \cdot 0 + 1 = 1$	$(0, 1)$
1	$y = f(1) = 2 \cdot 1 + 1 = 3$	$(1, 3)$
2	$y = f(2) = 2 \cdot 2 + 1 = 5$	$(2, 5)$

Para esboçar o gráfico da função, primeiramente marcamos os pontos determinados no quadro e depois traçamos uma reta passando por eles, como mostra a Figura 1.9.

Para uma visualização mais dinâmica da construção do gráfico dessa função, acesse: <http://tube.geogebra.org/m/1980917>. Acesso em: 4 nov. 2015.

Da geometria, sabe-se que para determinar uma reta bastam dois pontos. Logo, para esboçar o gráfico do exemplo anterior (e o de qualquer função afim) basta determinarmos dois pares ordenados, e não mais que isso.



Faça você mesmo

1) Esboce o gráfico da função $f(x) = 3x - 2$.

Assim como podemos esboçar o gráfico de uma função afim a partir de sua lei de formação, também é possível determinar sua lei de formação a partir de seu gráfico. Para executar essa tarefa é necessário determinar a e b , de modo que a função $f(x) = ax + b$ possua o gráfico desejado. Veja um exemplo:



Exemplificando

Com base no gráfico da função afim f representado na Figura 1.10, determine sua lei de formação.

Resolução:

O primeiro detalhe importante a ser observado é que a função é afim, ou seja, seu gráfico é uma reta e sua lei de formação é $f(x) = ax + b$. Para determinar os valores de a e b , em que o gráfico dessa função passe pelos pontos destacados na Figura 1.10, podemos escolher dois pontos quaisquer (escolheremos os pontos de coordenadas $(1, -1)$ e $(-1, 3)$). Lembre-se de que o gráfico de uma função é formado pelos pontos (x, y) , em que $y = f(x)$ e $x \in D(f)$. Para o ponto de coordenadas:

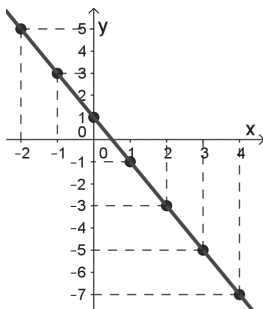
- $(1, -1)$, temos: $f(x) = ax + b \rightarrow f(1) = a \cdot 1 + b \rightarrow -1 = a + b$;
- $(-1, 3)$, temos: $f(x) = ax + b \rightarrow f(-1) = a \cdot (-1) + b \rightarrow 3 = -a + b$.

Observe que temos duas equações lineares, com duas incógnitas, ou seja, um sistema linear. Neste caso, podemos simplificar o sistema somando as duas equações, como segue:

$$\begin{cases} a + b = -1 \\ -a + b = 3 \end{cases}$$

$$a + (-a) + b + b = -1 + 3 \Rightarrow 0 + 2b = 2 \Rightarrow 2b = 2 \Rightarrow b = 1$$

Figura 1.10 | Gráfico de f



Fonte: O autor (2015).

Como $b = 1$ temos: $a + b = -1 \rightarrow a + 1 = -1 \rightarrow a = -1 - 1 = -2$.
Portanto, a função procurada é $f(x) = -2x + 1$.



Faça você mesmo

2) Determine a lei de formação da função afim cujo gráfico passa pelos pontos $(-2, 8)$ e $(2, -4)$.

Função afim crescente e função afim decrescente

Uma característica interessante de ser observada em uma função afim é se ela é crescente ou decrescente. Como essa característica é estudada para qualquer função, podemos compreendê-la de modo geral e, depois, ver como ela se aplica à função afim. De acordo com Thomas, Weir e Hass (2012, p. 6):



Assimile

Seja f uma função definida em um intervalo I e sejam x_1 e x_2 dois pontos em I .

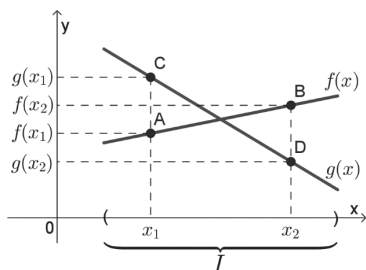
- 1) Se $f(x_2) > f(x_1)$ sempre que $x_1 < x_2$, então f é **crescente** em I .
- 2) Se $f(x_2) < f(x_1)$ sempre que $x_1 < x_2$, então f é **decrescente** em I .

Essa definição pode ser facilmente visualizada na Figura 1.11. No caso, $f(x)$ é crescente e $g(x)$ é decrescente em I . Decorre da definição anterior que, dado $x_1 < x_2$, a função:

- $f(x)$ é crescente, pois $\frac{f(x_2) - f(x_1)}{x_2 - x_1} > 0$

- $g(x)$ é decrescente, pois $\frac{g(x_2) - g(x_1)}{x_2 - x_1} < 0$.

Figura 1.11 | Função crescente e função decrescente



Fonte: O autor (2015).

Simplificadamente, $f(x)$ é crescente porque seus valores aumentam com o aumento dos valores de x ; e $g(x)$ é decrescente porque seus valores diminuem conforme os valores de x aumentam. Observe as inclinações das funções $f(x)$ e $g(x)$.

Podemos denotar $\Delta y = f(x_2) - f(x_1)$ (ou $\Delta y = g(x_2) - g(x_1)$, variação de y) e $\Delta x = (x_2) - (x_1)$ (variação de x) e utilizar a razão $\Delta y / \Delta x$ para avaliar se a função é crescente ou decrescente.

Uma grande vantagem de utilizar a razão $\Delta y / \Delta x$ é que ela está diretamente relacionada à lei de formação da função afim, sendo inclusive muito utilizada para determinar a lei de formação a partir do gráfico. Mais precisamente, dada uma função afim $f(x) = ax + b$, em relação aos seus coeficientes, temos:



Assimile

$$a = \Delta y / \Delta x;$$

se $a > 0$ a função é crescente e se $a < 0$ a função é decrescente;

$$f(0) = a \cdot 0 + b = b.$$

Você pode encontrar a demonstração da igualdade $a = 3y / 3x$ disponível em: <http://www.professores.uff.br/hjbortol/disciplinas/2010.1/gma00116/aulas/gma00116-aula-12-4-up-color.pdf>. Acesso em: 6 nov. 2015.



Exemplificando

Sabendo que os pontos de coordenadas (1,3) e (2,5) pertencem ao gráfico de uma função afim, qual é a lei de formação dessa função?

Resolução:

Primeiramente calculamos as diferenças Δy e Δx e o coeficiente $a = \Delta y / \Delta x$:

$$\Delta y = f(x_2) - f(x_1) = 5 - 3 = 2;$$

$$\Delta x = x_2 - x_1 = 2 - 1 = 1;$$

$$a = \Delta y / \Delta x = 2/1 = 2.$$

Substituindo, $f(x) = 2x + b$ e, além disso, $f(1) = 3 \rightarrow 2 \cdot 1 + b = 3 \rightarrow 2 + b = 3 \rightarrow b = 1$. Portanto, a lei de formação da função é $f(x) = 2x + 1$



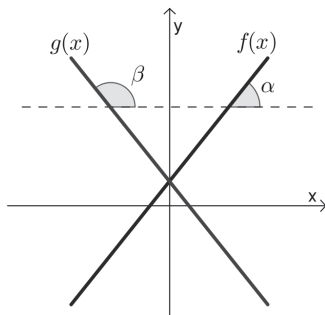
Faça você mesmo

3) Volte ao exemplo da Figura 1.10 e determine a lei de formação da função f utilizando as igualdades $a = \Delta y / \Delta x$ e $b = f(0)$.

Ângulo associado a uma função afim

A toda função afim podemos associar um ângulo θ que está diretamente relacionado ao seu gráfico. Esse ângulo pode ser medido a partir da horizontal, no sentido anti-horário, como ilustra a Figura 1.12.

Figura 1.12 | Ângulo relacionado a uma função afim



Fonte: O autor (2015).



Dica

Para visualizar a localização desse ângulo de forma mais dinâmica, acesse: <<http://tube.geogebra.org/m/1995699>>. Acesso em: 06 nov. 2015.

Quando o gráfico é de uma função afim, há apenas duas possibilidades para o ângulo θ formado com a horizontal: $0^\circ < \theta < 90^\circ$ (a exemplo

do ângulo α da Figura 1.12); ou $90^\circ < \theta < 180^\circ$ (a exemplo do ângulo β da Figura 1.12). Se $\theta = 0^\circ$, ou seja, se o gráfico for horizontal, a função é denominada constante e sua lei de formação é $f(x) = b$, em que b pertence a \mathbb{R} (conjunto dos números reais). Se $\theta = 90^\circ$, ou seja, se o gráfico for vertical, não se trata de uma função, mas de uma relação.

Zero e sinal da função afim

Observe na Figura 1.13 que o gráfico de $f(x) = ax + b$ cruza o eixo horizontal (eixo x) no ponto P . É perceptível que a ordenada de P é igual a 0, ou seja, $y = 0$. Mas e a abscissa de P , qual é seu valor? A abscissa de P é o que denominamos **zero da função**.



Assimile

O zero de uma função $f(x)$ é o valor x_0 tal que $f(x_0) = 0$.



Atenção

Alguns livros utilizam a denominação raiz no lugar de zero. Contudo, o mais comum é dizer que funções possuem zeros e equações possuem raízes.

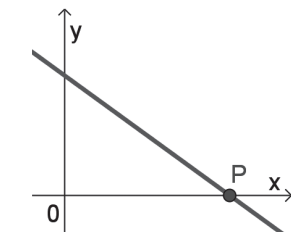
Para uma função afim, se x_0 é o seu zero, temos: Figura 1.13 | Ponto de interseção com o eixo x

$$f(x_0) = ax_0 + b = 0 \Rightarrow ax_0 = -b \Rightarrow x_0 = -\frac{b}{a}$$

Na linguagem matemática, para $f(x)$ crescente, temos:

(a) $\frac{f(x) - f(x_0)}{x - x_0} > 0$ quando $x_0 < x$ ou, ainda, $f(x) - f(x_0) > 0 \rightarrow f(x) > f(x_0) = 0$;

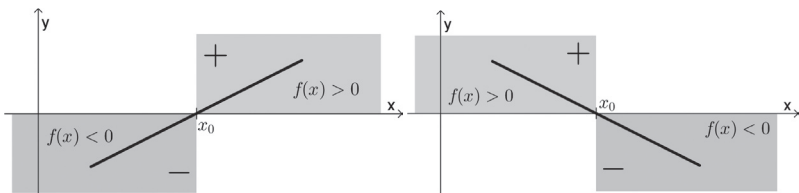
(b) $\frac{f(x_0) - f(x)}{x_0 - x} > 0$ quando $x < x_0$ ou, ainda, $f(x_0) - f(x) > 0 \rightarrow f(x) < f(x_0) = 0$. Simplificadamente, se $f(x)$ é crescente e $f(x_0) = 0$, $f(x) > 0$ para $x > x_0$ e $f(x) < 0$ para $x < x_0$. A mesma análise pode ser feita para o caso de $f(x)$ decrescente e ambos os casos estão ilustrados na Figura 1.14.



Fonte: O autor (2015).

De modo mais simples, para a região do plano cartesiano em que o gráfico de $f(x)$ está acima do eixo das abscissas, isto é, $f(x)$ tem valores maiores que zero, diz-se que o sinal da função é positivo. E para regiões em que $f(x) < 0$, diz-se que a função tem sinal negativo.

Figura 1.14 | Sinal da função afim: (a) $f(x)$ crescente; (b) $f(x)$ decrescente



Fonte: O autor (2015).



Exemplificando

Dada a função $f(x) = 5x - 10$, determine:

- o zero;
- os valores de x para os quais $f(x) > 0$;
- os valores de x para os quais $f(x) < 0$.

Resolução:

Lembre-se de que o zero da função é um valor x_0 tal que $f(x_0) = 0$. Além disso, se a função é crescente, $f(x) > 0$ para $x > x_0$ e $f(x) < 0$ para $x < x_0$. Aplicando estes conceitos, temos:

- $f(x_0) = 0 \rightarrow 5x_0 - 10 = 0 \rightarrow 5x_0 = 10 \rightarrow x_0 = 10/5 = 2$. Logo, 2 é o zero de $f(x)$.
- Como a função é crescente (pois $a = 5 > 0$), $f(x) > 0$ para todos os valores $x > x_0 = 2$.
- $f(x) < 0$ para todos os valores $x < x_0 = 2$.



Dica

Esboce o gráfico da função e verifique as respostas graficamente.



Pesquise mais

Veja mais sobre funções e, em especial, funções afim em: <http://cejarj.cecierj.edu.br/material_impreso/matematica/ceja_matematica_unidade_6.pdf>. Acesso em: 10 nov. 2015. E acesse também este link: <http://cejarj.cecierj.edu.br/material_impreso/matematica/ceja_matematica_unidade_9.pdf>. Acesso em: 10 nov. 2015.

Sem medo de errar!

Vamos retomar o problema proposto no início desta seção: imagine-se como o dono da fábrica de bonés e suponha que você deva convencer seu sócio a expandir o negócio. Para isso, você deve fazer uma apresentação contendo:

a) Um gráfico com os lucros/prejuízos para cada quantidade produzida;

b) Determinar intervalos de produção para os quais há lucro ou prejuízo;

c) O lucro do trimestre com o aumento da produção dos atuais 600 bonés para 1200 bonés ao mês, com acréscimo de produção de 200 bonés mensais.

Primeiramente, para esboçar um gráfico com o lucro/prejuízo, é necessário construir a função lucro $L(x) = R(x) - C(x)$, ou seja, a diferença entre a receita e o custo de produção.

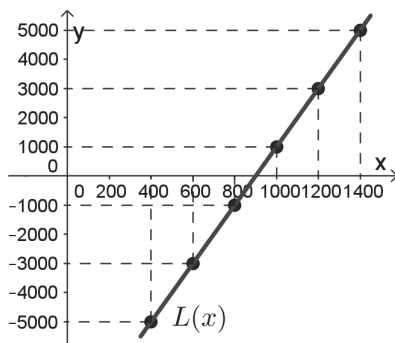


Lembre-se

Na seção anterior (Seção 1.1) você estudou que a função receita era $R(x) = 30 \cdot x$ e a função custo $C(x) = 9000 + 20 \cdot x$, em que x é a quantidade de bonés.

Logo, dado $R(x) = 30 \cdot x$ e $C(x) = 9000 + 20 \cdot x$, a função lucro é $L(x) = 30 \cdot x - (9000 + 20 \cdot x) = 10x - 9000$. Podemos construir uma tabela com alguns valores de x e os respectivos lucros/prejuízos para esboçar o gráfico, como na Figura 1.15. Com isso resolvemos o item (a).

Figura 1.15 | Gráfico de $L(x) = 10x - 9000$



Fonte: O autor (2015).

Foi traçada uma linha junto ao gráfico de $L(x)$ para melhorar a visualização. Entretanto, o correto, nesse caso, seriam somente pontos isolados, pois só faz sentido para essa função a atribuição de valores inteiros para x , pois se trata da quantidade de bonés produzida.

Observe que o gráfico de $L(x)$ cruza o eixo x no ponto de coordenadas $(x_0, 0)$, em que x_0 é o zero da função. Para este problema o zero da função indica a quantidade produzida para a qual não há lucro nem prejuízo. Para quantidades maiores que x_0 há lucro e para quantidades menores, prejuízo. Para determinar x_0 resolvemos a equação $L(x_0) = 0$, como segue: $L(x_0) = 0 \rightarrow 10x_0 - 9000 = 0 \rightarrow 10x_0 = 9000 \rightarrow x_0 = 9000/10 = 900$. Portanto, ao produzir 900 bonés o lucro é zero, ao produzir menos de 900 há prejuízo e, ao produzir mais, há lucro, ficando resolvido o item (b).

Para chegar a 1200 bonés ao mês, a produção deve aumentar 200 bonés por mês nos próximos três meses, sendo produzidos um total de: 800 bonés no primeiro mês; 1000 bonés no segundo mês; 1200 bonés no terceiro mês. Logo, o lucro no trimestre será dado pela expressão $L(800) + L(1000) + L(1200)$. Temos:

$$L(800) + L(1000) + L(1200) = 10 \cdot 800 - 9000 + 10 \cdot 1000 - 9000 + 10 \cdot 1200 - 9000 = -1000 + 1000 + 3000 = 3000$$

Portanto, respondendo o item (c), haverá um lucro de R\$ 3000,00 no trimestre.



Dica

Pense no fato de um dia você estar em uma empresa e ter de convencer alguém a concordar com suas ideias. Uma demonstração com embasamento matemático, como a apresentada, não seria muito mais convincente? Pense em como mostrar suas ideias na forma de uma apresentação com dados, tabelas e gráficos!

Avançando na prática

Pratique mais

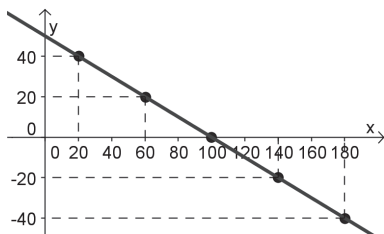
Instrução

Desafiamos você a praticar o que aprendeu transferindo seus conhecimentos para novas situações que pode encontrar no ambiente de trabalho. Realize as atividades e depois as compare com as de seus colegas.

Melhor Negócio	
1. Competências de fundamentos de Área	Conhecer os conceitos matemáticos básicos e proporcionar o desenvolvimento do raciocínio lógico e quantitativo.
2. Objetivos de aprendizagem	Determinar uma função cuja análise de sinal resolva o problema proposto.
3. Conteúdos relacionados	Sinal da função afim.
4. Descrição da SP	<p>Uma empresa de aluguel de veículos possui duas opções de locação:</p> <p>1ª) R\$ 90,00 a diária livre de quilometragem.</p> <p>2ª) R\$ 40,00 a diária mais R\$ 0,50 por quilômetro rodado.</p> <p>Um cliente vai até essa empresa para saber as seguintes informações:</p> <p>a) Para quais distâncias é mais vantajosa a 1ª opção? E a 2ª opção?</p> <p>b) Para qual distância percorrida no dia ambas as opções geram o mesmo custo?</p> <p>Imagine que você seja o funcionário dessa empresa. Como orientar o cliente?</p>
5. Resolução da SP	<p>Perceba que há uma semelhança entre esse problema e o da fábrica de bonés. A primeira pergunta que você deve se fazer é: quais funções relacionam a distância percorrida e o preço a pagar para ambas as opções de locação?</p> <p>Vamos denotar por f e g as funções para a 1ª e 2ª opções, respectivamente, e por x a distância percorrida. Temos:</p> <p>$f(x) = 90,00$ (função constante, pois independe da quilometragem);</p> <p>$g(x) = 40,00 + 0,50x$ (custo fixo de R\$ 40,00 mais custo variável de R\$ 0,50).</p> <p>Agora considere a função diferença $d(x) = f(x) - g(x) = 90,00 - (40,00 + 0,50x) = -0,50x + 50,00$. Se para dado x a diferença for:</p> <ul style="list-style-type: none"> - negativa, é mais vantajosa a 1ª opção, pois $d(x) < 0 \rightarrow f(x) - g(x) < 0 \rightarrow f(x) < g(x)$; - positiva, é mais vantajosa a 2ª opção, pois $d(x) > 0 \rightarrow f(x) - g(x) > 0 \rightarrow f(x) > g(x)$; - nula, ou seja, igual a zero, ambas as opções geram o mesmo custo, pois $d(x) = 0 \rightarrow f(x) - g(x) = 0 \rightarrow f(x) = g(x)$. <p>Sendo x_0 o zero de $d(x)$, temos: $d(x_0) = 0 \rightarrow -0,50x_0 + 50,00 = 0 \rightarrow 0,50x_0 = 50,00 \rightarrow x_0 = \frac{50,00}{0,50} = 100$ Portanto, para 100 quilômetros percorridos no dia, o custo é o mesmo em ambas as opções (ficando respondido o item (b)).</p> <p>Como o coeficiente angular de $d(x)$ é $a = -0,50 < 0$, a função é decrescente e, como consequência, positiva à esquerda de $x_0 = 100$ e negativa à direita desse mesmo valor. Podemos concluir a partir disso que para distâncias menores que 100 quilômetros ($x < x_0 = 100$) é mais vantajosa a 2ª opção, e para distâncias maiores ($x > x_0 = 100$) é mais vantajosa a 1ª opção (ficando respondido o item (a)). Essa conclusão pode ser observada na Figura 1.16.</p>

5. Resolução da SP

Figura 1.16 | Gráfico de $d(x) = -0,50x + 50,00$



Fonte: O autor (2015).

Faça valer a pena

1. Estimou-se que em 22 dias foram desperdiçados 57,2 litros de água por uma torneira pingando. A partir dessa estimativa pode ser desejado saber o quanto é desperdiçado em 4 dias, em 37 dias ou em x dias. Pensando nisso, assinale a alternativa que relaciona a quantidade de dias (x) e o volume de água ($V(x)$) desperdiçado por essa torneira:

- a) $V(x) = 4x$. d) $V(x) = 3,4x$.
b) $V(x) = 22x$. e) $V(x) = 37x$.
c) $V(x) = 2,6x$.

2. Lembre-se de que função afim é aquela cuja lei de formação é $f(x) = ax + b$, em que a e b são os coeficientes. Sendo o coeficiente linear igual a 2, o coeficiente angular igual a -1 e dado $x = 4$, assinale a alternativa que contém as coordenadas de um ponto pertencente ao gráfico de f :

- a) $(4,3)$. d) $(4,-2)$.
b) $(4,-3)$. e) $(4,0)$.
c) $(4,1)$.

3. O preço de uma corrida de táxi é composto pelo valor da bandeirada (R\$ 5,00) mais um valor variável que depende da distância percorrida (R\$ 3,00/km). Considerando essas informações e que por determinada corrida foram pagos R\$ 29,00, qual foi a distância percorrida?

- a) 5 km. d) 10 km.
b) 8 km. e) 12 km.
c) 9 km.

Seção 1.3

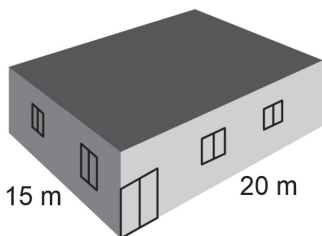
Função quadrática

Diálogo aberto

Lembra-se que na aula anterior você precisava convencer seu sócio a aumentar o investimento na fábrica de bonés e ampliar os negócios? Pois é, o resultado foi melhor que o esperado. Vocês saíram do prejuízo de quando produziam 600 bonés ao mês e começaram a ganhar dinheiro ao produzir 1200. Seu sócio ficou tão feliz que vocês aumentaram ainda mais a produção, chegando a 2400 bonés por mês.

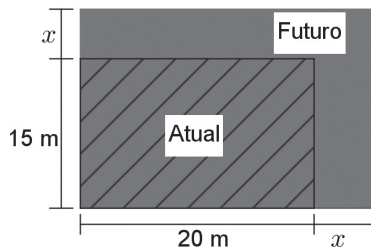
Com uma boa margem de lucro, agora é seu sócio quem quer convencê-lo a ampliar o negócio ainda mais aumentando o espaço físico, indo dos atuais 300 m^2 (como mostra a Figura 1.17) para 750 m^2 futuramente. Devido aos equipamentos que estão instalados e o terreno onde o galpão se encontra, o plano é aumentar tanto o comprimento quanto a largura em um valor x ainda desconhecido, conforme Figura 1.18. Como seu sócio não entende tanto do assunto, pediu para que você determinasse a medida x que deve ser acrescida e o custo desse investimento, uma vez que se estima o valor de R\$ 725,85 por metro quadrado a ser construído.

Figura 1.17 | Galpão



Fonte: O autor

Figura 1.18 | Esboço do projeto



Fonte: O autor

Aqui vão algumas dicas: para resolver este problema você precisa estudar um novo tipo de função, a quadrática. Além disso, para facilitar todo o processo, você pode se focar em responder as seguintes perguntas:

a) Que função relaciona a medida x e a área total do galpão, incluindo a atual? E qual função relaciona x com o valor do investimento? Quais os gráficos dessas funções?

b) Qual medida x proporcionará uma área total de 750 m^2 ?

Bons estudos e sucesso neste planejamento!

Não pode faltar!

As funções quadráticas são uma classe de funções muito utilizadas em problemas de cálculo de área, em cálculos de erro, no estudo do movimento de projéteis, entre outros. Assim como a função afim, essa também é uma função polinomial, mas de grau 2, motivo pelo qual é conhecida popularmente como de 2º grau. Segundo lezzi et al. (1977, p. 123):



Assimile

Uma aplicação (ou relação) f de \mathbb{R} em \mathbb{R} recebe o nome de função quadrática ou do 2º grau quando associa a cada $x \in \mathbb{R}$ o elemento $(ax^2 + bx + c) \in \mathbb{R}$, em que $a \neq 0$.

Alternativamente, podemos dizer que uma função quadrática $f: \mathbb{R} \rightarrow \mathbb{R}$ é aquela cuja lei de formação é $f(x) = ax^2 + bx + c$ com $a \neq 0$. Os valores a , b e c são denominados coeficientes e ax^2 é o **termo dominante**.

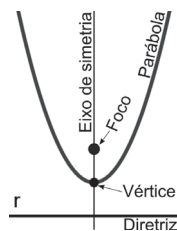


Refleta

Por que para definir a função quadrática é especificado que $a \neq 0$?

Uma característica importante das funções quadráticas é seu gráfico, que apresenta uma curva plana denominada parábola (SODRÉ, 2010, p. 1). Para definir uma parábola são necessários dois objetos, uma reta diretriz e um ponto que chamamos de foco, conforme Figura 1.19. Não abordaremos aspectos formais da construção de uma parábola, mas você pode se aprofundar neste assunto acessando: <http://mathworld.wolfram.com/Parabola.html>. Acesso em: 14 nov. 2015.

Figura 1.19 | Parábola



Fonte: O autor (2015).

Para compreender melhor o gráfico de uma função quadrática, veja o exemplo a seguir.



Esboce o gráfico da função $f(x) = x^2 - 4x + 5$. Resolução:

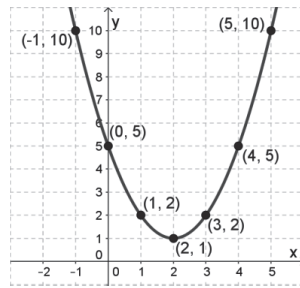
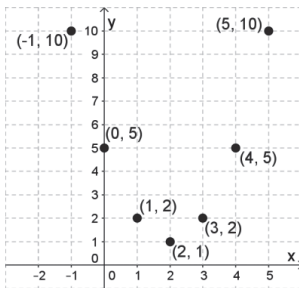
Primeiramente construímos um quadro com alguns valores de x , os respectivos $y = f(x)$ e as coordenadas (x,y) . Observe:

x	$y = f(x) = x^2 - 4x + 5$	(x,y)
-1	$(-1)^2 - 4 \cdot (-1) + 5 = 10$	(-1, 10)
0	$0^2 - 4 \cdot 0 + 5 = 5$	(0,5)
1	$1^2 - 4 \cdot 1 + 5 = 2$	(1,2)
2	$2^2 - 4 \cdot 2 + 5 = 1$	(2,1)

x	$y = f(x) = x^2 - 4x + 5$	(x,y)
3	$3^2 - 4 \cdot 3 + 5 = 2$	(3,2)
4	$4^2 - 4 \cdot 4 + 5 = 5$	(4,5)
5	$5^2 - 4 \cdot 5 + 5 = 10$	(5,10)

Com base nas coordenadas calculadas, marcamos os pontos e traçamos a parábola, conforme Figura 1.20.

Figura 1.20 | Gráfico de $f(x) = x^2 - 4x + 5$

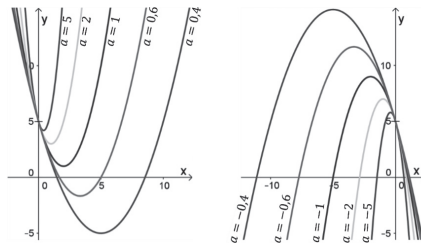


Fonte: O autor (2015).

Observando a Figura 1.20, há alguns elementos importantes: o ponto de coordenadas $(2,1)$ é o **vértice** e a linha vertical $x = 2$ é o **eixo de simetria** da parábola.

No caso do exemplo anterior, dizemos que a parábola tem concavidade para cima, e isso é controlado pelo coeficiente do termo dominante, ou seja, o valor de a . Veja a seguir alguns gráficos de funções quadráticas da forma $f(x) = ax^2 - 4x + 5$ para $a > 0$ (Figura 1.21 (a)) e para $a < 0$ (Figura 1.21 (b)).

Figura 1.21 | Gráficos de $f(x) = ax^2 - 4x + 5$ para vários valores de a : (a) $a > 0$ (b) $a < 0$



Fonte: O autor (2015).

Perceba na Figura 1.21 que, quanto mais próximo de zero está o valor de a , mais “aberta” é a parábola e, quanto mais distante, mais “fechada”. Além disso:



Assimile

Se o valor de a é:

- **Positivo**, $a > 0$, a concavidade da parábola é voltada para cima;
- **Negativo**, $a < 0$, a concavidade da parábola é voltada para baixo.

Observe ainda na Figura 1.21 que, em todos os casos, o ponto de coordenadas $(0,5)$ pertence ao gráfico de $f(x) = ax^2 - 4x + 5$ e que isso se deve ao fato de o coeficiente c ser igual a 5. Veja: se $x = 0$, temos $f(0) = a \cdot 0^2 - 4 \cdot 0 + 5 = 5$, não importando o valor de a ou b .



Assimile

O coeficiente c é igual à ordenada do ponto de interseção do gráfico de $f(x) = ax^2 + bx + c$ com o eixo y .

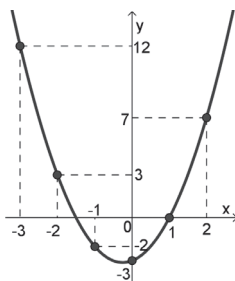
Assim como podemos determinar a lei de formação de uma função afim observando seu gráfico, também é possível fazer o mesmo com uma função quadrática. Veja um exemplo.



Exemplificando

Determine a lei de formação da função quadrática cujo gráfico é apresentado na Figura 1.22.

Figura 1.22 | Gráfico de $f(x)$



Fonte: O autor (2015).

Resolução:

Observe que o ponto de interseção do gráfico de $f(x) = ax^2 + bx + c$ com o eixo y possui coordenadas $(0, -3)$. Logo, $c = -3$ e $f(x) = ax^2 + bx - 3$.

Além disso, como os pontos de coordenadas $(1,0)$ e $(-1,-2)$ pertencem ao gráfico de $f(x)$, temos:

$$f(1)=0 \rightarrow a \cdot 1^2 + b \cdot 1 - 3 = 0 \rightarrow a + b - 3 = 0 \rightarrow a + b = 3;$$

$$f(-1)=-2 \rightarrow a \cdot (-1)^2 + b \cdot (-1) - 3 = -2 \rightarrow a - b - 3 = -2 \rightarrow a - b = 3 - 2 = 1.$$

Segue que a e b são tais que
$$\begin{cases} a + b = 3 \\ a - b = 1 \end{cases}$$

Adicionando as equações, temos: $(a + b) + (a - b) = 3 + 1 \rightarrow 2a = 4 \rightarrow a = 4/2 = 2$. Com $a = 2$ obtemos: $a + b = 3 \rightarrow 2 + b = 3 \rightarrow b = 1$.

Por fim, concluímos que $f(x) = 2x^2 + x - 3$.

Para compreender melhor a relação entre os coeficientes da função quadrática e seu gráfico, acesse o objeto disponível no link: <http://tube.geogebra.org/m/2078515>. Acesso em: 16 nov. 2015.

Zeros da função quadrática

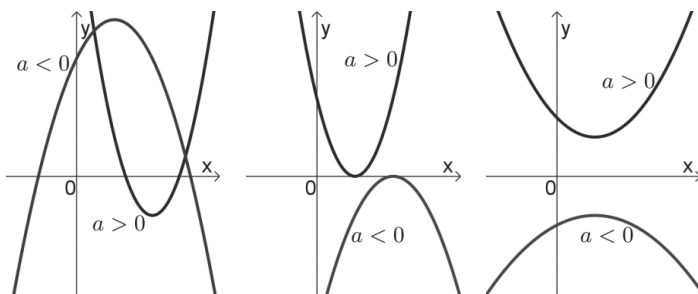


Lembre-se

Na seção anterior você aprendeu que denominamos zero da função o valor da abscissa do ponto de interseção do gráfico com o eixo x . No caso, x_0 será um zero de $f(x)$ se $f(x_0) = 0$. Além disso, para uma função afim $f(x) = ax + b$, o único zero era $x_0 = -b/a$.

Observe agora na Figura 1.22 que o gráfico de $f(x) = 2x^2 + x - 3$ corta o eixo em dois pontos, e não somente em um, como na função afim. Entretanto, nem sempre isso ocorre. O gráfico de uma função quadrática pode tocar o eixo das abscissas em dois, em um ponto ou até não o tocar, como mostra a Figura 1.23.

Figura 1.23 | Zeros de uma função quadrática: (a) dois zeros; (b) um zero; (c) nenhum zero



Fonte: O autor (2015).

Para obter os zeros de uma função quadrática, quando existem, utilizamos a **fórmula do discriminante**, popularmente conhecida como Fórmula de **Bhaskara**:



Assimile

Dada uma função quadrática $f(x) = ax^2 + bx + c$, os valores de x para os quais $f(x) = 0$ são:

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a},$$

ou ainda:

$$x_1 = \frac{-b + \sqrt{b^2 - 4ac}}{2a} \text{ e } x_2 = \frac{-b - \sqrt{b^2 - 4ac}}{2a}.$$

O valor $\Delta = b^2 - 4ac$ é denominado discriminante ou "delta".

Veja um exemplo de como utilizar a fórmula do discriminante:



Exemplificando

Dada as funções a seguir, determine seus zeros, caso existam:

a) $f(x) = x^2 - 6x + 5$ b) $g(x) = 2x^2 + 12x + 18$ c) $h(x) = x^2 - 2x + 3$

Resolução:

a) Para esta função, os coeficientes são $a = 1$, $b = -6$ e $c = 5$. Logo o discriminante será $\Delta = b^2 - 4ac = (-6)^2 - 4 \cdot 1 \cdot 5 = 36 - 20 = 16$. Substituindo o valor $\Delta = 16$, temos:

$$x = \frac{-b \pm \sqrt{\Delta}}{2a} = \frac{-(-6) \pm \sqrt{16}}{2 \cdot 1} = \frac{6 \pm 4}{2} \left\{ \begin{array}{l} x_1 = \frac{6+4}{2} = 5 \\ x_2 = \frac{6-4}{2} = 1 \end{array} \right.$$

Portanto, os zeros de f são $x_1 = 5$ e $x_2 = 1$.

b) No caso da função g , os coeficientes são $a = 2$, $b = 12$ e $c = 18$. Assim, o discriminante será $\Delta = b^2 - 4ac = 12^2 - 4 \cdot 2 \cdot 18 = 144 - 144 = 0$ e:

$$x = \frac{-b \pm \sqrt{\Delta}}{2a} = \frac{-12 \pm \sqrt{0}}{2 \cdot 2} = \frac{-12}{4} = -3$$

Portanto, g possui um único zero e este é $x = -3$.

c) Para a função h , os coeficientes são $a = 1$, $b = -2$ e $c = 3$. Com isso, segue que $\Delta = b^2 - 4ac = (-2)^2 - 4 \cdot 1 \cdot 3 = 4 - 12 = -8$ e:

$$x = \frac{-b \pm \sqrt{\Delta}}{2a} = \frac{-(-2) \pm \sqrt{-8}}{2 \cdot 1}$$

Como $\sqrt{-8} \notin \mathbb{R}$, $-8 \notin \mathbb{R}$, isto é, não é um número real, a expressão anterior não faz sentido para os números reais e, em consequência, a função h não possui zeros reais.



Atenção

No exemplo anterior a função $f(x) = x^2 - 6x + 5$ possui discriminante positivo, $\Delta = 16 > 0$, e dois zeros. Já a função $g(x) = 2x^2 + 12x + 18$ possui discriminante nulo, $\Delta = 0$, e um único zero. Por fim, o discriminante da função $h(x) = x^2 - 2x + 3$ é negativo, $\Delta = -8 < 0$, e esta não possui zeros reais.

Esta observação é válida para toda função quadrática e pode ser compreendida geometricamente com a Figura 1.23. Em: (a) o discriminante é positivo; (b) o discriminante é nulo; (c) o discriminante é negativo.



Faça você mesmo

1) Determine os zeros e esboce o gráfico das funções a seguir:

a) $f(x) = x^2 - 8x + 12$

b) $g(x) = x^2 + 6x - 12$



Pesquise mais

Para saber mais sobre as funções quadráticas, acesse o material disponível no link: http://bit.profmtat-sbm.org.br/xmlui/bitstream/handle/123456789/465/2011_00355_FABIO_ANTONIO_LEAO_SOUSA.pdf?sequence=1. Acesso em: 17 nov. 2015.

Além disso, você pode encontrar uma demonstração simples da fórmula do discriminante em: http://www.ufrgs.br/espmat/disciplinas/funcoes_modelagem/modulo_IV/fundamentos4f.htm. Acesso em: 17 nov. 2015.

Sem medo de errar!

Agora que já tratamos de vários detalhes acerca da função quadrática, vamos retomar o problema proposto no início desta seção?

Uma das perguntas que você deveria responder era: qual função relaciona a medida x e a área total do galpão, incluindo a atual? Para começar, a área de um retângulo é obtida multiplicando as medidas de dois lados consecutivos. No caso da área atual, a medida 300 m^2 é obtida multiplicando 20 m por 15 m . Para calcular a área futura, multiplicamos $(20 + x) \text{ m}$ por $(15 + x) \text{ m}$. Logo, a função que relaciona a medida x , em metros, e a área futura, em metros quadrados, é:

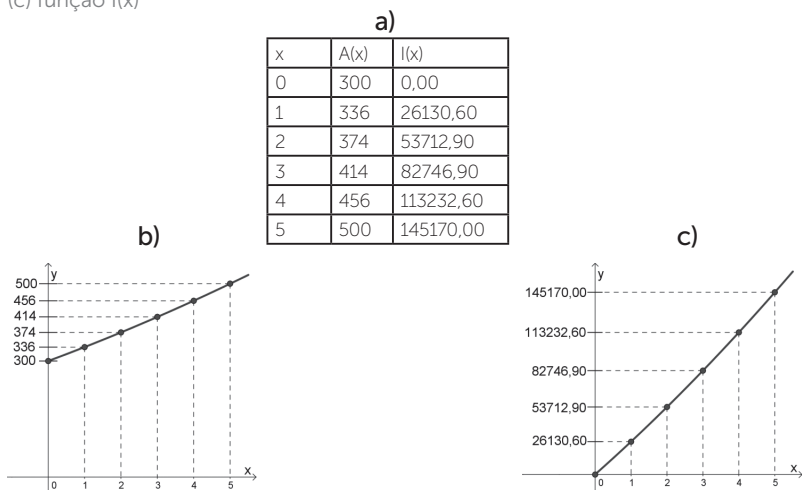
$$A(x) = (20 + x)(15 + x) = (20 + x)15 + (20 + x)x = 20 \cdot 15 + x \cdot 15 + 20 \cdot x + x \cdot x = x^2 + 35x + 300.$$

Você também deveria obter a função que relaciona a medida x com o valor do investimento. Para construir determinada área, o investimento realizado pode ser calculado multiplicando a área correspondente pelo valor do metro quadrado, que é R\$ 725,85. Logo, a função investimento $I(x)$ é obtida multiplicando 725,85 (valor do metro quadrado) pela área que será acrescida. Veja:

$$I(x) = \underbrace{\overbrace{A(x) - 300}^{\text{Área acrescida}}}_{\substack{\text{Área} \\ \text{futura}}} \cdot \underbrace{725,85}_{\substack{\text{Valor do metro} \\ \text{quadrado}}} = (x^2 + 35x + 300 - 300) \cdot 725,85 = 725,85x^2 + 25404,75x$$

Para esboçar os gráficos de $A(x)$ e $I(x)$, calculamos alguns pares ordenados, os marcamos no plano cartesiano e traçamos a parábola, como na Figura 1.24.

Figura 1.24 | Área acrescida e investimento: (a) quadro de valores; (b) função $A(x)$; (c) função $I(x)$



Fonte: O autor (2015).

Por fim, a última informação que você deveria obter é a medida x que proporcionará uma área total de 750 m^2 . Como temos a função área $A(x)$, basta igualar:

$$A(x) = 750 \Rightarrow x^2 + 35x + 300 = 750 \Rightarrow x^2 + 35x + 300 - 750 = 0 \Rightarrow \frac{x^2 + 35x - 450}{(f(x))} = 0$$

Se definirmos $f(x) = x^2 + 35x - 450$, determinar x para o qual $A(x) = 450$ é equivalente a calcular o zero de f . Logo:

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} = \frac{-35 \pm \sqrt{35^2 - 4 \cdot 1 \cdot (-450)}}{2 \cdot 1} = \frac{-35 \pm \sqrt{1225 + 1800}}{2} =$$

$$\frac{-35 \pm 55}{2} \left\{ \begin{array}{l} x_1 = \frac{-35 + 55}{2} = 10 \\ x_2 = \frac{-35 - 55}{2} = -45 \end{array} \right.$$

Observe que f possui dois zeros e, portanto, há também dois valores de x para os quais $A(x) = 450$. Contudo, para o problema prático, só faz sentido utilizarmos valores positivos, pois x é uma medida de comprimento.

Concluimos deste modo que, para a área futura do galpão ser de 750 m^2 , tanto a largura quanto o comprimento devem ser acrescidos em 10 m .



Faça você mesmo

2) Para $x = 10 \text{ m}$, qual é o valor do investimento na reforma do galpão?

Avançando na prática

Pratique mais

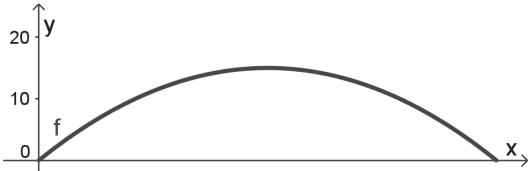
Instrução

Desafiamos você a praticar o que aprendeu transferindo seus conhecimentos para novas situações que pode encontrar no ambiente de trabalho. Realize as atividades e depois as compare com as de seus colegas.

Movimento de projéteis

1. Competências de fundamentos de Área

Conhecer os conceitos matemáticos básicos e proporcionar o desenvolvimento do raciocínio lógico e quantitativo.

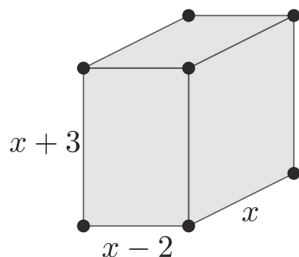
<p>2. Objetivos de aprendizagem</p>	<p>Aplicar os conhecimentos sobre função quadrática no estudo do movimento de projéteis.</p>
<p>3. Conteúdos relacionados</p>	<p>Função quadrática; zero.</p>
<p>4. Descrição da SP</p>	<p>Determinado projétil é lançado para o alto e para frente, descrevendo uma trajetória parabólica. A equação que fornece a altura do projétil em função da distância horizontal x a que ele se encontra do ponto de lançamento é $f(x) = -\frac{1}{15}x^2 + 2x$. Com base nessas informações, que distância horizontal o projétil percorrerá até que toque o solo?</p>
<p>5. Resolução da SP</p>	<p>Vamos primeiramente observar o gráfico dessa função na Figura 1.25.</p> <p>Figura 1.25 Gráfico de $f(x)$</p>  <p>Fonte: O autor (2015).</p> <p>Note que, após o lançamento, o objeto sobe até certa altura e cai novamente até atingir o solo num ponto P, sendo a abscissa desse ponto o zero da função. Calculando o zero, temos:</p> $x = \frac{-2 \pm \sqrt{2^2 - 4 \cdot (-1/15) \cdot 0}}{2 \cdot (-1/15)} = \frac{-2 \pm \sqrt{4}}{-2/15} = \frac{-2 \pm 2}{-2/15} \left\{ \begin{array}{l} x_1 = \frac{-2 + 2}{-2/15} = 0 \\ x_2 = \frac{-2 - 2}{-2/15} = 30 \end{array} \right.$ <p>O valor de x_1 corresponde ao ponto de partida e o valor de x_2 é a abscissa do ponto P. Portanto, o projétil percorrerá 30 m até atingir o solo.</p>

Faça valer a pena

1. Um bloco retangular de concreto tem dimensões $x + 3$, $x - 2$ e x , conforme Figura 1.26. A função $A(x)$ que fornece a área total da superfície do bloco é:

- a) $A(x) = 4x^2 + 4x - 12$.
- b) $A(x) = 6x^2 + 4x - 12$.
- c) $A(x) = 6x^2 + 4x + 12$.
- d) $A(x) = 4x^2 + 4x + 12$.
- e) $A(x) = 8x^2 + 4x - 12$.

Figura 1.26 | Bloco

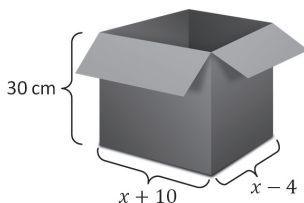


Fonte: O autor (2015).

2. Uma caixa de papelão tem suas dimensões representadas na Figura 1.27. A função $V(x)$ que relaciona x com o volume da caixa e o respectivo volume para $x = 20$ cm são:

- a) $V(x) = 30x^2 + 180x - 1200$ e 12400 cm^3 .
- b) $V(x) = 30x^2 + 160x - 1200$ e 14400 cm^3 .
- c) $V(x) = 30x^2 + 180x - 1200$ e 14400 cm^3 .
- d) $V(x) = 30x^2 + 160x - 1200$ e 12400 cm^3 .
- e) $V(x) = 30x^2 + 180x + 1200$ e 14400 cm^3 .

Figura 1.27 | Caixa de papelão



Fonte: adaptada de <<https://pixabay.com/p-152428>>. Acesso em: 17 nov. 2015.

3. Uma revendedora de cosméticos estima que para um preço de x reais são vendidas $5000 - 2x$ unidades de certo produto mensalmente. Para este produto há um custo de R\$ 10,00 por unidade. Nestas condições, qual é o lucro obtido em um mês em que o preço de venda deste produto era R\$ 16,00?

- a) R\$ 28618,00.
- b) R\$ 16168,00.
- c) R\$ 50000,00.
- d) R\$ 29808,00.
- e) R\$ 48861,00.

Seção 1.4

Sinal, mínimo e máximo da função quadrática

Diálogo aberto

Na seção anterior você estudou a função quadrática, cuja aplicação proporcionou uma solução para o problema da ampliação do galpão da empresa. Dos 300 m² que havia de espaço físico, passou-se para 750 m² com a ampliação, sendo acrescidos 10 m tanto no comprimento quanto na largura. O galpão atualmente possui 30 m de comprimento por 25 m de largura.

Você ainda pôde calcular o investimento com a reforma por meio da função $I(x) = 725,85 \cdot x^2 + 25404,75 \cdot x$. Para o valor x acrescido nas dimensões do galpão, temos: $I(10) = 725,85 \cdot 10^2 + 25404,75 \cdot 10 = 72585 + 254047,5 = 326632,5$ R\$ \rightarrow 326632,50, isto é, o investimento com a reforma foi de R\$ 326632,50.

Após todos esses gastos, seu sócio quer agora recuperar parte do investimento aumentando o preço de venda dos bonés. Atualmente, são produzidos e comercializados 2400 bonés por mês, vendidos por R\$ 30,00 cada. Para que tudo ocorra de modo planejado, ele se adiantou e fez uma pesquisa junto aos consumidores estimando que para cada x reais acrescidos no preço de cada boné são vendidas $(2400 - 60x)$ unidades por mês.

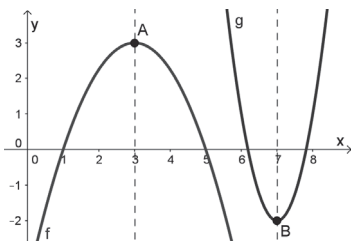
Considerando as informações anteriores, qual deve ser o preço de cada boné para que a receita seja a maior possível?

Não pode faltar!

Máximos e mínimos

Você viu na seção anterior alguns elementos da parábola, entre eles o vértice, como ilustrado na Figura 1.28. O ponto A é o vértice do gráfico de $f(x) = -075x^2 + 4,5x - 3,75$ e o ponto B é o vértice do gráfico de $g(x) = 3x^2 - 42x + 145$. Ambos os gráficos possuem eixo de simetria (linha tracejada) que passa pelo vértice.

Figura 1.28 | Gráficos de f e g



Fonte: O autor (2015).

O fato de uma parábola ter eixo de simetria significa que o lado direito da curva é o reflexo do lado esquerdo, ou seja, se desenhássemos uma parábola em um papel e o dobrássemos sobre o eixo de simetria, os lados da curva se sobreporiam. Observe que o coeficiente do termo dominante de $f(x) = -0,75x^2 + 4,5x - 3,75$ é negativo e que o coeficiente do termo dominante de $g(x) = 3x^2 - 42x + 145$ é positivo. Como já abordado na seção anterior, isso influencia na concavidade da parábola: o gráfico de f tem concavidade para baixo e o gráfico de g tem concavidade para cima. Em decorrência disso, há algo interessante em relação ao vértice: no caso do gráfico de f, o vértice A é o ponto mais alto da parábola e, no caso do gráfico de g, o vértice B é o ponto mais baixo da parábola. Isso pode ser observado para toda função quadrática e está de acordo com o exposto a seguir:



Assimile

Seja $f(x) = ax^2 + bx + c$ uma função quadrática. Se:

- $a > 0$, o gráfico tem concavidade voltada **para cima**, e o vértice é seu ponto **mais baixo**;
- $a < 0$, o gráfico tem concavidade voltada **para baixo**, e o vértice é seu ponto **mais alto**.

Essa percepção gráfica em relação à função quadrática auxilia no entendimento de um conceito estudado para qualquer função:



Assimile

Uma função $f(x)$ possui um máximo em x_v pertencente a um intervalo I, se $f(x_v) \geq f(x)$ para todo $x \in I$. Nesse caso, $f(x_v)$ será o maior valor alcançado (valor máximo) pela função nesse intervalo.

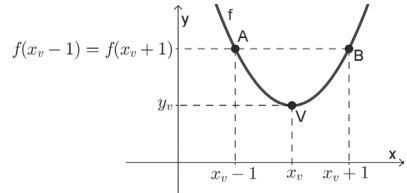
De modo semelhante, uma função $f(x)$ possui um mínimo em x_v pertencente a um intervalo I, se $f(x_v) \leq f(x)$ para todo $x \in I$. Nesse caso, $f(x_v)$

será o menor valor alcançado (valor mínimo) pela função nesse intervalo.

Em ambos os casos, dizemos que os valores são extremos da função.

No exemplo da Figura 1.28, A é um ponto de máximo e B é um ponto de mínimo. Para uma função quadrática, as coordenadas do vértice são (x_v, y_v) , em que x_v é o "x do vértice" e y_v o "y do vértice".

Figura 1.29 | Simetria da parábola



Fonte: O autor (2015).

Como a parábola é simétrica em relação ao seu vértice, segue que $f(x_v - 1) = f(x_v + 1)$, como mostra a Figura 1.29. Com base nessa igualdade, temos:

$$\begin{aligned}
 a(x_v - 1)^2 + b(x_v - 1) + c &= a(x_v + 1)^2 + b(x_v + 1) + c \\
 \cancel{ax_v^2} - 2ax_v + \cancel{a} + \cancel{bx_v} - b + \cancel{c} &= \cancel{ax_v^2} + 2ax_v + \cancel{a} + \cancel{bx_v} + b + \cancel{c} \\
 -2ax_v - b &= 2ax_v + b \\
 -b - b &= 2ax_v + 2ax_v \\
 -2b &= 4ax_v
 \end{aligned}$$

Da última igualdade, segue que $x_v = -\frac{2b}{4a} = -\frac{b}{2a}$. Com essa propriedade e as observações anteriores, podemos enunciar o seguinte:



Assimile

Dada uma função quadrática $f(x) = ax^2 + bx + c$, o vértice de seu gráfico tem coordenadas $(-b/2a, f(-b/2a))$.

Não entraremos em detalhes, mas pode ser demonstrado que $x_v = -b/2a$ e $y_v = -\Delta/4a$.



Refleta

Como podemos deduzir $y_v = -\Delta/4a$ a partir de $x_v = -b/2a$ e $f(x) = ax^2 + bx + c$?



Dada a função quadrática $f(x) = 2x^2 - 4x + 8$, determine as coordenadas do vértice de seu gráfico e se este é um ponto de máximo ou de mínimo.

Resolução:

Para esta função temos $a = 2$, $b = -4$ e $c = 8$. Logo:

$$x_v = -\frac{b}{2a} = -\frac{(-4)}{2 \cdot 2} = 1;$$

$$y_v = -\frac{\Delta}{4a} = -\frac{b^2 - 4ac}{4a} = -\frac{(-4)^2 - 4 \cdot 2 \cdot 8}{4 \cdot 2} = -\frac{16 - 64}{8} = -\frac{16 - 64}{8} = -\frac{(-48)}{8} = 6.$$

Portanto, as coordenadas do vértice são $(1, 6)$.

Como $a = 2 > 0$ o gráfico de f possui concavidade voltada para cima, o que implica que seu vértice é um ponto de mínimo. Nesse caso, $f(1) = 6$ é o menor valor (mínimo) assumido pela função.

Sinal da função quadrática

Observe na Figura 1.30 as funções f , g , h , p , q , r . A partir do exposto na seção anterior e analisando os gráficos, segue que as funções f e p possuem dois zeros reais cada ($\Delta > 0$), as funções g e q possuem um único zero cada ($\Delta = 0$) e as funções h e r não possuem zeros reais ($\Delta < 0$). A partir de uma análise gráfica, podemos ainda afirmar que:

$h(x) > 0$ (é positiva) no intervalo $(-3, +3) = \mathbb{R}$, pois seu gráfico está totalmente acima do eixo x ;

$r(x) < 0$ (é negativa) no intervalo $(-3, +3) = \mathbb{R}$, pois seu gráfico está totalmente abaixo do eixo x ;

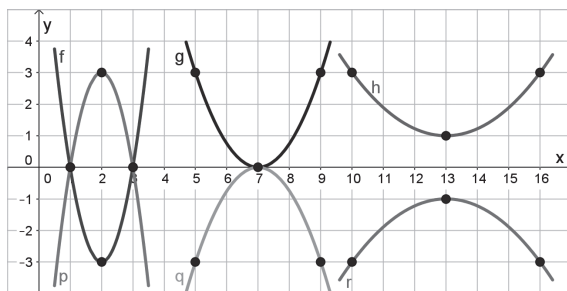
$g(x) > 0$ nos intervalos $(-3, x_1)$ e $(x_1, +3)$, em que $g(x_1) = 0$ (na Figura 1.30, $x_1 = 7$);

$q(x) < 0$ nos intervalos $(-3, x_1)$ e $(x_1, +3)$, em que $q(x_1) = 0$ (na Figura 1.30, $x_1 = 7$);

$f(x) > 0$ em $(-3, x_1)$ e $(x_2, +3)$, $f(x) < 0$ em (x_1, x_2) e $f(x_1) = f(x_2) = 0$ (na Figura 1.30, $x_1 = 1$ e $x_2 = 3$);

$p(x) > 0$ em $(-3, x_1)$ e $(x_2, +3)$, $p(x) < 0$ em (x_1, x_2) e $p(x_1) = p(x_2) = 0$ (na Figura 1.30, $x_1 = 1$ e $x_2 = 3$).

Figura 1.30 | Funções quadráticas



Fonte: O autor (2015).



Exemplificando

Dada a função $f(x) = -x^2 + 2x + 3$, faça o estudo dos sinais e determine se f possui um valor máximo ou um mínimo e especifique esse valor.

Como para esta função $a = -1 < 0$, a concavidade de seu gráfico é voltada para baixo. Em consequência, o vértice é o ponto mais alto do gráfico, tornando-o um ponto de máximo. Além disso, como $b = 2$ e $c = 3$, temos:

$$\Delta = b^2 - 4ac = 2^2 - 4 \cdot (-1) \cdot 3 = 4 - (-12) = 16 \rightarrow \Delta = 16 > 0.$$

Como o discriminante é positivo, a função possui dois zeros reais, além de seu gráfico interceptar o eixo da ordenadas no ponto de coordenadas $(0, 3)$, pois $c = 3$. Com essas informações, podemos inferir que o gráfico da função é semelhante ao esboço da Figura 1.31. Calculando os zeros de f , temos:

$$x_1 = \frac{-b + \sqrt{\Delta}}{2a} = \frac{-2 + \sqrt{16}}{2 \cdot (-1)} = \frac{-2 + 4}{-2} = \frac{2}{-2} = -1;$$

$$x_2 = \frac{-b - \sqrt{\Delta}}{2a} = \frac{-2 - \sqrt{16}}{2 \cdot (-1)} = \frac{-2 - 4}{-2} = \frac{-6}{-2} = 3.$$

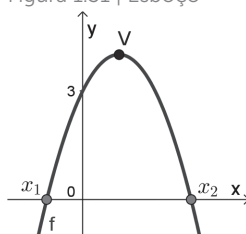
Logo, $f(x) > 0$ em $(-3, -1)$ e $(3, +3)$, $f(x) < 0$ em $(-1, 3)$ e $f(-1) = f(3) = 0$.

Para determinar o máximo de f , precisamos primeiramente do valor de x_v :

$$x_v = -\frac{b}{2a} = -\frac{2}{2 \cdot (-1)} = -\frac{2}{(-2)} = 1.$$

Com isso, o valor máximo de f será $f(x_v) = f(1) = -1^2 + 2 \cdot 1 + 3 = -1 + 2 + 3 = 4$.

Figura 1.31 | Esboço



Fonte: O autor (2015).



Faça você mesmo

1) Dada a função $f(x) = x^2 + 6x + 5$, faça o estudo dos sinais e determine se f possui um valor máximo ou um mínimo e especifique esse valor.



Pesquise mais

Você pode investigar de forma mais dinâmica a relação entre os coeficientes da função quadrática e seu sinal com o objeto disponível no link: <https://www.geogebra.org/m/171465>. Acesso em: 24 nov. 2015.

Além disso, para ver mais sobre as funções quadráticas, principalmente quanto a máximos e mínimos e ao sinal, acesse: http://www.fund198.ufba.br/apos_cnf/funcao4.pdf. Acesso em: 24 nov. 2015.

Sem medo de errar!

Vamos retomar o problema proposto no início da seção: atualmente são produzidos e comercializados 2400 bonés por mês e estes são vendidos por R\$ 30,00 cada. Além disso, seu sócio estimou que para cada x reais acrescidos no preço de cada boné são vendidas $(2400 - 60x)$ unidades por mês. Com todas essas informações, como calcular o preço de cada boné para que a receita seja a maior possível?

Vamos interpretar o problema: obter a maior receita possível é o mesmo que obter a **receita máxima**. Desse modo, se conseguirmos construir uma função receita que modele toda essa dinâmica, obter a receita máxima é o mesmo que calcular o valor **máximo da função**. Considere que o preço do boné, que atualmente é de R\$ 30,00, seja acrescido em x reais. O novo preço será:

$$\underbrace{30,00}_{\text{Preço atual}} + \underbrace{x}_{\text{Acréscimo}}$$

Com o boné nessa faixa de preço, são vendidas $(2400 - 60x)$ unidades. Lembre-se de que a função receita é obtida multiplicando a quantidade vendida pelo preço, logo:

$$\underbrace{R(x)}_{\text{Receita}} = \underbrace{(2400 - 60x)}_{\text{Quantidade}} \cdot \underbrace{(30 + x)}_{\text{Preço}}$$

Desenvolvendo os cálculos, temos:

$$R(x) = (2400 - 60x)(30 + x) = (2400 - 60x)30 + (2400 - 60x)x \\ = 72000 - 1800x + 2400x - 60x^2$$

$$\text{Portanto, } R(x) = -60x^2 + 600x + 72000.$$

Depois de interpretar o problema, podemos resolvê-lo com o auxílio da função receita: para essa função, temos $a = -60 < 0$ e, conseqüentemente, essa função possui um valor máximo atingido em $x_v = b/2a = 600 / (2 \cdot (-60)) = -600/(-120) = 5$. Esse é o valor que pode ser acrescentado no preço atual do boné para alcançar a receita máxima. Como o preço atual é R\$ 30,00, o novo valor será R\$ 35,00, ficando resolvido o problema.



Faça você mesmo

2) Qual será a receita máxima?

Avançando na prática

Pratique mais

Instrução

Desafiamos você a praticar o que aprendeu transferindo seus conhecimentos para novas situações que pode encontrar no ambiente de trabalho. Realize as atividades e depois as compare com as de seus colegas.

Área máxima

1. Competências de fundamentos de Área	Conhecer os conceitos matemáticos básicos e proporcionar o desenvolvimento do raciocínio lógico e quantitativo.
2. Objetivos de aprendizagem	Utilizar o conceito de máximo e mínimo de uma função na resolução de problemas de otimização.
3. Conteúdos relacionados	Máximos e mínimos.
4. Descrição da SP	Uma área retangular será cercada com tela em três lados, sendo que no quarto lado será utilizado um muro já existente, conforme Figura 1.32.

Figura 1.32 | Área a ser cercada

Fonte: O autor (2015).

Se há 40 metros de tela disponível, quais serão as dimensões do cercado que possui área máxima?

Faça valer a pena

1. Um aspecto muito interessante em relação às funções consiste em seus valores extremos, que podem ser mínimos ou máximos. Para as funções quadráticas, sabemos se um valor extremo será um mínimo ou um máximo apenas observando seus coeficientes.

Em relação aos valores extremos, as funções $f(x) = x^2 + 2x$, $g(x) = -2x^2 + 3$ e $h(x) = 4x^2 - 5x - 8$ possuem, respectivamente:

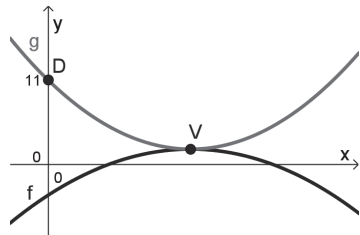
- a) máximo, mínimo e máximo.
- b) mínimo, máximo e mínimo.
- c) máximo, máximo e mínimo.
- d) mínimo, mínimo e máximo.
- e) mínimo, máximo e máximo.

2. Os gráficos das funções

$f(x) = -\frac{2}{3}x^2 + 4x + c$ e $g(x) = x^2 - 6x + 11$ possuem o mesmo vértice, conforme Figura 1.33. Nesse caso, qual é o valor do coeficiente c da função f ?

- a) -4.
- b) -2.
- c) -1.
- d) -3.
- e) -5.

Figura 1.33 | Funções f e g



Fonte: O autor (2015).

3. Determinado trecho de uma montanha-russa tem seu trilho a uma altura $f(x) = 0,1x^2 - 2x + 14$, com x pertencente ao intervalo $(0,20)$, em metros. Nesse trecho, qual é a altura do trilho no seu ponto mais baixo, considerando o eixo das abscissas como sendo o solo?

- a) 1 m.
- b) 2 m.
- c) 3 m.
- d) 4 m.
- e) 5 m.

Referências

ANTON, Howard; BIVENS, Ir; DAVIS, Stephen. **Cálculo**. 10. ed. Porto Alegre: Bookman, 2014.

IEZZI, Gelson et al. **Fundamentos de matemática elementar**: conjuntos e funções. 3. ed. São Paulo: Atual, 1977.

LARSON, Ron. **Cálculo aplicado**: curso rápido. 8. ed. São Paulo: Cengage Learning, 2011.

ROGAWSKI, Jon. **Cálculo**. Porto Alegre: Bookman, 2009.

SIMMONS, George F. **Cálculo com geometria analítica**. São Paulo: McGraw-Hill, 1987.

SODRÉ, Ulysses. **Funções quadráticas**. 2010. Disponível em: <<http://www.uel.br/projetos/matessencial/superior/matzoo/quadratica.pdf>>. Acesso em: 14 nov. 2015.

STEWART, James. **Cálculo**. 7. ed. São Paulo: Cengage Learning, 2013, 1. v.

THOMAS, George B.; WEIR, Maurice D.; HASS, Joel. **Cálculo**. 12. ed. São Paulo: Pearson, 2012.

Estatística descritiva

Convite ao estudo

Você já tomou conhecimento na Unidade 1 de alguns termos utilizados na estatística, entre eles, a própria palavra **estatística**, que simplificada nós poderíamos definir como a ciência que cuida da coleta, descrição e interpretação de dados. Contudo, essa não é a única maneira de se definir estatística. Segundo Johnson e Kuby (2013), "a palavra estatística possui significados diferentes para pessoas de diferentes áreas e interesses". Veja, por exemplo, a seguinte definição para estatística:



Refleta

"A estatística moderna é uma tecnologia quantitativa para a ciência experimental e observacional que permite avaliar e estudar as incertezas e os seus efeitos no planejamento e interpretação de experiências e de observações de fenômenos da natureza e da sociedade".

Raul Yukihiro Matsushita, professor assistente do Departamento de Estatística da Universidade de Brasília

Sugestão: pesquise outras definições para estatística e faça um comparativo.

Ainda na Unidade 1 citamos a **estatística descritiva** e a **estatística inferencial**. A primeira, mais frequentemente utilizada, cuida da coleta, análise e sintetização de dados. Já a segunda se utiliza dos resultados obtidos pela primeira para realizar a interpretação das informações e, posteriormente, auxiliar na tomada de decisão.

Para iniciar seus estudos em estatística descritiva, imagine que você é um funcionário de uma grande empresa (que

denominaremos de M) e foi incumbido de realizar uma pesquisa para determinar o perfil dos 30 mil funcionários. O relatório final dessa pesquisa deverá conter informações pessoais como idade, peso, altura, sexo, cor dos olhos, raça e também informações sobre a satisfação em relação às condições de trabalho e à remuneração. Considere que o prazo estipulado para a realização dessa tarefa seja uma semana. Nesse ponto algumas dúvidas devem ter surgido. Entre elas podemos mencionar:

- O que exatamente devo pesquisar?
- Como fazer essa pesquisa?
- O tempo será suficiente para pesquisar todos os funcionários? Em caso negativo, o que fazer?
- Como apresentar os resultados obtidos com a pesquisa?

No decorrer dessa unidade, pouco a pouco, algumas dessas perguntas serão respondidas e você poderá ter uma visão geral de todo o processo. Ao final, esperamos que você:

- Compreenda as principais técnicas de amostragem;
- Interprete informações apresentadas em tabelas e gráficos;
- Entenda as medidas de posição e sua representatividade;
- Compreenda as medidas de dispersão e sua representatividade.



Atenção

Observe que o termo "peso" foi empregado incorretamente. Nesse contexto, o correto seria "massa". O peso é uma grandeza física, mais especificamente, uma força. Apesar disso, como esse termo é de uso frequente no dia a dia, manteremos o sentido coloquial da palavra peso.

Seção 2.1

Amostragem

Diálogo aberto

Para iniciarmos os estudos em estatística, vamos retomar a situação proposta anteriormente: imagine que você é um funcionário da empresa M e que foi incumbido de realizar uma pesquisa para determinar o perfil dos 30 mil funcionários. O relatório final dessa pesquisa deverá conter informações pessoais como idade, peso, altura, sexo, cor dos olhos, raça e também informações sobre a satisfação em relação às condições de trabalho e à remuneração. Considere que o prazo estipulado para a realização dessa tarefa seja uma semana.

Para que essa tarefa seja executada, o primeiro passo é planejar a coleta de dados, assunto que será estudado nesta seção de autoestudo.

Não Pode Faltar!

Conceitos básicos

Antes de iniciar o planejamento da coleta de dados, é essencial que você consiga identificar alguns objetos de estudo da estatística, tais como **população** e **amostra**. Uma **população** é o conjunto de todos os elementos que possuem determinada característica em comum. Para o nosso exemplo, a população corresponde aos 30 mil funcionários da empresa M. Além de pessoas, populações podem ser compostas por animais, objetos, substâncias químicas etc.



Exemplificando

Suponha que se queira analisar o:

- Comportamento das formigas cortadeiras no Brasil. Nesse caso, a população corresponderia à totalidade das formigas dessa espécie no país.
- Número de peças defeituosas fabricadas por determinada máquina. Nesse caso, a população corresponderia a todas as peças fabricadas por essa máquina.

A população pode ser **finita**, quando é possível listar fisicamente todos os seus elementos, ou **infinita**, quando não há essa possibilidade. No caso dos funcionários da empresa M, a população é finita, pois poderíamos, por exemplo, solicitar ao departamento de pessoal que fornecesse uma lista com os nomes de todos os funcionários que constam na folha de pagamento. Para o exemplo do estudo do comportamento das formigas cortadeiras, apesar de haver um número finito dessas formigas, podemos considerar essa população como sendo infinita, pois esse número é muito grande e jamais conseguiríamos observar todas elas.

Uma **amostra** é qualquer subconjunto de uma população. Geralmente, amostras são finitas e utilizadas quando a população é muito numerosa ou infinita, o que dificulta ou até impossibilita a observação de todos os seus elementos.



Assimile

População é o conjunto de todos os elementos que possuem determinada característica em comum.

Amostra é qualquer subconjunto de uma população.

Outros objetos de estudo da estatística são o **censo** e a **amostragem**. Um **censo** corresponde ao processo de coleta de dados de toda a população, enquanto que uma **amostragem** é o processo de coleta de dados de uma amostra, ou seja, de apenas parte da população. Censos são raramente feitos, pois são muito demorados e caros quando comparados a uma amostragem.



Pesquise mais


No Brasil o IBGE (Instituto Brasileiro de Geografia e Estatística) se encarrega de realizar um censo a cada 10 anos. Nele são coletadas diversas informações sobre a população e os domicílios. Para mais informações, acesse: <www.ibge.gov.br>. Acesso em: 8 jul. 2015.

Variáveis

Quando estudamos uma população, estamos interessados em suas **v** e nos possíveis valores que elas podem assumir. Idade, peso, altura, sexo, cor dos olhos e raça são exemplos de variáveis.

Veja na Figura 2.1 algumas informações a respeito do jogador Pelé e mais alguns exemplos de variáveis.

Figura 2.1 | Dados pessoais do jogador Pelé em 2015



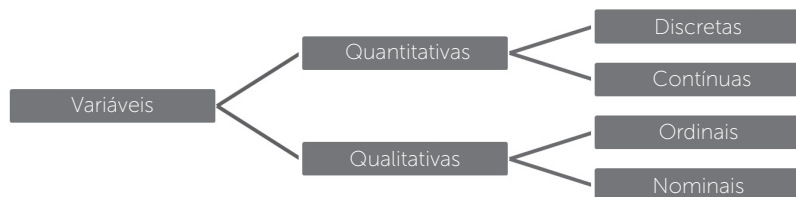
Nome completo	Édson Arantes do Nascimento
Posição	Meia-atacante
Ano de nascimento	1940
Idade (anos)	74
Nacionalidade	Brasileira
Local de nascimento	Três Corações (MG)
Altura (metros)	1,73
Peso (kg)	75
Formação acadêmica	Superior completo

Fontes: © 2010-2015 Graphic Resources LLC. Disponível em: <http://br.freepik.com/fotos-gratis/pele--jogador-de-futebol-lendas_566595.htm> e <<http://esporte.uol.com.br/futebol/biografias/559/pele>>. Acesso em: 28 abr. 2015.

Na Figura 2.1 podemos identificar as variáveis "nome", "posição", "ano de nascimento", "idade", "nacionalidade", "local de nascimento", "altura", "peso" e "formação acadêmica". Além disso, "Édson Arantes do Nascimento", "meia-atacante", "1940", "74", "brasileira", "Três Corações (MG)", "1,73", "75" e "superior completo" são, respectivamente, os valores que elas assumem para o jogador Pelé. Observe que algumas dessas variáveis retornaram valores numéricos e outras, não numéricos.

Quando uma variável retorna valores numéricos, nós a denominamos **variável quantitativa**. Já aquela que retorna valores não numéricos, nós a denominamos **variável qualitativa**. Essa diferença é fácil de ser assimilada, pois a palavra quantitativa lembra "quantidade", ou seja, números, enquanto a palavra qualitativa lembra "qualidade", isto é, atributos. As variáveis quantitativas e as qualitativas podem ainda ser subdivididas em dois subgrupos, conforme ilustra a Figura 2.2.

Figura 2.2 | Tipos de variáveis



Fonte: Os autores (2015)

Uma **variável quantitativa discreta** é aquela que, em geral, assume valores inteiros ou um número finito de valores bem definidos. Na Figura 2.1 podemos observar duas variáveis com essa característica: “ano de nascimento” e “idade”. Já uma **variável quantitativa contínua** é aquela que pode assumir qualquer valor (inteiro ou não) dentro de um intervalo. Na Figura 2.1 podemos observar também duas variáveis com essa característica: “altura” e “peso”.

Uma **variável qualitativa ordinal** é aquela não numérica que apresenta uma ordenação entre seus valores, a exemplo da variável “formação acadêmica”. Veja que Pelé possui ensino superior completo. Entretanto, caso observássemos os valores dessa variável para outras pessoas, poderíamos ter como resposta “ensino fundamental” ou “ensino médio”, por exemplo. Uma ordenação natural para nós é que o “ensino fundamental” antecede o “ensino médio”, que, por sua vez, antecede o “ensino superior”. Por fim, uma **variável qualitativa nominal** é aquela, não numérica, que não possui ordenação entre seus valores, como “nome”, “posição”, “nacionalidade” e “local de nascimento”.



Assimile

Variável:

- **quantitativa discreta**: aquela que, em geral, assume valores inteiros ou um número finito de valores bem definidos;
- **quantitativa contínua**: aquela que pode assumir qualquer valor (inteiro ou não) dentro de um intervalo;
- **qualitativa ordinal**: aquela, não numérica, que apresenta uma ordenação entre seus valores;
- **qualitativa nominal**: aquela, não numérica, que não possui ordenação entre seus valores.

Para verificar se você compreendeu as diferenças entre os diversos tipos de variáveis, classifique as da Figura 2.3 em discretas, contínuas, nominais ou ordinais.

Figura 2.3 | Exemplos de variáveis

altura	formação acadêmica	peso	cor dos olhos	número de filhos
doente/sadio	fumante/não fumante	estágio de uma doença (inicial, intermediário, terminal)	mês de observação (janeiro, fevereiro, ..., dezembro)	sexo
tempo	número de bactérias por litro de leite	número de cigarros fumados por dia	pressão arterial	idade (anos)

Fonte: O autor (2015)

Confira sua classificação com a proposta no apêndice da Seção 2.1. Quando aferimos um valor a partir de uma análise de determinada variável em uma amostra, o denominamos **estatística**. Já se o referido valor é obtido a partir de uma análise de uma variável na população como um todo, o denominamos **parâmetro**. A média de altura dos funcionários do setor administrativo da empresa M, por exemplo, corresponde a uma estatística. Tal estatística busca estimar a verdadeira média da altura de todos os funcionários da empresa M, a qual corresponde a um parâmetro.

Grande parte das pesquisas é feita a partir de amostras. Tais pesquisas obtêm estatísticas que buscam estimar os parâmetros da população.

Tipos de amostragem

Antes de atingirmos o objetivo dessa seção de autoestudo, que é o de planejar a coleta de dados, precisamos ainda compreender os principais tipos de amostragem. A escolha adequada do método é de fundamental importância para a confiabilidade dos dados a serem coletados.

Um grande desafio de quem está planejando fazer uma pesquisa é saber como coletar uma amostra confiável, ou seja, como conseguir selecionar na população um subconjunto que seja representativo do todo. Observe que essa é uma etapa de grande importância, já que pode impactar todo o restante do trabalho.

Uma coleta mal planejada pode provocar impressões erradas acerca da população, fornecendo valores que não a representam. A distorção de uma estatística em comparação com um parâmetro populacional é denominada **viés**. Você poderá notar um exemplo clássico de amostragens enviesadas na época das eleições. Vários candidatos apresentam resultados de pesquisas de intenção de voto, sendo que cada uma tem um resultado diferente. Fique atento!

Na literatura sobre o assunto são diversos os métodos de amostragem. Dentre eles, os mais conhecidos são:

- Amostragem de conveniência
- Amostragem voluntária
- Amostragem aleatória simples
- Amostragem sistemática
- Amostragem aleatória estratificada
- Amostragem por conglomerado

Uma **amostragem por conveniência** geralmente ocorre quando o indivíduo seleciona na população elementos que considera pertinentes, os quais imagina serem representativos do todo. Essa conduta, estatisticamente falha, muitas vezes é a causadora de resultados muito divergentes dos verdadeiros parâmetros da população. Vide exemplo das pesquisas eleitorais.

Na **amostragem voluntária** a amostra é obtida por seleção de voluntários. Frequentemente vemos esse tipo de pesquisa sendo feita pela internet ou por telefone. Pense um pouco... você já respondeu a alguma enquete realizada por esses canais? Foi sincero na resposta dada à enquete? Se suas respostas foram "sim" e "não" você acaba de perceber a origem de um dos problemas desse tipo de amostragem, a saber, o nível de confiança nos dados coletados. Geralmente as pessoas não estão dispostas a responder a pesquisas. Portanto, quando estas são feitas com voluntários os resultados obtidos devem ser tratados com muito cuidado. Pode parecer então que esse tipo de amostragem não deve nunca ser empregado, contudo, em muitos casos, essa é a única opção. Imagine que uma empresa farmacêutica queira testar um novo fármaco destinado à prevenção e ao tratamento do HIV. Você concordaria em fazer parte da pesquisa (considerando que não

possua a doença)? Imaginamos que não. Portanto, em casos como este, não há outra opção senão amostragem por voluntários.

A **amostragem aleatória simples** é aquela realizada por meio de sorteio. Esse tipo de amostragem tem a vantagem em relação às anteriores de garantir que todos os elementos da população tenham a mesma probabilidade de pertencer à amostra. Para realizar uma amostragem desse tipo também se pode utilizar uma tabela de números aleatórios, como a apresentada na página 146 do arquivo disponível em <<http://www.est.ufpr.br/ce003/material/apostilace003.pdf>>. (Acesso em: 29 abr. 2015). Para obter orientações de como utilizar uma tabela de números aleatórios, assista ao vídeo disponível em <<https://www.youtube.com/watch?v=UxgLkk-XuRQ>>. Outra maneira de realizar uma amostragem aleatória simples é por meio de uma planilha eletrônica. Leia um pequeno tutorial de como gerar números aleatórios em planilhas no *link* <<http://dicasdeexcel.com.br/2009/05/26/como-gerar-numeros-aleatorios>>. Há ainda a possibilidade de utilizar uma calculadora científica. Para obter orientações de como gerar números aleatórios em uma calculadora, assista a um vídeo sobre o assunto em: <<https://www.youtube.com/watch?v=2fW92PRPwfQ>>. Acesso em: 29 abr. 2015.

Uma **amostra sistemática** pode ser feita facilmente quando há uma ordenação natural dos elementos da população, como a ordem alfabética ou a sequência de casas em uma rua. Para retirar uma amostra sistemática de tamanho n de uma população com N elementos, ordenados de 1 até N , seguimos os seguintes passos:

1. Dividimos a população em n subgrupos de tamanho $k = \frac{N}{n}$;
2. No primeiro subgrupo realizamos um sorteio (amostragem aleatória simples) para determinar o primeiro elemento pertencente à amostra. Suponha que ele esteja na posição $p \leq k$;
3. A partir do sorteio do passo anterior, os demais $n - 1$ elementos pertencentes à amostra ficam determinados. Serão aqueles que estiverem nas posições: $p + k$, $p + 2k$, $p + 3k$, ..., $p + (n - 1)k$.



Considere uma população de 20 alunos da disciplina de Métodos Quantitativos, os quais estão listados a seguir em ordem alfabética.

1	Alice	8	Isabella	15	Matheus
2	Arthur	9	Júlia	16	Miguel
3	Bernardo	10	Laura	17	Pedro
4	Davi	11	Lucas	18	Rafael
5	Gabriel	12	Luíza	19	Sophia
6	Giovanna	13	Manuela	20	Valentina
7	Heitor	14	Maria		

Selecione uma amostra sistemática de tamanho 4 dessa população.

Resolução:

Observe que essa população tem tamanho $N = 20$ e a amostra solicitada tem tamanho $n = 4$. Portanto, devemos dividir a população em 4 subgrupos de tamanho $k = 20/4 = 5$, como segue:

1	Alice	6	Giovanna	11	Lucas	16	Miguel
2	Arthur	7	Heitor	12	Luíza	17	Pedro
3	Bernardo	8	Isabella	13	Manuela	18	Rafael
4	Davi	9	Júlia	14	Maria	19	Sophia
5	Gabriel	10	Laura	15	Matheus	20	Valentina

Nessa etapa é necessário que façamos um sorteio no primeiro grupo para determinar o primeiro a pertencer à amostra. Suponha que o sorteado tenha sido o número $p = 2$, ou seja, Arthur. Desse modo, os próximos a pertencerem à amostra serão:

$$p + k = 2 + 5 = 7 \rightarrow \text{Heitor}$$

$$p + 2k = 2 + 2 \cdot 5 = 12 \rightarrow \text{Luíza}$$

$$p + 3k = 2 + 3 \cdot 5 = 17 \rightarrow \text{Pedro}$$

A **amostragem aleatória estratificada** difere das anteriores, principalmente, por envolver mais de uma etapa. Esse tipo de amostragem é utilizado geralmente nos casos em que a população possui subgrupos com características próprias que podem ser pertinentes à pesquisa. Imagine que se queira pesquisar o gênero

musical preferido de uma população. Convém supor que a preferência possa ser diferente de acordo com a idade da pessoa, pois em épocas diferentes as tendências musicais são outras e considerar toda a população como um grupo homogêneo pode ser um erro para a coleta de dados. Desse modo, talvez seja prudente dividir a população em vários grupos por faixa etária, por exemplo, de 0 a 9 anos, de 10 a 19 anos, de 20 a 40 anos e mais de 40 anos. Atenção: essa é apenas uma sugestão. Para determinar quais subdivisões da população devemos considerar é necessário um estudo mais aprofundado.

Cada subgrupo considerado na amostragem aleatória estratificada recebe o nome de **estrato**. A definição desses estratos, primeira etapa da amostragem, é feita de modo a se obter maior homogeneidade entre os seus elementos e maior heterogeneidade entre os estratos. Na segunda etapa, retira-se uma amostra em cada estrato, podendo este procedimento ser realizado por amostragem aleatória simples, sistemática ou outra que for mais adequada.

Geralmente, na amostragem aleatória estratificada, o tamanho da amostra retirada de cada estrato é correspondente ao percentual que o estrato representa em relação à população.



Exemplificando

Suponha que para determinada pesquisa seja necessário dividir a população de 100 indivíduos em dois estratos: os homens (45 indivíduos) e as mulheres (55 indivíduos). Se quisermos retirar uma amostra estratificada de tamanho 20 dessa população, quantos homens e quantas mulheres teremos?

Resolução:

Inicialmente calculamos a porcentagem que cada estrato representa em relação ao total:

- Estrato 1 (mulheres): $\frac{55}{100} = 55\%$
- Estrato 2 (homens): $\frac{45}{100} = 45\%$

Desse modo, a amostra deve ser composta em 55% de mulheres e 45% de homens, ou seja:

- Amostra do estrato 1: $55\% \cdot 20 = 11$ mulheres
- Amostra do estrato 2: $45\% \cdot 20 = 9$ homens

Observe no exemplo anterior que, pelo fato de termos dividido a população em dois estratos (homens e mulheres), dentro de cada um os elementos são homogêneos (todos os elementos são do mesmo sexo), e, quando comparamos os estratos entre si, eles são significativamente heterogêneos, pois em um há só mulheres e no outro, apenas homens.

A **amostragem por conglomerado** (também denominada amostragem por *cluster*) é um processo que, assim como a amostragem estratificada, envolve mais de uma etapa. A diferença básica entre essas duas é que, enquanto a estratificada busca dividir a população em subgrupos cujos elementos sejam homogêneos, a por conglomerado divide a população em subgrupos cujos elementos sejam heterogêneos. Cada subgrupo definido nesse tipo de amostragem, denominado **conglomerado** (ou *cluster*), será semelhante à população, o que implica a semelhança entre os conglomerados.

Após definir os conglomerados (primeira etapa), geralmente se utiliza amostragem aleatória simples para escolher quais farão parte da amostra (segunda etapa). Em seguida, realiza-se um censo em cada conglomerado selecionado (terceira etapa).



Exemplificando

A amostragem por conglomerado pode ser utilizada no caso de uma empresa que possua várias filiais. Espera-se que as filiais sejam semelhantes entre si e semelhantes à empresa como um todo. Considerando que dentro de cada filial possa ser observada a mesma heterogeneidade que no restante da empresa, temos uma situação semelhante à teorizada para esse tipo de amostragem.

Um procedimento padrão seria considerar cada filial da empresa como um conglomerado, realizando-se uma amostragem aleatória simples para definir quais conglomerados serão recenseados.

Agora que você já conhece alguns métodos de amostragem, elabore um roteiro para realizar a coleta de dados proposta na situação-problema do tópico DIÁLOGO ABERTO dessa seção de

autoestudo. Após a elaboração do roteiro, compare sua proposta com a apresentada a seguir.

Roteiro para uma coleta de dados

Uma das etapas mais importantes de toda coleta de dados é o planejamento. Geralmente, ele pode ser feito por meio da determinação de um roteiro ou um *checklist*. Para ter eficiência, esse roteiro deve ser elaborado e revisado a fim de evitar falhas. Ao final, o pesquisador deve conferir se todas as etapas previstas no roteiro foram concluídas. Veja a seguir um possível roteiro para uma coleta de dados, exemplificado para o caso da empresa M, apresentada no início dessa seção de autoestudo:

1. Definir o objetivo da pesquisa. Exemplo: determinar o perfil dos funcionários da empresa M.
2. Definir as variáveis e a população de interesse. Exemplo: idade, peso, altura, sexo, cor dos olhos, raça, satisfação em relação às condições de trabalho e à remuneração. A população corresponde aos 30 mil funcionários da empresa M.
3. Definir o sistema de coleta. Exemplo: será realizado um censo ou uma amostragem? No caso de uma amostragem, qual método será utilizado? Qual é o tamanho da amostra? Quais são os meios de obtenção dos dados (telefonemas, questionários, entrevistas, etc.)?
4. Coletar os dados. Nessa etapa é necessário que o pesquisador tome o cuidado de não criar um viés. Exemplo: é possível que se tenha respostas enviesadas realizando perguntas como: você NÃO está feliz com o seu trabalho? Você acha que está ganhando POUCO? Perguntas com negativas ou com ênfase em determinados termos podem influenciar as respostas dos entrevistados.
5. Revisar os dados coletados. Essa etapa é muito importante para a coleta, pois é possível que sejam identificados erros que podem impactar todo o restante do trabalho. Exemplo: determinado funcionário da empresa M pode ter respondido que seu nome é João da Silva e também que é do sexo feminino. Será que

essa resposta é verdadeira? Vale a pena conferir o processo para verificar possíveis erros de coleta.

Esperamos que até o momento você tenha tido uma visão geral de como é feita a amostragem e sua importância para a realização de uma pesquisa. Vale ressaltar que o explicitado aqui é apenas uma noção básica do processo. Existem livros inteiros dedicados ao estudo desse tema e muitos materiais disponíveis na internet.



Pesquise mais

Para se aprofundar nas técnicas de amostragem, faça uma pesquisa sobre o assunto. Algumas sugestões são:

Livros

- BOLFARINE, Heleno; BUSSAB, Wilton de O. **Elementos de amostragem**. São Paulo: Edgard Blucher, 2005.
- SILVA, Nilza N. da. **Amostragem probabilística**: um curso introdutório. 2. ed. São Paulo: Editora da Universidade de São Paulo, 2004.
- Internet
- **Receita Federal do Brasil**: <<http://www.receita.fazenda.gov.br/manuaisweb/exportacao/topicos/conferencia-aduaneira/verificacao-fisica/amostragem.htm>>. Acesso em: 29 abr. 2015.
- **Tribunal de Contas da União**: <<http://portal2.tcu.gov.br/portal/pls/portal/docs/2064402.PDF>>. Acesso em: 29 abr. 2015.

Sem Medo de Errar!

Vamos retomar a situação-problema do início dessa seção de autoestudo e personalizar um roteiro para a coleta de dados, incluindo detalhes específicos como os propostos no tópico ROTEIRO PARA UMA COLETA DE DADOS.

1. Definir o objetivo da pesquisa.

O objetivo da pesquisa pode ser identificado na situação-problema proposta no início dessa seção de autoestudo: determinar o perfil dos 30 mil funcionários da empresa M.

2. Definir as variáveis e a população de interesse.

A tarefa à qual você foi incumbido especificava que o objetivo era determinar o perfil dos funcionários da empresa M. Para realizar essa tarefa, você deveria pesquisar idade, peso, altura, sexo, cor dos olhos, raça, satisfação em relação às condições de trabalho e à remuneração. Portanto, essas serão as variáveis de estudo. Para cada uma delas você deve elaborar uma pergunta de forma imparcial, por exemplo:

- a) Qual é a sua idade?
- b) Qual é o seu peso?
- c) Qual é a sua altura?
- d) Qual é o seu sexo?
- e) Qual é a cor de seus olhos?
- f) Qual é a sua raça?
- g) Qual é a sua satisfação em relação às condições de trabalho?
- h) Qual é a sua satisfação em relação à sua remuneração?

Observe que algumas das perguntas anteriores deixam espaço para respostas muito amplas, a exemplo do item (g). Alguns funcionários poderiam responder "me sinto bem", outros poderiam dizer apenas "nota 10". Como comparar essas respostas posteriormente? Nessa etapa do planejamento é melhor incluir algumas restrições para as respostas para facilitar a análise *a posteriori*. Algumas sugestões são:

- a) Qual é a sua idade? _____ anos
- b) Qual é o seu peso? _____ kg (coloque valores inteiros)
- c) Qual é a sua altura? _____ centímetros
- d) Sexo: () Masculino () Feminino
- e) Cor dos olhos: () Castanhos () Azuis () Verdes
- f) Raça: () Amarela () Branca () Indígena () Parda () Preta

g) De 0 (insatisfeito) a 10 (muito satisfeito), qual é a sua satisfação em relação às condições de trabalho?

() 0 () 1 () 2 () 3 () 4 () 5 () 6 () 7 () 8 ()
9 () 10

h) De 0 (insatisfeito) a 10 (muito satisfeito), qual é a sua satisfação em relação à sua remuneração?

() 0 () 1 () 2 () 3 () 4 () 5 () 6 () 7 () 8 ()
9 () 10

A população corresponde aos 30 mil funcionários da empresa M.

3. Definir o sistema de coleta. Lembre-se de que para a situação proposta no início dessa seção de autoestudo o prazo estipulado para a pesquisa é de uma semana. Nesse caso, consideramos que fazer uma amostragem é o mais prudente, pois, devido ao tempo de coleta e àquele que ainda será gasto no tratamento das informações, realizar um censo seria inviável. Entretanto, essa escolha irá depender de quais recursos estão disponíveis, ficando a cargo do pesquisador definir entre amostragem ou censo.

Em relação ao tipo de amostragem, vamos supor que na empresa M, 5% dos funcionários sejam gerentes, 15% possuam cargos administrativos e os 80% restantes, cargos operacionais. Como a pesquisa pode ser influenciada pela variável "cargo", o mais adequado é realizar uma amostragem estratificada em que os estratos "gerentes", "cargos administrativos" e "cargos operacionais" possuam a mesma representatividade na amostra tal qual é observada na população.

Não entraremos em detalhes sobre o cálculo do tamanho da amostra, pois ele envolve conceitos ainda não trabalhados. Entretanto, para este exemplo, vamos utilizar uma sugestão disponível em <<https://pt.surveymonkey.com/mp/sample-size>> com uma margem de erro de 10% para os resultados obtidos. Você poderá verificar nesse *link* que o tamanho sugerido para a amostra é 96.

Resta ainda definir como serão coletados os dados. Novamente, isso irá depender de quais recursos estão disponíveis. Uma sugestão para pequenas amostras é a utilização de um formulário a ser enviado via internet. Um exemplo de ferramenta gratuita para criação desse tipo de formulário é o Formulários Google. Veja mais informações sobre este recurso em <<https://www.google.com/intl/pt-BR/forms/about>>.

4. Coletar os dados. No caso da utilização da ferramenta sugerida anteriormente, os formulários podem ser enviados por *e-mail*. Entretanto, nessa etapa, o pesquisador deve ficar atento se os indivíduos selecionados para a amostra realmente estão respondendo à pesquisa.
5. Revisar os dados coletados. Com a ferramenta sugerida anteriormente, essa etapa tende a ser ágil, visto também que o tamanho da amostra é pequeno.

Avançando na Prática

Pratique mais!	
Instrução	
Desafiamos você a praticar o que aprendeu transferindo seus conhecimentos para novas situações que pode encontrar no ambiente de trabalho. Realize as atividades e depois as compare com a de seus colegas.	
1. Competência de fundamentos de área	Conhecer os conceitos matemáticos básicos e proporcionar o desenvolvimento do raciocínio lógico e quantitativo.
2. Objetivos de aprendizagem	Elaborar um roteiro para a realização de uma pesquisa.
3. Conteúdos relacionados	Amostragem.
4. Descrição da situação problema	Suponha que uma companhia telefônica o contratou para elaborar uma pesquisa de satisfação com os clientes acerca do serviço prestado. Elabore um roteiro para coleta de dados sabendo que a companhia possui 1 milhão de clientes e deseja realizar a pesquisa por amostragem com 1000 clientes.

<p>5. Resolução da situação problema</p>	<ol style="list-style-type: none"> 1. Definir o objetivo da pesquisa. Determinar a satisfação dos clientes acerca do serviço prestado pela companhia telefônica. 2. Definir as variáveis e a população de interesse. Como não foram indicadas restrições e o detalhamento acerca da pesquisa foi pequeno, assumiremos que a empresa queira saber apenas a satisfação de seus clientes (variável de interesse), sem outras informações agregadas. A população corresponde a um milhão de clientes da companhia. 3. Definir o sistema de coleta. Como os clientes devem constar na base de dados da empresa, podemos considerar que é possível gerar uma lista em ordem alfabética com os nomes e os telefones de cada um. Sendo assim, uma amostragem sistemática pode atender às necessidades, e o meio de coleta das informações será o contato por telefone (algo natural para uma empresa de telefonia). O tamanho da amostra já foi definido anteriormente (1000 indivíduos). Resta determinar o que será perguntado aos clientes. Novamente, pela ausência de detalhes, assumiremos que uma pergunta como "Em uma escala de 0 a 10, sendo 0 ruim e 10 ótimo, qual nota o(a) senhor(a) atribuiria ao serviço prestado por esta companhia?" seja suficiente para obter as informações requeridas. 4. Coletar os dados. Empregar pessoal devidamente treinado para realizar os contatos por telefone. Nessa etapa é interessante ressaltar que o treinamento dos entrevistadores é muito importante, pois se pode causar um viés. 5. Revisar os dados coletados. Havendo peculiaridade nos dados coletados, eles podem ser revisados com base nas gravações telefônicas das entrevistas, caso esse recurso esteja disponível.
--	--



Lembre-se

População: conjunto de todos os elementos que possuem determinada característica em comum.

Amostra: qualquer subconjunto de uma população.

Variável: determinada característica que se deseja estudar em uma população ou amostra. Subdivide-se em: quantitativa discreta, quantitativa contínua, qualitativa ordinal e qualitativa nominal.

Amostragem: processo de coleta de dados em um subconjunto da população, denominado amostra. Os principais tipos são: amostragem de conveniência, amostragem voluntária, amostragem aleatória simples, amostragem sistemática, amostragem aleatória estratificada, e amostragem por conglomerado.



Faça você mesmo

Reúna-se com seus colegas e planejem a criação de uma eleição (mesmo que fictícia) para presidente de turma. Após a definição dos candidatos, elaborem roteiros para pesquisas de intenção de voto, explorando vários métodos de amostragem.

Realizem pelos menos duas pesquisas por amostragem (sistemática e aleatória simples, por exemplo) e depois realizem a eleição, em que todos devem votar.

Ao final, comparem o resultado da eleição com os obtidos nas pesquisas de intenção de voto, verificando qual método obteve o resultado mais representativo. Redijam um pequeno texto com as conclusões.

Faça Valer a Pena!

1. Assinale a alternativa que apresenta características de censo.

- a) Em uma linha de produção, uma em cada 50 peças é inspecionada para controle de qualidade.
- b) A contagem da população realizada pelo IBGE, em 2007, por questões de custos, envolveu a coleta de dados em municípios de até 170 mil habitantes e em mais 21 municípios com população acima desta quantidade.
- c) O Teste Rápido de HIV é feito com a retirada de uma gota de sangue do paciente. O sangue é colocado em dois dispositivos de testagem. Em caso de resultado positivo nos dois, o diagnóstico já é dado como certo.
- d) Nas eleições municipais, a cada quatro anos, todos os eleitores são obrigados a ir às urnas para votar em dois representantes, um candidato a vereador e um candidato a prefeito.
- e) Ultimamente algumas lojas têm instalado painéis eletrônicos ou urnas para que os clientes, caso queiram, possam deixar suas opiniões sobre o atendimento.

2. Assinale a alternativa que apresenta uma variável quantitativa discreta.

- a) Altura.
- b) Sexo.
- c) Peso.

- d) Velocidade.
- e) Número de filhos.

3. Assinale a alternativa que apresenta uma variável qualitativa nominal.

- a) Cor dos olhos.
- b) Escolaridade (Ensino Fundamental, Ensino Médio, Ensino Superior).
- c) Idade.
- d) Classificação em uma corrida (Primeiro, Segundo, Terceiro, ...).
- e) Cor da faixa de um judoca (Cinza, Azul, Amarela, Laranja, ...).

4. Para controle de qualidade, 10% das peças que saem de uma linha de produção são inspecionadas. Para compor a amostra seleciona-se 1 em cada 10, sempre na ordem em que são produzidas, ou seja, produzem-se 9 peças e retira-se a décima; produzem-se mais 9 e a vigésima é selecionada para a amostra; e assim por diante. Essa amostragem é do tipo:

- a) conveniência.
- b) aleatória simples.
- c) sistemática.
- d) aleatória estratificada.
- e) conglomerado.

5. Uma grande rede de lojas pretende consultar os consumidores de determinada região para determinar suas preferências na hora de comprar roupas. O foco dessa rede é o mercado feminino, que corresponde a 70% de seu faturamento. Os 30% restantes correspondem ao público masculino. Para fazer essa consulta, que tipo de amostragem essa rede de lojas deve utilizar?

- a) Amostragem por conveniência.
- b) Amostragem aleatória simples.
- c) Amostragem sistemática.
- d) Amostragem aleatória estratificada.
- e) Amostragem por conglomerado.

6. Conceitue população e amostra.

7. Descreva as características de uma amostragem aleatória estratificada.

Seção 2.2

Métodos tabulares e métodos gráficos

Diálogo aberto

Nessa seção vamos dar continuidade ao estudo da estatística descritiva. A primeira etapa foi trabalhada na Seção 2.1, em que você aprendeu alguns conceitos básicos de estatística e os principais tipos de amostragem. Além disso, foi abordada a seguinte situação-problema: imagine que você é um funcionário da empresa M e que foi incumbido de realizar uma pesquisa para determinar o perfil dos 30 mil funcionários. O relatório final dessa pesquisa deverá conter informações pessoais como idade, peso, altura, sexo, cor dos olhos, raça e também informações sobre a satisfação em relação às condições de trabalho e à remuneração. Considere que o prazo estipulado para a realização dessa tarefa seja uma semana.

Fizemos um planejamento da coleta de dados para a pesquisa da empresa M e elaboramos um roteiro para a realização de tal tarefa. Para continuar o trabalho, vamos considerar que os dados já foram coletados, sendo a nossa tarefa agora organizá-los, resumir-los e apresentá-los.



Refleta

"Dados são fatos; em si não trazem grande significado; só depois que eles forem, de alguma forma, agrupados ou processados é que poderemos ver o significado ser revelado".

Marcello Martinelli, professor associado da Universidade de São Paulo

Não Pode Faltar!

Dados brutos

A coleta de dados planejada na Seção 2.1 terá como resultado fichas como a exemplificada na Figura 2.4.

Figura 2.4 | Exemplo de ficha-resposta da coleta de dados

Nº 1

a) Qual é a sua idade? 26 anos.

b) Qual é o seu peso? 74 kg (coloque valores inteiros).

c) Qual é a sua altura? 174 centímetros.

d) Sexo: (X) Masculino () Feminino.

e) Cor dos olhos: (X) Castanhos () Azuis () Verdes.

f) Raça: () Amarela () Branca () Indígena (X) Parda () Preta.

g) De 0 (insatisfeito) a 10 (muito satisfeito), qual é a sua satisfação em relação às condições de trabalho?

() 0 () 1 () 2 () 3 () 4 () 5 () 6 () 7 (X) 8 () 9 () 10

h) De 0 (insatisfeito) a 10 (muito satisfeito), qual é a sua satisfação em relação à sua remuneração?

() 0 () 1 () 2 () 3 () 4 () 5 () 6 (X) 7 () 8 () 9 () 10

Fonte: O autor (2015)

Lembre-se de que no planejamento da coleta de dados com os funcionários da empresa M o tamanho pretendido para a amostra era 96. Com essa quantidade de fichas em mãos fica difícil inferir algo sobre o perfil dos funcionários, pois os dados não estão organizados, ou seja, estão de **forma bruta**. Para facilitar a visualização, precisamos agrupar os dados que se referem à mesma variável e dispor as informações de um modo mais agradável e intuitivo para leitura e apresentação.

Figura 2.5 | Registro da ficha Nº 1 em uma planilha eletrônica

	A	B	C	D	E	F	G	H	I
1	Nº	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)
2	1	21	74	174	M	C	Parda	8	7
3									

Fonte: O autor (2015).

Uma maneira interessante de organizar os dados registrados nas fichas é em uma planilha eletrônica, pois ela permite uma manipulação simples e rápida dos dados. Observe na Figura 2.5 uma possível maneira de registrar as respostas obtidas na ficha da Figura 2.1.

Observe na Figura 2.5 que indicamos o número da ficha para eventuais conferências, pois, em caso de dúvidas, fica simples localizá-la. Veja também que, para ficar sucinto, indicamos somente as letras para fazer correspondência às perguntas realizadas. Outra tática utilizada foi registrar apenas as letras M e C para representar as respostas “Masculino” e “Castanhos”. Note que na pergunta (f) não pudemos utilizar essa estratégia, pois causaríamos confusão entre as respostas “Parda” e “Preta”.

Para tornar o processo de aprendizagem mais simples, vamos supor que o tamanho da nossa amostra seja apenas 20 indivíduos, cujas respostas dadas ao questionário estão apresentadas na Tabela 2.1.

Tabela 2.1 | Respostas recebidas na coleta de dados

N°	Questão							
	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)
1	21	74	174	M	C	Parda	8	7
2	21	93	176	F	A	Parda	5	4
3	25	86	171	F	C	Amarela	1	5
4	27	83	179	F	C	Preta	4	5
5	28	88	185	M	C	Preta	10	7
6	29	63	181	M	C	Amarela	5	4
7	31	60	177	M	V	Parda	5	5
8	35	58	163	F	C	Branca	5	4
9	37	84	180	F	C	Preta	4	4
10	37	81	165	M	C	Branca	5	4
11	39	74	175	M	C	Parda	9	8
12	42	85	162	M	C	Indígena	7	7
13	43	60	165	M	C	Parda	3	4
14	47	67	170	M	C	Parda	4	3
15	48	81	162	M	A	Branca	2	4
16	51	85	165	F	C	Preta	5	3
17	51	86	170	M	A	Branca	1	5
18	53	57	170	F	C	Branca	7	6
19	55	88	179	F	C	Parda	10	6
20	59	68	188	F	A	Branca	9	8

Fonte: O autor (2015).

Distribuição de frequências

Geralmente, quando se estuda determinado problema, procura-se conhecer o comportamento das variáveis envolvidas, observando a ocorrência de seus valores na amostra para poder inferir algo a respeito da população. Figura 2.4 | Exemplo de ficha-resposta da coleta de dados.1 apresenta os dados de forma bruta, de modo

que não conseguimos extrair muita informação ao observá-la. Dessa maneira, para podermos fazer algum tipo de inferência, precisamos organizar os valores obtidos para cada variável de forma mais simples e de modo que uma rápida leitura possa fornecer informações que ainda estão ocultas.

Tabela 2.2 | Distribuição de frequências da variável idade na amostra

Faixa etária	Frequência	Proporção	Porcentagem
20 – 30	6	0,30	30
30 – 40	5	0,25	25
40 – 50	4	0,20	20
50 – 60	5	0,25	25
Total	20	1,00	100

Fonte: O autor (2015).

Uma das maneiras mais utilizadas para organizar dados são as **tabelas de distribuição de frequências**. Veja na Tabela 2.2 como podemos dispor os dados relativos à idade da amostra de funcionários da empresa M.

A notação 20 |– 30 indica que estão sendo considerados os valores maiores ou iguais a 20 e menores que 30. Os demais são interpretados de modo semelhante.

A coluna “Frequência” da Tabela 2.2 (também denominada **frequência absoluta**) é construída por contagem direta dos valores dispostos na coluna (a) da Tabela 2.1. As demais colunas exigem alguns cálculos. Para calcular as:

- **Proporções** (também conhecidas como **frequências relativas**), dividimos as frequências absolutas pela soma de todas as frequências.

$$\text{Exemplo: } 0,30 = \frac{6}{20}$$

- **Porcentagens**, multiplicamos as proporções por 100%.

$$\text{Exemplo: } 30\% = 0,30 \cdot 100\%$$

Ao observarmos a distribuição de frequências da variável “idade”, conseguimos obter informações que antes não estavam visíveis. Por exemplo, podemos afirmar que a faixa etária de 20 anos ou mais e menos de 30 anos tem 30% dos funcionários, ou, ainda, que mais da metade dos funcionários (30%+25%) têm menos de 40 anos.



Para calcularmos uma **proporção**, dividimos a frequência absoluta correspondente pela soma de todas as frequências.

Uma **porcentagem** é calculada multiplicando uma proporção por 100%.

A Tabela 2.2 está organizada em **intervalos de classe**. Essa estratégia é bastante utilizada quando os valores obtidos são muito variados (não há repetição).



Algumas vezes não é necessário agrupar os valores obtidos de uma variável em intervalos de classe, como é o caso da variável "número de títulos em mundiais de futebol". Até 2015, os campeões das 20 edições do torneio eram: Brasil (5 títulos), Itália (4 títulos), Alemanha (4 títulos), Uruguai (2 títulos), Argentina (2 títulos), França (1 título), Inglaterra (1 título) e Espanha (1 título). A distribuição de frequências dessa variável pode ser observada na Tabela 2.3.

Tabela 2.3 | Distribuição de frequências da variável "número de títulos em mundiais de futebol"

Número de títulos	Frequência	Proporção	Porcentagem
1	3	0,375	37,5
2	2	0,250	25,0
3	1	0,125	12,5
4	1	0,125	12,5
5	1	0,125	12,5
Total	8	1,000	100,0

Fonte: FUTPÉDIA (2015)

Desafio: construa as distribuições de frequências das demais variáveis presentes na Tabela 2.1 e compare sua resposta com a esperada no apêndice da Seção 2.2.

Rol

Você deve ter tido um pouco mais de trabalho para construir a tabela de distribuição de frequências da variável "peso" se compararmos com o empenho que seria necessário para construir a Tabela 2.2. Isso se deve ao fato de os dados referentes a essa variável não estarem organizados do modo como estão os referentes à variável "idade" (vide Tabela 2.1).

Como já foi mencionado, quando os dados estão desordenados dizemos que eles estão de forma bruta, a exemplo dos dados referentes

à variável "peso". Quando os dados estão organizados em uma ordem crescente ou decrescente, dizemos que eles estão em **rol**, como é o caso dos dados relativos à variável "idade".



Exemplificando

Organize o seguinte conjunto de dados em rol crescente e em rol decrescente:

18 – 11 – 35 – 26 – 22 – 16

Resolução:

Rol crescente: 11 – 16 – 18 – 22 – 26 – 35

Rol decrescente: 35 – 26 – 22 – 18 – 16 – 11



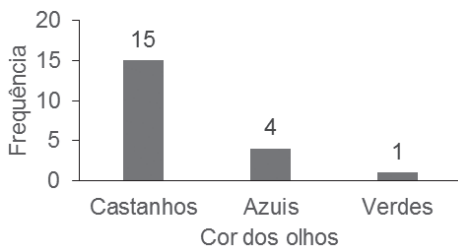
Assimile

Um **rol** é uma organização crescente ou decrescente de dados.

Gráficos estatísticos

Além de tabelas, também podemos representar informações em gráficos. Esse tipo de representação tem forte apelo visual, sendo atrativo aos olhos dos leitores, e, muitas vezes, pode dar uma ideia melhor da variabilidade de um conjunto de dados do que uma tabela. São inúmeros os tipos de representações gráficas tanto para variáveis qualitativas quanto para quantitativas. Abordaremos apenas os tipos mais simples e caberá a você pesquisar outras representações e suas particularidades.

Figura 2.6 | Cor dos olhos da amostra de funcionários da empresa M



Fonte: Os autores (2015)

- **Gráficos para variáveis qualitativas**

Um tipo particular de gráfico é o de barras. Observe na Figura 2.6 um exemplo desse tipo de gráfico, elaborado a partir da Tabela 2.1.

Num gráfico de barras, o eixo horizontal é denominado **eixo das categorias** e o vertical, **eixo das quantidades**. Cada barra representa um dos valores assumidos pela variável (uma categoria). A altura da barra é proporcional à frequência com que determinada resposta aparece na amostra e a largura é constante para todas as barras.

A construção de gráficos de barras pode ser facilitada em papel milimetrado ou em uma planilha eletrônica. Veja um vídeo tutorial de como utilizar uma planilha para construção de um gráfico de barras em <<https://www.youtube.com/watch?v=PrqDotQ7B54>>.

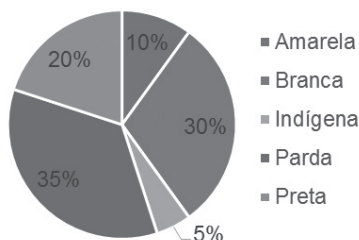
Outro tipo bastante utilizado de gráfico é o de setores, também conhecido como gráfico de "pizza". São muito úteis para visualizar a participação (frequência) de determinada categoria em relação ao todo. Esse tipo de gráfico é construído com base em um círculo dividido, a partir de seu centro, em quantas partes for o número de valores possíveis para a variável em questão.

Na Figura 2.7 é apresentado um gráfico de setores construído a partir da tabela de distribuição de frequências para a variável "raça", cuja construção foi proposta anteriormente e que consta no apêndice da Seção 2.2.

Cada setor de um gráfico de "pizza" corresponde a um possível valor da variável. Esse setor terá tamanho proporcional à participação desse valor em relação ao todo. Essa participação fica facilmente visualizável quando observamos a coluna "Porcentagem" da tabela de distribuição de frequências. Veja que os valores percentuais calculados para essa coluna são apresentados no interior do gráfico de setores.

Em determinadas situações, a participação de certas categorias em relação ao todo é tão pequena que convém agrupar seus valores em um único setor para facilitar a visualização.

Figura 2.7 | Distribuição dos funcionários da empresa M



Fonte: O autor (2015).



Observe na Tabela 2.4 a quantidade de mandioca (em kg) produzida pelos estados da região Nordeste em 2006.

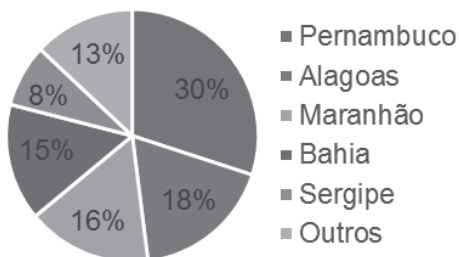
Tabela 2.4 | Produção de mandioca (em kg) na região Nordeste em 2006

UF	Quantidade (em kg)	Porcentagem
Pernambuco	2401684	30%
Alagoas	1479204	18%
Maranhão	1315186	16%
Bahia	1246801	15%
Sergipe	685133	8%
Piauí	394665	5%
Ceará	426183	5%
Rio Grande do Norte	139452	2%
Paraíba	82627	1%

Fonte: O autor (2015).

Para construir um gráfico de setores a partir dessa tabela teríamos problemas de visualização, pois alguns setores ficariam muito pequenos. Em casos como este, geralmente, agrupam-se os setores menores em um só, denominado "Outros". Veja um exemplo na Figura 2.8.

Figura 2.8 | Participação dos estados na produção de mandioca da região Nordeste em 2006



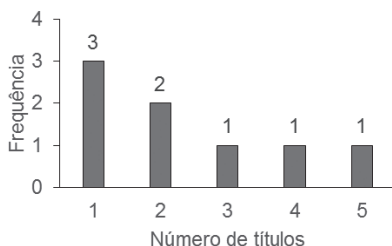
Fonte: IBGE - Produção vegetal (2015)

• Gráficos para variáveis quantitativas

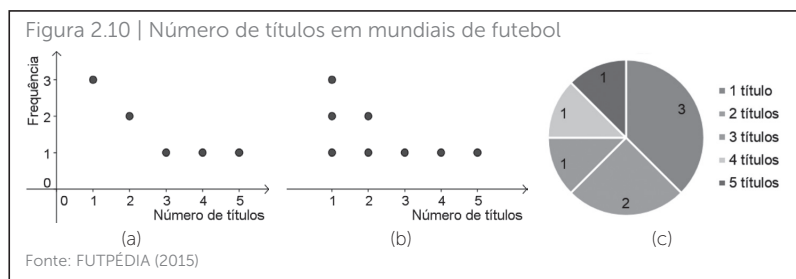
Gráficos de barra também são bastante utilizados para representar variáveis quantitativas. Veja um exemplo na Figura 2.9, construído a partir da Tabela 2.3.

Vale ressaltar que, quando se trata de variáveis quantitativas, a variedade de representações gráficas é maior. A mesma informação apresentada na Figura 2.9 também poderia ser representada de outras formas, como mostra a Figura 2.10.

Figura 2.9 | Número de títulos em mundiais de futebol



Fonte: FUTPÉDIA (2015)



As representações da Figura 2.10 (a) e da Figura 2.10 (b) são denominadas gráficos de dispersão. Nesses exemplos, a variável em questão é discreta. Quando trabalhamos com variáveis quantitativas contínuas precisamos fazer uma adaptação para construir representações semelhantes. Considere a distribuição de frequências para a variável “altura”, cuja construção foi proposta anteriormente e é apresentada na Tabela 2.5 com alguns acréscimos.

Veja na Tabela 2.5 que foi acrescentada a coluna “Ponto médio”. Nessa coluna temos os:

- **pontos médios** das classes, que são obtidos somando-se os valores extremos e dividindo-se por 2.

O ponto médio da classe servirá como representante na hora de construir uma representação gráfica. Veja na Figura 2.11 um gráfico de barras construído a partir da Tabela 2.5.

Tabela 2.5 | Distribuição de frequências da variável “altura” apresentada na Tabela 2.1

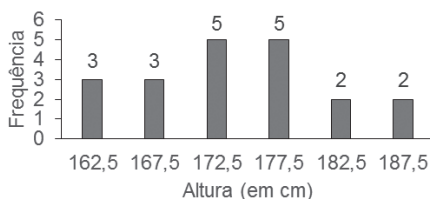
Altura (em cm)	Ponto médio	Frequência	Proporção	Porcentagem
160 – 165	162,5	3	0,15	15
165 – 170	167,5	3	0,15	15
170 – 175	172,5	5	0,25	25
175 – 180	177,5	5	0,25	25
180 – 185	182,5	2	0,10	10
185 – 190	187,5	2	0,10	10
Total	-	20	1,00	100

Fonte: O autor (2015).

Outra maneira de representar os dados da Tabela 2.5 é por meio de um **histograma**, o qual pode ser visto na Figura 2.12.

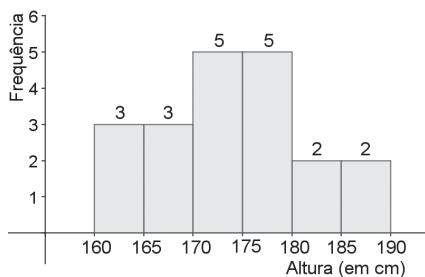
Em um histograma, cada barra possui a largura do intervalo da classe e altura proporcional à frequência correspondente. Uma das principais características do histograma é que suas barras são justapostas, ou seja, “grudadas” umas às outras.

Figura 2.11 | Altura dos funcionários da amostra da empresa M



Fonte: O autor (2015).

Figura 2.12 | Altura dos funcionários da amostra da empresa M



Fonte: O autor (2015).

Para criar um histograma de modo rápido e fácil, acesse o *site* <<http://www.socscistatistics.com/descriptive/histograms/Default.aspx>> e insira os dados relativos às alturas da amostra, disponíveis na Tabela 2.1.

Ramos-e-folhas

A ideia geral ao se construir um gráfico, seja de barras, dispersão ou histograma, é determinar como os dados se distribuem. A linguagem simplificada e resumida de um gráfico é de grande ajuda nessa tarefa.

Uma grande desvantagem ao construir uma representação gráfica é a perda de informação dos dados originais. Para facilitar na visualização da distribuição dos dados e ainda não perder muita informação dos dados brutos, podemos utilizar um diagrama de **ramos-e-folhas**.

Veja na Figura 2.13 um diagrama construído a partir da Tabela 2.1 para a variável “altura”.

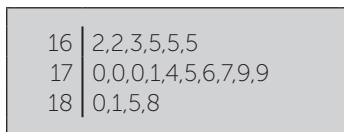
162	162	163	165	165	165	170
170	170	171	174	175	176	177
179	179	180	181	185	188	

Para compreender como é construído um diagrama como esse, vamos repetir no quadro ao lado os dados relativos às alturas dos funcionários da amostra, organizados em rol e destacado o último algarismo em cada observação.

Note que para construir o diagrama, traçamos uma linha vertical e: à esquerda dessa linha dispomos os dois primeiros algarismos das observações; à direita dessa linha dispomos os últimos algarismos das observações, separados por vírgula.

Analisando o diagrama da Figura 2.13 podemos observar a distribuição dos valores (há prevalência de valores na classe do meio) e ainda mantivemos as informações originais.

Figura 2.13| Diagrama de ramos-e-folhas para a variável "altura"



Fonte: O autor (2015).



Exemplificando

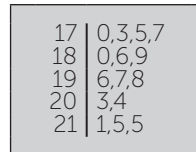
Observe o diagrama de ramos-e-folhas da Figura 2.14.

Como poderíamos escrever os dados desse diagrama por extenso?

Resolução:

170, 173, 175, 177, 180, 186, 189, 196, 197, 198, 203, 204, 211, 215, 215

Figura 2.14 | Peso de uma amostra de 15 novilhos



Fonte: O autor (2015).

Em estatística existem diversas normas para apresentação de dados em tabelas e gráficos. Determinados elementos são considerados "essenciais", e, quando compõem um desses objetos, é preciso se atentar às normas. Há também uma maneira padrão para a elaboração de uma tabela de distribuição de frequências agrupadas em intervalos de classe. Veja no *link* <<http://www.ee.usp.br/graduacao/ens435/modulo4/modulo4l.html>> alguns dos elementos considerados essenciais na apresentação de tabelas e gráficos e no *link* <<http://www.fernandokb.pro.br/?p=201>> a forma padrão para a construção de uma tabela de distribuição de frequências.



Pesquise mais

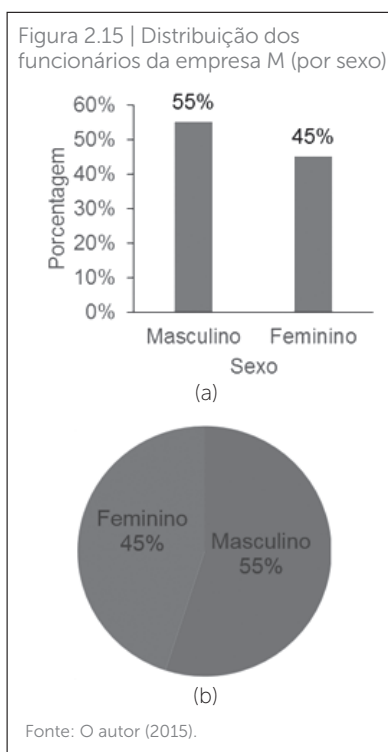
Além dos tipos de gráficos apresentados, existem diversos outros, sendo que cada um possui uma característica própria que o faz mais adequado para determinada situação. Para se aprofundar nesse assunto, segue uma sugestão de leitura:

- Suporte Office. Disponível em: <<https://support.office.com/pt-br/article/Tipos-de-gr%C3%A1ficos-dispon%C3%ADveis-b22a8bb9-a673-4d7f-b481-aa747c48eb3d?ui=pt-BR&rs=pt-BR&ad=BR>>. Acesso em: 15 jul. 2015.

Sem Medo de Errar!

Relembrando a situação-problema proposta no início da seção, vamos agora nos preocupar com a apresentação dos dados da amostra apresentada na Tabela 2.1 e, para termos maior riqueza de detalhes, vamos tratar apenas das variáveis “sexo” e “peso”. As demais podem ser trabalhadas de modo semelhante. Considerando essas duas variáveis, qual seria a melhor forma de apresentá-las? Com um gráfico ou uma tabela? Como seria essa apresentação?

Métodos tabulares são aplicáveis na maioria das vezes, tanto para variáveis qualitativas quanto para quantitativas. Têm a vantagem de, em geral, possuir maior riqueza de detalhes do que uma representação gráfica. Contudo, tabelas podem ser de difícil leitura, trabalho que pode ser simplificado por meio de um gráfico. Como a construção das tabelas de distribuição de frequências para essas variáveis já foi proposta anteriormente (e consta no apêndice dessa seção), vamos apresentar os dados de forma gráfica. Para isso, resta decidir qual tipo de gráfico é mais adequado para cada uma.



Primeiramente, vamos tratar da variável “sexo”, que é qualitativa nominal. Nessa seção foram mostrados dois tipos de gráficos para apresentação de dados relativos a esse tipo de variável: o de barras e o de setores. Ambos estão representados na Figura 2.15.

Para que a representação gráfica seja adequada, sempre devemos considerar o objetivo do pesquisador. Não existe uma rigidez quanto à escolha do tipo de gráfico, mas vale observar as sugestões apresentadas nas leituras sugeridas no “Pesquise Mais!”. Nesse exemplo, por questões estéticas, consideramos como mais adequado o gráfico de setores, pois ele deixa visível a participação da cada categoria com relação ao todo.

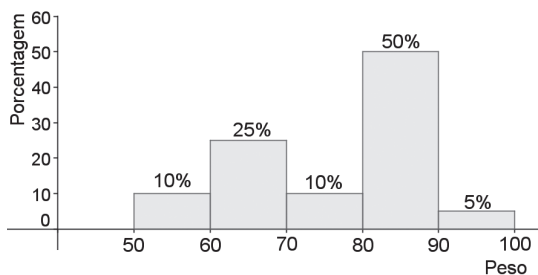
Dando continuidade, a variável “peso” é quantitativa contínua. Para escolher uma entre as representações gráficas apresentadas, são válidas algumas observações:

- **Gráfico de barras:** para elaborar um gráfico desse tipo, temos que considerar a mesma estratégia utilizada para elaborar o gráfico da Figura 2.8, ou seja, determinar o ponto médio de cada intervalo de classe. Acreditamos que com isso perderíamos um pouco de informação.

- **Gráfico de dispersão:** como não há muitos valores repetidos para a variável “peso”, um gráfico desse tipo poderia ser de difícil leitura.

- **Gráfico de setores:** os dados estão agrupados na tabela de distribuição de frequências (veja apêndice) em classes justapostas. Logo, um gráfico de setores pode não ser adequado ou pouco informativo para esse exemplo.

Figura 2.16 | Peso dos funcionários da empresa M



Fonte: O autor (2015).

Considerando as observações anteriores e a semelhança entre esse exemplo e o apresentado na Figura 2.12, concluímos que a representação gráfica mais adequada é um histograma, que está representado na Figura 2.16.

AVANÇANDO NA PRÁTICA

Pratique mais!

Instrução

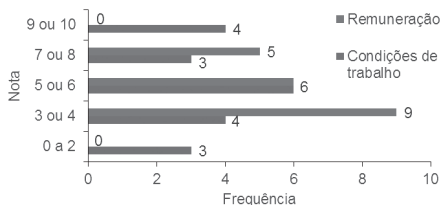
Desafiamos você a praticar o que aprendeu transferindo seus conhecimentos para novas situações que pode encontrar no ambiente de trabalho. Realize as atividades e depois as compare com as de seus colegas.

1. Competências de fundamentos de área	Conhecer os conceitos matemáticos básicos e proporcionar o desenvolvimento do raciocínio lógico e quantitativo.								
2. Objetivos de aprendizagem	Expor dados da amostra por meio de gráficos e tabelas.								
3. Conteúdos relacionados	Métodos tabulares e métodos gráficos para apresentação de dados.								
4. Descrição da situação problema	<p>O IMC (Índice de massa corpórea) é um indicador reconhecido pela Organização Mundial da Saúde para dimensionar a relação entre peso e altura de um indivíduo. Para calculá-lo, dividimos o peso (em kg) pela altura (em metros) ao quadrado, ou seja, $IMC = \frac{\text{Peso}}{\text{Altura}^2}$. Conhecendo-se o IMC, podemos utilizar a seguinte classificação¹:</p>								
	<table border="1"> <tbody> <tr> <td>Menos de 18,5</td> <td>Abaixo do peso</td> </tr> <tr> <td>18,5 – 25,0</td> <td>Peso adequado</td> </tr> <tr> <td>25,0 – 30,0</td> <td>Sobrepeso</td> </tr> <tr> <td>30,0 ou mais</td> <td>Obeso</td> </tr> </tbody> </table>	Menos de 18,5	Abaixo do peso	18,5 – 25,0	Peso adequado	25,0 – 30,0	Sobrepeso	30,0 ou mais	Obeso
	Menos de 18,5	Abaixo do peso							
	18,5 – 25,0	Peso adequado							
	25,0 – 30,0	Sobrepeso							
30,0 ou mais	Obeso								
De acordo com essas informações e com base nos dados da Tabela 2.1, construa:									
<ul style="list-style-type: none"> Um gráfico de barras duplas (horizontais) para as variáveis "satisfação em relação às condições de trabalho" e "remuneração". Uma tabela de distribuição de frequências para o IMC. 									

5. Resolução da situação problema

Para construir o gráfico de barras, vamos considerar os dados da tabela de distribuição de frequências para as variáveis "satisfação em relação às condições de trabalho" e "remuneração", cuja construção foi proposta anteriormente e cujos dados constam no apêndice da seção 2.2. Veja esse gráfico na Figura 2.17.

Figura 2.17 | Satisfação em relação às condições de trabalho e remuneração da amostra de funcionários da empresa M



Fonte: O autor (2015).

Para construir a tabela de distribuição de frequências, primeiro precisamos calcular os IMCs.

Nº	Peso (em kg)	Altura (em metros)	IMC	Nº	Peso (em kg)	Altura (em metros)	IMC
1	74	1,74	24,4	11	74	1,75	24,2
2	93	1,76	30,0	12	85	1,62	32,4
3	86	1,71	29,4	13	60	1,65	22,0
4	83	1,79	25,9	14	67	1,70	23,2
5	88	1,85	25,7	15	81	1,62	30,9
6	63	1,81	19,2	16	85	1,65	31,2
7	60	1,77	19,2	17	86	1,70	29,8
8	58	1,63	21,8	18	57	1,70	19,7
9	84	1,80	25,9	19	88	1,79	27,5
10	81	1,65	29,8	20	68	1,88	19,2

Agora, a partir desses dados, podemos construir a distribuição de frequências apresentada na Tabela 2.6.

Tabela 2.6 | Distribuição de frequências do IMC

Classificação	Frequência	Proporção	Porcentagem
Abaixo do peso	0	0,00	0
Peso adequado	9	0,45	45
Sobrepeso	8	0,40	40
Obeso	3	0,15	15
Total	20	1,00	100

Fonte: O autor (2015).



Lembre-se

Dados brutos são aqueles que não estão ordenados.

Uma **tabela de distribuição de frequências** é uma organização tabular que busca relacionar os possíveis valores de uma variável às frequências com que aparecem numa amostra.

Rol é uma organização crescente ou decrescente dos dados de uma amostra.



Faça você mesmo

Organize-se com seus colegas e pesquisem por amostragem a altura dos alunos da turma. Depois, construa uma tabela de distribuição de frequências agrupadas em intervalos de classe e um histograma.

Por fim, compare sua resposta com a dos colegas.

Faça valer a pena!

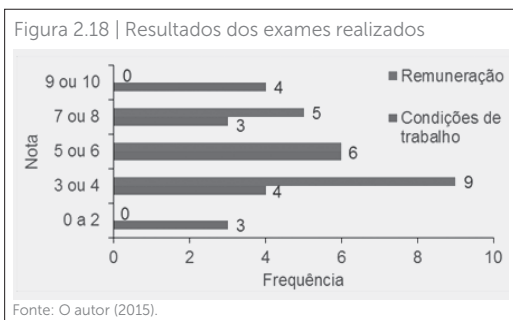
1. Assinale a alternativa que contém dados organizados em rol.

- a) 71, 88, 56, 65, 99 b) 95, 92, 75, 60, 44 c) 20, 56, 67, 53, 70
d) 32, 81, 59, 17, 38 e) 87, 62, 37, 76, 61

2. Com base nas informações da Tabela 2.7, assinale a alternativa que contém a sequência de valores x , y , z , w , nessa ordem.

- a) 64, 24, 9, 3
b) 62, 24, 9, 5
c) 64, 22, 9, 5
d) 60, 28, 9, 5
e) 66, 21, 8, 5

3. Em determinado hospital foram realizados exames em diversos pacientes com suspeita de uma doença. Os resultados foram compilados no gráfico da Figura 2.18.



Assinale a alternativa que contém os dados brutos que deram origem a esse gráfico.

- a) N,N,N,N,N,N,P,P,P,P,P,P,P,I,I
- b) N,N,N,N,P,P,P,P,P,I
- c) N,N,N,N,N,N,N,P,P,P,P,P,P,P,P,P,I
- d) N,N,N,N,N,N,N,P,P,P,P,P,P,P,I,I,I
- e) N,N,N,N,N,N,P,P,P,P,P,P,P,P,P,I,I,I

4. A Tabela 2.8 apresenta a distribuição da população brasileira em 2010 por faixa etária.

Tabela 2.8 | Distribuição da população brasileira em 2010

Faixa etária	Frequência (em milhões)	Porcentagem
0 – 20	62,89	33,10
20 – 40	x	33,30
40 – 60	y	22,60
60 ou mais	z	w
Total	190,00	100,00

Fonte: O autor (2015).

Com base nos dados da tabela, assinale a alternativa correta.

- a) Metade da população tinha mais 40 anos.
- b) Aproximadamente 30 milhões de brasileiros tinham 60 anos ou mais.
- c) Um terço da população tinha menos de 40 anos.
- d) 18% da população tinha 60 anos ou mais.
- e) Mais de 63 milhões de brasileiros possuíam entre 20 e 40 anos.

5. O histograma da Figura 2.19 mostra a distribuição dos salários ganhos pelos funcionários de uma empresa.

Tabela 2.9 | População brasileira por região em 2010

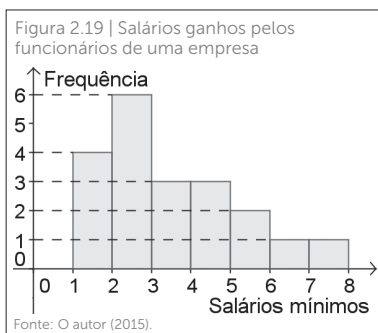
Região	População (em milhões)
Norte	15,9
Nordeste	53,1
Sudeste	80,4
Sul	27,4
Centro-Oeste	14,1

Fonte: O autor (2015).

Com base no histograma, assinale a alternativa correta:

- Seis funcionários ganham até 3 salários mínimos.
- Sete funcionários ganham 5 salários mínimos ou mais.
- 10% dos funcionários ganham 6 salários mínimos ou mais.
- 25% dos funcionários ganham menos de 2 salários mínimos.
- Metade dos funcionários ganha mais de 4 salários mínimos.

6. A Tabela 2.9 mostra a quantidade de habitantes no Brasil em 2010.



Elabore um gráfico de setores a partir da Tabela 2.9 e em cada setor indique a porcentagem correspondente.

7. Elabore um diagrama de ramos-e-folhas a partir do conjunto de dados a seguir.

132 259 188 573 540 458 663 780 614 937 872
170 312 223 601 559 535 687 782 645 953 895

Seção 2.3

Medidas de posição

Diálogo aberto

Na Seção 2.2, você aprendeu sobre diferentes formas de apresentar informações, como as tabelas e os gráficos. Vimos que as tabelas são úteis para organizar e resumir dados, mas que, em alguns casos, podem ser de difícil leitura e interpretação. Nesse sentido, uma representação gráfica pode ser mais fácil de interpretar, resume ainda mais os dados e dá uma ideia melhor da distribuição.

Existem ainda outras maneiras de resumir conjuntos de dados, que vão além de uma tabela ou um gráfico. Ferramentas para esse fim são denominadas **medidas de posição**, as quais buscam sintetizar um conjunto com um único valor. São exemplos de medidas de posição: a **média aritmética**, a **mediana** e a **moda**.

Considere a Tabela 2.1 e imagine maneiras de representar os dados referentes a cada variável por um único valor. Como você faria isso? Que valor seria mais adequado para cada conjunto?

Essas perguntas serão respondidas ao final dessa seção de autoestudo. Leia o texto e se aprofunde no assunto com a sugestão de leitura do "Pesquise mais!". Não deixe também de consultar outras bibliografias.



Reflita

"Estatística: a ciência que diz que se eu comi um frango e tu não comeste nenhum, teremos comido, em média, meio frango cada um."

Dino Segrè, escritor italiano

Não pode faltar!

Você aprendeu anteriormente que **dados brutos** são aqueles que se apresentam da maneira como foram coletados, ou seja,

fora de ordem. Também vimos que ao ordenar esses dados em ordem crescente ou decrescente estamos construindo um **rol**. Veja um exemplo:

- Dados brutos: 18 – 42 – 31 – 26 – 21 – 24 – 20 – 90
- Rol crescente: 18 – 20 – 21 – 24 – 26 – 31 – 42 – 90
- Rol decrescente: 90 – 42 – 31 – 26 – 24 – 21 – 20 – 18

A construção de um rol é o primeiro passo para a confecção de tabelas e gráficos. Entretanto, essa não é sua única utilidade. O rol será de grande ajuda na obtenção de algumas **medidas de posição**. Estas, por sua vez, são valores que buscam resumir ainda mais um conjunto de dados, mais até do que as tabelas e os gráficos. As medidas mais conhecidas que possuem essa finalidade são: a **média aritmética**, a **mediana** e a **moda**. Trataremos de cada uma mais adiante, mas antes temos que adotar algumas notações para simplificar a escrita.

Na Seção 2.2 atribuímos valores às variáveis estudadas na amostra de funcionários. Eram elas: idade, peso, altura, sexo, cor dos olhos, raça, satisfação em relação às condições de trabalho e satisfação em relação à remuneração. Para não ser necessário reescrever repetidamente os nomes dessas variáveis, utilizamos letras maiúsculas para representá-las, como a seguir:

- | | | |
|---|------------------|-----------|
| A: idade | B: peso | C: altura |
| D: sexo | E: cor dos olhos | F: raça |
| G: satisfação em relação às condições de trabalho | | |
| H: satisfação em relação à remuneração | | |

Com essa padronização, sempre que quisermos nos referir à variável “satisfação em relação às condições de trabalho”, por exemplo, podemos escrever simplesmente variável G. Bem mais simples, não concorda?

Outro procedimento bastante utilizado é enumerar os elementos de um conjunto de dados, geralmente quando eles já estão organizados em rol. Para exemplificarmos, considere a variável definida a seguir:

X: idade dos leitores de uma revista

Admita que em uma pesquisa realizada para estudar a variável X se tenham sido obtidos os seguintes valores, já organizados em rol:

Dados da amostra: 18 – 20 – 21 – 24 – 26 – 31 – 42 – 90

Como os dados se referem à variável X, comumente simbolizamos cada um pela letra x (minúscula) acompanhada do índice i , que indica a posição que o valor aparece no rol. O quadro a seguir resume essa associação.

Posição no rol (i)	1	2	3	4	5	6	7	8
x_i	18	20	21	24	26	31	42	90

Ao escrevermos o símbolo x_3 , por exemplo, estamos nos referindo ao valor obtido para a variável X que ocupa a posição $i = 3$ (terceira posição) do rol, isto é, $x_3 = 21$.

No quadro anterior foram apresentados oito valores obtidos a partir de uma amostra. Essa quantidade geralmente é simbolizada pela letra n (minúscula). Nesse caso, temos que $n = 8$ é o tamanho da amostra.



Atenção

Na continuidade dessa seção, com frequência os conjuntos de dados serão apresentados em **rol crescente** para facilitar a compreensão das medidas de posição. Entretanto, fique atento ao trabalhar com os dados brutos em outras situações, pois você terá, primeiro, que organizá-los em ordem crescente ou decrescente para calcular algumas dessas medidas.

Média aritmética

A **média aritmética** (ou simplesmente média) corresponde à divisão da soma de todos os valores de um conjunto de dados pela quantidade de valores desse conjunto. Se um conjunto tiver n valores, $\{x_1, x_2, x_3, \dots, x_n\}$, sua média será simbolizada por:

$$\bar{x} = \frac{\sum x}{n} = \frac{x_1 + x_2 + x_3 + \dots + x_n}{n}$$



Calcule a média do seguinte conjunto de dados:

18 – 20 – 21 – 24 – 26 – 31 – 42 – 90

Resolução:

$$\bar{x} = \frac{\sum x}{n} = \frac{x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7 + x_8}{8}$$
$$\bar{x} = \frac{18 + 20 + 21 + 24 + 26 + 31 + 42 + 90}{8} = \frac{272}{8} = 34$$

Leitura dos símbolos:

\bar{x} : x barra

$\sum x$: soma dos valores de x

A média é uma das mais importantes medidas de posição. Veja que o conjunto de dados do exemplo anterior corresponde ao da amostra da variável X apresentada anteriormente. O resultado obtido nesse exemplo nos indica que, **em média**, os leitores da referida revista possuem 34 anos. Pergunta: o valor 34 descreveu bem o conjunto de dados? Por quê?

A resposta esperada para essa pergunta seria não. A média aritmética é uma medida fortemente influenciada por valores extremos (muito baixos ou muito altos), motivo que nos leva, quase sempre, a não a utilizar sozinha para descrever um conjunto. Em inúmeros casos essa medida é utilizada conjuntamente com outras, como a mediana e medidas de dispersão (que medem a variabilidade de um conjunto). Quando o conjunto de valores é mais homogêneo, a média cumpre eficientemente o papel de descrevê-lo, como no exemplo a seguir.

Conjunto: 82 – 82 – 83 – 83 – 83 – 84 – 84; Média: $\bar{x} = 83$

Dizer que nesse conjunto os valores são, em média, 83 não é longe do esperado.

Média aritmética ponderada

Considere a Tabela 2.10 que apresenta as notas de um aluno nas 4 avaliações de uma disciplina.

Tabela 2.10 | Notas de um aluno

Avaliação	Trabalho 1	Prova 1	Trabalho 2	Prova 2
i	1	2	3	4
Peso (p_i)	3	7	4	6
Nota (x_i)	9,0	8,0	8,5	7,0

Fonte: O autor (2015).

Como determinar a média final do aluno, visto que cada avaliação tem importância diferente? Em situações como essa, em que alguns elementos de um conjunto têm mais importância do que outros, costuma-se utilizar a média ponderada para resumir os dados.

A **média aritmética ponderada** (\bar{x}_p) de um conjunto de dados é calculada ao multiplicarmos os números por seus respectivos pesos e dividirmos a soma desses produtos pela soma dos pesos. Para o exemplo anterior, temos:

$$\bar{x}_p = \frac{9,0 \cdot 3 + 8,0 \cdot 7 + 8,5 \cdot 4 + 7,0 \cdot 6}{3 + 7 + 4 + 6} = \frac{159}{20} = 7,95$$

Portanto, a média final do aluno na disciplina é 7,95 pontos.

Simbolicamente, representamos a média ponderada de x_1, x_2, \dots, x_n valores cujos pesos são, respectivamente, p_1, p_2, \dots, p_n , por

$$\bar{x}_p = \frac{x_1 \cdot p_1 + x_2 \cdot p_2 + \dots + x_n \cdot p_n}{p_1 + p_2 + \dots + p_n} = \frac{\sum x \cdot p}{\sum p}$$



Exemplificando

Considere que em um concurso os candidatos devam realizar três testes: conhecimentos gerais (CG), conhecimentos específicos (CE) e aptidão física (AF), sendo que cada etapa possui um peso diferente. Na Tabela 2.11 estão os resultados obtidos por dois candidatos.

Se para ser aprovado é necessário obter nota final igual a 8 ou superior, qual dos candidatos foi aprovado?

Tabela 2.11 | Pontuação dos candidatos em um concurso

Teste	CG	CE	AF
Peso	2	5	3
Marcio	10,0	5,0	9,0
Lucas	8,0	9,0	7,0

Fonte: O autor (2015).

Resolução:

Efetuamos os cálculos das médias:

$$\text{Marcio: } \bar{x}_p = \frac{\sum x \cdot p}{\sum p} = \frac{10,0 \cdot 2 + 5,0 \cdot 5 + 9,0 \cdot 3}{2 + 5 + 3} = 7,2$$

$$\text{Lucas: } \bar{x}_p = \frac{\sum x \cdot p}{\sum p} = \frac{8,0 \cdot 2 + 9,0 \cdot 5 + 7,0 \cdot 3}{2 + 5 + 3} = 8,2$$

Portanto, o candidato aprovado é Lucas.

Mediana

Considere a variável Y como sendo "os salários dos funcionários de uma empresa" e que os valores amostrados para essa variável foram: 840 – 860 – 790 – 780 – 1800 – 880 – 2800. A média dos salários é $\bar{y} = 1250$, contudo, ao afirmar isso não descrevemos o quadro de salários satisfatoriamente, visto que grande parte dos valores é próxima de 800 reais. Em casos como este, a mediana pode ser uma boa opção para descrever o conjunto.

A **mediana** (ou valor mediano) de um conjunto de dados corresponde ao valor central de um rol. Para calculá-la temos de considerar dois casos: (1º) quantidade ímpar de valores no conjunto; (2º) quantidade par de valores no conjunto.

1º caso: quantidade ímpar de valores no conjunto

No caso da amostra colhida para a variável Y , considere o seguinte rol:

i	1	2	3	4	5	6	7
y_i	780	790	840	860	880	1800	2800

Como $n = 7$ (ímpar), a mediana (simbolizada por Md) corresponde ao valor que ocupa a posição $i = (n + 1)/2 = (7 + 1)/2 = 4$, ou seja, $Md = 860$. Veja que abaixo e acima de 860 temos 3 valores, isto é, a mediana divide o rol ao meio, em que metade dos valores é menor ou igual à mediana e a outra metade é maior ou igual à mediana.

Afirmar que o salário mediano dos trabalhadores da referida empresa é 860 reais corresponde melhor a uma descrição do conjunto do que a média.

2º caso: quantidade par de valores no conjunto

Considere Z o "número diário de visitantes em um museu" e a amostra coletada para essa variável como sendo o conjunto: 80 – 73 – 92 – 98 – 160 – 77. Nesse caso, temos $n = 6$ (par) elementos na amostra. Ao organizar os dados em rol, obtemos:

i	1	2	3	4	5	6
z_i	73	77	80	92	98	160

Observe que agora não temos um único valor no centro do rol, mas dois deles. Um dos valores está localizado na posição $i = n/2 = 6/2 = 3$ e o outro na posição $i = n/2 + 1 = 6/2 + 1 = 3 + 1 = 4$. Para representar a mediana nesse caso, utilizamos a média aritmética dos dois valores centrais, ou seja, $Md = (z_3 + z_4)/2 = (80 + 92)/2 = 86$.



Exemplificando

Calcule a mediana dos valores amostrados das variáveis X e Y apresentados a seguir.

i	1	2	3	4	5	6	7	8
x_i	102	103	135	144	148	159	160	166
y_i	413	484	495	543	558	565	580	-

O símbolo (-) indica que o dado é ausente.

Resolução:

Observe que o tamanho da amostra da variável X é $n_x = 8$ (par) e o tamanho da amostra para Y é $n_y = 7$ (ímpar). Assim, para calcular a mediana de X (Md_x) temos que utilizar o 2º caso e para a mediana de Y (Md_y) o 1º caso, como segue:

- mediana de X :

$$i = n/2 = 8/2 = 4 \rightarrow \text{primeira posição central}$$

$$i = n/2 + 1 = 8/2 + 1 = 4 + 1 = 5 \rightarrow \text{segunda posição central}$$

$$Md_x = (x_4 + x_5)/2 = (144 + 148)/2 = 146$$

- mediana de Y:

$$i = (n+1)/2 = (7+1)/2 = 8/2 = 4 \rightarrow \text{posição central}$$

$$Md_y = y_4 = 543$$

Moda

O que lhe vem à cabeça quando falamos a palavra moda? Aquela roupa descolada que bastante gente está usando? Pois bem, em estatística essa palavra tem um sentido semelhante.

A **moda**, simbolizada por **Mo**, é o valor com maior frequência em um conjunto de dados. Lembra-se da tabela apresentada na seção 2.2 que continha os dados referentes à amostra de 20 funcionários da empresa M (veja Tabela 2.1)? Os dados referentes à variável raça (F) estão reproduzidos a seguir.

Parda – Parda – Amarela – Preta – Preta – Amarela – Parda – Branca – Preta – Branca – Parda – Indígena – Parda – Parda – Branca – Preta – Branca – Branca – Parda – Branca

A distribuição de frequências desse conjunto de dados é apresentada na Tabela 2.11. Como pode ser observado na coluna “Frequência”, o valor que teve a maior frequência foi “Parda”. Portanto, para este exemplo, **Mo = Parda**.

Tabela 2.12 | Distribuição de frequências da variável F

Raça	Frequência	Proporção	Porcentagem
Amarela	2	0,10	10
Branca	6	0,30	30
Indígena	1	0,05	5
Parda	7	0,35	35
Preta	4	0,20	20
Total	20	1,00	100

Fonte: O autor (2015).

A moda é uma medida de posição indicada tanto para variáveis quantitativas quanto para qualitativas, como é o caso da variável “raça”. Isso não ocorre, por exemplo, com a média e a mediana, que são medidas indicadas somente para variáveis quantitativas.

Quando os possíveis valores de uma variável aparecem em igual número de vezes em uma amostra (têm a mesma frequên-

cia), dizemos que o conjunto é **amodal**, isto é, não possui moda. Também podemos classificar os conjuntos em **unimodais** (uma moda), **bimodais** (duas modas), **trimodais** (três modas) e **multimodais** (quatro ou mais modas).



Assimile

A média aritmética, a mediana e a moda são medidas que buscam resumir um conjunto de dados em um único valor.

- Para calcular a **média aritmética**, adicionamos todos os valores e dividimos o resultado pela quantidade de valores adicionados. Se a média aritmética for **ponderada**, devemos levar em consideração os respectivos pesos.
- A **mediana** divide o conjunto de dados ao meio. Ela corresponde ao valor central em um rol, se a quantidade de valores for ímpar, e à média aritmética dos dois valores centrais, se a quantidade for par.
- A **moda** é o valor com maior frequência em um conjunto.



Pesquise mais

Apresentamos aqui os cálculos de algumas medidas de posição para dados não agrupados em classes. Essas mesmas medidas podem ser utilizadas também para dados agrupados (em tabelas de distribuição de frequências). Para conhecer o cálculo dessas medidas para dados agrupados e outras medidas de posição, como as separatrizes, sugerimos a seguinte leitura complementar:

- Descrição de amostras. Disponível em: <<http://www.ufpa.br/dicas/biome/bioamos.htm>>. Acesso em: 15 jul. 2015.

Sem Medo de Errar!

Vamos relembrar as questões feitas no início dessa seção de autoestudo: considerando a Tabela 2.1, como representar os dados referentes a cada variável por um único valor? Que valor seria mais adequado para cada conjunto?

Trataremos aqui das variáveis A (idade) e D (sexo). O tratamento das demais será proposto na seção "AVANÇANDO NA PRÁTICA".

Variável A

Iremos utilizar a média aritmética e a mediana para descrever as idades. Para calcular a média, efetuamos:

$$\bar{a} = \frac{\sum a}{n} = \frac{21+21+25+\dots+53+55+59}{20} = \frac{779}{20} = 38,95 \cong 39$$

Logo, os funcionários da empresa têm, em média, 39 anos, aproximadamente. Para o cálculo da mediana, observe que o conjunto possui $n = 20$ (par) valores. Portanto, a mediana corresponderá à média aritmética dos elementos nas posições:

$$i = n/2 = 20/2 = 10;$$

$$i = n/2 + 1 = 20/2 + 1 = 10 + 1 = 11.$$

Assim, $Md = (a_{10} + a_{11})/2 = (37 + 39)/2 = 38$. Concluímos então que a mediana das idades é 38 anos. Observe que a mediana e a média, sozinhas, não descrevem muito bem o conjunto de dados. Isso ocorre porque há muita variabilidade. Temos funcionários com 21 anos e outros com 59. Veremos na seção 2.4 que há ferramentas que auxiliam as medidas de posição a descrever o conjunto. Por enquanto nos limitaremos a apresentar estes valores como descrição.

Variável D

Para descrever os dados referentes à variável D não podemos utilizar a média e a mediana, visto que essa variável é qualitativa. Nos resta então utilizar a moda. Repetimos aqui os dados para contabilizar as frequências:

M – F – F – F – M – M – M – F – F – M – M – M – M – M – M – F
– M – F – F – F

Observe que temos 11 funcionários do sexo masculino e 9 do sexo feminino, resultando em $Mo = M$, ou seja, a maioria dos funcionários é do sexo masculino.

Pratique mais!

Instrução

Desafiamos você a praticar o que aprendeu transferindo seus conhecimentos para novas situações que pode encontrar no ambiente de trabalho. Realize as atividades e depois as compare com as de seus colegas.

1. Competências de fundamentos de área	Conhecer os conceitos matemáticos básicos e proporcionar o desenvolvimento do raciocínio lógico e quantitativo.
2. Objetivos de aprendizagem	Resumir dados por meio de uma medida de posição
3. Conteúdos relacionados	Medidas de posição
4. Descrição da situação problema	Considerando os dados apresentados na Tabela 2.1, calcule: - o peso médio; - a altura mediana; - a cor dos olhos e a raça de maior frequência; - a moda dos dados referentes às variáveis G e H.
5. Resolução da situação problema	<p>- Peso médio</p> <p>Para calcular a média dos pesos efetuamos:</p> $\bar{b} = \frac{\sum b}{n} = \frac{74+93+86+\dots+57+88+68}{20} = \frac{1521}{20} = 76,05 \cong 76$ <p>Portanto, o peso médio dos funcionários é aproximadamente 76 quilogramas.</p> <p>- Altura mediana</p> <p>Para calcular a mediana, primeiro precisamos construir um rol, como a seguir:</p> <p>162 – 162 – 163 – 165 – 165 – 170 – 170 – 170 – 171 – 174 – 175 – 176 – 177 – 179 – 179 – 180 – 181 – 185 – 188</p> <p>Como $n=20$ (par), a mediana corresponderá à média dos valores nas posições:</p> $i = n / 2 = 20 / 2 = 10 ;$ $i = n / 2 + 1 = 20 / 2 + 1 = 10 + 1 = 11 .$ <p>Assim, $Md = (c_{10} + c_{11}) / 2 = (171 + 174) / 2 = 172,5 \cong 173$. Concluímos que a altura mediana é aproximadamente 173 centímetros.</p> <p>- Cor dos olhos e raça de maior frequência</p>

	<p>Para determinar a moda de cada conjunto de dados utilizaremos as tabelas de distribuição de frequências, cuja construção foi proposta na seção 2.2 e cujo gabarito consta no apêndice.</p> <p>O valor modal para a: cor dos olhos foi "Castanhos" com porcentagem igual a 75%; raça foi "Parda" com porcentagem igual a 35%.</p> <p>- Moda dos dados referentes às variáveis G e H</p> <p>Para determinarmos a moda de cada conjunto, primeiro construímos um rol, como a seguir:</p> <p>Variável G: 1 - 1 - 2 - 3 - 4 - 4 - 5 - 5 - 5 - 5 - 5 - 5 - 7 - 7 - 8 - 9 - 9 - 10 - 10</p> <p>Variável H: 3 - 3 - 4 - 4 - 4 - 4 - 4 - 4 - 4 - 5 - 5 - 5 - 5 - 6 - 6 - 7 - 7 - 7 - 8 - 8</p> <p>Com o rol disponível podemos observar que o valor de maior frequência em cada caso é 5 e 4, para as variáveis G e H, respectivamente.</p>
--	---



Lembre-se

Para calcular a **média aritmética**, adicionamos todos os valores e dividimos o resultado pela quantidade de valores adicionados. Se a média aritmética for **ponderada**, devemos levar em consideração os respectivos pesos.

A **mediana** divide o conjunto de dados ao meio. Ela corresponde ao valor central em um rol, se a quantidade de valores for ímpar, e à média aritmética dos dois valores centrais, se a quantidade for par.

A **moda** é o valor com maior frequência em um conjunto.



Faça você mesmo

Na Seção 2.2, no tópico "Faça você mesmo", foi proposto que, junto com seus colegas, você pesquisasse a altura dos alunos da turma. Aproveite os dados coletados nessa pesquisa e calcule a média e a mediana do conjunto. Também classifique o conjunto com relação à quantidade de modas e indique os valores modais.

Faça Valer a Pena!

1. Assinale a alternativa que contém o conjunto com a maior média.

- a) $409 - 337 - 104$
- b) $131 - 115 - 302$
- c) $395 - 404 - 369$
- d) $250 - 432 - 1562$
- e) $258 - 156 - 223$

2. Assinale a alternativa que contém a média aritmética do conjunto de dados sintetizado no diagrama de ramos-e-folhas ao lado.

- a) 138
- b) 139
- c) 140
- d) 141
- e) 142

10	8,8
11	6
12	0,7
13	1,1
14	4
15	0,8,8,9,9
16	4,7

3. Observe o seguinte conjunto de dados.

$$12 - 27 - 16 - 42 - 16 - 23 - 41 - 25$$

Com relação à média, à mediana e à moda do conjunto anterior, assinale a alternativa correta:

- a) $Mo < \bar{x} < Md$
- b) $Mo < Md < \bar{x}$
- c) $Md < Mo < \bar{x}$
- d) $Md < \bar{x} < Mo$
- e) $\bar{x} < Mo < Md$

4. Observe a Tabela 2.13 em que constam as idades de 20 crianças que participam de um projeto social.

Assinale a alternativa que contém a média e a mediana das idades das crianças.

Dica: escreva o rol correspondente à distribuição de frequências.

a) $\bar{x} = 9$ e $Md = 7$

b) $\bar{x} = 9$ e $Md = 8$

c) $\bar{x} = 9$ e $Md = 9$

d) $\bar{x} = 8$ e $Md = 9$

e) $\bar{x} = 7$ e $Md = 9$

Tabela 2.13 | Distribuição de frequências

Idade	Frequência	Proporção	Porcentagem
7	4	0,20	20
8	4	0,20	20
9	5	0,25	25
10	3	0,15	15
11	3	0,15	15
12	1	0,05	5
Total	20	1,00	100

Fonte: O autor (2015).

5. Em uma turma de 40 alunos há alguns com 20 anos, 21 anos, 24 anos e ainda outros com 28 anos. A quantidade de alunos com cada uma dessas idades é apresentada na Figura 2.20.

Assinale a alternativa que contém a média das idades dos alunos dessa turma.

a) 24,2 anos

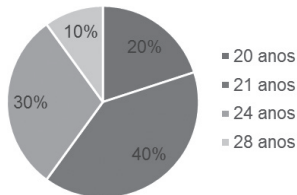
b) 24,0 anos

c) 22,0 anos

d) 24,4 anos

e) 22,4 anos

Figura 2.20 | Distribuição das idades



Fonte: O autor (2015).

6. Calcule a média, a mediana e a moda do seguinte conjunto de dados.

1 – 1 – 2 – 2 – 2 – 3 – 3 – 3 – 3 – 4

4 – 4 – 5 – 5 – 6 – 6 – 6 – 6 – 6 – 7

7 – 7 – 7 – 7 – 8 – 8 – 9 – 9 – 9 – 9

Depois, classifique o conjunto com relação à quantidade de modas.

7. Considere o seguinte conjunto de dados:

1932 – y – 1596 – 1649 – 1597

Calcule o valor de y sabendo que a média do conjunto é 1646.

Seção 2.4

Medidas de dispersão

Diálogo aberto

Vimos na unidade anterior que existem maneiras ainda mais sintéticas de resumir um conjunto de dados do que as tabelas e os gráficos. Tais métodos envolvem a obtenção de um único valor (ou poucos valores) para representar todo o conjunto, valor esse que denominamos medida de posição. As medidas que estudamos foram a média aritmética, a média aritmética ponderada, a mediana e a moda.

No exemplo apresentado na seção SEM MEDO DE ERRAR! constatamos que nem sempre uma medida como a média representa significativamente um conjunto. Isso também pode ser observado nos conjuntos a seguir:

1° conjunto: 90 – 90 – 90 – 90 – 90

2° conjunto: 86 – 88 – 90 – 92 – 94

3° conjunto: 30 – 60 – 90 – 120 – 150

Os conjuntos possuem média e mediana iguais a 90 (calcule!), entretanto, apenas para os dois primeiros esse valor é representativo. Aqui surgem alguns questionamentos: quando uma média é representativa em um conjunto? Quais ferramentas podem ser utilizadas para auxiliar as medidas de posição na descrição de um conjunto de dados?

Para auxiliar as medidas de posição na descrição de um conjunto utilizamos as **medidas de dispersão**. Essas medidas buscam dimensionar quanto os dados estão distantes da média, por exemplo. Com o auxílio delas podemos decidir, por exemplo, se a média pode ser utilizada como representante de um conjunto.

No decorrer dessa seção buscaremos responder aos questionamentos anteriores e, mais especificamente, decidir se a média é adequada para resumir os dados referentes aos funcionários da empresa M (apresentados na Tabela 2.1) e quantificar a variabilidade de cada conjunto de dados.



"Dispersão (ou variabilidade) de um conjunto refere-se à maior ou menor diversificação dos valores de uma variável em torno de um valor de tendência central tomado como ponto de comparação".

Carlos Augusto de Medeiros, chefe da Unidade de Administração Geral da Fundação Universidade Aberta do Distrito Federal

Não Pode Faltar!

Desvio

Vamos considerar os dados do 1º, 2º e 3º conjuntos apresentados anteriormente como sendo provenientes de censo das variáveis X , Y e Z , respectivamente. Denominamos **desvio** a diferença de um valor do conjunto com relação à média. Para os conjuntos de dados apresentados anteriormente, temos os desvios calculados na Tabela 2.14 (lembre-se de que $\bar{x} = \bar{y} = \bar{z} = 90$).

Tabela 2.14 | Desvios dos conjuntos de dados

i	Valores do conjunto			Desvios		
	x_i	y_i	z_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$z_i - \bar{z}$
1	90	86	30	$90 - 90 = 0$	$86 - 90 = -4$	$30 - 90 = -60$
2	90	88	60	$90 - 90 = 0$	$88 - 90 = -2$	$60 - 90 = -30$
3	90	90	90	$90 - 90 = 0$	$90 - 90 = 0$	$90 - 90 = 0$
4	90	92	120	$90 - 90 = 0$	$92 - 90 = 2$	$120 - 90 = 30$
5	90	94	150	$90 - 90 = 0$	$94 - 90 = 4$	$150 - 90 = 60$
Total	$\Sigma x = 450$	$\Sigma y = 450$	$\Sigma z = 450$	$\Sigma(x_i - \bar{x}) = 0$	$\Sigma(y_i - \bar{y}) = 0$	$\Sigma(z_i - \bar{z}) = 0$

Fonte: O autor (2015).

Observe que para as amostras das variáveis X , Y e Z a soma de todos os desvios é igual a zero. Isso não ocorre somente para estes conjuntos, mas para todos os conjuntos de dados. Desse modo, qualquer tentativa de utilizar a soma dos desvios $\Sigma(x_i - \bar{x})$ para dimensionar a variabilidade dos dados será frustrada. Isso ocorre, pois os desvios negativos neutralizam os positivos, tornando o total igual a zero.

Para driblar esse contratempo, os estatísticos se utilizam de um artifício matemático, o **valor absoluto**.



O **valor absoluto** de um número corresponde à distância que este se encontra do 0 (zero). A distância é sempre um valor positivo ou zero. Na prática, o valor absoluto de um número: (a) negativo é ele próprio com sinal trocado; (b) não negativo é ele próprio. Exemplos:

O valor absoluto de:

-1, simbolizado por $|-1|$ é igual a 1, ou seja, $|-1| = 1$;

2, simbolizado por $|2|$ é igual a 2, ou seja, $|2| = 2$;

0, simbolizado por $|0|$ é igual a 0, ou seja, $|0| = 0$.

Utilizando o valor absoluto, podemos refazer os cálculos como na Tabela 2.15.

Tabela 2.15 | Valores absolutos dos desvios

<i>i</i>	Valores do conjunto			Valor absoluto dos desvios		
	x_i	y_i	z_i	$ x_i - \bar{x} $	$ y_i - \bar{y} $	$ z_i - \bar{z} $
1	90	86	30	0	4	60
2	90	88	60	0	2	30
3	90	90	90	0	0	0
4	90	92	120	0	2	30
5	90	94	150	0	4	60
Total	$\Sigma x = 450$	$\Sigma y = 450$	$\Sigma z = 450$	$\Sigma x_i - \bar{x} = 0$	$\Sigma y_i - \bar{y} = 12$	$\Sigma z_i - \bar{z} = 180$

Fonte: O autor (2015).

Também podemos simbolizar a soma dos valores absolutos dos desvios por $\Sigma |x - \bar{x}|$, sem o acréscimo do índice *i*. Com a construção da Tabela 2.15, definimos nossa primeira medida de dispersão.

Desvio médio

Desvio médio, simbolizado por *Dm*, é uma medida de dispersão calculada por meio da média aritmética dos valores absolutos dos desvios. Para as variáveis X, Y e Z, temos:

$$Dm(X) = \frac{\Sigma |x - \bar{x}|}{n} = \frac{0}{5} = 0 \quad Dm(Y) = \frac{\Sigma |y - \bar{y}|}{n} = \frac{12}{5} = 2,4 \quad Dm(Z) = \frac{\Sigma |z - \bar{z}|}{n} = \frac{180}{5} = 36$$



Atenção

Quanto menor o desvio médio, menor a dispersão; quanto maior o desvio médio, maior a dispersão dos dados. O menor desvio médio possível é 0 (zero) e ocorre quando os dados são totalmente homogêneos.

Outra maneira de neutralizar o efeito do sinal negativo ocorrido na Tabela 2.14 é elevar cada desvio ao quadrado, como mostra a Tabela 2.16.

Tabela 2.16 | Quadrado dos desvios

i	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$(z_i - \bar{z})^2$
1	$0^2 = 0$	$(-4)^2 = 16$	$(-60)^2 = 3600$
2	$0^2 = 0$	$(-2)^2 = 4$	$(-30)^2 = 900$
3	$0^2 = 0$	$0^2 = 0$	$0^2 = 0$
4	$0^2 = 0$	$2^2 = 4$	$30^2 = 900$
5	$0^2 = 0$	$4^2 = 16$	$60^2 = 3600$
Total	$\sum(x_i - \bar{x})^2 = 0$	$\sum(y_i - \bar{y})^2 = 40$	$\sum(z_i - \bar{z})^2 = 9000$

Fonte: O autor (2015).

A partir da Tabela 2.16 definimos nossa segunda medida de dispersão.

Variância

A **variância**, simbolizada por **Var**, é uma medida de dispersão calculada por meio da média aritmética dos quadrados dos desvios. Para as variáveis X, Y e Z, temos:

$$Var(X) = \frac{\sum(x - \bar{x})^2}{n} = \frac{0}{5} = 0$$

$$Var(Y) = \frac{\sum(y - \bar{y})^2}{n} = \frac{40}{5} = 8$$

$$Var(Z) = \frac{\sum(z - \bar{z})^2}{n} = \frac{9000}{5} = 1800$$

Imagine que os valores observados para as variáveis X, Y e Z sejam idades. Quando elevamos os desvios ao quadrado para o cálculo da variância, obtemos um valor que, teoricamente, tem unidade de medida **idade²** (idade ao quadrado). Como isso pode causar confusão e dificuldade de interpretação, definimos a terceira medida de dispersão.

Atenção

A fórmula apresentada para o cálculo da variância é utilizada somente quando os dados são provenientes da população, ou seja, quando a coleta de dados é feita por meio de censo. No caso de uma

amostragem, a variância do conjunto é calculada por meio da fórmula

$$Var(X) = \frac{\sum(x - \bar{x})^2}{(n-1)}.$$

Desvio padrão

O **desvio padrão**, simbolizado por Dp , é uma medida de dispersão definida como a raiz quadrada da variância. Para as variáveis X , Y e Z , temos:

$$Dp(X) = \sqrt{Var(X)} = \sqrt{0} = 0$$

$$Dp(Y) = \sqrt{Var(Y)} = \sqrt{8} \cong 2,8$$

$$Dp(Z) = \sqrt{Var(Z)} = \sqrt{1800} \cong 42,4$$



Atenção

Ao calcularmos o desvio padrão retornamos à unidade de medida do conjunto de dados, ou seja, se o conjunto de dados é medido em:

- **idade**, a variância é medida em **idade²** e o desvio padrão é medido em **idade**;
- **m** (metros), a variância é medida em **m²** e o desvio padrão é medido em **m**;
- **R\$** (reais), a variância é medida em **R\$²** e o desvio padrão é medido em **R\$**.

As medidas apresentadas até aqui estão de forma absoluta (não percentual). Por esse motivo, ao calculá-las nem sempre conseguimos inferir muita coisa sobre a dispersão de um conjunto de dados. Por exemplo, o valor $Dp(Y) \cong 2,8$ é muito ou pouco? Se não tivermos outro valor para que possamos compará-lo fica difícil fazer alguma afirmação. Por causa disso, definimos nossa quarta medida de dispersão.

Coefficiente de variação

O **coeficiente de variação**, simbolizado por CV , é uma medida de dispersão definida como a razão entre o desvio padrão e a média de um conjunto de dados. Para as variáveis X , Y e Z , temos:

$$CV(X) = \frac{Dp(X)}{\bar{x}} = \frac{0}{90} = 0 \qquad CV(Y) = \frac{Dp(Y)}{\bar{y}} = \frac{2,8}{90} \cong 0,031 \qquad CV(Z) = \frac{Dp(Z)}{\bar{z}} = \frac{42,4}{90} \cong 0,471$$

Também podemos indicar os valores de forma percentual, como a seguir:

$$CV(X) = 0 \cdot 100\% = 0\% \qquad CV(Y) = 0,031 \cdot 100\% = 3,1\% \qquad CV(Z) = 0,471 \cdot 100\% = 47,1\%$$

O coeficiente de variação permite uma comparação do desvio padrão com a média do conjunto de dados. Por exemplo, o desvio padrão de Y corresponde a 3,1% do valor médio do conjunto; o desvio padrão de Z corresponde a 47,1% do valor médio do conjunto. Alguns autores costumam utilizar o coeficiente de variação para classificar um conjunto de dados quanto à dispersão dos valores em torno da média. Essa classificação é feita conforme Tabela 2.17.

O coeficiente de variação também permite comparar conjuntos totalmente distintos quanto à variabilidade dos dados. Veja o exemplo a seguir.

Tabela 2.17 | Classificação de um conjunto de dados

Classificação	Critério
Baixa dispersão	$CV \leq 15\%$
Média dispersão	$15\% < CV < 30\%$
Alta dispersão	$CV \geq 30\%$

Fonte: Os autores (2015)



Exemplificando

Considerando os conjuntos $A = \{2, 3, 6, 9\}$ e $B = \{959, 1065, 1090\}$, qual deles possui os dados mais dispersos em torno da média?

Resolução:

Primeiramente calculamos \bar{a} , \bar{b} , $Var(A)$, $Var(B)$, $Dp(A)$, $Dp(B)$, $CV(A)$ e $CV(B)$.

$$\bar{a} = \frac{\sum a}{n_a} = \frac{2+3+6+9}{4} = 5; \quad \bar{b} = \frac{\sum b}{n_b} = \frac{959+1065+1090}{3} = 1038$$

$$Var(A) = \frac{\sum (a - \bar{a})^2}{n_a} = \frac{(2-5)^2 + (3-5)^2 + (6-5)^2 + (9-5)^2}{4} = 7,5$$

$$Var(B) = \frac{\sum(b - \bar{b})^2}{n_b} = \frac{(959 - 1038)^2 + (1065 - 1038)^2 + (1090 - 1038)^2}{3}$$

$$Var(B) = \frac{6241 + 729 + 2704}{3} \cong 3224,67$$

$$Dp(A) = \sqrt{Var(A)} = \sqrt{7,5} \cong 2,74; \quad Dp(B) = \sqrt{Var(B)} = \sqrt{3224,67} \cong 56,79$$

$$CV(A) = \frac{Dp(A)}{\bar{a}} = \frac{2,74}{5} = 0,548 = 54,8\%; \quad CV(B) = \frac{Dp(B)}{\bar{b}} = \frac{56,79}{1038} \cong 0,055 = 5,5\%$$

Como $CV(A) > CV(B)$, concluímos que o conjunto A é mais disperso que o conjunto B. Além disso, poderíamos acrescentar que A possui alta dispersão e B, baixa dispersão.



Pesquise mais

Existe uma maneira alternativa (mais rápida) para calcular a variância. Para conhecer essa forma alternativa consulte o documento no link a seguir, na página 35. Além disso, existem outras medidas de dispersão além das apresentadas aqui. Para saber mais sobre elas, leia a seção 2.3 do mesmo material indicado a seguir.

- Estatística Descritiva. Disponível em: <http://www.uff.br/ieeanamariafarias/estdesc_2006.pdf>. Acesso em: 27 maio 2015.

Sem Medo de Errar!

Vamos relembrar os questionamentos feitos no início dessa seção:

1. Quando a média é representativa em um conjunto?
2. Quais ferramentas podem ser utilizadas para auxiliar as medidas de posição na descrição de um conjunto de dados?
3. A média é adequada para resumir os dados referentes aos funcionários da empresa M?
4. Como quantificar a variabilidade dos dados referentes a cada variável?

A resposta para a primeira pergunta é: depende dos critérios estabelecidos pelo pesquisador. Geralmente, ao elaborar um

relatório, são definidas determinadas regras/normas, as quais o pesquisador segue fielmente, deixando-as explícitas para os leitores. De modo semelhante, para adotarmos certa padronização, iremos recorrer à Tabela 2.17. Consideraremos a média representativa de um conjunto de dados quando este tiver baixa dispersão.

Em relação à segunda pergunta, esperamos que tenha ficado claro que as medidas de posição são ferramentas que devem ser utilizadas em conjunto com as medidas de dispersão, pois, se um conjunto possui alta variabilidade, pouca informação será fornecida por uma medida pontual.

Para responder à terceira pergunta são necessários alguns dados (os quais podem ser obtidos a partir da Tabela 2.1):

Variável A (idade)	Variável B (peso)	Variável C (altura)
$\bar{a} = 38,95$	$\bar{b} = 76,05$	$\bar{c} = 172,85$
$Md_a = 38$	$Md_b = 81$	$Md_c = 172,5$

Agora, calculamos a variância, o desvio padrão e o coeficiente de variação para cada variável:

$$Var(A) = \frac{\sum(a - \bar{a})^2}{n_a - 1} = \frac{(21 - 38,95)^2 + (21 - 38,95)^2 + \dots + (55 - 38,95)^2 + (59 - 38,95)^2}{20 - 1}$$

$$Var(A) = \frac{2606,95}{19} \cong 137,21 \Rightarrow Dp(A) = \sqrt{137,21} \cong 11,7 \Rightarrow CV(A) = \frac{11,7}{38,95} \cong 0,30 = 30\%$$

$$Var(B) = \frac{\sum(b - \bar{b})^2}{n_b - 1} = \frac{(74 - 76,05)^2 + (93 - 76,05)^2 + \dots + (88 - 76,05)^2 + (68 - 76,05)^2}{20 - 1}$$

$$Var(B) = \frac{2620,95}{19} \cong 137,94 \Rightarrow Dp(B) = \sqrt{137,94} \cong 11,7 \Rightarrow CV(B) = \frac{11,7}{76,05} \cong 0,154 = 15,4\%$$

$$Var(C) = \frac{\sum(c - \bar{c})^2}{n_c - 1} = \frac{(174 - 172,85)^2 + (176 - 172,85)^2 + \dots + (179 - 172,85)^2 + (188 - 172,85)^2}{20 - 1}$$

$$Var(C) = \frac{1148,55}{19} = 60,45 \Rightarrow Dp(C) = \sqrt{60,45} \cong 7,8 \Rightarrow CV(C) = \frac{7,8}{172,85} \cong 0,045 = 4,5\%$$

Como podemos perceber, somente a amostra da variável C possui baixa dispersão. De acordo com o critério estabelecido anteriormente, a média não é representativa das amostras das variáveis A e B, apenas da amostra de C. No caso das variáveis A e B, como a média e a mediana estão muito próximas, também assumiremos que a mediana não é representativa do conjunto, sendo necessário um método gráfico ou tabular para sintetizar os dados.

A mediana seria representativa nos casos em que apenas poucos valores do conjunto se distanciam consideravelmente da média. Quando isso ocorre, geralmente, esses valores são denominados *outliers* (ou valores atípicos).

Avançando na Prática

Pratique mais!	
Instrução	
Desafiamos você a praticar o que aprendeu transferindo seus conhecimentos para novas situações que pode encontrar no ambiente de trabalho. Realize as atividades e depois ascompare com as de seus colegas.	
1. Competências de fundamentos de área	Conhecer os conceitos matemáticos básicos e proporcionar o desenvolvimento do raciocínio lógico e quantitativo.
2. Objetivos de aprendizagem	Compreender a utilização das medidas de dispersão.
3. Conteúdos relacionados	Variância; Desvio padrão; Coeficiente de variação.
4. Descrição da situação problema	<p>Uma área em que a estatística está muito presente é a de controle de qualidade. Geralmente processos industriais procuram uniformidade nos produtos que saem de uma linha de produção. Imagine que uma fábrica de refrigerantes, que envasa embalagens de 1 litro e de 600 mililitros, utilize os seguintes critérios para realizar o controle de qualidade:</p> <ul style="list-style-type: none"> • Serão amostradas sistematicamente para controle de qualidade 5% da produção: a cada 20 embalagens de cada tipo, será retirada 1 para conferência do volume de refrigerante. • Se uma amostra de tamanho $n = 20$ apresentar coeficiente de variação superior a 4%, todo o lote de 400 embalagens correspondente a essa amostra será rejeitado. <p>Com base nesses critérios, analise as amostras a seguir e decida:</p> <ol style="list-style-type: none"> a) Qual das amostras é mais homogênea? b) Qual das amostras causará rejeição do lote de refrigerantes?

	<p>Amostra de embalagens de 1 litro</p> <table border="1"> <tr><td>0,983</td><td>1,025</td><td>1,047</td><td>1,027</td><td>1,013</td><td>0,958</td><td>0,996</td></tr> <tr><td>0,991</td><td>0,960</td><td>1,036</td><td>0,987</td><td>0,971</td><td>0,972</td><td>1,016</td></tr> <tr><td>0,996</td><td>1,013</td><td>0,951</td><td>1,024</td><td>1,050</td><td>0,969</td><td></td></tr> </table> <p>Amostra de embalagens de 600 mililitros</p> <table border="1"> <tr><td>627</td><td>641</td><td>556</td><td>591</td><td>590</td><td>613</td><td>646</td></tr> <tr><td>565</td><td>614</td><td>592</td><td>584</td><td>627</td><td>600</td><td>597</td></tr> <tr><td>620</td><td>660</td><td>601</td><td>627</td><td>586</td><td>578</td><td></td></tr> </table>	0,983	1,025	1,047	1,027	1,013	0,958	0,996	0,991	0,960	1,036	0,987	0,971	0,972	1,016	0,996	1,013	0,951	1,024	1,050	0,969		627	641	556	591	590	613	646	565	614	592	584	627	600	597	620	660	601	627	586	578	
0,983	1,025	1,047	1,027	1,013	0,958	0,996																																					
0,991	0,960	1,036	0,987	0,971	0,972	1,016																																					
0,996	1,013	0,951	1,024	1,050	0,969																																						
627	641	556	591	590	613	646																																					
565	614	592	584	627	600	597																																					
620	660	601	627	586	578																																						
5. Resolução da situação problema	<p>Sejam:</p> <p>X: volume das embalagens de 1 litro</p> <p>Y: volume das embalagens de 600 mililitros</p> <p>Temos:</p> $\bar{x} = \frac{\sum x}{n_x} = \frac{0,983+1,025+\dots+1,050+0,969}{20} = 0,999$ $Var(X) = \frac{\sum (x - \bar{x})^2}{n_x - 1} = \frac{(0,983 - 0,999)^2 + \dots + (0,969 - 0,999)^2}{20 - 1} \cong 0,000912$ $Dp(X) = \sqrt{0,000912} \cong 0,030199$ $CV(X) = \frac{Dp(X)}{\bar{x}} = \frac{0,030199}{0,999} \cong 0,03 = 3\%$ $\bar{y} = \frac{\sum y}{n_y} = \frac{627+641+\dots+586+578}{20} = 605,75$ $Var(Y) = \frac{\sum (y - \bar{y})^2}{n_y - 1} = \frac{(627 - 605,75)^2 + \dots + (578 - 605,75)^2}{20 - 1} \cong 738,93$ $Dp(Y) = \sqrt{738,93} \cong 27,18$ $CV(Y) = \frac{Dp(Y)}{\bar{y}} = \frac{27,18}{605,75} \cong 0,045 = 4,5\%$ <p>Em relação à pergunta (a), como $CV(Y) > CV(X)$, segue que a amostra de X é mais homogênea, ou seja, a amostra de refrigerantes de 1 litro é mais homogênea que a de 600 mililitros.</p>																																										
	<p>Com relação à pergunta (b), como $CV(X) < 4\% < CV(Y)$, a amostra de Y causará rejeição do lote, enquanto a amostra de X está dentro das conformidades.</p>																																										



Lembre-se

O **desvio** é a diferença de um valor do conjunto com relação à média.

O **desvio médio**, simbolizado por Dm , é uma medida de dispersão calculada por meio da média aritmética dos valores absolutos dos desvios.

A **variância**, simbolizada por Var , é uma medida de dispersão calculada por meio da média aritmética dos quadrados dos desvios.

O **desvio padrão**, simbolizado por Dp , é uma medida de dispersão definida como a raiz quadrada da variância.

O **coeficiente de variação**, simbolizado por CV , é uma medida de dispersão definida como a razão entre o desvio padrão e a média de um conjunto de dados.

Um conjunto de dados é classificado como de: baixa dispersão se $CV \leq 15\%$; média dispersão se $15\% < CV < 30\%$; alta dispersão se $CV \geq 30\%$.



Faça você mesmo

Na seção 2.2, no tópico "Faça você mesmo", foi proposto que, junto com seus colegas, você pesquisasse a altura dos alunos da turma. Verifique se a média é representativa do conjunto de dados de acordo com os critérios estabelecidos nesta seção.

Faça valer a Pena!

1. Assinale a alternativa que contém o desvio médio do conjunto de dados a seguir.

$$50 - 48 - 48 - 36 - 41 - 11 - 29 - 37$$

- a) 5,92 b) 9,52 c) 2,59 d) 9,25 e) 2,95

2. Assinale a alternativa que contém a variância e o desvio padrão da amostra a seguir.

$$118 - 104 - 124 - 116 - 117 - 105$$

a) $Var(X) = 63$ e $Dp(X) \cong 7,937$

b) $Var(X) = 62$ e $Dp(X) \cong 7,874$

- c) $Var(X) = 62$ e $Dp(X) \cong 7,937$
- d) $Var(X) = 63$ e $Dp(X) \cong 7,874$
- e) $Var(X) = 66$ e $Dp(X) \cong 8,124$

3. O conjunto de dados a seguir, obtido a partir da população, possui média $\bar{x} = 33$. Assinale a alternativa que contém o desvio padrão do conjunto.

$$y - 20 - 40 - 60$$

- a) 18,44
- b) 18,46
- c) 18,63
- d) 18,02
- e) 17,74

4. Observe os conjuntos $A=\{1,2,3\}$, $B=\{2,3,4\}$ e $C=\{5,6,7\}$. Assinale a alternativa que apresenta, respectivamente, a classificação desses conjuntos quanto à dispersão.

- a) alta dispersão; média dispersão; média dispersão.
- b) média dispersão; média dispersão; baixa dispersão.
- c) alta dispersão; alta dispersão; média dispersão.
- d) alta dispersão; baixa dispersão; média dispersão.
- e) alta dispersão; média dispersão, baixa dispersão.

5. Considerando o apresentado nessa seção e os conjuntos $A=\{1,2,3\}$, $B=\{2,3,4\}$ e $C=\{5,6,7\}$, assinale a alternativa que completa a frase: "A média é uma medida representativa..."

- a) somente para o conjunto C.
- b) para os conjuntos B e C.
- c) somente para o conjunto B.
- d) para os conjuntos A e C.
- e) somente para o conjunto A.

6. Os dados a seguir referem-se às alturas dos atletas das seleções masculina e feminina do vôlei brasileiro que participaram das Olimpíadas de Atenas, em 2004.

Seleção masculina (X)

1,99 – 1,99 – 2,01 – 1,84 – 1,92 – 1,96 – 2,03 – 1,84 – 1,95 – 1,91 – 2,05 – 1,90

Seleção feminina (Y)

1,77 – 1,79 – 1,84 – 1,80 – 1,94 – 1,80 – 1,73 – 1,88 – 1,79 – 1,80 – 1,85 – 1,90

Calcule a média, a variância, o desvio padrão e o coeficiente de variação de cada conjunto e conclua em qual deles há maior variabilidade na altura dos atletas.

7. Observe os dados a seguir.

1000 – 1260 – 1320 – 1380 – 1410 – 1645 – 1980 – 2106 – 2230 – 2239 – 2379 – 2760 – 3060 – 3120 – 3460 – 4030 – 4260 – 5050 – 5120 – 6460

Esse conjunto refere-se aos salários amostrados de alguns funcionários de uma grande empresa. Calcule a média e justifique por que ela não é representativa para esse conjunto. Em seguida, construa um histograma para sintetizar os dados. Os intervalos de classes devem ser 1000 |-- 2000, 2000 |-- 3000, 3000 |-- 4000, 4000 |-- 5000, 5000 |-- 6000, 6000 |-- 7000.

Referências

ANDERSON, David R.; SWEENEY, Dennis J.; WILLIAMS, Thomas A. **Estatística aplicada à administração e economia**. Trad. José Carlos Barbosa dos Santos. 2. ed. São Paulo: Cengage Learning, 2011.

CRESPO, Antônio A. **Estatística fácil**. 17. ed. São Paulo: Saraiva, 2002.

FREUND, John E. **Estatística aplicada**: economia, administração e contabilidade. Trad. Claus Ivo Doering. 11. ed. Porto Alegre: Bookman, 2006.

FUTPÉDIA. Disponível em: <<http://futpedia.globo.com/campeonato/copa-do-mundo>>. Acesso em: 13 maio 2015.

IBGE – Instituto Brasileiro de Geografia e Estatística. **População presente e residente**. Disponível em: <www.ibge.gov.br>. Acesso em: 14 maio 2015.

IBGE – Instituto Brasileiro de Geografia e Estatística. **Produção vegetal**. Disponível em: <www.ibge.gov.br>. Acesso em: 14 maio 2015.

JOHNSON, Robert; KUBY, Patrícia. **Estatística**. São Paulo: Cengage Learning, 2013.

MEDEIROS, Valéria Z. (Coord.). **Métodos quantitativos com excel**. São Paulo: Cengage Learning, 2008.

MORETTIN, Luiz G. **Estatística básica**: probabilidade e inferência. São Paulo: Pearson Prentice Hall, 2010.

UOL Esporte. Disponível em: <<http://esporte.uol.com.br/futebol/biografias/559/pele>>. Acesso em: 28 abr. 2015.

Estatística inferencial (parte I)

Convite ao estudo

Tomamos nossas decisões com base em conhecimentos prévios e experiências vivenciadas em situações semelhantes. Fazemos isso com o objetivo de potencializar os benefícios ou minimizar os efeitos negativos; entretanto, sempre estamos sujeitos a erros. Sabendo disso, uma área da estatística denominada *estatística inferencial* tenta mensurar as chances de ocorrerem erros e acertos nas tomadas de decisão.

Você aprendeu na unidade anterior a coletar, classificar, organizar e apresentar dados. No que se refere à apresentação, foram discutidos métodos pontuais que denominamos medidas de posição e vimos que nem sempre uma medida como a média é capaz de representar um conjunto. Mais adiante você verá que a chance de uma média estimada corresponder exatamente ao parâmetro populacional é muito pequena ou até impossível de ocorrer. Desse modo, o que a estatística inferencial tenta fazer, por exemplo, é mensurar a chance de determinado estimador pertencer a um intervalo.

Outro objeto de estudo da estatística inferencial é o levantamento e o teste de hipóteses. Por exemplo, podemos considerar como hipótese a afirmativa "o produto A pertence ao lote 1". O que a estatística inferencial irá fazer nesse caso é aceitar ou refutar essa afirmação com certa margem de incerteza, algo que sempre está presente na estatística.

Para trabalharmos todas essas ideias faz-se necessário abordar alguns pontos importantes: probabilidade; distribuições amostrais; intervalos de confiança; e testes de hipóteses. Um a um esses pontos serão discutidos no decorrer desta unidade e, ao final, esperamos que você seja capaz de estimar probabilidades, construir intervalos de confiança para estimadores e testar hipóteses estatísticas, objetivando a tomada de decisão.

Bons estudos!

Seção 3.1

Noções de probabilidade

Diálogo aberto

Qual é a chance de você ser atingido por um raio? E de ganhar na Mega Sena? Pode parecer piada, mas é mais fácil ocorrer o primeiro do que o segundo acontecimento. As chances de acertar as seis dezenas são de uma para cada 50 milhões (aproximadamente). Já as chances de ser atingido por um raio durante sua vida são de uma para cada 6250, de acordo com a **National Oceanic and Atmospheric Administration**.

A chance de ocorrência de determinado acontecimento é mensurada pela **probabilidade**, uma subárea da matemática que se tornou o pilar da estatística inferencial. Com base nessa mensuração, podemos tomar decisões apoiados em certos níveis de segurança do que pode vir a ocorrer.

Lembra-se de que na Unidade 2 fizemos uma coleta de dados com base em uma amostra de funcionários da empresa M? Com base nessa amostra, é possível medir a chance de sortear um funcionário na empresa e este ser do sexo masculino? Ou então, qual é a chance de ele pesar 70 quilogramas ou mais? Essas e outras questões serão respondidas ao longo desta seção. Bons estudos!

Não pode faltar!

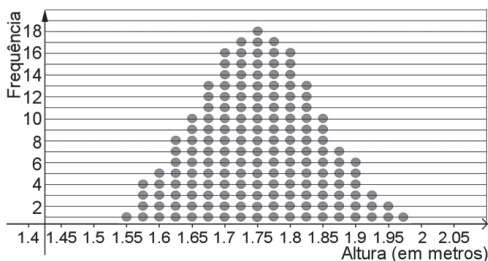
Noção de probabilidade

Para ilustrarmos a ideia de probabilidade, considere o diagrama de dispersão representado na Figura 3.1, o qual se refere a uma amostragem de funcionários da empresa M.

Nesse diagrama, pontos marcados sobre as marcas de escala no eixo horizontal referem-se àquele valor específico (por exemplo, exatamente 1 funcionário declarou ter exatamente 1,55 m). Já os pontos marcados entre duas marcas de escala referem-se a funcionários que declararam ter altura entre esses valores e não

iguais a eles (por exemplo, exatamente 4 funcionários declararam ter mais de 1,55 m e menos de 1,60 m).

Figura 3.1 | Frequência das alturas de uma amostra de 167 da empresa M



Fonte: Os autores (2015)

Como já foi descrito na Unidade 2, o diagrama de dispersão tenta dar uma ideia da distribuição dos valores de uma variável. Observando a Figura 3.1, por exemplo, podemos perceber que os valores estão concentrados em torno de 1,75 m, e as frequências diminuem conforme nos afastamos desse valor. Intuitivamente temos a impressão de que, ao selecionarmos aleatoriamente um funcionário dessa amostra, as chances de que ele tenha por volta de 1,75 m são maiores que as chances de que ele tenha por volta de 1,55 m.

Antes de continuarmos, faz-se necessário introduzir alguns conceitos:



Assimile

- Denominamos **experimento** todo e qualquer ato de experimentação (ou experiência) e investigação de determinado fenômeno sob condições controladas, a fim de observá-lo e classificá-lo. Como exemplo de experimento, temos a investigação da altura dos funcionários da empresa M.
- O conjunto de todos os resultados possíveis na investigação de uma variável em um experimento é denominado **espaço amostral**, o qual denotamos por Ω (ômega). O espaço amostral da variável altura é o intervalo $\Omega = (0, \infty) = \{t \in \mathbb{R} \mid t > 0\}$ que contempla os valores maiores que zero.
- Um valor específico pertencente a um espaço amostral é denominado **ponto amostral**. A altura 1,75 m é um exemplo de ponto amostral de Ω .
- Qualquer subconjunto de um espaço amostral é denominado **evento**. As alturas compreendidas entre 1,55 m e 1,75 m, por exemplo, compõem um evento.

Medimos a chance de ocorrência de determinado evento utilizando a **probabilidade**. Simplificadamente, a probabilidade é um valor numérico, compreendido no intervalo $[0,1] = \{t \in \mathbb{R} \mid 0 \leq t \leq 1\}$ e calculado por meio da razão entre o número de resultados favoráveis ao evento em questão pelo total de resultados possíveis no espaço amostral. Quanto mais próximo de 0, menor é a chance de ocorrência de um evento; quanto mais próximo de 1, maior é a chance de ocorrência.

Vamos compreender melhor o conceito de probabilidade por meio do exemplo a seguir.



Exemplificando

Considerando a Figura 3.1, qual é a probabilidade de, em um sorteio ao acaso, selecionarmos um funcionário da empresa M que possua altura maior ou igual a 1,85 m e menor que 1,90 m?

Resolução:

Considere o evento $A = \{\text{alturas maiores ou iguais a 1,85 m e menores que 1,90 m}\}$. Denotamos por $n(A)$ o número de elementos do conjunto A , ou seja, o número de ocorrências de alturas no intervalo citado. Observando o diagrama de dispersão, vemos que $n(A) = 17 (= 10 + 7)$. Além disso, o espaço amostral Ω possui 167 elementos, ou seja, $n(\Omega) = 167$.

Desse modo, a probabilidade de ocorrência do evento A é igual a:

$$P(A) = \frac{n(A)}{n(\Omega)} = \frac{17}{167} \cong 0,102 = 10,2\%$$

No exemplo anterior, denotando por X a variável altura e por x um ponto amostral qualquer, podemos simbolizar a probabilidade de ocorrência do evento A por $P(A) = P(1,85 \leq X < 1,90)$.



Refleta

Dados dois eventos B e C , sendo $P(B) = 1 = 100\%$ e $P(C) = 0 = 0\%$, dizemos que B é um **evento certo** e que C é um **evento impossível**.



Faça você mesmo

Ainda considerando a Figura 3.1, calcule:

a) $P(1,60 \leq X < 1,70)$

c) $P(X \geq 2,00)$

$$b) P(1,65 \leq X < 1,90)$$

$$d) P(1,50 \leq X < 2,00)$$

Curva normal

Observando a Figura 3.1, você notou alguma peculiaridade? A forma como os pontos se distribuem se assemelha a algum objeto conhecido do mundo real?

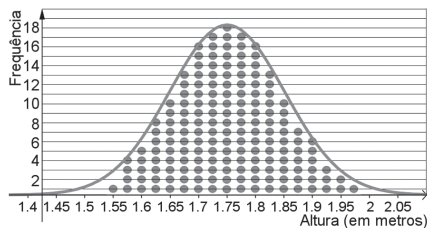
Esperamos que você tenha notado que a forma como os pontos se distribuem se assemelha a um sino. Observe novamente esse diagrama na Figura 3.2, na qual adicionamos uma linha contínua contornando os pontos.

A linha contornando os pontos (denominada curva normal) obedece a uma regra matemática dada por uma função do tipo exponencial, descrita por

$$f(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < +\infty,$$

em que x corresponde a um ponto amostral, μ (mu) é a média da população, σ^2 é a variância populacional e σ (sigma) é o desvio padrão populacional.

Figura 3.2 | Frequência das alturas de uma amostra de 167 da empresa M - Curva normal



Fonte: Os autores (2015)

Atenção

Na Unidade 2, com exceção da variância e do desvio padrão, não fizemos distinção simbólica entre medidas calculadas a partir de uma amostra e medidas calculadas a partir de dados populacionais. Naquele momento, não havia necessidade de abordar essa diferença. Entretanto, agora podemos ampliar a simbologia:

\bar{x} : média amostral

μ : média populacional

$Var(X)$: variância amostral¹

σ^2 : variância populacional

$Dp(X)$: desvio padrão amostral

σ : desvio padrão populacional

As demais medidas, por serem utilizadas em menor frequência, não serão simbolizadas de forma diferente para amostras ou populações.

¹ Alguns autores também denotam a variância amostral por s^2 e o desvio padrão amostral por s .

A função f descrita anteriormente, chamada de **função densidade de probabilidade (f.d.p.)**, é determinada pelos valores de μ e σ^2 . Sendo X uma variável que possui distribuição dos dados com formato de sino (caracterizada por μ e σ^2), simbolizamos $X \sim N(\mu, \sigma^2)$ para descrever que **X possui distribuição normal, com média μ e variância σ^2 .**

Variáveis com distribuição normal são muito comuns na natureza. Um dos principais estudiosos a observá-las foi Carl Friedrich Gauss (1777-1855) em seus trabalhos sobre astronomia por volta de 1810. Motivo pelo qual alguns autores também denominam gaussiana essa distribuição..

A probabilidade de ocorrência de um evento está diretamente ligada aos parâmetros μ e σ^2 provenientes da população. Conhecendo esses valores, considerando dada variável com distribuição normal e um evento A , podemos calcular a probabilidade de ocorrência de A por meio do cálculo de uma área.



Exemplificando

Identifique a área correspondente à probabilidade de ocorrência de $A = \{Z > 0,5 \text{ e } Z < 2,1\}$, sendo $Z \sim N(0,1)$.

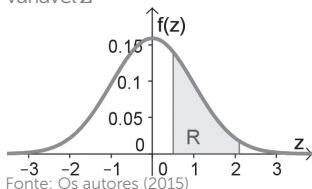
Resolução:

Observe que, para esse exemplo, $\mu = 0$ e $\sigma^2 = 1$ (e $\sigma = 1$). Com isso, a f.d.p. fica

$$f(z; 0, 1) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \text{ cujo gráfico}$$

está representado na Figura 3.3. A área R destacada corresponde à probabilidade de ocorrência de A , ou seja, $P(A) = R$.

Figura 3.3 | Distribuição da variável Z



Fonte: Os autores (2015)

No exemplo anterior temos $Z \sim N(0,1)$. Pelo fato de $\mu = 0$ e $\sigma^2 = 1$, essa distribuição recebe uma denominação especial, **normal padrão** (ou **normal padronizada**). Veja outras curvas normais em <http://www.ufpa.br/dicas/biome/biofig/curnor02.gif> para diferentes valores dos parâmetros μ e σ^2 .

O cálculo da área R destacada no exemplo é feito por meio de técnicas que não serão detalhadas aqui, pois não é o objetivo do nosso estudo. Uma maneira alternativa (e mais simples) para o cálculo dessa área é a utilização da **Tabela da Distribuição Normal Padrão** (ou tabela Z). Para compreendermos a utilização dessa tabela fazem-se necessárias algumas observações:

- A área limitada pela curva normal e pelo eixo horizontal ($f(z) = 0$), de $Z = -\infty$ até $Z = +\infty$, é igual a 1 (no exemplo anterior, temos $P(-\infty < Z < +\infty) = P(\Omega) = 1$);
- $P(a \leq X \leq b) = P(X \leq b) - P(X \leq a)$ = área sob a curva entre a e b (no exemplo anterior, temos $P(0,5 \leq Z \leq 2,1) = R$);
- $P(X = x_0) = 0$, para x_0 fixo. Na prática, a probabilidade de ocorrência de um valor específico é igual a zero, o que nos força a calcular a probabilidade para intervalos e não para valores particulares. (No exemplo anterior, temos $P(Z = 0,5) = 0 = 0\%$);
- $P(X \leq \mu) = P(X \geq \mu) = 0,5$, ou seja, a probabilidade de X ser menor que a média é igual a 50%, assim como a probabilidade de X ser maior que a média (no exemplo anterior, temos $P(Z \leq 0) = P(Z \geq 0) = 0,5$);
- $P(X \geq x) = 1 - P(X \leq x)$.

Entendidas essas observações, vamos então ao cálculo da área R . A Tabela 3.1 apresenta o valor da área abaixo da curva $f(z)$, acima do eixo horizontal ($f(z) = 0$) entre $Z = -\infty$ e $Z = z$, como mostra a Figura 3.4. Simbolizamos o valor dessa área por $P(Z \leq z)$ (ou $P(Z < z)$).

Figura 3.4 | Área representada por $P(Z \leq z)$

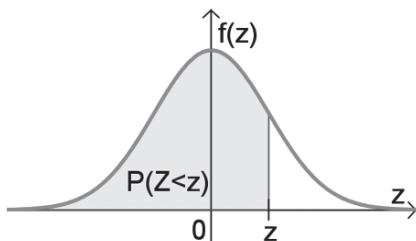


Tabela 3.1 | Tabela da Distribuição Normal Padrão Acumulada

z	-0,0	-0,1	-0,2	-0,3	-0,4	-0,5	-0,6	-0,7	-0,8	-0,9
-3	0,001	0,001	0,001	0,000	0,000	0,000	0,000	0,000	0,000	0,000
-2	0,023	0,018	0,014	0,011	0,008	0,006	0,005	0,003	0,003	0,002
-1	0,159	0,136	0,115	0,097	0,081	0,067	0,055	0,045	0,036	0,029
-0	0,500	0,460	0,421	0,382	0,345	0,309	0,274	0,242	0,212	0,184

z	+0,0	+0,1	+0,2	+0,3	+0,4	+0,5	+0,6	+0,7	+0,8	+0,9
+0	0,500	0,540	0,579	0,618	0,655	0,691	0,726	0,758	0,788	0,816
+1	0,841	0,864	0,885	0,903	0,919	0,933	0,945	0,955	0,964	0,971
+2	0,977	0,982	0,986	0,989	0,992	0,994	0,995	0,997	0,997	0,998
+3	0,999	0,999	0,999	1,000	1,000	1,000	1,000	1,000	1,000	1,000

Fonte: Os autores (2015)

Para calcularmos $P(A) = P(0,5 \leq Z \leq 2,1)$, efetuaremos $P(0,5 \leq Z \leq 2,1) = P(Z \leq 2,1) - P(Z \leq 0,5)$, pois os valores à direita da igualdade podem ser consultados na Tabela 3.1 (em destaque). Para calcularmos $P(Z \leq 2,1)$ consultamos a primeira coluna da tabela onde há o valor $z = +2$. Em seguida, percorremos essa linha até alcançarmos a coluna $z = +0,1$. Como $2,1 = 2 + 0,1$, temos que $P(Z \leq 2,1) = 0,982$. De modo semelhante chegamos a $P(Z \leq 0,5) = 0,691$. Logo, $P(A) = R = P(0,5 \leq Z \leq 2,1) = P(Z \leq 2,1) - P(Z \leq 0,5) = 0,982 - 0,691 = 0,291 = 29,1\%$. Portanto, o evento $A = \{Z > 0,5 \text{ e } Z < 2,1\}$ tem 29,1% de chance de ocorrência.

Normalização de variáveis

Como você deve ter notado na indicação que fizemos anteriormente, uma distribuição normal depende dos parâmetros μ e σ^2 . Se formos considerar todas as possibilidades de μ e σ^2 , teríamos que ter infinitas tabelas para consultar as probabilidades correspondentes. Para contornar essa dificuldade, "normalizamos" a variável em questão. Considere $X \sim N(\mu, \sigma^2)$ e a transformação $Z = (X - \mu)/\sigma$. Nessas condições é possível demonstrar que:

- $Z \sim N(0,1)$, ou seja, Z é uma variável normal padronizada;
- $P(X \leq x) = P(Z \leq z)$, em que $z = \frac{x - \mu}{\sigma}$.

Com o auxílio dessa transformação, podemos utilizar a Tabela 3.1 para calcularmos $P(X \leq x)$ para quaisquer μ e σ^2 .



Seendo $X \sim N(10,4)$, calcule:

a) $P(X \geq 6,4)$

b) $P(8,8 < X \leq 11,6)$

Resolução:

a) $P(X \geq 6,4) = 1 - P(X \leq 6,4) = 1 - P(Z \leq z)$, em que

$$z = \frac{6,4 - 10}{\sqrt{4}} = \frac{-3,6}{2} = -1,8.$$

Consultando a Tabela 3.1, vemos que $P(Z \leq -1,8) = 0,036$ (linha $z = -1$, coluna $z = -0,8$).

Logo

$$P(X \geq 6,4) = 1 - P(Z \leq -1,8) = 1 - 0,036 = 0,964 = 96,4\%.$$

b) $P(8,8 < X \leq 11,6) = P(X \leq 11,6) - P(X \leq 8,8)$

Calculamos separadamente $P(X \leq 11,6)$ e $P(X \leq 8,8)$.

$$P(X \leq 11,6) = P(Z \leq z), \text{ em que } z = \frac{11,6 - 10}{\sqrt{4}} = \frac{1,6}{2} = 0,8.$$

Consultando a tabela, vemos que $P(Z \leq 0,8) = 0,788$. Logo

$$P(X \leq 11,6) = P(Z \leq 0,8) = 0,788 = 78,8\%.$$

$$P(X \leq 8,8) = P(Z \leq z), \text{ em que } z = \frac{8,8 - 10}{\sqrt{4}} = \frac{-1,2}{2} = -0,6.$$

Consultando a tabela, vemos que $P(Z \leq -0,6) = 0,274$ (linha $z = -0$,

coluna $z = -0,6$). Logo $P(X \leq 8,8) = P(Z \leq -0,6) = 0,274 = 27,4\%$.

Portanto, $P(8,8 < X \leq 11,6) = P(X \leq 11,6) - P(X \leq 8,8) = 0,788 - 0,274 = 0,514 = 51,4\%$.



Leia mais sobre a distribuição normal e sobre outras distribuições de probabilidade no *link* indicado a seguir.

- A distribuição normal. Disponível em: <<http://www.ufpa.br/dicas/biome/bionor.htm>>.

Para consultar uma tabela de distribuição normal mais completa que a Tabela 3.1, acesse o *link* a seguir.

- Tabela normal padrão. Disponível em: <<http://www.leg.ufpr.br/~silvia/CE001/tabela-normal.pdf>>.²

² No decorrer deste livro sempre serão utilizados os valores desta tabela.

Sem medo de erro

Vamos relembrar a situação-problema proposta no início desta seção: Com base na amostra de funcionários apresentada na Tabela 2.1, é possível medir a chance de sortear um funcionário na empresa e este ser do sexo masculino? Ou então, qual é a chance de ele pesar 70 quilogramas ou mais?

Para responder a essas questões, vamos representar as variáveis sexo e peso, respectivamente, por X e Y e considerar os eventos $A = \{\text{funcionário sorteado ser do sexo masculino}\}$ e $B = \{\text{funcionário sorteado ter 70 quilogramas ou mais}\}$. As perguntas anteriores podem ser representadas simbolicamente por $P(A) = P(X = \text{masculino})$ e $P(B) = P(Y \geq 70)$, respectivamente.

Para o cálculo de $P(A)$ e $P(B)$ vamos utilizar os dados amostrais e supor que os verdadeiros parâmetros populacionais sejam próximos.

O número de homens na amostra é igual a 11 e o total de elementos amostrados foi 20. Logo, $P(A) = 11/20 = 0,55 = 55\%$.

Para o cálculo de $P(B)$ vamos supor que $Y \sim N(\mu, \sigma^2)$. Você pode verificar a partir da Tabela 2.1 que $\bar{y} = 76,05$, $Var(Y) \cong 137,9$ e $Dp(Y) \cong 11,7$. Sendo assim, consideraremos $\mu \cong \bar{y} = 76,05$, $\sigma^2 \cong Var(Y) \cong 137,9$ e $\sigma \cong Dp(Y) \cong 11,7$.

Assim, $P(B) = P(Y \geq 70) = 1 - P(Y \leq 70) = 1 - P(Z \leq z)$
, em que $z = \frac{70 - 76,05}{11,7} = \frac{-6,05}{11,7} \cong -0,5$. Consultando

a tabela Z, temos $P(Z \leq -0,5) = 0,309$. Logo,

$$P(B) = 1 - P(Z \leq z) = 1 - 0,309 = 0,691 = 69,1\%.$$

Para finalizar, a probabilidade de sortear um funcionário do sexo masculino na empresa M é de 55% e a de ele ter 70 quilogramas ou mais é de 69,1%.

Avançando na prática

Pratique mais!	
Instrução	
Desafiamos você a praticar o que aprendeu transferindo seus conhecimentos para novas situações que pode encontrar no ambiente de trabalho. Realize as atividades e depois as compare com as de seus colegas.	
1. Competências técnicas	Não se aplica.
2. Objetivos de aprendizagem	Compreender o conceito de probabilidade e suas aplicações.
3. Conteúdos relacionados	Probabilidade. Distribuição normal.
4. Descrição da SP	<p>Em uma indústria, para garantir a qualidade, são inspecionadas amostras de 15 unidades a cada lote de 200 produzidas. A seguir constam as medições, em mililitros, dos conteúdos de duas amostras, uma do lote 1 e outra do lote 2.</p> <p>Lote 1: 104 – 95 – 96 – 104 – 96 – 104 – 101 – 104 – 104 – 103 – 100 – 100 – 102 – 102 – 95</p> <p>Lote 2: 105 – 104 – 100 – 96 – 97 – 105 – 100 – 100 – 94 – 97 – 99 – 97 – 104 – 102 – 101</p> <p>O lote deve ser descartado se a probabilidade de nele conter uma unidade do produto com menos de 95 mililitros for maior que 6%.</p> <p>Pergunta: qual lote deve ser descartado, 1 ou 2?</p> <p>Utilize os estimadores como aproximações para os parâmetros populacionais.</p>

<p>5. Resolução da SP:</p>	<p>Façamos algumas considerações:</p> <p>X: quantidade em mililitros de cada unidade do produto no lote 1</p> <p>Y: quantidade em mililitros de cada unidade do produto no lote 2</p> <p>$A = \{\text{uma unidade do produto conter menos de 95 mililitros}\}$</p> <p>Você pode verificar que: $\bar{x} \cong 100,67$; $Dp(X) \cong 3,52$; $\bar{y} \cong 100,07$; $Dp(Y) \cong 3,45$</p> <p>Assim:</p> <p>Lote 1: $P(A) = P(X < 95) = P(Z < z)$, em que $z = \frac{95 - 100,67}{3,52} \cong -1,6$ Consultando a tabela Z, temos $P(X < 95) = P(Z < -1,6) = 0,055 = 5,5\%$.</p> <p>Lote 2: $P(A) = P(Y < 95) = P(Z < z)$, em que $z = \frac{95 - 100,07}{3,45} \cong -1,5$ Consultando a tabela Z, temos $P(Y < 95) = P(Z < -1,5) = 0,067 = 6,7\%$.</p> <p>Como $P(X < 95) = 6\% < P(Y < 95)$, segue que o lote 2 deve ser descartado.</p>
----------------------------	--



Lembre-se

Experimento: todo e qualquer ato de experimentação (ou experiência) e investigação de determinado fenômeno sob condições controladas, a fim de observá-lo e classificá-lo.

Espaço amostral: conjunto de todos os resultados possíveis na investigação de uma variável em um experimento, o qual denotamos por Ω .

Ponto amostral: valor específico pertencente a um espaço amostral.

Evento: qualquer subconjunto de um espaço amostral.

Probabilidade: valor numérico, compreendido no intervalo $[0,1] = \{t \in \mathbb{R} \mid 0 \leq t \leq 1\}$ e calculado por meio da razão entre o número de resultados favoráveis a um evento pelo total de resultados possíveis no espaço amostral.



Faça você mesmo

Sendo X a altura, em metros, dos alunos de graduação no Brasil, faça uma estimativa para $P(X \leq 1,45)$. Para isso, calcule as estatísticas \bar{x} e $Dp(X)$ a partir da sua turma e utilize esses valores como aproximação para os verdadeiros parâmetros populacionais.

Faça valer a pena

1. Assinale a alternativa INCORRETA.

- a) A probabilidade é igual à razão entre o número de resultados favoráveis a um evento pelo total de resultados possíveis no espaço amostral.
- b) Denominamos evento qualquer subconjunto de um espaço amostral.
- c) Um ponto amostral é um valor específico de Ω .
- d) Quando a probabilidade de ocorrência de um evento é igual a zero, dizemos que o evento é certo.
- e) Quanto mais próxima de 1, maior a probabilidade de ocorrência de um evento.

2. Considere $\Omega = \{a, b, c, d, e, f, g, h, i, j, k, l\}$ e um evento $A = \{b, d, f\}$. Assinale a alternativa que contém $P(A)$.

- a) 0,20
- b) 0,25
- c) 0,30
- d) 0,35
- e) 0,40

3. Sendo $Y \sim N(0,1)$, assinale a alternativa que contém o valor de $P(Y > 1,6)$.

- a) 0,945

- b) 0,055
- c) 1,000
- d) 0,000
- e) 0,726

4. Considere $Z \sim N(0,1)$ e um ponto amostral $z > 0$ tal que $P(-z < Z < z) = 95,4\%$. Assinale a alternativa que contém o valor de z .

- a) 1,0
- b) 1,5
- c) 2,0
- d) 2,5
- e) 3,0

5. Sendo $X \sim N(15,9)$, assinale a alternativa que contém o valor de $P(12 < X < 18)$.

- a) 15,9%
- b) 84,1%
- c) 62,8%
- d) 42,9%
- e) 68,2%

6. Considerando $X \sim N(50,16)$ e $Y \sim N(100,25)$, qual o evento mais provável: sortear um valor de X menor que 48 ou um valor de Y maior que 102?

7. Em determinada linha de produção, um produto é descartado se seu peso for menor que 4,9 kg. Sabe-se que a variável peso (X) nessa linha de produção possui distribuição normal com média de 5 kg e desvio padrão de 0,06 kg. Nessas condições, qual é a probabilidade de se descartar um produto?

Seção 3.2

Distribuição dos estimadores

Diálogo aberto

Você aprendeu na seção anterior o conceito de probabilidade e como ele pode ser aplicado a situações reais, tais como no exemplo do controle de qualidade. Nesse exemplo e em outros momentos utilizamos os valores \bar{x} e $Var(X)$ para estimar os verdadeiros parâmetros μ e σ^2 e solicitamos a você que fizesse o mesmo.

Como é de se esperar, sempre que utilizamos \bar{x} e $Var(X)$ para estimar μ e σ^2 estamos cometendo erros. Muitas vezes, esses erros são tão importantes quanto o valor que se pretende estimar. Você confiaria, por exemplo, em uma estimativa para a altura média da população brasileira de 1,90 m, considerando que o dado não seja acompanhado de nenhuma informação sobre os erros de estimativa? Essa informação é um tanto quanto suspeita.

Quando apresentamos estimadores como a média e a variância, seja em relatórios, artigos, obrigatoriamente temos de apresentar informações acerca dos erros de estimativa, pois isso dá credibilidade. Nesse contexto, podemos nos perguntar: fixada certa probabilidade de acerto e dado \bar{x} calculado a partir de uma amostra, qual é o erro que estamos cometendo ao aproximar μ por \bar{x} ? Ou, ainda, fixada uma probabilidade de acerto, qual é o tamanho da amostra que temos de coletar para cometer um erro máximo predeterminado?

Essa última pergunta nos faz lembrar uma informação dada na Unidade 2 de que, quanto maior a amostra, melhor a estimativa feita dos verdadeiros parâmetros. Nesta seção, você verá uma justificativa concreta para essa afirmação. Entretanto, para iniciarmos os estudos, propomos a seguinte situação-problema: com uma probabilidade de 95% de acerto, qual é o erro máximo que estamos cometendo ao aproximar a média do peso dos funcionários da empresa M por $\bar{y} = 76,05$? Qual deveria ser o tamanho da amostra para que o erro fosse de, no máximo, 2 kg?

Para que possamos responder a essas perguntas, precisamos entender melhor a distribuição de probabilidade da média amostral.

Não pode faltar

Teorema do Limite Central

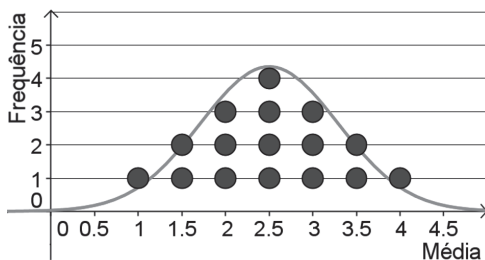
Para entendermos o que significa distribuição de probabilidade da média, considere que ao observar uma variável X na população tenhamos obtido $\Omega = \{1,2,3,4\}$. Qual o valor de μ ? Lembre-se de que μ é a média populacional e um cálculo simples mostra que $\mu = \frac{1+2+3+4}{4} = 2,5$.

Ao retirar uma amostra de tamanho 2 dessa população, conseguiríamos estimar precisamente μ por \bar{x} ? Ou, ainda, em todas as amostras o valor de \bar{x} seria o mesmo? As respostas para essas perguntas são, respectivamente, “pouco provável” e “não”. Veja a seguir todas amostras possíveis de tamanho 2 e suas respectivas médias.

Amostra	\bar{x}	Amostra	\bar{x}	Amostra	\bar{x}	Amostra	\bar{x}
{1,1}	1,0	{2,1}	1,5	{3,1}	2,0	{4,1}	2,5
{1,2}	1,5	{2,2}	2,0	{3,2}	2,5	{4,2}	3,0
{1,3}	2,0	{2,3}	2,5	{3,3}	3,0	{4,3}	3,5
{1,4}	2,5	{2,4}	3,0	{3,4}	3,5	{4,4}	4,0

Podemos montar um diagrama de dispersão com os valores das médias amostrais, como na Figura 3.5.

Figura 3.5 | Frequências das médias amostrais



Fonte: Os autores (2015)

Observou algo de curioso na forma como os dados se distribuíram? A linha ajudou, mas esperamos que você tenha notado que os dados se distribuíram de forma semelhante a uma

curva normal. A média amostral também pode ser considerada uma variável. Vamos calcular a média das médias amostrais ($\mu_{\bar{x}}$) e a variância das médias amostrais ($\sigma_{\bar{x}}^2$) para termos uma ideia quantitativa da distribuição?

$$\mu_{\bar{x}} = \frac{1,0+1,5+2,0+\dots+3,0+3,5+4,0}{16} = \frac{40}{16} = 2,5$$

$$\sigma_{\bar{x}}^2 = \frac{(1,0-2,5)^2 + (1,5-2,5)^2 + \dots + (3,5-2,5)^2 + (4,0-2,5)^2}{16} = \frac{10}{16} = 0,625$$

Observe que a média das médias amostrais é exatamente igual à média da população, ou seja, $\mu_{\bar{x}} = \mu$. E quanto à variância, será que $\sigma_{\bar{x}}^2 = \sigma^2$? Vejamos:

$$\sigma^2 = \frac{(1,0-2,5)^2 + (2,0-2,5)^2 + (3,0-2,5)^2 + (4,0-2,5)^2}{4} = \frac{5}{4} = 1,25$$

Note que $\sigma_{\bar{x}}^2 < \sigma^2$, resultado que pode ser mais bem compreendido com a leitura do **Teorema do Limite Central (TLC)**.

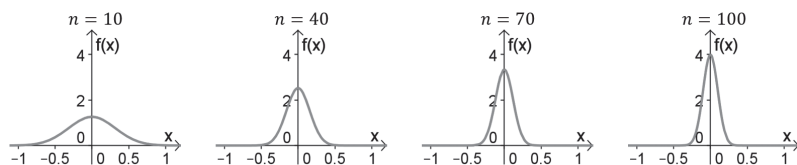


Assimile

De acordo com Morettin (2010), "o TLC diz que para n amostras aleatórias simples, retiradas de uma população com média μ e variância σ^2 finita, a distribuição amostral da média aproxima-se, para n grande, de uma distribuição normal, com média μ e variância σ^2/n ".

O TLC é de extrema importância para a estatística inferencial e tem implicações muito interessantes. Observe que, apesar de ele não dizer nada a respeito da distribuição da população, afirma que a distribuição amostral da média aproxima-se de uma curva normal, e, além disso, essa distribuição tem a mesma média que a população e variância σ^2/n , isto é, a mesma variância que a população, mas dividida por n . A partir desse resultado, concluímos que, quanto maior o número de amostras, mais precisão teremos para a média, pois σ^2/n diminui conforme n aumenta. Podemos visualizar esse resultado na Figura 3.6.

Figura 3.6 | Distribuição amostral da média \bar{x} de uma população $X \sim N(0,1)$ para vários valores de n



Fonte: Os autores (2015)

Se $X \sim N(0,1)$, a f.d.p. da variável \bar{x} pode ser escrita como

$$f(x; 0,1/n) = \sqrt{\frac{n}{2\pi}} e^{-nx^2/2}$$

Com base no TLC há ainda dois resultados interessantes que podemos enunciar.



Assimile

De acordo com Morettin (2010), "sendo X uma variável com média μ e variância σ^2 finita, e \bar{x} a variável média amostral, então a variável

$Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = \frac{\sqrt{n}(\bar{x} - \mu)}{\sigma}$ tem distribuição normal com média 0 e variância 1, ou seja, $Z \sim N(0,1)$ ".

Podemos ainda definir a variável e como a diferença entre o estimador \bar{x} e o parâmetro μ , ou seja, $e = \bar{x} - \mu$.

Determinando o tamanho de uma amostra

Vamos relembrar um questionamento feito no início desta seção: fixada certa probabilidade de acerto e dado \bar{x} calculada a partir de uma amostra, qual é o erro que estamos cometendo ao aproximar μ por \bar{x} ? Ou, ainda, fixada uma probabilidade de acerto, qual é o tamanho da amostra que temos de coletar para cometer um erro máximo predeterminado?

Vamos supor que o erro máximo que estipulamos para estimar a média populacional seja ε . Desse modo, qualquer valor \bar{x} no intervalo $[\mu - \varepsilon, \mu + \varepsilon]$ nos deixará satisfeitos para essa estimativa. Para assimilar melhor, suponha que queiramos estimar a verdadeira média populacional $\mu = 1,70$ m da altura de certo grupo de atletas e, para isso, queiramos cometer um erro máximo de $\varepsilon = 2$ cm.

Portanto, qualquer valor de \bar{x} pertencente ao intervalo [1,68 m; 1,72 m] servirá. Além disso, para acompanhar essa estimativa, suponha que queiramos ter uma probabilidade de acerto de γ (95%, por exemplo), uma margem de segurança.

Matematicamente, afirmar que \bar{x} pertence ao intervalo $[\mu - \varepsilon, \mu + \varepsilon]$ implica $\mu - \varepsilon \leq \bar{x} \leq \mu + \varepsilon$ ou, $|\bar{x} - \mu| \leq \varepsilon$. Além disso, ter uma probabilidade de acerto de γ que $|\bar{x} - \mu| \leq \varepsilon$ pode ser traduzido matematicamente por $P(|\bar{x} - \mu| \leq \varepsilon) \geq \gamma$.

Com base nos resultados obtidos do TLC, temos:

$$P(|\bar{x} - \mu| \leq \varepsilon) \geq \gamma \Rightarrow P(-\varepsilon \leq \bar{x} - \mu \leq +\varepsilon) \geq \gamma \Rightarrow P\left(-\frac{\sqrt{n}\varepsilon}{\sigma} \leq \frac{(\bar{x} - \mu)\sqrt{n}}{\sigma} \leq +\frac{\sqrt{n}\varepsilon}{\sigma}\right) \geq \gamma$$

Lembre-se de que $Z = \frac{\sqrt{n}(\bar{x} - \mu)}{\sigma}$, logo:

$$P\left(-\frac{\sqrt{n}\varepsilon}{\sigma} \leq \frac{(\bar{x} - \mu)\sqrt{n}}{\sigma} \leq +\frac{\sqrt{n}\varepsilon}{\sigma}\right) \geq \gamma \Rightarrow P\left(-\frac{\sqrt{n}\varepsilon}{\sigma} \leq Z \leq +\frac{\sqrt{n}\varepsilon}{\sigma}\right) \geq \gamma$$

Dado um valor γ podemos obter na tabela Z um valor z_γ tal que $P(-z_\gamma \leq Z \leq +z_\gamma) \geq \gamma$ e ainda:

$$z_\gamma = \frac{\sqrt{n}\varepsilon}{\sigma} \Rightarrow z_\gamma\sigma = \sqrt{n}\varepsilon \Rightarrow \sqrt{n} = \frac{z_\gamma\sigma}{\varepsilon} \Rightarrow n = \frac{z_\gamma^2\sigma^2}{\varepsilon^2}$$

Observe que, se tivermos o conhecimento de σ^2 , podemos estimar n em função de γ e ε , prefixados, ou estimar ε em função de γ e n . Com base na última igualdade podemos justificar a afirmativa feita na Unidade 2 de que o erro diminui à medida que o tamanho da amostra aumenta, pois:

$$n = \frac{z_\gamma^2\sigma^2}{\varepsilon^2} \Rightarrow \varepsilon^2 = \frac{z_\gamma^2\sigma^2}{n} \Rightarrow \varepsilon = \frac{\sqrt{z_\gamma^2\sigma^2}}{\sqrt{n}}$$

Podemos agora, observando a última igualdade, ver claramente que, se n aumenta ($n \rightarrow \infty$), o erro diminui ($\varepsilon \rightarrow 0$).



Seja uma variável $X \sim N(\mu, 4)$ observada em dada população. Com precisão de:

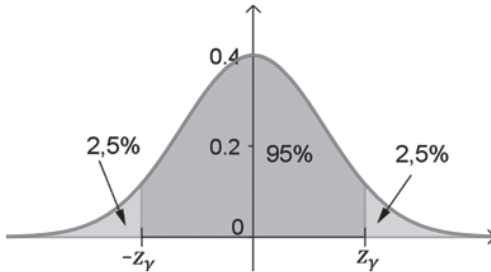
95%, qual o erro máximo que cometeremos ao estimar a verdadeira média dessa população com base em uma amostra de tamanho $n = 30$?

90%, qual o tamanho da amostra que deve ser coletada para que o erro seja de, no máximo, $\varepsilon = 1$?

Resolução:

a) Observe que a fórmula do erro $\varepsilon = \frac{\sqrt{z_\gamma^2 \sigma^2}}{\sqrt{n}}$ depende de z_γ , σ^2 e n . O parâmetro $\sigma^2 = 4$ foi dado e $n = 30$. Resta determinar z_γ , em que $\gamma = 95\% = 0,95$, para que tenhamos $P(-z_\gamma \leq Z \leq +z_\gamma) \geq \gamma = 0,95$. Observe a Figura 3.7.

Figura 3.7 | Região correspondente a $P(-z_\gamma \leq Z \leq +z_\gamma) \geq 0,95$



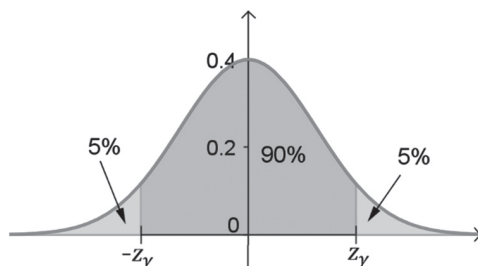
Fonte: Os autores (2015)

Veja que o valor z_γ deve ser tal que $P(Z \leq z_\gamma) \geq 0,025 + 0,95 = 0,975$. Consultando a tabela Z, temos $z_\gamma = 1,96$. Logo

$$\varepsilon = \frac{\sqrt{z_\gamma^2 \sigma^2}}{\sqrt{n}} = \frac{\sqrt{1,96^2 \cdot 4}}{\sqrt{30}} \cong 0,72$$

Portanto, com precisão de 95%, o erro máximo que cometeremos ao estimar a verdadeira média dessa população com base em uma amostra de tamanho $n = 30$ é $\varepsilon = 0,72$.

Figura 3.8 | Região correspondente a $P(-z_\gamma \leq Z \leq +z_\gamma) \geq 0,90$



Fonte: Os autores (2015)

b) Observe que, para determinar o tamanho da amostra, devemos utilizar a fórmula $n = \frac{z_\gamma^2 \sigma^2}{\varepsilon^2}$, em que σ^2 e ε foram dados, e z_γ deve ser consultado na tabela Z para $\gamma = 90\% = 0,90$. Veja a Figura 3.8.

Veja que o valor z_γ deve ser tal que $P(Z \leq z_\gamma) \geq 0,05 + 0,90 = 0,95$

. Consultando a tabela Z, temos $z_\gamma = 1,65$. Logo

$$n = \frac{z_\gamma^2 \sigma^2}{\varepsilon^2} = \frac{1,65^2 \cdot 4}{1^2} = 10,89 \cong 11.$$

Portanto, com precisão de 90%, para ter erro máximo $\varepsilon = 1$, temos de obter uma amostra de tamanho $n = 11$ para estimar a verdadeira média da população.

Observe que para calcular o erro e o tamanho da amostra ficamos dependentes de conhecer o valor de σ^2 , isto é, a variância populacional. Dificilmente conhecemos esse valor com exatidão, mas em certas situações ele pode ser conhecido de pesquisas anteriores. O IBGE, por exemplo, a cada dez anos realiza um censo e obtém todos os parâmetros populacionais. Entre um censo e outro é óbvio que os valores sofrem alterações, mas utilizar σ^2 obtido no censo anterior não é muito distante da realidade e é considerado aceitável. Caso esse valor seja desconhecido, comumente se utiliza $Var(X)$ em seu lugar.



Atenção

A desigualdade $P(-\varepsilon \leq \bar{x} - \mu \leq +\varepsilon) \geq \gamma$ pode dar origem a **intervalo de confiança** para a média populacional. Para a construção do mesmo, efetuamos:

$$P(-\varepsilon \leq \bar{x} - \mu \leq +\varepsilon) \geq \gamma \Rightarrow P(-\bar{x} - \varepsilon \leq -\mu \leq -\bar{x} + \varepsilon) \geq \gamma \Rightarrow \\ \Rightarrow P(\bar{x} - \varepsilon \leq \mu \leq \bar{x} + \varepsilon) \geq \gamma$$

Portanto, um intervalo de confiança para a média populacional, com nível de confiança γ , é definido como $IC(\mu, \gamma) = (\mu_1, \mu_2)$, em que $\mu_1 = \bar{x} - \varepsilon$ e $\mu_2 = \bar{x} + \varepsilon$.



Pesquise mais

Observe que não tratamos da distribuição amostral de $Var(X)$. Não entraremos em detalhes sobre essa distribuição, pois ela demanda maior detalhamento. Citaremos apenas que a distribuição da variância amostral é conhecida como **distribuição de qui-quadrado**, a qual simbolizamos por X^2 .

Veja mais detalhes sobre a distribuição amostral da média e a distribuição de qui-quadrado no *link* indicado a seguir.

- Inferência Estatística. Disponível em: <http://www.professores.uff.br/patricia/images/stories/arquivos/5_inferencia.pdf>.

Sem medo de errar!

Vamos relembrar a situação-problema proposta no início desta seção: com uma probabilidade de 95% de acerto, qual é o erro máximo que estamos cometendo ao aproximar a média do peso dos funcionários da empresa M por $\bar{y} = 76,05$? Qual deveria ser o tamanho da amostra para que o erro fosse de, no máximo, 2 kg?

Como não temos a variância populacional, iremos utilizar $Var(X)$ como estimativa para σ^2 . Temos:

$$Var(X) = \frac{(74 - 76,05)^2 + (93 - 76,05)^2 + \dots + (88 - 76,05)^2 + (68 - 76,05)^2}{20 - 1} \cong 137,9$$

Observando a fórmula do erro $\varepsilon = \frac{\sqrt{z_\gamma^2 \sigma^2}}{\sqrt{n}}$, vemos que nos resta determinar z_γ para $\gamma = 95\% = 0,95$, uma vez que $n = 20$ e $\sigma^2 \cong 137,9$ são conhecidos. A interpretação geométrica dessa probabilidade pode ser vista na Figura 3.7, no mesmo exemplo em que determinamos $z_\gamma = z_{95\%} = 1,96$. Assim:

$$\varepsilon = \frac{\sqrt{z_{\gamma}^2 \sigma^2}}{\sqrt{n}} = \frac{\sqrt{1,96^2 \cdot 137,9}}{\sqrt{20}} \cong \frac{23,02}{4,47} \cong 5,15$$

Portanto, com uma amostra de tamanho $n = 20$ estamos cometendo um erro máximo $\varepsilon = 5,15$ kg, com 95% de probabilidade.

Se desejarmos um erro máximo $\varepsilon = 2$ kg temos, por substituição direta na fórmula $n = \frac{z_{\gamma}^2 \sigma^2}{\varepsilon^2}$, uma amostra de tamanho:

$$n = \frac{z_{\gamma}^2 \sigma^2}{\varepsilon^2} = \frac{1,96^2 \cdot 137,9}{2^2} = \frac{529,75664}{4} = 132,43916 \cong 133$$

Portanto, se coletarmos uma amostra de 133 indivíduos, cometeremos um erro máximo de 2 kg para a estimativa de μ .

Avançando na prática

Pratique mais!	
Instrução	
Desafiamos você a praticar o que aprendeu transferindo seus conhecimentos para novas situações que pode encontrar no ambiente de trabalho. Realize as atividades e depois as compare com as de seus colegas.	
1. Competências técnicas	Não se aplica.
2. Objetivos de aprendizagem	Mensurar o erro de estimação da média populacional e dimensionar o tamanho de uma amostra para determinado erro máximo estipulado.
3. Conteúdos relacionados	Erro amostral da média. Dimensionamento de amostra.
4. Descrição da situação-problema	<p>Determinada linha de produção, que envasa leite em embalagens de 3 L, possui as seguintes regras para o controle de qualidade:</p> <ol style="list-style-type: none"> 1) Retiram-se 10 unidades de cada lote de 200 para compor a amostra de controle. 2) Utiliza-se como estimativa de σ^2 o maior valor calculado para as amostras dos três últimos lotes que saíram da linha de produção.

<p>4. Descrição da situação-problema</p>	<p>3) Se, com 98% de probabilidade de acerto, o erro amostral da média for superior a 0,05 L, a linha de produção é pausada para verificações nos equipamentos.</p> <p>Os valores a seguir correspondem às amostras dos três últimos lotes que saíram da linha de produção:</p> <p>Lote 1 (X_1): 3,006 – 2,935 – 2,976 – 3,018 – 2,996 – 2,978 – 3,045 – 3,075 – 2,857 – 2,953</p> <p>Lote 2 (X_2): 2,973 – 3,108 – 2,894 – 3,053 – 3,031 – 2,968 – 3,051 – 2,956 – 3,109 – 2,971</p> <p>Lote 3 (X_3): 2,864 – 3,005 – 3,065 – 2,901 – 2,94 – 3,059 – 3,005 – 3,025 – 3,152 – 3,112</p> <p>Com base nessas amostras:</p> <p>a) A linha de produção deveria ser pausada?</p> <p>b) Se modificarmos para 90,1% a probabilidade de acerto, a decisão seria a mesma?</p>
<p>5. Resolução da Situação-Problema</p>	<p>a) Primeiramente calculamos $Var(X_1)$, $Var(X_2)$ e $Var(X_3)$. Para isso é necessário conhecer também as respectivas médias. Você pode verificar que $\bar{x}_1 = 2,9839$, $\bar{x}_2 = 3,0114$, $\bar{x}_3 = 3,0128$, $Var(X_1) \cong 0,00371$, $Var(X_2) \cong 0,00493$ e $Var(X_3) \cong 0,00825$.</p> <p>De acordo com a regra (2), temos que $\sigma^2 \cong 0,00825$. Além disso, observando a fórmula do erro $\varepsilon = \frac{\sqrt{z_\gamma^2 \sigma^2}}{\sqrt{n}}$, temos de determinar z_γ para $\gamma = 98\% = 0,98$. Esse valor deve ser tal que $P(Z \leq z_\gamma) \geq 0,01 + 0,98 = 0,99$. Consultando a tabela Z, temos $z_\gamma = 2,33$, o que implica:</p> $\varepsilon = \frac{\sqrt{z_\gamma^2 \sigma^2}}{\sqrt{n}} = \frac{\sqrt{2,33^2 \cdot 0,00825}}{\sqrt{10}} \cong \frac{0,21163}{3,16228} \cong 0,06692$ <p>De acordo com a regra (3), devemos pausar a linha de produção, pois $\varepsilon > 0,05$ L.</p> <p>b) Se $\gamma = 90,1\% = 0,901$, temos de determinar z_γ para o qual $P(Z \leq z_\gamma) \geq 0,0495 + 0,901 = 0,9505$. Consultando a tabela Z, temos $z_\gamma = 1,65$. Logo:</p>

$$\varepsilon = \frac{\sqrt{z_{\gamma}^2 \sigma^2}}{\sqrt{n}} = \frac{\sqrt{1,65^2 \cdot 0,00825}}{\sqrt{10}} \cong \frac{0,14987}{3,16228} \cong 0,04739$$

Como $\varepsilon < 0,05$, a decisão seria diferente, ou seja, se $\gamma = 90,1\%$ a linha de produção não seria pausada.



Refleta

Qual interpretação temos da redução, no exemplo anterior, de $\gamma = 98\%$ para $\gamma = 90,1\%$?



Lembre-se

Teorema do Limite Central (TLC): para n amostras aleatórias simples, retiradas de uma população com média μ e variância σ^2 finita, a distribuição amostral da média aproxima-se, para n grande, de uma distribuição normal, com média μ e variância σ^2/n .

Consequência do TLC: sendo X uma variável com média μ e variância σ^2 finita, e \bar{x} a variável média amostral, então a variável

$$Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = \frac{\sqrt{n}(\bar{x} - \mu)}{\sigma}$$

tem distribuição normal com média 0 e variância

1, ou seja, $Z \sim N(0,1)$.

Erro amostral da média: Definido como $e = \bar{x} - \mu$, permite reescrever Z da seguinte forma, $Z = \frac{\sqrt{n}(\bar{x} - \mu)}{\sigma} = \frac{\sqrt{n}e}{\sigma}$, em que $e \sim N(0, \sigma^2/n)$.



Faça você mesmo

Acesse o link <<http://www.de.ufpb.br/~tarciana/CPEI/Aula3.pdf>> e estime o erro amostral da média para os dados apresentados na página 15. Considere diferentes valores de γ , como 90%, 95% e 98%.

Faça valer a pena!

1. Seja uma variável $X \sim N(\mu, 9)$ observada em dada população. Com precisão de 90%, assinale a alternativa que contém o erro máximo que cometemos ao estimar a verdadeira média dessa população com base em uma amostra de tamanho $n = 25$.

- a) 1,099.
- b) 0,099.
- c) 2,909.
- d) 2,970.
- e) 0,990.

2. Seja uma variável $X \sim N(\mu, 16)$ observada em dada população. Supondo que queiramos um erro amostral da média máximo de $\varepsilon = 1$, com 94% de probabilidade, entre as alternativas a seguir, assinale aquela que contém o menor tamanho de amostra que possibilite esse erro máximo estabelecido.

- a) 46.
- b) 55.
- c) 59.
- d) 62.
- e) 68.

3. Os conjuntos de dados a seguir são obtidos a partir de amostragem. Eles representam as idades de determinado grupo de frequentadores de um estabelecimento.

Grupo 1 (X_1): 31 – 27 – 33 – 33 – 24 – 25 – 28 – 29 – 24 – 31.

Grupo 2 (X_2): 31 – 28 – 28 – 30 – 29 – 31 – 31 – 28.

Grupo 3 (X_3): 30 – 28 – 31 – 28 – 31 – 30 – 28 – 31 – 29 – 32.

Sendo ε_1 , ε_2 e ε_3 os erros amostrais dos grupos 1, 2 e 3, respectivamente, assinale a alternativa correta.

- a) $\varepsilon_1 = \varepsilon_2 < \varepsilon_3$,
- b) $\varepsilon_1 > \varepsilon_2 > \varepsilon_3$,
- c) $\varepsilon_1 > \varepsilon_2 = \varepsilon_3$,

d) $\varepsilon_1 = \varepsilon_2 = \varepsilon_3$.

e) $\varepsilon_1 < \varepsilon_2 < \varepsilon_3$.

4. As variáveis $X \sim N(\mu_X, 49)$, $Y \sim N(\mu_Y, 45)$ e $W \sim N(\mu_W, 30)$ são observadas em uma população. Deseja-se coletar uma única amostra para estimar a média populacional de ambas as variáveis. Para os estudos que serão realizados é necessário que o erro amostral da média seja, no máximo, $\varepsilon = 2$ com confiança de 90,30%, 88,12% e 97,96% para as variáveis X , Y e Z , respectivamente. Desse modo, assinale a alternativa que contém o menor tamanho de amostra que atenda a essas exigências.

a) $n = 41$.

b) $n = 34$.

c) $n = 28$.

d) $n = 26$.

e) $n = 49$.

5. Para a realização de certo estudo, coletou-se a seguinte amostra:

1075 – 979 – 1034 – 1090 – 904 – 920 – 908 – 1026 – 963

Foi constatado, com 95% de probabilidade, que o erro amostral da média era de, no máximo, 46,38, valor que foi considerado alto. Com base nisso, estabeleceu-se um novo erro máximo tolerado, $\varepsilon = 15$, sendo necessário coletar uma nova amostra que será dimensionada com base na variância $Var(X)$ da amostra que será descartada. Assinale a alternativa que contém a dimensão da nova amostra.

a) 43.

b) 94.

c) 72.

d) 87.

e) 112.

6. Enuncie o Teorema do Limite Central e elenque duas de suas consequências.

7. As duas amostras a seguir foram retiradas de uma mesma população e são referentes a uma mesma variável $X \sim N(\mu, \sigma^2)$.

Amostra 1 (X_1): 61,6 – 63,8 – 61,7 – 59,7 – 66,5 – 64,1 – 58,6 – 59,0

Amostra 2 (X_2): 59,4 – 59,4 – 63,0 – 58,8 – 63,6 – 59,6 – 59,2 – 64,5 – 61,6 – 60,3

Faça uma estimativa pontual para μ calculando \bar{x} a partir da amostra que apresentar o menor erro amostral para a média. Calcule o erro amostral com precisão de 93,86%.

Seção 3.3

Testes de hipóteses para a média (σ^2 conhecido)

Diálogo aberto

Você aprendeu anteriormente sobre o Teorema do Limite Central (TLC) e algumas de suas implicações. Esse teorema é de extrema importância para a estatística inferencial e existem diversas situações em que pode ser utilizado, sendo que uma delas é no **teste estatístico de hipóteses**. Mas o que significa isso? Segundo Morettin e Bussab (2010, p. 330):

[...] feita determinada afirmação sobre uma população, usualmente sobre um parâmetro dessa, desejamos saber se os resultados experimentais provenientes de uma amostra contrariam ou não tal afirmação. Muitas vezes, essa afirmação sobre a população é derivada de teorias desenvolvidas no campo substantivo do conhecimento. A adequação ou não dessa teoria ao universo real pode ser verificada ou refutada pela amostra. O objetivo do teste estatístico de hipóteses é, então, fornecer uma metodologia que nos permita verificar se os dados amostrais trazem evidências que apoiem ou não uma hipótese (estatística) formulada.

Para avançarmos um pouco neste assunto, considere a seguinte situação-problema: suponha que a empresa M seja uma prestadora de serviços e que irá concorrer com outras para ser contratada para determinado projeto. A empresa contratante (empresa N) afirma que, para a execução das tarefas pertinentes ao projeto, é desejável que os funcionários possuam, em média, 80 kg e altura média maior ou igual a 170 cm, para utilizarem os Equipamentos de Proteção Individual (EPIs) de que a empresa dispõe. Em vista disso, a empresa M declara que seus funcionários se encaixam nesses padrões e acrescenta que, em medições feitas recentemente, constatou-se que o desvio padrão do peso de seus funcionários era 12 kg e que o desvio padrão da altura era 8 cm. Considerando que a empresa N tenha acesso aos dados

amostrados na Tabela 2.1, ela consegue constatar se a afirmação da empresa M é verdadeira?

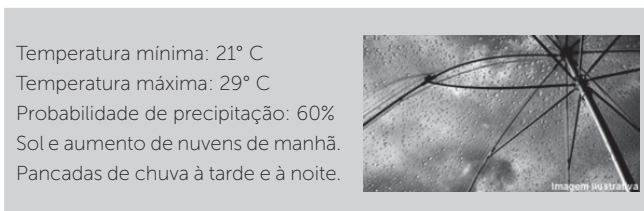
Para que possamos verificar essas afirmações, precisamos compreender melhor como formular hipóteses, adotar algumas simbologias que nos auxiliarão no processo e compreender os erros a que estamos sujeitos. No decorrer desta seção apresentaremos um roteiro para que você possa testar as hipóteses apresentadas e, ao final, verificaremos as afirmações.

Não pode faltar!

Formulando hipóteses

As situações abordadas em testes estatísticos de hipóteses podem nos parecer bem familiares. Considere, por exemplo, a afirmação (A) “vai chover hoje”. Essa afirmação pode ser considerada uma hipótese, cuja negativa é outra hipótese, (B) “não vai chover hoje”. Observe que as duas hipóteses levantadas são complementares, isto é, ocorre a primeira ou ocorre a segunda, não há outra possibilidade.

Figura 3.9 | Previsão do Tempo em Natal – RN para o dia 18/06/2015



Fonte: Climatempo

Como verificar a veracidade da hipótese (A)? É possível ter certeza absoluta da ocorrência de (A) ou (B)? Para respondermos a essas perguntas, observe a Figura 3.9.

Veja que a previsão do tempo para Natal traz uma informação muito importante, a probabilidade de precipitação, ou seja, a chance de chover. Para facilitar nossa discussão, vamos denotar as hipóteses A e B como a seguir:

H_0 : vai chover hoje

H_1 : não vai chover hoje



Assimile

A hipótese H_0 , denominada **hipótese nula**, geralmente é afirmativa ou, no caso de uma variável quantitativa, uma hipótese de igualdade. Ela é nossa principal hipótese, o foco da nossa análise e a que queremos

pôr à prova. A hipótese H_1 , denominada **hipótese alternativa**, é aquela que será aceita se rejeitarmos a hipótese nula.



Atenção

Alguns autores também denotam a hipótese alternativa por H_a .

Em relação a nossa decisão de aceitar ou rejeitar H_0 , podemos ter quatro resultados possíveis, elencados na Tabela 3.2.

Tabela 3.2 | Resultados possíveis para um teste de hipóteses

Decisão	Possibilidades para H_0	
	Verdadeira	Falsa
Não rejeitar H_0	Decisão correta	Erro tipo II
Rejeitar H_0	Erro tipo I	Decisão correta

Fonte: Morettin (2010)

Para o nosso exemplo, a ocorrência do erro tipo I seria rejeitar a hipótese “vai chover hoje” e, ao final do dia, constatarmos que choveu. Denotamos por α a probabilidade de ocorrência desse erro e, nesse caso, $\alpha = 60\%$. A ocorrência do erro tipo II, nesse caso, seria não rejeitar a hipótese “vai chover hoje” e, no final do dia, constatarmos que não choveu. A probabilidade de ocorrência desse erro é $\beta = 40\%$. Podemos escrever $P(\text{erro tipo I}) = \alpha$ e $P(\text{erro tipo II}) = \beta$. Portanto, respondendo às perguntas feitas anteriormente, para verificar a veracidade da hipótese (A) temos de realizar um teste de hipóteses. Contudo, nunca teremos certeza absoluta da ocorrência de uma hipótese, pois sempre estamos sujeitos a cometer um dos erros apresentados na Tabela 3.2.

Testando hipóteses

Para fixarmos um procedimento para o teste de uma hipótese nula, considere o seguinte exemplo.



Exemplificando

Uma variável $X \sim N(\mu, 18)$ é estudada em determinada população. Parte dos pesquisadores suspeita que $\mu = \mu_1 = 55$ e outros que $\mu = \mu_2 = 50$. No intuito de pôr à prova essas suspeitas eles decidiram fazer testes para identificar qual delas é a correta. Para isso foi retirada

uma amostra da população, a qual é apresentada a seguir.

49 – 50 – 48 – 51 – 47 – 48 – 55 – 50 – 55 – 49 – 51 – 53

Com 95% de confiança, qual é a verdadeira média da população, $\mu_1 = 55$ ou $\mu_2 = 50$?

Resolução:

Vamos inicialmente testar se $\mu = \mu_1 = 55$.

Passo 1 (elaborar as hipóteses): precisamos estipular duas hipóteses, a nula e a alternativa. Como a hipótese nula é sempre de igualdade, como foi descrito anteriormente, determinamos:

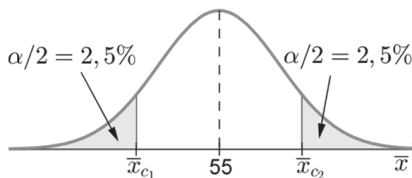
$$H_0: \mu = 55$$

$$H_1: \mu \neq 55$$

Passo 2 (determinar a estatística de teste): Como nosso objetivo é testar a média populacional da variável $X \sim N(\mu, 18)$, pelo TLC nossa estatística de teste será $\bar{x} \sim N(55, 18/12)$ ou $\bar{x} \sim N(55; 1,5)$, caso a hipótese nula seja verdadeira.

Passo 3 (fixar o nível de significância): Como queremos 95% de confiança, a probabilidade de cometermos o erro tipo I deve ser $\alpha = 100\% - 95\% = 5\%$. Essa probabilidade também é denominada **nível de significância**.

Figura 3.10 | Região crítica para $H_0: \mu = 55$ e $H_1: \mu \neq 55$, com $\alpha = 5\%$



Fonte: Os autores (2015)

Rejeitaremos a hipótese H_0 caso o valor \bar{x} obtido a partir da amostra seja muito maior ou muito menor que $\mu_1 = 55$ ou, ainda, quando \bar{x} pertencer à **região crítica** (RC), ilustrada na Figura 3.10.

A região crítica pode ser denotada por $RC = \{\bar{x} \in \mathbb{R} \mid \bar{x} \leq \bar{x}_{c_1} \text{ ou } \bar{x} \geq \bar{x}_{c_2}\}$. Observando a tabela Z e lembrando que $z = \frac{\bar{x} - \mu}{\sqrt{\sigma^2 / n}}$, temos:

$$z_1 = -1,96 = \frac{\bar{x}_{c_1} - 55}{\sqrt{1,5}} \Rightarrow -1,96\sqrt{1,5} = \bar{x}_{c_1} - 55 \Rightarrow \bar{x}_{c_1} = 55 - 1,96\sqrt{1,5} \cong 52,6;$$

$$z_2 = 1,96 = \frac{\bar{x}_{c_2} - 55}{\sqrt{1,5}} \Rightarrow 1,96\sqrt{1,5} = \bar{x}_{c_2} - 55 \Rightarrow \bar{x}_{c_2} = 55 + 1,96\sqrt{1,5} \cong 57,4;$$

$$RC = \{\bar{x} \in \mathbb{R} \mid \bar{x} \leq 52,6 \text{ ou } \bar{x} \geq 57,4\}.$$

Passo 4 (calcular a estatística a partir da amostra): a média amostral é $\bar{x}_0 = 50,5$.

Passo 5 (tomar uma decisão): como $\bar{x}_0 \in RC$, decidimos rejeitar H_0 , isto é, há indícios suficientes que nos permitem refutar a possibilidade de a média populacional ser $\mu = 55$.

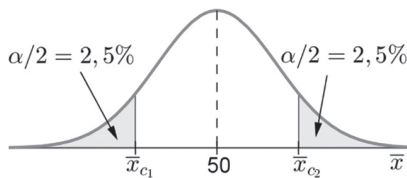
Vamos testar agora se $\mu = \mu_2 = 50$.

Passo 1 (elaborar as hipóteses):

$$H_0: \mu = 50$$

$$H_1: \mu \neq 50$$

Figura 3.11 | Região crítica para $H_0: \mu = 50$ e $H_1: \mu \neq 50$, com $\alpha = 5\%$



Fonte: Os autores (2015)

Passo 2 (determinar a estatística de teste): $\bar{x} \sim N(50, 18/12)$ ou $\bar{x} \sim N(50; 1,5)$, caso a hipótese nula seja verdadeira.

Passo 3 (fixar o nível de significância): $\alpha = 5\%$

Rejeitaremos a hipótese H_0 caso o valor \bar{x} obtido a partir da amostra pertença à **região crítica** (RC), ilustrada na Figura 3.11.

Observando a tabela Z e lembrando que $z = \frac{\bar{x} - \mu}{\sqrt{\sigma^2 / n}}$, temos:

$$z_1 = -1,96 = \frac{\bar{x}_{c_1} - 50}{\sqrt{1,5}} \Rightarrow$$

$$\Rightarrow -1,96\sqrt{1,5} = \bar{x}_{c_1} - 50 \Rightarrow \bar{x}_{c_1} = 50 - 1,96\sqrt{1,5} \cong 47,6;$$

$$z_2 = 1,96 = \frac{\bar{x}_{c_2} - 50}{\sqrt{1,5}} \Rightarrow 1,96\sqrt{1,5} = \bar{x}_{c_2} - 50 \Rightarrow \bar{x}_{c_2} = 50 + 1,96\sqrt{1,5} \cong 52,4;$$

$$RC = \{\bar{x} \in \mathbb{R} \mid \bar{x} \leq 47,6 \text{ ou } \bar{x} \geq 52,4\}.$$

Passo 4 (calcular a estatística a partir da amostra): a média amostral é $\bar{x}_0 = 50,5$.

Passo 5 (tomar uma decisão): como $\bar{x}_0 \notin RC$, não podemos rejeitar H_0 , isto é, não há indícios suficientes que nos permitam refutar a possibilidade de a média populacional ser $\mu_2 = 50$.

Desse modo, em concordância com o problema apresentado, devemos concluir que a verdadeira média da população é $\mu_2 = 50$.



Assimile

Região crítica: região de rejeição da hipótese nula.



Atenção

Testes de hipóteses como o do exemplo anterior são ditos **bilaterais**, pois a região crítica tem parte à esquerda e parte à direita do valor que está sendo testado.

Caso a região crítica estivesse somente à esquerda do valor que está sendo testado, o teste seria **unilateral à esquerda**; caso estivesse somente à direita, o teste seria **unilateral à direita**.

Veja a seguir um exemplo de teste unilateral à esquerda.



Exemplificando

Uma empresa de telefonia fixa oferece um pacote de acesso à internet com franquia ilimitada e velocidade média mensal de $\mu = 50$ Mbps com $\sigma^2 = 6$ Mbps². Paulo contratou o serviço e anda desconfiado de que a velocidade média é menor que a anunciada. Para testar se está sendo trapaceado pela empresa de telefonia, ele mediu a velocidade de sua

conexão durante um mês, em 15 diferentes dias e horários, obtendo a seguinte amostra:

47,7 - 47,9 - 49,2 - 48,5 - 47,5 - 48,3 - 50,5 - 51,1 - 48,0 - 48,9 - 47,9 - 47,9 - 47,9 - 50,2 - 51,4

Considerando que o valor $\sigma^2 = 6$ esteja correto, há evidências de que a velocidade fornecida é menor que a contratada?

Resolução:

Passo 1 (elaborar as hipóteses):

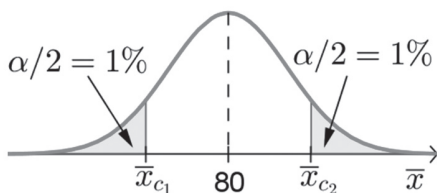
$$H_0: \mu = 50$$

$$H_1: \mu < 50$$

Passo 2 (determinar a estatística de teste): $\bar{x} \sim N(50, 6/15)$ ou $\bar{x} \sim N(50, 0,4)$

Passo 3 (fixar o nível de significância): suponha $\alpha = 5\%$. Como nos interessa $H_1: \mu < 50$, a região crítica é $RC = \{\bar{x} \in \mathbb{R} \mid \bar{x} \leq \bar{x}_c\}$ com $P(\bar{x} \leq \bar{x}_c) = 5\%$, como mostra a Figura 3.12.

Figura 3.14 | Região crítica para $H_0: \mu_x = 80$ e $H_1: \mu_x \neq 80$, com $\alpha = 2\%$



Fonte: Os autores (2015)

Mas $P(\bar{x} \leq \bar{x}_c) = P(Z \leq z)$ em que $z = \frac{\bar{x}_c - \mu}{\sqrt{\sigma^2/n}}$. Observando a tabela

Z , temos:

$$z = -1,64 = \frac{\bar{x}_c - 50}{\sqrt{0,4}} \Rightarrow -1,64\sqrt{0,4} = \bar{x}_c - 50 \Rightarrow \bar{x}_c = 50 - 1,64\sqrt{0,4} \cong 48,96;$$

$$RC = \{\bar{x} \in \mathbb{R} \mid \bar{x} \leq 48,96\}$$

Passo 4 (calcular a estatística a partir da amostra): $\bar{x} = 48,9$

Passo 5 (tomar uma decisão): como $\bar{x} \in RC$, optamos por rejeitar H_0 , ou seja, existem indícios suficientes de que a média populacional é menor que $\mu = 50$.

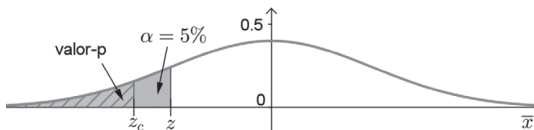
Concluimos que Paulo está correto ao afirmar que a velocidade média fornecida é menor que a velocidade média contratada.

Valor-p

Se efetuarmos $z_c = \frac{\bar{x} - \mu}{\sqrt{\sigma^2/n}}$ obtemos um valor denominado "z calculado" ou, ainda, "z estrela" (z_*) como alguns autores preferem denotar. Retomando o exemplo anterior, temos $z_c = \frac{\bar{x} - \mu}{\sqrt{\sigma^2/n}} = \frac{48,9 - 50}{\sqrt{0,4}} \cong -1,74$. Veja a representação de $z_c = -1,74$ e de $z = -1,64$ (determinado a partir do nível de significância $\alpha = 5\%$) na Figura 3.13.

Lembre-se de que, ao fixarmos o nível de significância $\alpha = 5\%$, obtivemos $z = -1,64$ a partir da tabela Z. Além disso, com a relação $P(Z \leq z) = P(\bar{x} \leq \bar{x}_c)$ calculamos o valor de \bar{x}_c que serviu

Figura 3.13 | Representação de $z_c = -1,74$ e de $z = -1,64$



Fonte: Os autores (2015)

de base para analisar a hipótese nula. Além dessa metodologia de análise existe outra bastante utilizada, a qual envolve o cálculo do **valor-p**. No caso do exemplo anterior, representado pela Figura 3.13, o valor-p corresponde à área que se apresenta à esquerda de z_c , abaixo da curva normal e acima do eixo horizontal (região hachurada). Mais formalmente, se o teste de hipóteses for:

- **unilateral à esquerda**, o valor-p é igual a $P(Z \leq z_c)$;
- **unilateral à direita**, o valor-p é igual a $P(Z \geq z_c)$;
- **bilateral**, o valor-p é igual a $P(Z \leq -|z_c|) + P(Z \geq |z_c|) = 2 P(Z \leq -|z_c|)$.

De acordo com Robert Johnson e Patrícia Kuby (2013), uma vez calculado o valor-p, podemos adotar a seguinte regra de decisão:

- Se o valor-p é menor ou igual ao nível de significância α , então a decisão deve ser **rejeitar H_0** .
- Se o valor-p é maior que o nível de significância α , então a decisão deve ser **não rejeitar H_0** .



Leia mais sobre os testes de hipóteses no Capítulo 5 do material disponível em: <<http://www.est.ufpr.br/ce003/material/apostilace003.pdf>>.

Sem medo de errar!

Observe que cada afirmação feita pela empresa M trata de uma suposição: (1) o peso médio dos funcionários é 80 kg; (2) a altura média é maior ou igual a 170 cm. Denotando por X e Y , respectivamente, o peso e a altura, temos que as afirmações anteriores podem ser traduzidas matematicamente como $\mu_X = 80$ e $\mu_Y \geq 170$. Sendo assim, temos duas hipóteses nulas a serem testadas:

Problema 1

$$H_0: \mu_X = 80$$

$$H_1: \mu_X \neq 80$$

Problema 2

$$H_0: \mu_Y = 170$$

$$H_1: \mu_Y < 170$$

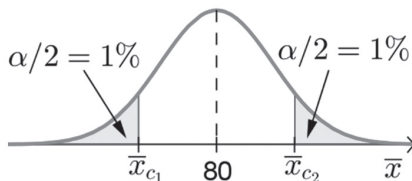
Para ambos os problemas, o passo 1 já foi realizado, ou seja, as hipóteses já foram fixadas.

Problema 1: testar $\mu_X = 80$

Passo 2 (determinar a estatística de teste): $\bar{x} \sim N(80, 12^2 / 20)$ ou $\bar{x} \sim N(80; 7,2)$, caso a hipótese nula seja verdadeira.

Passo 3 (fixar o nível de significância): suponha $\alpha = 2\%$ e RC como mostra a Figura 3.14. Consultando a tabela Z, e lembrando que $z = \frac{\bar{x} - \mu}{\sqrt{\sigma^2 / n}}$, temos:

Figura 3.14 | Região crítica para $H_0: \mu_X = 80$ e $H_1: \mu_X \neq 80$, com $\alpha = 2\%$



Fonte: Os autores (2015)

$$z_1 = -2,33 = \frac{\bar{x}_{c_1} - 80}{\sqrt{7,2}} \Rightarrow -2,33\sqrt{7,2} = \bar{x}_{c_1} - 80 \Rightarrow \bar{x}_{c_1} = 80 - 2,33\sqrt{7,2} \cong 73,7;$$

$$z_2 = 2,33 = \frac{\bar{x}_{c_2} - 80}{\sqrt{7,2}} \Rightarrow 2,33\sqrt{7,2} = \bar{x}_{c_2} - 80 \Rightarrow \bar{x}_{c_2} = 80 + 2,33\sqrt{7,2} \cong 86,3;$$

$$RC = \{\bar{x} \in \mathbb{R} \mid \bar{x} \leq 73,7 \text{ ou } \bar{x} \geq 86,3\}.$$

Passo 4 (calcular a estatística a partir da amostra): $\bar{x}_0 = 76,05$

Passo 5 (tomar uma decisão): como $\bar{x}_0 \notin RC$, não podemos rejeitar H_0 , isto é, não há indícios suficientes que nos permitam refutar a possibilidade de a média populacional ser $\mu_X = 80$.

Problema 2: testar $\mu_Y = 170$

Passo 2 (determinar a estatística de teste): $\bar{y} \sim N(170, 8^2/20)$ ou $\bar{y} \sim N(170; 3,2)$, caso a hipótese nula seja verdadeira.

Passo 3 (fixar o nível de significância): suponha $\alpha = 2\%$ e RC como mostra a Figura 3.15. Consultando a tabela Z, e lembrando que $z = \frac{\bar{x} - \mu}{\sqrt{\sigma^2/n}}$, temos:

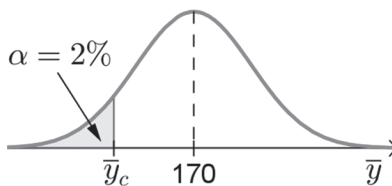
$$z = -2,05 = \frac{\bar{y}_c - 170}{\sqrt{3,2}} \Rightarrow -2,05\sqrt{3,2} = \bar{y}_c - 170 \Rightarrow \bar{y}_c = 170 - 2,05\sqrt{3,2} \cong 166,3;$$

$$RC = \{\bar{y} \in \mathbb{R} \mid \bar{y} \leq 166,3\}$$

Passo 4 (calcular a estatística a partir da amostra): $\bar{y}_0 = 172,85$

Passo 5 (tomar uma decisão): como $\bar{y}_0 \notin RC$, não podemos rejeitar H_0 , isto é, não há indícios suficientes que nos permitam refutar a possibilidade de a média populacional ser $\mu_Y = 170$.

Figura 3.15 | Região crítica para $H_0: \mu_Y = 170$ e $H_1: \mu_Y < 170$, com $\alpha = 2\%$



Fonte: Os autores (2015)

Portanto, considerando que a empresa N tenha acesso aos dados amostrados na Tabela 2.1 e o nível de significância $\alpha = 2\%$, não há indícios suficientes para que ela consiga refutar a afirmação da empresa M de que o peso médio de seus funcionários é 80 kg e que a altura média é maior ou igual a 170 cm.

Pratique mais!	
Instrução	
Desafiamos você a praticar o que aprendeu transferindo seus conhecimentos para novas situações que pode encontrar no ambiente de trabalho. Realize as atividades e depois as compare com as de seus colegas.	
1. Competências técnicas	Não se aplica.
2. Objetivos de aprendizagem	Realizar testes de hipóteses pela abordagem valor-p.
3. Conteúdos relacionados	Testes de hipóteses para a média com variância conhecida.
4. Descrição da Situação-Problema	<p>Determinada máquina corta barras de metal com 50 cm, em média, sendo o comprimento dessas barras uma variável $X \sim (\mu, 25 \text{ cm}^2)$. Caso a média dos comprimentos seja superior a 50 cm, há prejuízo para a empresa.</p> <p>Alguns funcionários suspeitam que a máquina esteja desregulada e que isso tem causado prejuízo. Para verificarem a suspeita, coletaram uma amostra de tamanho $n = 36$ e obtiveram $\bar{x} = 52$ cm.</p> <p>Utilizando a abordagem valor-p e o nível de significância $\alpha = 2\%$, verifique se há indícios suficientes para confirmar a suspeita dos funcionários.</p>
5. Resolução da Situação-Problema	<p>Passo 1 (elaborar as hipóteses):</p> $H_0: \mu = 50 \qquad H_1: \mu > 50$ <p>Passo 2 (determinar a estatística de teste):</p> $\bar{x} \sim N(50, 25/36) \text{ ou } \bar{x} \sim N(50; 0,69)$ <p>Passo 3 (fixar o nível de significância): $\alpha = 2\%$ (dado)</p> <p>Passo 4 (calcular a estatística a partir da amostra):</p> $\bar{x} = 52$ $z_c = \frac{\bar{x} - \mu}{\sqrt{\sigma^2 / n}} = \frac{52 - 50}{\sqrt{25 / 36}} = 2,4$

	<p>valor-p =</p> $P(Z \geq z_c) = P(Z \geq 2,4) = 1 - P(Z \leq 2,4) = 0,82\%$ <p>Passo 5 (tomar uma decisão): como $\text{valor-p} < \alpha$ optamos por rejeitar H_0, ou seja, existem indícios suficientes de que a média populacional é maior que $\mu = 50$.</p> <p>Concluimos que há indícios suficientes de que a suspeita dos funcionários pode ser confirmada.</p>
--	---



Lembre-se

Hipótese nula (H_0): geralmente é afirmativa ou, no caso de uma variável quantitativa, uma hipótese de igualdade. Ela é nossa principal hipótese, o foco da nossa análise e a que queremos pôr à prova.

Hipótese alternativa (H_1): é aquela que será aceita se rejeitarmos a hipótese nula.

Região crítica (RC): região de rejeição da hipótese nula.

Regra de decisão (abordagem valor-p): se o valor-p é menor ou igual ao nível de significância α , então a decisão deve ser **rejeitar H_0** ; se o valor-p é maior que o nível de significância α , então a decisão deve ser **não rejeitar H_0** .



Faça você mesmo

Junto a um colega, colete as informações sobre a altura de todos os alunos da turma. Um de vocês (primeiro) irá calcular a média μ e a variância σ^2 , sem que o outro (segundo) veja o resultado de μ . O primeiro irá fazer ao segundo uma afirmação sobre a média, por exemplo, "a média é $\mu = 1,70$ m" (não necessariamente o verdadeiro valor de μ). O segundo, por sua vez, irá coletar uma amostra e formular uma hipótese alternativa, por exemplo, "a média μ é menor que 1,70 m". Em seguida, conhecendo-se o valor de σ^2 e estipulando um nível de significância, o segundo irá testar as hipóteses para refutar ou não a afirmação do primeiro.

Faça valer a pena

1. Considere as hipóteses $H_0: \mu = 100$ e $H_1: \mu \neq 100$ elaboradas para a média de uma variável $X \sim N(\mu, 9)$. Para testar essas hipóteses coletou-se uma amostra de tamanho $n = 36$ e obteve-se $\bar{x} = 98$. Supondo um nível de significância $\alpha = 5\%$, assinale a alternativa que contém a região crítica, ou seja, a região de rejeição da hipótese nula:

- a) $RC = \{\bar{x} \in \mathbb{R} \mid \bar{x} \leq 99,18 \text{ ou } \bar{x} \geq 100,82\}$
- b) $RC = \{\bar{x} \in \mathbb{R} \mid \bar{x} \leq 99,02 \text{ ou } \bar{x} \geq 100,98\}$
- c) $RC = \{\bar{x} \in \mathbb{R} \mid \bar{x} \leq 99,175 \text{ ou } \bar{x} \geq 100,825\}$
- d) $RC = \{\bar{x} \in \mathbb{R} \mid \bar{x} \leq 95,0 \text{ ou } \bar{x} \geq 105,0\}$
- e) $RC = \{\bar{x} \in \mathbb{R} \mid \bar{x} \leq 97,5 \text{ ou } \bar{x} \geq 102,5\}$

2. Considere as hipóteses $H_0: \mu = 150$ e $H_1: \mu > 150$ elaboradas para a média de uma variável $X \sim N(\mu, 16)$. Para testar essas hipóteses coletou-se uma amostra de tamanho $n = 49$ e obteve-se $\bar{x} = 154$. Supondo um nível de significância $\alpha = 6,3\%$, assinale a alternativa que contém a região crítica, ou seja, a região de rejeição da hipótese nula:

- a) $RC = \{\bar{x} \in \mathbb{R} \mid \bar{x} \geq 150,66\}$
- b) $RC = \{\bar{x} \in \mathbb{R} \mid \bar{x} \leq 150,66\}$
- c) $RC = \{\bar{x} \in \mathbb{R} \mid \bar{x} \geq 150,87\}$
- d) $RC = \{\bar{x} \in \mathbb{R} \mid \bar{x} \leq 150,87\}$
- e) $RC = \{\bar{x} \in \mathbb{R} \mid \bar{x} \leq 150,76\}$

3. Considere as hipóteses $H_0: \mu = 200$ e $H_1: \mu < 200$ elaboradas para a média de uma variável $X \sim N(\mu, 25)$. Para testar essas hipóteses coletou-se uma amostra de tamanho $n = 36$ e obteve-se $\bar{x} = 199,033$. Assinale a alternativa que contém o menor nível de significância para o qual a hipótese nula seja rejeitada:

- a) 12,3%
- b) 10,2%

- c) 7,5%
- d) 5,0%
- e) 2,5%

4. Considere as hipóteses $H_0: \mu = 85$ e $H_1: \mu \neq 85$ elaboradas para a média de uma variável $X \sim N(\mu, 16)$ e as amostras $A = \{80, 86, 88, 90, 85\}$, $B = \{81, 81, 87, 87, 81\}$ e $C = \{88, 89, 85, 92, 87\}$, coletadas para testar as hipóteses. Sendo a , b e c , respectivamente, os valores-p correspondentes às amostras A , B e C , assinale a alternativa correta:

- a) $a = b = c$
- b) $a = b > c$
- c) $a < b = c$
- d) $a < b < c$
- e) $a > b > c$

5. Os salários dos funcionários de determinado setor da indústria correspondem a uma variável $X \sim N(1500, 22500)$. Em uma pesquisa foram selecionadas três indústrias desse setor e 30 funcionários de cada uma para verificar a média salarial, sendo obtidos os seguintes resultados:

Indústria 1	Indústria 2	Indústria 3
$\bar{x}_1 = 1400$	$\bar{x}_2 = 1550$	$\bar{x}_3 = 1470$

Com base nesses resultados e considerando um nível de significância $\alpha = 2\%$, assinale a alternativa correta:

- a) A indústria 1 paga salários abaixo da média; e não se pode afirmar que as indústrias 2 e 3 paguem salários diferentes da média.
- b) As indústrias 1 e 3 pagam salários abaixo da média; e a indústria 3 paga salários acima da média.
- c) As três indústrias pagam salários diferentes da média.
- d) Não se pode afirmar que essas indústrias paguem salários diferentes da média.

e) A probabilidade de se selecionar um funcionário desse setor da indústria e este receber mais de R\$ 1500,00 é menor que 10%.

6. Considere uma fábrica de refrigerantes que envasa embalagens de 2 L, sendo a quantidade de refrigerante nas garrafas uma variável $X \sim N(\mu; 0,01)$.

Para controle de qualidade são coletadas periodicamente amostras de 20 unidades e mensuradas respectivas quantidades. Se, ao nível de significância de 2%, a hipótese de a média das quantidades ser igual a 2 L for refutada, a linha de produção é pausada para verificações e ajustes nos equipamentos.

Com base na amostra a seguir, a linha de produção deve ser pausada?

1,90 – 2,09 – 2,07 – 1,89 – 1,94 – 1,89 – 2,15 – 2,10 – 2,06 – 2,13 – 2,05 – 2,03 – 2,04 – 2,11 – 2,12 – 2,15 – 1,86 – 2,10 – 1,98 – 1,90

7. Os parafusos fabricados por uma empresa têm resistência média à tração de 120 kg, com desvio padrão de 5 kg. Um depósito possui uma caixa com parafusos que o proprietário afirma ser desse fabricante. Entretanto, a informação não pode ser confirmada, pois algum funcionário descuidado estragou a embalagem e perdeu-se a informação sobre a origem. Na tentativa de vender para um comprador interessado nos parafusos desse fabricante, ou de melhor qualidade, o proprietário do depósito disse que faria um desconto no produto e daria 15 unidades para que o comprador pudesse testar a resistência média à tração e confirmar a origem. Da amostra testada o comprador constatou que a resistência média foi de 117,5 kg.

Com essas informações e um nível de significância de 2%, é possível confirmar a informação dada pelo proprietário do depósito?

Seção 3.4

Testes de hipóteses para a média (σ^2 desconhecido)

Diálogo aberto

Na seção anterior, você aprendeu a formular e testar hipóteses. Entretanto, há um detalhe que deve ser acrescentado: nós supusemos que a variância populacional era conhecida. Essa suposição também foi feita nas seções 3.1 e 3.2 e, em alguns casos, utilizamos $Var(X)$ como aproximação de σ^2 . Diante disso surgem alguns questionamentos: (1) em situações reais, com que frequência conhecemos o verdadeiro valor de σ^2 ? (2) é correto utilizarmos $Var(X)$ no lugar de σ^2 ?

Em relação ao primeiro questionamento, a resposta é “quase nunca”. Somente em raras situações isso ocorre. Um exemplo, inclusive descrito anteriormente, é quando o IBGE, a cada dez anos, realiza um censo e obtém os verdadeiros parâmetros populacionais. Se utilizamos no ano seguinte ao censo o valor de σ^2 , de certa forma estaremos lidando com um parâmetro real, mas com um pequeno atraso; devemos esperar que ele esteja desatualizado, mas podemos supor que o verdadeiro valor seja próximo. Essa mesma suposição teria de ser feita com cautela se utilizássemos σ^2 muito tempo depois do censo.

A resposta para o segundo questionamento é “sim, desde que de forma adequada”. A distribuição normal padrão (ou distribuição z) é utilizada para os casos em que a variância populacional é conhecida ou quando temos grandes amostras. Para pequenas amostras e variância populacional desconhecida, o correto é utilizarmos a **distribuição de Student** (ou distribuição t).

Para nos aprofundarmos nesse assunto iremos propor a mesma situação-problema da seção anterior, mas com uma pequena modificação, supor as variâncias populacionais desconhecidas. Desse modo, questionamos novamente: considerando que a empresa N tenha acesso aos dados amostrados na Tabela 2.1, ela consegue constatar se a afirmação da empresa M é verdadeira, isto é, que os funcionários possuem, em média, 80 kg e altura média maior ou igual a 170 cm?

Não pode faltar!

Distribuição de Student

A distribuição t de Student é uma distribuição de probabilidade proposta pelo irlandês W. S. Gosset, em 1908. Gosset era funcionário de uma cervejaria e escreveu sobre essa distribuição em um trabalho publicado com o pseudônimo "Student", daí a justificativa para o nome atribuído. Nesse trabalho, Student supôs que as amostras eram retiradas de populações normalmente distribuídas. Mesmo sem essa suposição, mais tarde se constatou que são obtidos resultados satisfatórios para quaisquer populações (normais ou não) quando são utilizadas grandes amostras.

Lembre-se de que na seção anterior utilizamos para os testes de hipóteses a estatística $z = \frac{\bar{x} - \mu}{\sqrt{\sigma^2/n}}$, a qual é normalmente distribuída com média 0 e variância 1, ou seja, $Z \sim N(0,1)$. Além disso, segundo Johnson e Kuby (2013):

[...] quando um σ^2 conhecido é usado para fazer uma inferência sobre a média μ , a amostra fornece um valor para aplicar nas fórmulas. Esse valor é \bar{x} . Quando o desvio padrão da amostra $Dp(X)$ também é usado, esta fornece dois valores: a média amostral \bar{x} e o erro padrão estimado $\sqrt{Var(X)/n}$. Como resultado, a estatística z será substituída por uma estatística que representa o uso de um erro padrão estimado. Essa nova estatística é conhecida como a estatística t de Student.

Desse modo, substituindo σ^2 por $Var(X)$, temos:



Assimile

Estatística z

$$z = \frac{\bar{x} - \mu}{\sqrt{\sigma^2/n}}$$

→

Estatística t

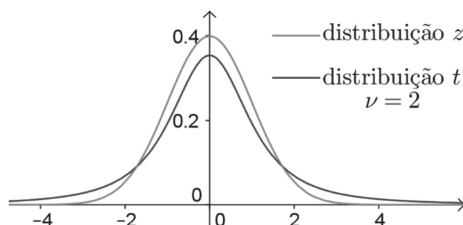
$$t = \frac{\bar{x} - \mu}{\sqrt{Var(X)/n}}$$

Diremos que uma variável T possui distribuição t de Student, e denotaremos por $t(\nu)$, se sua f.d.p. é dada por:

$$f(t; \nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\nu\pi}} \left(1+t^2/\nu\right)^{-(\nu+1)/2}, \text{ com } -\infty < t < \infty,$$

em que Γ é denominada função Gama e ν são os **graus de liberdade**. Veja na Figura 3.16 as distribuições normal padrão e t de Student, com $\nu = 2$. Para calcularmos os graus de liberdade efetuamos $\nu = n - 1$, em que n é o tamanho da amostra que estamos trabalhando. O número de graus de liberdade é equivalente ao número de desvios em relação à média que não estão relacionados. Para compreender melhor, lembre-se de que foi descrito na seção 2.4 que a soma dos desvios $\sum(x_i - \bar{x})$ é igual a zero. Portanto, quando temos n desvios $x_i - \bar{x}$ somente $n - 1$ destes têm liberdade de valor, pois o último desvio fica determinado pela relação $\sum(x_i - \bar{x}) = 0$.

Figura 3.16 | Distribuição z e distribuição t



Fonte: Os autores (2015)

Não entraremos em detalhes acerca da f.d.p. da distribuição de Student ou da função Gama, pois nosso intuito é utilizar valores tabelados para as probabilidades relacionadas a essa distribuição.



Assimile

Graus de liberdade (ν): número de desvios em relação à média que não estão relacionados entre si. Para calcular os graus de liberdade, efetuamos $\nu = n - 1$.

Tabela para a distribuição t

Diferentemente da distribuição z , em que possuímos uma única tabela, para a distribuição t teríamos de ter uma grande variedade, uma para cada grau de liberdade. Entretanto, para o nosso trabalho não é necessária uma tabela tão completa quanto a tabela Z ,

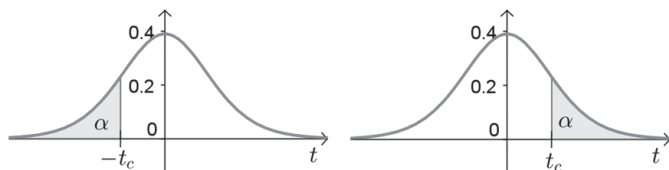
disponível em <<http://www.leg.ufpr.br/~silvia/CE001/tabela-normal.pdf>>, por exemplo. Basta uma que possua probabilidades-chave, as quais são muito utilizadas, como mostra a Tabela 3.3. O cabeçalho dessa tabela fornece os níveis de significância ou, em outras palavras, as probabilidades de ocorrência. A coluna da esquerda apresenta os graus de liberdade, e o corpo da tabela apresenta os valores t_c (t calculado), tais que a área à esquerda de $-t_c$, abaixo da curva de densidade de probabilidade e acima do eixo horizontal, é igual a α . Essa tabela pode ser mais bem interpretada com o auxílio da Figura 3.17.

Tabela 3.3 | Distribuição t de Student: valores de t_c tais que $P(t \leq -t_c) = \alpha$ (ou $P(t \geq t_c) = \alpha$)

v	Nível de significância (α) unilateral						
	0,1%	0,2%	0,5%	1%	2%	2,5%	5%
1	318,309	159,153	63,657	31,821	15,895	12,706	6,314
2	22,327	15,764	9,925	6,965	4,849	4,303	2,920
3	10,215	8,053	5,841	4,541	3,482	3,182	2,353
4	7,173	5,951	4,604	3,747	2,999	2,776	2,132
5	5,893	5,030	4,032	3,365	2,757	2,571	2,015
6	5,208	4,524	3,707	3,143	2,612	2,447	1,943
7	4,785	4,207	3,499	2,998	2,517	2,365	1,895
8	4,501	3,991	3,355	2,896	2,449	2,306	1,860
9	4,297	3,835	3,250	2,821	2,398	2,262	1,833
10	4,144	3,716	3,169	2,764	2,359	2,228	1,812
11	4,025	3,624	3,106	2,718	2,328	2,201	1,796
12	3,930	3,550	3,055	2,681	2,303	2,179	1,782
13	3,852	3,489	3,012	2,650	2,282	2,160	1,771
14	3,787	3,438	2,977	2,624	2,264	2,145	1,761
15	3,733	3,395	2,947	2,602	2,249	2,131	1,753
16	3,686	3,358	2,921	2,583	2,235	2,120	1,746
17	3,646	3,326	2,898	2,567	2,224	2,110	1,740
18	3,610	3,298	2,878	2,552	2,214	2,101	1,734
19	3,579	3,273	2,861	2,539	2,205	2,093	1,729
20	3,552	3,251	2,845	2,528	2,197	2,086	1,725
21	3,527	3,231	2,831	2,518	2,189	2,080	1,721
22	3,505	3,214	2,819	2,508	2,183	2,074	1,717
23	3,485	3,198	2,807	2,500	2,177	2,069	1,714
24	3,467	3,183	2,797	2,492	2,172	2,064	1,711
25	3,450	3,170	2,787	2,485	2,167	2,060	1,708
26	3,435	3,158	2,779	2,479	2,162	2,056	1,706
27	3,421	3,147	2,771	2,473	2,158	2,052	1,703
28	3,408	3,136	2,763	2,467	2,154	2,048	1,701
29	3,396	3,127	2,756	2,462	2,150	2,045	1,699
30	3,385	3,118	2,750	2,457	2,147	2,042	1,697

Fonte: Os autores (2015)

Figura 3.17 | Área α correspondente a: (a) $P(t \leq -t_c)$; (b) $P(t \geq t_c)$



Fonte: Os autores (2015)

Alguns valores de probabilidades que não constam na tabela T também podem ser obtidos por meio das propriedades apresentadas na seção 3.1, também válidas para a distribuição t , entre as quais destacamos:

- $P(a \leq t \leq b) = P(t \leq b) - P(t \leq a)$
- $P(t \leq \mu = 0) = P(t \geq \mu = 0) = 0,5$
- $P(t \geq t_0) = 1 - P(t \leq t_0)$

Observe que a Tabela 3.3 tem valores de t_c para graus de liberdade variando de 1 a 30. Algumas tabelas, como a disponível no *link* <<http://www.ime.unicamp.br/~cnaber/Tabela%20t.pdf>>, apresentam valores de t_c para graus de liberdade acima de 30, contudo, de modo mais espaçado e até, no máximo, 130 graus de liberdade. Pergunta: por que não construir também uma tabela para $\nu > 130$? A resposta é simples: quanto mais graus de liberdade temos, mais a curva de densidade da distribuição t se aproxima da curva normal padrão. Logo, quando tivermos muitos graus de liberdade, podemos utilizar a tabela Z em vez da tabela T.

Veja um exemplo de aplicação da tabela T.



Exemplificando

Uma variável $X \sim N(\mu, \sigma^2)$ é estudada em determinada população. Parte dos pesquisadores suspeita que $\mu = \mu_1 = 55$ e outros que $\mu = \mu_2 = 50$. No intuito de pôr à prova essas suspeitas eles decidiram fazer testes para identificar qual delas é a correta. Para isso, foi retirada uma amostra da população, a qual é apresentada a seguir.

49 – 50 – 48 – 51 – 47 – 48 – 55 – 50 – 55 – 49 – 51 – 53

Com 95% de confiança, qual é a verdadeira média da população, $\mu_1 = 55$ ou $\mu_2 = 50$?

Resolução:

Observe que este é o mesmo exemplo apresentado na Seção 3.3, com a diferença de que agora não conhecemos a variância populacional. Apesar disso, os passos a serem seguidos são os mesmos:

Vamos inicialmente testar se $\mu = \mu_1 = 55$.

Passo 1 (elaborar as hipóteses):

$$H_0: \mu = 55$$

$$H_1: \mu \neq 55$$

Passo 2 (determinar a estatística de teste): como a variância populacional é desconhecida, a estatística será $t = \frac{\bar{x} - \mu}{\sqrt{\text{Var}(X)/n}}$ com $v = 11$ graus de liberdade.

Passo 3 (fixar o nível de significância): $\alpha = 100\% - 95\% = 5\%$

Passo 4 (calcular a estatística a partir da amostra): a média amostral é $\bar{x}_0 = 50,5$ e $\text{Var}(X) = 7$.

Rejeitaremos a hipótese H_0 caso o valor \bar{x} obtido a partir da amostra seja muito maior ou muito menor que $\mu_1 = 55$ ou, ainda, quando \bar{x} pertencer à região crítica

$RC = \{\bar{x} \in \mathbb{R} \mid \bar{x} \leq \bar{x}_{c_1} \text{ ou } \bar{x} \geq \bar{x}_{c_2}\}$. Observando a tabela T na linha $v = 11$ e coluna correspondente à probabilidade 2,5% (pois o teste é bilateral), temos:

$$t_1 = -2,201 = \frac{\bar{x}_{c_1} - 55}{\sqrt{\frac{7}{12}}} \Rightarrow -2,201\sqrt{0,583} = \bar{x}_{c_1} - 55 \Rightarrow \bar{x}_{c_1} = 55 - 2,201\sqrt{0,583} \Rightarrow \bar{x}_{c_1} \cong 53,3;$$

$$t_2 = 2,201 = \frac{\bar{x}_{c_2} - 55}{\sqrt{\frac{7}{12}}} \Rightarrow 2,201\sqrt{0,583} = \bar{x}_{c_2} - 55 \Rightarrow \bar{x}_{c_2} = 55 + 2,201\sqrt{0,583} \Rightarrow \bar{x}_{c_2} \cong 56,7;$$

$$RC = \{\bar{x} \in \mathbb{R} \mid \bar{x} \leq 53,3 \text{ ou } \bar{x} \geq 56,7\}.$$

Passo 5 (tomar uma decisão): como $\bar{x}_0 \in RC$, decidimos rejeitar H_0 , isto é, há indícios suficientes que nos permitem refutar a possibilidade de a média populacional ser $\mu_1 = 55$.

Vamos testar agora se $\mu = \mu_2 = 50$.

Passo 1 (elaborar as hipóteses):

$$H_0: \mu = 50$$

$$H_1: \mu \neq 50$$

Passo 2 (determinar a estatística de teste): como a variância populacional é desconhecida, a estatística será $t = \frac{\bar{x} - \mu}{\sqrt{\text{Var}(X)/n}}$ com $v = 11$ graus de liberdade.

Passo 3 (fixar o nível de significância): $\alpha = 100\% - 95\% = 5\%$

Passo 4 (calcular a estatística a partir da amostra): a média amostral é $\bar{x}_0 = 50,5$ e $\text{Var}(X) = 7$. Logo:

$$t_c = \frac{50,5 - 50}{\sqrt{7/12}} = 0,655.$$

Observe na tabela T que, à medida que percorremos suas colunas da esquerda para a direita, o valor α aumenta e os valores de t_c diminuem. Assim, $t_c = 0,655$ (menor que todos os valores da tabela T) deve corresponder a um valor de α_{t_c} maior que 5%. Em consequência disso, temos:

$$\text{valor-p} = 2 \cdot P(t \leq -|t_c|) = 2 \cdot \alpha_{t_c} > 2 \cdot 5\% = 10\%.$$

Mais precisamente, podemos verificar utilizando o computador ou uma tabela T mais completa que o valor-p, nesse caso, é igual a 52,62%.

Passo 5 (tomar uma decisão): como o valor-p é maior que o nível de significância $\alpha = 5\%$ estipulado, não podemos rejeitar H_0 , isto é, não há indícios suficientes que nos permitam refutar a possibilidade de a média populacional ser $\mu_2 = 50$.

Desse modo, em concordância com o problema apresentado, devemos concluir que a verdadeira média da população é $\mu_2 = 50$.

Vejamos agora um exemplo de teste unilateral à direita.



Exemplificando

Seja uma variável $X \sim N(\mu, \sigma^2)$ de dada população. Foram levantadas duas hipóteses para a média populacional μ_x :

$$H_0: \mu = 15$$

$$H_1: \mu > 15$$

Para testar essas hipóteses, foi coletada uma amostra de tamanho $n = 30$ da qual se extraiu $\bar{x} = 15,265$ e $Var(X) = 0,5$.

Com 95% de confiança, é possível refutar a hipótese nula?

Resolução:

Passo 1 (elaborar as hipóteses):

$$H_0: \mu = 15$$

$$H_1: \mu > 15$$

Passo 2 (determinar a estatística de teste): como a variância populacional é desconhecida, a estatística será $t = \frac{\bar{x} - \mu}{\sqrt{Var(X)/n}}$ com $v = 29$ graus de liberdade.

Passo 3 (fixar o nível de significância): $\alpha = 100\% - 95\% = 5\%$

Passo 4 (calcular a estatística a partir da amostra): $\bar{x} = 15,265$ e $Var(X) = 0,5$. Logo:

$$t_c = \frac{15,265 - 15}{\sqrt{0,5/30}} \cong 2,053.$$

Observe na tabela T, na linha correspondente a $v = 29$, que o valor que mais se aproxima de $t_c = 2,053$ é 2,045, que corresponde a 2,5%. Assim, temos:

$$\text{valor-p} = P(t \geq t_c) = 2,5\%.$$

Mais precisamente, podemos verificar utilizando o computador que o valor-p, nesse caso, é igual a 2,46%.

Passo 5 (tomar uma decisão): como o valor-p é menor que o nível de significância $\alpha = 5\%$ estipulado, podemos rejeitar H_0 , isto é, há indícios suficientes que nos permitem refutar a possibilidade de a média populacional ser $\mu = 15$.



Leia mais sobre os testes de hipóteses com variância desconhecida no capítulo 5 do material disponível em: <http://www.est.ufpr.br/ce003/material/apostilace003.pdf>.

Sem medo de errar!

Observe que queremos novamente pôr à prova as afirmações feitas pela empresa M: (1) o peso médio dos funcionários é 80 kg; (2) a altura média é maior ou igual a 170 cm. A principal diferença com relação aos testes da seção anterior se apresenta na distribuição que será utilizada, pois não iremos mais supor que a variância populacional é conhecida. Denotando por X e Y , respectivamente, temos as seguintes hipóteses a serem testadas.

Problema 1

$$H_0: \mu_X = 80$$

$$H_1: \mu_X \neq 80$$

Problema 2

$$H_0: \mu_Y = 170$$

$$H_1: \mu_Y < 170$$

Para ambos os problemas, o passo 1 já foi realizado, ou seja, as hipóteses já foram fixadas.

Problema 1: testar $\mu_X = 80$

Passo 2 (determinar a estatística de teste): $t = \frac{\bar{x} - \mu}{\sqrt{\text{Var}(X)/n}}$ com $v = 19$ graus de liberdade (veja Tabela 2.1).

Passo 3 (fixar o nível de significância): $\alpha = 2\%$

Passo 4 (calcular a estatística a partir da amostra): $\bar{x}_0 = 76,05$ e $\text{Var}(X) = 137,94$.

$$t_c = \frac{76,05 - 80}{\sqrt{137,94 / 20}} \cong -1,504$$

Observe na tabela T, na linha correspondente a $v = 19$, que o valor que mais se aproxima de $t_c = -1,504$ é $-1,729$, que corresponde a 5%. Assim, temos:

$$\text{valor-p} = 2 \cdot P(t \leq -|t_c|) \geq 2 \cdot 5\% = 10\%$$

Mais precisamente, podemos verificar utilizando o computador que o valor-p, nesse caso, é igual a 14,9%.

Passo 5 (tomar uma decisão): como o valor-p é maior que o nível de significância estipulado, não podemos rejeitar H_0 , isto é, não há indícios suficientes que nos permitam refutar a possibilidade de a média populacional ser $\mu_x = 80$.

Problema 2: testar $\mu_y = 170$

Passo 2 (determinar a estatística de teste): $t = \frac{\bar{y} - \mu}{\sqrt{\text{Var}(Y)/n}}$ com $v = 19$ graus de liberdade.

Passo 3 (fixar o nível de significância): $\alpha = 2\%$

Passo 4 (calcular a estatística a partir da amostra): $\bar{y}_0 = 172,85$ e $\text{Var}(X) = 60,45$.

$$t_c = \frac{172,85 - 170}{\sqrt{60,45/20}} \cong 1,639$$

Observe na tabela T, na linha correspondente a $v = 19$, que o valor que mais se aproxima de $t_c = 1,639$ é $1,729$, que corresponde a 5%. Assim, temos:

$$\text{valor-p} = P(t \geq t_c) \geq 5\%.$$

Mais precisamente, podemos verificar utilizando o computador que o valor-p, nesse caso, é igual a 5,88%.

Passo 5 (tomar uma decisão): como o valor-p é maior que o nível de significância estipulado, não podemos rejeitar H_0 , isto é, não há indícios suficientes que nos permitam refutar a possibilidade de a média populacional ser $\mu_y = 170$.

Portanto, considerando que a empresa N tenha acesso aos dados amostrados na Tabela 2.1 e o nível de significância $\alpha = 2\%$, não há indícios suficientes para que ela consiga refutar a afirmação da empresa M de que o peso médio de seus funcionários é 80 kg e que a altura média deles é maior ou igual a 170 cm.

Avançando na prática

Pratique mais!

Instrução

Desafiamos você a praticar o que aprendeu transferindo seus conhecimentos para novas situações que pode encontrar no ambiente de trabalho. Realize as atividades e depois as compare com as de seus colegas.

1. Competências técnicas	Não se aplica.
2. Objetivos de aprendizagem	Realizar testes de hipóteses pela abordagem valor-p.
3. Conteúdos relacionados	Testes de hipóteses para a média com variância desconhecida.
4. Descrição da situação-problema	<p>Determinada máquina corta barras de metal com 50 cm, em média, sendo o comprimento dessas barras uma variável $X \sim (\mu, \sigma^2)$. Caso a média dos comprimentos seja diferente de 50 cm, há prejuízo para a empresa.</p> <p>Alguns funcionários suspeitam que a máquina esteja desregulada e que isso tem causado prejuízo. Para verificarem a suspeita, coletaram uma amostra de tamanho $n = 28$ e obtiveram $\bar{x} = 52,04$ cm e $\text{Var}(X) = 25$ cm².</p> <p>Utilizando a abordagem valor-p e o nível de significância $\alpha = 5\%$, verifique se há indícios suficientes para confirmar a suspeita dos funcionários.</p>
5. Resolução da Situação Problema:	<p>Passo 1 (elaborar as hipóteses): $H_0: \mu = 50$ $H_1: \mu \neq 50$</p> <p>Passo 2 (determinar a estatística de teste): $t = \frac{\bar{x} - \mu}{\sqrt{\text{Var}(X)/n}}$ com $v = 27$ graus de liberdade.</p> <p>Passo 3 (fixar o nível de significância): $\alpha = 5\%$</p> <p>Passo 4 (calcular a estatística a partir da amostra): $\bar{x} = 52,04$ e $\text{Var}(X) = 25$.</p> $t_c = \frac{52,04 - 50}{\sqrt{25/28}} \cong 2,159$ <p>Observe na tabela T, na linha correspondente a $v = 27$, que o valor $t_c = 2,159$, é maior que $2,158$ que corresponde a 2%. Assim, temos:</p>

$$\text{valor-p} = 2 \cdot P(t \leq -|t_c|) < 2 \cdot 2\% = 4\% .$$

Mais precisamente, podemos verificar utilizando o computador que o valor-p, nesse caso, é igual a 3,99%.

Passo 5 (tomar uma decisão): como $\text{valor-p} < \alpha = 5\%$ optamos por rejeitar H_0 , ou seja, existem indícios suficientes de que a média populacional é diferente de $\mu = 50$.

Concluimos que há indícios suficientes de que a suspeita dos funcionários pode ser confirmada.



Lembre-se

A distribuição normal padrão (ou distribuição z) é utilizada para os casos em que a variância populacional é conhecida ou quando temos grandes amostras (geralmente $n > 120$). Para pequenas amostras e variância populacional desconhecida, o correto é utilizarmos a **distribuição de Student** (ou distribuição t).

Graus de liberdade (v): número de desvios em relação à média que não estão relacionados entre si. Para calcular os graus de liberdade, efetuamos $v = n - 1$.



Faça você mesmo

Junto a um colega, colete as informações sobre o peso de todos os alunos da turma. Um de vocês (primeiro) irá calcular a média μ e a variância σ^2 , sem que o outro (segundo) veja os resultados. O primeiro irá fazer ao segundo uma afirmação sobre a média, por exemplo, "a média é $\mu = 70$ kg" (não necessariamente o verdadeiro valor de μ). O segundo, por sua vez, irá coletar uma amostra e formular uma hipótese alternativa, por exemplo, "a média μ é diferente de 70 kg". Em seguida, estipulando um nível de significância, o segundo irá testar as hipóteses para refutar ou não a afirmação do primeiro.

Faça valer a pena!

1. Em determinado teste de hipóteses temos $H_0: \mu = 20$ e $H_1: \mu \neq 20$. Sabendo que a partir de uma amostra de tamanho $n = 25$ obteve-se $\bar{x} = 19$ e $Var(X) = 4$, assinale a alternativa que contém a região crítica para $\alpha = 5\%$.

a) $RC = \{\bar{x} \in \mathbb{R} \mid \bar{x} \leq 19,3156 \text{ ou } \bar{x} \geq 20,6844\}$

b) $RC = \{\bar{x} \in \mathbb{R} \mid \bar{x} \leq 19,1760 \text{ ou } \bar{x} \geq 20,8240\}$

c) $RC = \{\bar{x} \in \mathbb{R} \mid \bar{x} \leq 19,1744 \text{ ou } \bar{x} \geq 20,8256\}$

d) $RC = \{\bar{x} \in \mathbb{R} \mid \bar{x} \leq 19,3168 \text{ ou } \bar{x} \geq 20,6832\}$

e) $RC = \{\bar{x} \in \mathbb{R} \mid \bar{x} \leq 19,3156 \text{ ou } \bar{x} \geq 20,6832\}$

2. Deseja-se testar as hipóteses $H_0: \mu = 60$ e $H_1: \mu > 60$. Sabendo que a partir de uma amostra de tamanho $n = 30$ obteve-se $\bar{x} = 61$ e $Var(X) = 9$, assinale a alternativa que contém a região crítica para $\alpha = 2,5\%$.

a) $RC = \{\bar{x} \in \mathbb{R} \mid \bar{x} \geq 61,120\}$

b) $RC = \{\bar{x} \in \mathbb{R} \mid \bar{x} \geq 61,118\}$

c) $RC = \{\bar{x} \in \mathbb{R} \mid \bar{x} \geq 60,931\}$

d) $RC = \{\bar{x} \in \mathbb{R} \mid \bar{x} \geq 60,929\}$

e) $RC = \{\bar{x} \in \mathbb{R} \mid \bar{x} \geq 61,178\}$

3. Para testar as hipóteses $H_0: \mu = 100$ e $H_1: \mu < 100$, coletou-se uma amostra de tamanho $n = 20$ e obteve-se $\bar{x} = 98,84$ e $Var(X) = 9$. Assinale a alternativa que contém o valor-p.

a) 0,5%

b) 1%

c) 2%

d) 2,5%

e) 5%

4. Suponha que para testar as hipóteses $H_0: \mu = 50$ e $H_1: \mu > 50$ tenha-se coletado uma amostra de tamanho $n = 25$, obtendo $\bar{x} = 51$ e $Var(X) = 16$. Assinale a alternativa correta.

a) $1\% < \text{valor-p} < 2\%$

b) $2\% < \text{valor-p} < 2,5\%$

c) $2,5\% < \text{valor-p} < 5\%$

d) valor-p > 5%

e) valor-p = 5%

5. Considere que para testar as hipóteses $H_0: \mu = 100$ e $H_1: \mu \neq 80$ tenha-se coletado a seguinte amostra:

83 – 83 – 82 – 80 – 79 – 81 – 80 – 79 – 84 – 80 – 82 – 82

Considerando $\alpha = 5\%$, assinale a alternativa correta:

a) Não se pode rejeitar a hipótese nula, pois valor-p é menor que 5%

b) Deve-se rejeitar a hipótese nula, pois valor-p é menor que 5%

c) Não se pode rejeitar a hipótese nula, pois valor-p é maior que 5%

d) Deve-se rejeitar a hipótese nula, pois valor-p é menor que 1%

e) Não se pode rejeitar a hipótese nula, pois valor-p é maior que 1%

6. Sejam as hipóteses $H_0: \mu = 500$ e $H_1: \mu < 500$. Determine a região crítica para $\alpha = 5\%$, sabendo que de uma amostra de tamanho $n = 28$ obteve-se $\bar{x} = 498$ e $Var(X) = 100$. Por fim, conclua se a hipótese nula deve ser rejeitada ou não.

7. Para testar as hipóteses $H_0: \mu = 150$ e $H_1: \mu \neq 150$ coletou-se uma amostra de tamanho $n = 200$, obtendo-se $\bar{x} = 151,46$ e $Var(X) = 64$. Considerando $\alpha = 2\%$, verifique se a hipótese nula deve ser rejeitada ou não.

Referências

ANDERSON, David R.; SWEENEY, Dennis J.; WILLIAMS, Thomas A. **Estatística aplicada à administração e economia**. 2. ed. São Paulo: Cengage Learning, 2011.

CLIMATEMPO. Disponível em: <<http://www.climatempo.com.br>>. Acesso em: 18 jun. 2015.

CRESPO, Antônio A. **Estatística fácil**. 17. ed. São Paulo: Saraiva, 2002.

FREUND, John E. *Estatística aplicada: economia, administração e contabilidade*. 11. ed. Porto Alegre: Bookman, 2006.

FUTPÉDIA. Disponível em: <<http://futpedia.globo.com/campeonato/copa-do-mundo>>. Acesso em: 13 maio 2015.

IBGE – Instituto Brasileiro de Geografia e Estatística. **População presente e residente**. Disponível em: <www.ibge.gov.br>. Acesso em: 14 maio 2015.

IBGE – Instituto Brasileiro de Geografia e Estatística. **Produção vegetal**. Disponível em: <www.ibge.gov.br>. Acesso em: 14 maio 2015.

JOHNSON, Robert; KUBY, Patricia. **Estatística**. São Paulo: Cengage Learning, 2013.

MEDEIROS, Valéria Z. (Coord.). **Métodos quantitativos com Excel**. São Paulo: Cengage Learning, 2008.

MORETTIN, Luiz G. **Estatística básica: probabilidade e inferência**. São Paulo: Pearson Prentice Hall, 2010.

MORETTIN, Luiz G.; BUSSAB, Wilton O. **Estatística básica**. São Paulo: Saraiva, 2010.

UOL Esporte. Disponível em: <<http://esporte.uol.com.br/futebol/biografias/559/pele>>. Acesso em: 28 abr. 2015.

Estatística Inferencial (parte II)

Convite ao estudo

Muitas das pesquisas e investigações que realizamos têm o objetivo de verificar a existência de relação entre duas variáveis. Você viu um exemplo disso na Unidade 1, quando foi descrita a lei da oferta e da demanda. Lembre-se de que o preço (X) e a quantidade ofertada (Y) possuem uma relação direta, ou seja, um aumento no preço implica um aumento na quantidade ofertada; já o preço (X) e a quantidade demandada (Z) possuem uma relação inversa, isto é, um aumento no preço ocasiona uma redução na quantidade demandada.

A lei da oferta e da demanda indica que as variáveis X e Y estão relacionadas, assim como as variáveis X e Z . Uma vez cientes da existência de relação entre duas variáveis, podemos fazer diversos questionamentos: (1) a relação entre as duas variáveis é forte ou fraca? (2) a relação é direta ou inversa? (3) como medimos a relação entre duas variáveis? (4) por que estudar a relação entre duas variáveis?

Para dar um direcionamento às possíveis respostas para essas perguntas, vamos retomar duas situações abordadas anteriormente: a da unidade 1, em que solicitamos que você se pusesse no papel de um vendedor que necessitava determinar a demanda de mercado de um produto; e a da unidade 2, em que foi sugerido que você supusesse que era um funcionário de uma grande empresa e deveria descrever o perfil dos funcionários.

Se soubermos que duas variáveis estão relacionadas, teremos a garantia de que, ao haver uma modificação em uma delas, a outra também será alterada. Com isso, saber se a relação é forte ou fraca, direta ou inversa, implica, na modificação de uma variável,

conhecer a magnitude da alteração que ocorrerá na outra variável e o sentido dessa alteração (positivo ou negativo). Vimos algo semelhante quando calculamos a elasticidade, na Unidade 1.

A principal motivação para estudarmos a relação entre duas variáveis é a possibilidade de prever resultados futuros ou inferir valores não amostrados de uma população. Lembre-se de que, na Unidade 2, foi perguntado aos funcionários da empresa M qual era a avaliação deles em relação às condições de trabalho e à remuneração. Imagine novamente que você é o funcionário citado na Unidade 2; será que essas variáveis estão relacionadas? Quanto maior a remuneração, maior a satisfação do funcionário?

Seção 4.1

Correlação entre variáveis quantitativas

Diálogo aberto

Nesta seção você aprenderá a medir o grau de associação entre duas variáveis. Mensuramos essa associação por meio do coeficiente de correlação. Para ilustrar esse conceito, imagine novamente que você é um funcionário da empresa M e que necessita avaliar a relação existente entre a satisfação em relação às condições de trabalho e a satisfação em relação à remuneração. Será que, quanto maior é a satisfação em relação à remuneração, mais satisfeitos ficam os funcionários em relação às condições de trabalho?

Para responder a essas perguntas você deverá elaborar um diagrama de dispersão e calcular o coeficiente de correlação.

Não pode faltar

Bastante ênfase foi dada até o momento para o tratamento de cada variável separadamente, estudada em dada população. Análises com essa característica são denominadas **univariadas**. O que ocorre é que nem sempre estamos interessados em estudar uma única variável de cada vez, mas sim duas ou mais e a relação entre elas. Tais análises são denominadas **multivariadas**. Neste livro nos limitaremos a estudar o caso **bivariado**, ou seja, a análise de duas variáveis simultaneamente.

Veja como exemplo os dados da Tabela 4.1, amostrados a partir da população de crianças de 0 a 5 anos em determinada cidade.

Tabela 4.1 | Idade e altura de uma amostra de 24 crianças

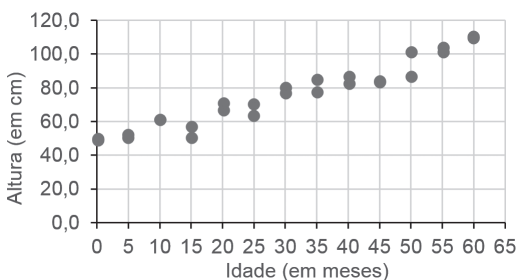
Idade (em meses)	0	0	5	5	10	10	15	15	20	20	25	25
Altura (em cm)	49,1	49,7	52,4	51,0	61,6	61,2	57,1	50,4	71,6	67,0	63,6	70,7
Idade (em meses)	35	35	40	40	45	45	50	50	55	55	60	60
Altura (em cm)	85,3	77,5	87,0	82,8	83,5	84,1	86,5	101,7	101,3	104,2	110,1	110,7

Fonte: Os autores (2015).

Observe que há um total de 24 observações, ou seja, uma amostra de 24 crianças. Além disso, de cada criança foram coletadas duas informações, a saber, a idade em meses e a altura em centímetros. Se denominarmos X a variável idade e Y a variável altura, também podemos escrever as informações anteriores da forma (X, Y) , em que o primeiro valor se refere à idade e o segundo à altura:

(0; 49,1), (0; 49,7), (5; 52,4), (5; 51), (10; 61,6), (10; 61,2), (15; 57,1), (15; 50,4), (20; 71,6), (20; 67), (25; 63,6), (25; 70,7), (30; 80,6), (30; 77,2), (35; 85,3), (35; 77,5), (40; 87), (40; 82,8), (45; 83,5), (45; 84,1), (50; 86,5), (50; 101,7), (55; 101,3), (55; 104,2), (60; 110,1), (60; 110,7)

Figura 4.1 | Idade e altura de uma amostra de 24 crianças



Fonte: Os autores (2015).

A escrita em **pares ordenados** (X, Y) – ou também $(X; Y)$ – é muito comum no âmbito da análise bivariada, pois deixa bem clara a associação do valor de X com o seu Y correspondente, na medida em que ambos foram coletados de um mesmo elemento da população (no caso, da mesma criança). Podemos representar essas informações em um gráfico de dispersão, como se observa na Figura 4.1.

Você aprendeu anteriormente que um gráfico tem o objetivo de facilitar a leitura e a interpretação dos dados, além de dar uma ideia da distribuição de uma variável. Quando a análise é bivariada, os gráficos também têm o objetivo de investigar a presença de uma relação entre as variáveis. Observando a Figura 4.1, o que você imagina em relação às variáveis X e Y ? Esperamos que você tenha percebido que, quanto maior a idade, maior a altura. Essa ideia nos parece óbvia, mas nem sempre a relação de dependência entre duas variáveis é tão clara assim.

Uma vez aceita a hipótese de relação de dependência entre duas variáveis, surgem duas perguntas básicas: (1ª) essa relação é forte ou fraca? (2ª) de que forma podemos mensurar essa relação?

Observando a Figura 4.1, imaginamos que se os pontos estivessem um pouco mais organizados quase daria para traçar uma linha reta passando por todos eles. Essa nossa percepção indica que a relação de dependência entre X e Y é forte e, além disso, linear. Quando isso ocorre, dizemos que existe uma **correlação linear** entre as variáveis. Veja mais alguns exemplos na Figura 4.2, em que no eixo horizontal é representada uma variável X e no eixo vertical uma variável Y .

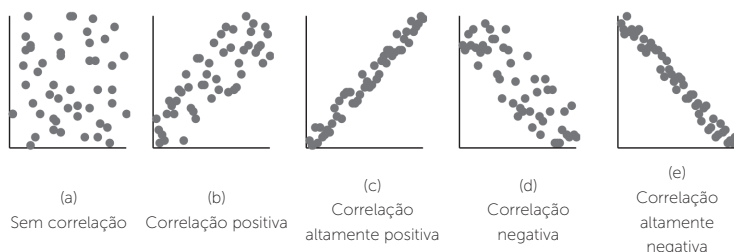


Assimile

Correlação: diz-se que duas variáveis estão correlacionadas quando existe uma relação de dependência entre elas.

Correlação linear: duas variáveis estão correlacionadas linearmente quando a relação entre elas pode ser representada geometricamente por meio de uma reta.

Figura 4.2 | Diagramas de dispersão e correlação linear



Fonte: Os autores (2015).

A Figura 4.2 (a) mostra um caso em que a variável X e a variável Y não estão correlacionadas, isto é, a variação de Y não é explicada pela variação de X . Já na Figura 4.2 (b) e (c), há uma correlação linear positiva entre as duas variáveis, e, além disso, a variação de Y é mais bem explicada pela variação de X em (c). Por fim, na Figura 4.2 (d) e (e), há uma correlação linear negativa entre as duas variáveis, e, além disso, a variação de Y é mais bem explicada pela variação de X em (e).



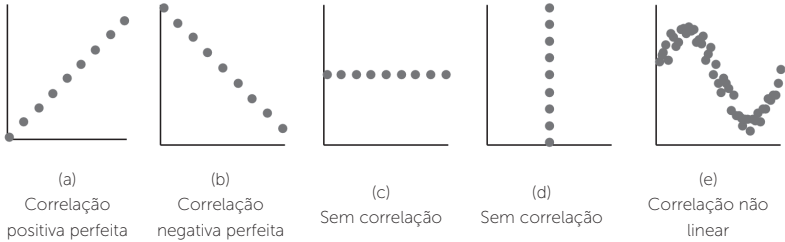
Assimile

Dizer que a correlação é **positiva** implica afirmar que, quando X aumenta, Y também aumenta; quando X diminui, Y também diminui. Se a correlação é **negativa** ocorre o contrário: se X aumenta, Y diminui; se X diminui, Y aumenta.

Há ainda outros casos interessantes, os quais podem ser observados na Figura 4.3 a seguir. Na figura, em (a) e (b) há a correlação linear perfeita,

em que todos os pontos se encontram sobre uma mesma reta. Apesar de em (c) e (d) os pontos estarem sobre uma mesma reta, não há correlação entre as variáveis, pois a variação de uma não é explicada pela variação da outra. Por fim, na Figura 4.3 (e) a correlação entre as variáveis existe, mas não é linear.

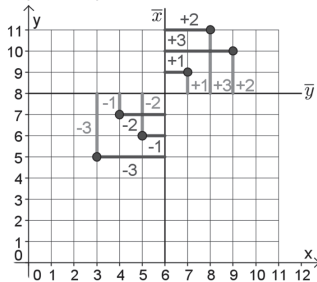
Figura 4.3 | Mais exemplos de diagrama de dispersão



Fonte: Os autores (2015).

Uma vez identificada uma correlação linear entre duas variáveis X e Y , como quantificá-la? Para responder a essa pergunta, observe o exemplo da Figura 4.4. Note que os pontos correspondem aos dados bivariados (3,5), (4,7), (5,6), (7,9), (8,11) e (9,10). Além disso, estão traçadas as retas \bar{x} e \bar{y} , em que $\bar{x} = \frac{\sum x}{6}$ e $\bar{y} = \frac{\sum y}{6}$, isto é, as médias aritméticas dos valores de X e Y , respectivamente. As retas \bar{x} e \bar{y} se cruzam no ponto (6,8), denominado centroide. Também estão representados no diagrama os desvios de cada valor em relação à média.

Figura 4.4 | Diagrama de dispersão para X e Y



Fonte: Os autores (2015).

Nesse exemplo, se multiplicarmos os desvios de X pelos desvios correspondentes de Y teremos somente valores positivos, como mostra a Tabela 4.2. Observe que $\sum(x_i - \bar{x})(y_i - \bar{y}) = 9 + 2 + 2 + 1 + 6 + 6 = 26 > 0$, o que define que a correlação entre as variáveis X e Y é positiva. Se obtivéssemos $\sum(x_i - \bar{x})(y_i - \bar{y}) < 0$, diríamos que a correlação

seria negativa; e se $\sum(x_i - \bar{x})(y_i - \bar{y}) = 0$, X e Y seriam variáveis não correlacionadas. Definimos, então, a covariância.

Tabela 4.2 | Produtos dos desvios

X	3	4	5	7	8	9
Y	5	7	6	9	11	10
$x_i - \bar{x}$	-3	-2	-1	+1	+2	+3
$y_i - \bar{y}$	-3	-1	-2	+1	+3	+2
$(x_i - \bar{x})(y_i - \bar{y})$	9	2	2	1	6	6

Fonte: Os autores (2015).



Assimile

Sendo X e Y duas variáveis contínuas, a covariância entre X e Y é dada por $Cov(X, Y) = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n-1}$, com $n \geq 2$.

Observe que ao dividir o número $\sum(x_i - \bar{x})(y_i - \bar{y})$ por $n - 1$ a interpretação feita anteriormente continua a mesma, pois $n - 1 > 0$. Podemos ainda escrever que duas variáveis X e Y :

- Estão correlacionadas positivamente se $Cov(X, Y) > 0$;
- Estão correlacionadas negativamente se $Cov(X, Y) < 0$;
- Não estão correlacionadas se $Cov(X, Y) = 0$.



Exemplificando

Observe os dados obtidos por amostragem para as variáveis X, Y, Z e W .

X	5	10	15	20	25	30
Y	9	15	18	26	29	31
Z	125	130	75	50	50	0
W	90	2	85	8	106	43

Calcule $Cov(X, Y)$, $Cov(X, Z)$ e $Cov(X, W)$ e classifique os pares de variáveis quanto à correlação.

Resolução:

$$\bar{x} = 17,5; \bar{y} \cong 21,33; \bar{z} \cong 71,67; \bar{w} \cong 55,67.$$

$$\text{Cov}(X, Y) = \frac{(5-17,5)(9-21,33) + \dots + (30-17,5)(31-21,33)}{6-1} \cong 80;$$

$$\text{Cov}(X, Z) = \frac{(5-17,5)(125-71,67) + \dots + (30-17,5)(0-71,67)}{6-1} \cong -445;$$

$$\text{Cov}(X, W) = \frac{(5-17,5)(90-55,67) + \dots + (30-17,5)(43-55,67)}{6-1} \cong 0.$$

Logo, X e Y estão correlacionadas positivamente, X e Z estão correlacionadas negativamente e X e W não estão correlacionadas.

Neste momento pode ter surgido uma dúvida: quanto maior é a magnitude da covariância, mais fortemente estão relacionadas as variáveis? A resposta é não. A covariância é influenciada pela escala, logo, quanto maiores os valores de um conjunto de dados, maiores as chances de a covariância assumir valores mais elevados. Uma maneira de corrigir isso é utilizar variáveis padronizadas $(x_i - \bar{x})/Dp(X)$ e $(y_i - \bar{y})/Dp(Y)$ e definir uma nova medida, o **coeficiente de correlação**:

$$r = \rho(X, Y) = \frac{\sum \frac{(x_i - \bar{x})}{Dp(X)} \cdot \frac{(y_i - \bar{y})}{Dp(Y)}}{n-1} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{(n-1) \cdot Dp(X) \cdot Dp(Y)} = \frac{\text{Cov}(X, Y)}{Dp(X) \cdot Dp(Y)}.$$

Com essa transformação, $-1 \leq r \leq +1$. Além disso, se:

- $r > 0$, as variáveis estão correlacionadas positivamente;
- $r < 0$, as variáveis estão correlacionadas negativamente;
- $r = 0$, as variáveis não estão correlacionadas;
- $r = +1$, temos uma correlação positiva perfeita;
- $r = -1$, temos uma correlação negativa perfeita.

Quanto mais próximo de 1 se encontra o valor de $|r|$, mais forte é a correlação; quanto mais próximo de 0 se encontra o valor de $|r|$, mais fraca é a correlação. Além disso, se r_{xy} e r_{zw} são os coeficientes de correlação das variáveis X e Y e das variáveis Z e W , respectivamente, $|r_{xy}| > |r_{zw}|$ implica que X e Y estão mais fortemente correlacionadas do que Z e W .



Exemplificando

Utilizando as variáveis do exemplo anterior, calcule os coeficientes de correlação $\rho(X, Y)$, $\rho(X, Z)$ e $\rho(X, W)$ e verifique quais variáveis estão mais fortemente correlacionadas.

Resolução:

Temos $Dp(X) \cong 9,35$, $Dp(Y) \cong 8,69$, $Dp(Z) \cong 49,67$ e $Dp(W) \cong 44,46$.

Logo:

$$\rho(X, Y) = \frac{Cov(X, Y)}{Dp(X) \cdot Dp(Y)} = \frac{80}{9,35 \cdot 8,69} \cong 0,98;$$

$$\rho(X, Z) = \frac{Cov(X, Z)}{Dp(X) \cdot Dp(Z)} = \frac{-445}{9,35 \cdot 49,67} \cong -0,96;$$

$$\rho(X, W) = \frac{Cov(X, W)}{Dp(X) \cdot Dp(W)} = \frac{0}{9,35 \cdot 44,46} \cong 0.$$

Portanto, as variáveis X e Y estão mais fortemente correlacionadas do que as variáveis X e Z e do que as variáveis X e W .

Existe uma forma alternativa (mais prática) de calcular o coeficiente de correlação. Para utilizá-la, é necessário definir $SQ(x)$, $SQ(y)$ e $SQ(xy)$:

- Soma dos quadrados para x : $SQ(x) = \sum x^2 - \frac{(\sum x)^2}{n}$;
- Soma dos quadrados para y : $SQ(y) = \sum y^2 - \frac{(\sum y)^2}{n}$;
- Soma dos quadrados para x e y : $SQ(xy) = \sum xy - \frac{(\sum x)(\sum y)}{n}$.

Com essa definição, temos:

$$r = \rho(X, Y) = \frac{SQ(xy)}{\sqrt{SQ(x) \cdot SQ(Y)}}.$$



Exemplificando

Utilizando a fórmula $r = \rho(X, Y) = \frac{SQ(xy)}{\sqrt{SQ(x) \cdot SQ(Y)}}$, calcule o coeficiente

de correlação para as variáveis X e Y , cujos dados amostrais foram apresentados na Tabela 4.1, e classifique as variáveis quanto à correlação.

Resolução:

$$SQ(x) = \sum x^2 - \frac{(\sum x)^2}{n} = (0^2 + 0^2 + \dots + 60^2 + 60^2) - \frac{(0 + 0 + \dots + 60 + 60)^2}{24};$$

$$SQ(x) = 30700 - \frac{720^2}{24} = 9100.$$

$$SQ(y) = \sum y^2 - \frac{(\sum y)^2}{n} = (49,1^2 + \dots + 110,7^2) - \frac{(49,1 + \dots + 110,7)^2}{24};$$

$$SQ(y) = 147300,45 - \frac{1820,1^2}{24} \cong 9268,62.$$

$$SQ(xy) = \sum xy - \frac{(\sum x)(\sum y)}{n};$$

$$SQ(xy) = (0 \cdot 49,1 + \dots + 60 \cdot 110,7) - \frac{(0 + \dots + 60)(49,1 + \dots + 110,7)}{24};$$

$$SQ(xy) = 63479,5 - \frac{720 \cdot 1820,1}{24} = 8876,5.$$

$$r = \rho(X, Y) = \frac{SQ(xy)}{\sqrt{SQ(x) \cdot SQ(y)}} = \frac{8876,5}{\sqrt{9100 \cdot 9268,62}} \cong 0,967.$$

Portanto, as variáveis X e Y estão positivamente correlacionadas.



Pesquise mais

Complemente e aprofunde seus estudos sobre covariância e coeficiente de correlação através do link: <http://www.cprm.gov.br/publique/media/cap9-correl_regres.pdf> (acesso em: 06 jul. 2015).

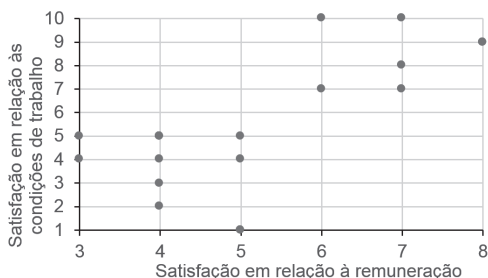
Sem medo de errar

Observe a seguir os dados referentes às variáveis G: satisfação em relação às condições de trabalho e H: satisfação em relação à remuneração.

H	7	4	5	5	7	4	5	4	4	4	8	7	4	3	4	3	5	6	6	8
G	8	5	1	4	10	5	5	5	4	5	9	7	3	4	2	5	1	7	10	9

O diagrama de dispersão para os dados pode ser observado na Figura 4.5.

Figura 4.5 | Diagrama de dispersão para G e H



Fonte: Os autores (2015).

Observa-se no diagrama que existe uma tendência positiva nos dados, ou seja, quanto maior a satisfação em relação à remuneração, maior a satisfação em relação às condições de trabalho. Vamos agora medir o grau de associação de H e G.

Temos:

$$SQ(h) = (7^2 + 4^2 + \dots + 6^2 + 8^2) - \frac{(7 + 4 + \dots + 6 + 8)^2}{20} = 577 - \frac{103^2}{20} = 46,55;$$

$$SQ(g) = (8^2 + 5^2 + \dots + 10^2 + 9^2) - \frac{(8 + 5 + \dots + 10 + 9)^2}{20} = 737 - \frac{109^2}{20} \cong 142,95;$$

$$SQ(hg) = (7 \cdot 8 + 4 \cdot 5 + \dots + 6 \cdot 10 + 8 \cdot 9) - \frac{(7 + 4 + \dots + 6 + 8)(8 + 5 + \dots + 10 + 9)}{20};$$

$$SQ(hg) = 619 - \frac{103 \cdot 109}{20} = 57,65.$$

$$r = \rho(H, G) = \frac{SQ(hg)}{\sqrt{SQ(h) \cdot SQ(g)}} = \frac{57,65}{\sqrt{46,55 \cdot 142,95}} \cong 0,707.$$

Portanto, as variáveis H e G estão correlacionadas positivamente e, além disso, como $r \cong 0,707$, essa correlação é forte.

Pratique mais!

Instrução

Desafiamos você a praticar o que aprendeu transferindo seus conhecimentos para novas situações que pode encontrar no ambiente de trabalho. Realize as atividades e depois as compare com as de seus colegas.

1. Competência de fundamentos de área	Conhecer os conceitos matemáticos básicos e proporcionar o desenvolvimento do raciocínio lógico e quantitativo.													
2. Objetivos de aprendizagem	Mensurar a relação entre duas variáveis por meio do coeficiente de correlação linear.													
3. Conteúdos relacionados	Correlação; covariância; coeficiente de correlação.													
4. Descrição da situação-problema	A seguir, consta o valor gasto com propaganda e a quantidade vendida de um produto no mesmo mês.													
	<table border="1"> <tbody> <tr> <td>Gastos com propaganda (× R\$ 1.000)</td> <td>10,0</td> <td>11,0</td> <td>12,2</td> <td>13,8</td> <td>14,4</td> <td>15,5</td> </tr> <tr> <td>Unidades vendidas (× 10.000)</td> <td>9,8</td> <td>9,7</td> <td>12,6</td> <td>14,4</td> <td>13,6</td> <td>16,2</td> </tr> </tbody> </table> <p>Verifique se essas variáveis estão correlacionadas linearmente.</p>	Gastos com propaganda (× R\$ 1.000)	10,0	11,0	12,2	13,8	14,4	15,5	Unidades vendidas (× 10.000)	9,8	9,7	12,6	14,4	13,6
Gastos com propaganda (× R\$ 1.000)	10,0	11,0	12,2	13,8	14,4	15,5								
Unidades vendidas (× 10.000)	9,8	9,7	12,6	14,4	13,6	16,2								
5. Resolução da situação-problema	<p>Considere X: gasto com propaganda e Y: unidades vendidas. Temos:</p> $SQ(x) = (10,0^2 + \dots + 15,5^2) - \frac{(10,0 + \dots + 15,5)^2}{6}$ $\cong 1007,89 - \frac{76,9^2}{6} \cong 22,288;$ $SQ(y) = (9,8^2 + \dots + 16,2^2) - \frac{(9,8 + \dots + 16,2)^2}{6}$ $\cong 1003,65 - \frac{76,3^2}{6} \cong 33,368;$ $SQ(xy)$ $= (10,0 \cdot 9,8 + \dots + 15,5 \cdot 16,2) - \frac{(10,0 + \dots + 15,5)(9,8 + \dots + 16,2)}{6} \cong 1004,08 - \frac{76,9 \cdot 76,3}{6}$ $\cong 26,168;$ $r = \rho(X, Y) = \frac{SQ(xy)}{\sqrt{SQ(x) \cdot SQ(y)}} = \frac{26,168}{\sqrt{22,288 \cdot 33,368}} \cong 0,960.$ <p>Como $r \cong 0,960$, as variáveis X e Y estão fortemente correlacionadas linearmente e positivamente.</p>													



Duas variáveis estão **correlacionadas** quando existe uma relação de dependência entre elas.

Dizer que a correlação é **positiva** implica afirmar que, quando X aumenta, Y também aumenta; quando X diminui, Y também diminui. Se a correlação é **negativa** ocorre o contrário: se X aumenta, Y diminui; se X diminui, Y aumenta.

Duas variáveis podem ser classificadas como: correlacionadas positivamente, se $r > 0$; correlacionadas negativamente, se $r < 0$; não correlacionadas, se $r = 0$; correlacionadas positiva e perfeitamente, se $r = +1$; e correlacionadas negativa e perfeitamente, se $r = -1$.



Faça uma amostragem com os seus colegas de classe sobre a satisfação em relação às condições de trabalho e a satisfação em relação à remuneração. Em seguida, verifique se essas variáveis estão correlacionadas.

Faça valer a pena

1. Considere o conjunto de dados bivariados (X, Y) , em que, por amostragem, coletou-se:

(5, 3), (14, 11), (15, 14), (5, 3), (9, 11), (13, 14), (7, 4)

Assinale a alternativa que contém o valor aproximado de $Cov(X, Y)$.

- a) 20,02 c) 22,22 e) 20,22
b) 22,02 d) 22,20

2. Assinale a alternativa que contém o coeficiente de correlação do conjunto:

(16, 59), (16, 39), (47, 68), (23, 22), (15, 55), (34, 48)

- a) 0,8543 c) 0,3584 e) 0,3845
b) 0,5834 d) 0,3485

3. Considere o seguinte conjunto de dados, obtidos por amostragem.

<i>X</i>	75	59	32	54	20
<i>Y</i>	78	63	39	59	26
<i>Z</i>	13	23	54	31	63
<i>W</i>	9	87	12	93	56

Assinale a alternativa correta:

- a) *X* e *Y* não estão correlacionadas.
- b) *X* e *W* estão correlacionadas positivamente.
- c) *X* e *Z* estão correlacionadas negativamente.
- d) *X* e *Y* estão positivamente correlacionadas, assim como *X* e *Z*.
- e) *X* e *Z* não estão correlacionadas.

4. Considere as variáveis *X*, *Y*, *Z* e *W*, para as quais temos $Cov(X, Y) = 10$, $Cov(X, Z) = 15$, $Cov(X, W) = 18$, $Dp(X) = 2$, $Dp(Y) = 6$, $Dp(Z) = 8$ e $Dp(W) = 10$. Assinale a alternativa correta:

- a) $\rho(X, Y) = \rho(X, Z) < \rho(X, W)$
- b) $\rho(X, Y) < \rho(X, Z) < \rho(X, W)$
- c) $\rho(X, Y) < \rho(X, W) < \rho(X, Z)$
- d) $\rho(X, Y) > \rho(X, Z) = \rho(X, W)$
- e) $\rho(X, Y) > \rho(X, Z) > \rho(X, W)$

5. Considere as variáveis *X*, *Y* e *Z*, tais que $Cov(X, Y) = 50$, $Cov(X, Z) = -60$, $Dp(X) = 10$, $Dp(Y) = 15$ e $Dp(Z) = 10$. Assinale a alternativa correta:

- a) $\rho(X, Y) > \rho(X, Z)$, o que indica que as variáveis *X* e *Y* estão mais fortemente relacionadas do que *X* e *Z*.
- b) $|\rho(X, Z)| > |\rho(X, Y)|$, o que indica que as variáveis *X* e *Z* estão mais fortemente relacionadas do que *X* e *Y*.
- c) $|\rho(X, Z)| = |\rho(X, Y)|$.
- d) $|\rho(X, Z)| < |\rho(X, Y)|$, o que indica que as variáveis *X* e *Y* estão mais fortemente relacionadas do que *X* e *Z*.
- e) $|\rho(X, Z)| = |\rho(X, Y)|$.

6. Classifique as variáveis X e Y como correlacionadas positivamente, correlacionadas negativamente ou não correlacionadas.

x	40	68	17	41	41	65
y	51	19	73	55	45	32

7. Considere os valores amostrados para as variáveis X , Y e Z a seguir.

X	118	122	139	119	127
Y	167	170	190	177	186
Z	189	193	177	191	190

Verifique quais variáveis estão mais fortemente correlacionadas: X e Y ou X e Z .

Seção 4.2

Teste de significância

Diálogo aberto

Você aprendeu na seção anterior a mensurar a correlação entre duas variáveis quantitativas por meio do coeficiente de correlação. Vale ressaltar que esse coeficiente mede o grau de associação **linear** entre duas variáveis, isto é, mede o quanto os pontos (X, Y) em um diagrama de dispersão se aproximam de uma reta.

As análises feitas para avaliar a força de associação entre as variáveis X e Y foram apenas subjetivas, considerando quão próximo o coeficiente de correlação se encontrava de -1 ou $+1$. Entretanto, em estatística, a ferramenta utilizada para comprovar algo é o teste estatístico de hipóteses. Logo, além de calcularmos o coeficiente de correlação r , precisamos verificar sua significância. Para prosseguirmos com essa análise, vamos relembrar a situação-problema proposta na seção anterior: imagine novamente que você é um funcionário da empresa M e que necessita avaliar a relação existente entre a satisfação em relação às condições de trabalho e a satisfação em relação à remuneração. Será que, quanto maior é a satisfação em relação à remuneração, mais satisfeitos ficam os funcionários em relação às condições de trabalho?

Verificamos na seção anterior que o coeficiente de correlação para a amostra apresentada na Tabela 2.1 é $r \cong 0,707$, e afirmamos que nesse caso a correlação é forte. A fim de sustentarmos essa afirmação, precisamos testá-la. Para isso, que procedimentos devemos adotar?

Não pode faltar

Você aprendeu que o coeficiente de correlação r é calculado a partir de dados bivariados (X, Y) e mede o grau de associação entre as variáveis X e Y . O coeficiente r varia no intervalo $[-1, +1]$, e, além disso:



- Se $r > 0$, a correlação entre X e Y é positiva, e, quanto mais próximo r estiver de $+1$, mais fortemente as variáveis estão correlacionadas.
- Se $r < 0$, a correlação entre X e Y é negativa, e, quanto mais próximo r estiver de -1 , mais fortemente as variáveis estão correlacionadas.
- Se $r = 0$, não há correlação entre X e Y .

Obviamente, se $r \cong 0$, mas não exatamente igual a zero, temos indícios de que as variáveis não estão correlacionadas.

O teste de hipóteses utilizado para testar a força de uma correlação por meio do coeficiente r é denominado **teste de significância**. Segundo Larson e Farber (2010, p. 403):

"[...] as hipóteses nula e alternativa para os testes são:

$\left\{ \begin{array}{l} H_0 : \rho \geq 0 \text{ (não há correlação negativa significativa)} \\ H_1 : \rho < 0 \text{ (correlação negativa significativa)} \end{array} \right.$	Teste unilateral à esquerda
$\left\{ \begin{array}{l} H_0 : \rho \leq 0 \text{ (não há correlação positiva significativa)} \\ H_1 : \rho > 0 \text{ (correlação positiva significativa)} \end{array} \right.$	Teste unilateral à direita
$\left\{ \begin{array}{l} H_0 : \rho = 0 \text{ (não há correlação significativa)} \\ H_1 : \rho \neq 0 \text{ (correlação significativa)} \end{array} \right.$	Teste bilateral

"[...]".

Além disso:



"[...] um teste t pode ser usado se a correlação entre duas variáveis for significativa. A estatística de teste é r e a estatística de teste padronizada

$$t_c = \frac{r}{\sigma_r} = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

segue uma distribuição t com $n - 2$ graus de liberdade."

Para facilitar a compreensão, vamos testar a significância do coeficiente de correlação para os dados apresentados na Figura 4.1.



Na seção anterior apresentamos o diagrama de dispersão para os dados correspondentes à idade (X) e à altura (Y) de uma amostra de 24 crianças de 0 a 5 anos. Ao final da seção, obtivemos $r = 0,967$ para a correlação entre essas variáveis. Com 95% de confiança, o valor $r = 0,967$ indica que a correlação é significativa?

Resolução:

Para testar a significância da correlação, executamos os seguintes passos:

Passo 1 (elaborar as hipóteses):

$H_0: \rho = 0$ (não há correlação significativa).

$H_1: \rho \neq 0$ (correlação significativa).

Passo 2 (determinar a estatística de teste):

$$t_c = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} \text{ com } v = n - 2 \text{ graus de liberdade.}$$

Passo 3 (fixar o nível de significância): $\alpha = 100\% - 95\% = 5\%$

Passo 4 (calcular a estatística a partir da amostra):

Rejeitaremos a hipótese H_0 caso o valor t_c obtido a partir da amostra seja muito maior ou muito menor que $\rho = 0$ ou, ainda, quando t_c pertencer à região crítica $RC = \{T = \mathbb{R} | T \leq -t \text{ ou } T \geq t\}$, em que t é obtido na tabela T. Observando a tabela na linha $v = 24 - 2 = 22$ e na coluna correspondente à probabilidade 2,5% (pois o teste é bilateral), temos $t = 2,074$. Logo, $RC = \{T \in \mathbb{R} | T \leq -2,074 \text{ ou } T \geq 2,074\}$.

Obtivemos $r = 0,967$ a partir de uma amostra de tamanho $n = 24$, logo, calculamos:

$$t_c = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{0,967}{\sqrt{\frac{1-0,967^2}{24-2}}} \cong 17,802 \in RC$$

Passo 5 (tomar uma decisão): como $t_c \in RC$, decidimos rejeitar H_0 , isto é, há indícios suficientes que nos permitem considerar a correlação entre idade e altura significativa.

Vejamos agora o caso de um teste de hipóteses unilateral à esquerda.



Exemplificando

Analise os dados bivariados na forma (X, Y) a seguir e verifique, para o nível de significância $\alpha = 2\%$, se a correlação entre X e Y é negativamente significativa.

$(54, 7), (60, 2), (48, 25), (57, 17), (57, 8)$

Resolução:

Temos:

$$SQ(x) = \sum x^2 - \frac{(\sum x)^2}{n} = (54^2 + \dots + 57^2) - \frac{(54 + \dots + 57)^2}{5} = 15318 - \frac{276^2}{5}$$

$$SQ(x) = 82,8$$

$$SQ(y) = \sum y^2 - \frac{(\sum y)^2}{n} = (7^2 + \dots + 8^2) - \frac{(7 + \dots + 8)^2}{5} = 1031 - \frac{59^2}{5}$$

$$SQ(y) = 334,8$$

$$SQ(xy) = \sum xy - \frac{(\sum x)(\sum y)}{n}$$

$$SQ(xy) = (54 \cdot 7 + \dots + 57 \cdot 8) - \frac{(54 + \dots + 57)(7 + \dots + 8)}{5}$$

$$SQ(xy) = 3123 - \frac{276 \cdot 59}{5} = -133,8$$

$$r = \frac{SQ(xy)}{\sqrt{SQ(x) \cdot SQ(y)}} = \frac{-133,8}{\sqrt{82,8 \cdot 334,8}} \cong -0,8036$$

Conhecendo-se o valor de r , podemos agora testar a significância.

Passo 1 (elaborar as hipóteses):

$H_0: \rho \geq 0$ (não há correlação negativa significativa).

$H_1: \rho < 0$ (correlação negativa significativa).

Passo 2 (determinar a estatística de teste):

$$t_c = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} \text{ com } v = n - 2 \text{ graus de liberdade.}$$

Passo 3 (fixar o nível de significância): $\alpha = 2\%$ (dado).

Passo 4 (calcular a estatística a partir da amostra):

Rejeitaremos a hipótese H_0 caso o valor t_c obtido a partir da amostra seja muito menor que $\rho = 0$ ou, ainda, quando t_c pertencer à região crítica $RC = \{T \in \mathbb{R} \mid T \leq -t\}$, em que t é obtido na tabela T. Observando a tabela na linha $v = 5 - 2 = 3$ e na coluna correspondente à probabilidade 2%, temos $t = 3,482$. Logo, $RC = \{T \in \mathbb{R} \mid T \leq -3,482\}$.

Obtivemos $r = -0,8036$ a partir de uma amostra de tamanho $n = 5$, logo, calculamos:

$$t_c = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{-0,8036}{\sqrt{\frac{1-(-0,8036)^2}{5-2}}} \cong -2,339 \notin RC$$

Passo 5 (tomar uma decisão): como $t_c \notin RC$, decidimos não rejeitar H_0 , isto é, não há indícios suficientes que nos permitam considerar a correlação entre X e Y negativamente significativa.

Por fim, vejamos o caso de um teste de hipóteses unilateral à direita.



Exemplificando

Analise os dados bivariados na forma (X, Y) a seguir e verifique, para o nível de significância $\alpha = 2\%$, se a correlação

entre X e Y é positivamente significativa.

(54, 49), (27, 35), (15, 6), (59, 64), (32, 42)

Resolução:

Temos:

$$SQ(x) = \sum x^2 - \frac{(\sum x)^2}{n} = (54^2 + \dots + 32^2) - \frac{(54 + \dots + 32)^2}{5} = 8375 - \frac{187^2}{5}$$

$$SQ(x) = 1381,2$$

$$SQ(y) = \sum y^2 - \frac{(\sum y)^2}{n} = (49^2 + \dots + 42^2) - \frac{(49 + \dots + 42)^2}{5} = 9522 - \frac{196^2}{5}$$

$$SQ(y) = 1838,8$$

$$SQ(xy) = \sum xy - \frac{(\sum x)(\sum y)}{n}$$

$$SQ(xy) = (54 \cdot 49 + \dots + 32 \cdot 42) - \frac{(54 + \dots + 32)(49 + \dots + 42)}{5}$$

$$SQ(xy) = 8801 - \frac{187 \cdot 196}{5} = 1470,6$$

$$r = \frac{SQ(xy)}{\sqrt{SQ(x) \cdot SQ(y)}} = \frac{1470,6}{\sqrt{1381,2 \cdot 1838,8}} \cong 0,923$$

Conhecendo-se o valor de r , podemos agora testar a significância.

Passo 1 (elaborar as hipóteses):

$H_0 : \rho \leq 0$ (não há correlação positiva significativa).

$H_1 : \rho > 0$ (correlação positiva significativa).

Passo 2 (determinar a estatística de teste):

$$t_c = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} \text{ com } \nu = n - 2 \text{ graus de liberdade.}$$

Passo 3 (fixar o nível de significância): $\alpha = 2\%$ (dado).

Passo 4 (calcular a estatística a partir da amostra):

Rejeitaremos a hipótese H_0 caso o valor t_c obtido a partir da amostra seja muito maior que $\rho = 0$ ou, ainda, quando t_c pertencer à região crítica $RC = \{T \in \mathbf{R} \mid T \geq t\}$, em que t é obtido na tabela T. Observando a tabela na linha $\nu = 5 - 2 = 3$ e na coluna correspondente à probabilidade 2%, temos $t = 3,482$. Logo, $RC = \{T \in \mathbf{R} \mid T \geq 3,482\}$.

Obtivemos $r = 0,923$ a partir de uma amostra de tamanho $n = 5$, logo, calculamos:

$$t_c = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{0,923}{\sqrt{\frac{1-(0,923)^2}{5-2}}} \cong 4,155 \in RC$$

Passo 5 (tomar uma decisão): como $t_c \in RC$, decidimos rejeitar H_0 , isto é, há indícios suficientes que nos permitem considerar a correlação entre X e Y positivamente significativa.



Pesquise mais

Leia mais sobre o teste de significância no material disponível no link: <http://people.ufpr.br/~jomarc/correlacao.pdf> (acesso em: 13 jul. 2015).

Sem medo de errar

Vamos relembra a situação-problema proposta no início desta seção: imagine novamente que você é um funcionário da empresa M e que necessita avaliar a relação existente entre a satisfação em relação às condições de trabalho e a satisfação em relação à remuneração. Será que quanto maior é a satisfação em relação à remuneração, mais satisfeitos ficam os funcionários em relação às condições de trabalho?

Na seção anterior foi verificado que o coeficiente de correlação entre G: satisfação em relação às condições de trabalho e H: satisfação em relação à remuneração é $r \cong 0,707$. Além disso, essa medida foi obtida a partir de uma amostra de tamanho $n = 20$, apresentada na Tabela 2.1. Para testar a significância de r , executamos os seguintes passos:

Passo 1 (elaborar as hipóteses):

$H_0 : \rho \leq 0$ (não há correlação positiva significativa).

$H_1 : \rho > 0$ (correlação positiva significativa).

Passo 2 (determinar a estatística de teste):

$$t_c = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} \text{ com } v = n - 2 \text{ graus de liberdade.}$$

Passo 3 (fixar o nível de significância): suponha $\alpha = 5\%$.

Passo 4 (calcular a estatística a partir da amostra):

Rejeitaremos a hipótese H_0 caso o valor t_c obtido a partir da amostra seja muito maior que $\rho = 0$ ou, ainda, quando t_c pertencer à região crítica $RC = \{T \in R \mid T \geq t\}$, em que t é obtido na tabela T. Observando a tabela na linha $v = 20 - 2 = 18$ e na coluna correspondente à probabilidade 5%, temos $t = 1,734$. Logo, $RC = \{T \in R \mid T \geq 1,734\}$.

Obtivemos $r \cong 0,707$ a partir de uma amostra de tamanho $n = 20$, logo, calculamos:

$$t_c = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{0,707}{\sqrt{\frac{1-(0,707)^2}{20-2}}} \cong 4,241 \in RC$$

Passo 5 (tomar uma decisão): como $t_c \in RC$, decidimos rejeitar H_0 , isto é, há indícios suficientes que nos permitem considerar a correlação entre G e H positivamente significante.

Avançando na prática

Pratique mais!															
Instrução															
Desafiamos você a praticar o que aprendeu transferindo seus conhecimentos para novas situações que pode encontrar no ambiente de trabalho. Realize as atividades e depois as compare com as de seus colegas.															
1. Competência de fundamentos de área	Conhecer os conceitos matemáticos básicos e proporcionar o desenvolvimento do raciocínio lógico e quantitativo.														
2. Objetivos de aprendizagem	Testar a significância da correlação entre duas variáveis.														
3. Conteúdos relacionados	Coeficiente de correlação; teste de significância.														
4. Descrição da situação-problema	A seguir, consta o valor gasto com propaganda e a quantidade vendida de um produto no mesmo mês.														
	<table border="1" style="width: 100%; border-collapse: collapse; text-align: center;"> <tbody> <tr> <td>Gastos com propaganda (x R\$ 1.000)</td> <td>10,0</td> <td>11,0</td> <td>12,2</td> <td>13,8</td> <td>14,4</td> <td>15,5</td> </tr> <tr> <td>Unidades vendidas (x 10.000)</td> <td>9,8</td> <td>9,7</td> <td>12,6</td> <td>14,4</td> <td>13,6</td> <td>16,2</td> </tr> </tbody> </table>	Gastos com propaganda (x R\$ 1.000)	10,0	11,0	12,2	13,8	14,4	15,5	Unidades vendidas (x 10.000)	9,8	9,7	12,6	14,4	13,6	16,2
	Gastos com propaganda (x R\$ 1.000)	10,0	11,0	12,2	13,8	14,4	15,5								
Unidades vendidas (x 10.000)	9,8	9,7	12,6	14,4	13,6	16,2									
Teste a significância da correlação linear entre as variáveis "gastos com propaganda" e "unidades vendidas".															

5. Resolução da situação-problema

Na seção anterior obtivemos para esses dados $r \cong 0,960$, e concluímos que as variáveis X : gastos com propaganda e Y : unidades vendidas estão correlacionadas linearmente e positivamente. Vamos pôr à prova essa afirmação testando a significância de r .

Passo 1 (elaborar as hipóteses):

$H_0 : \rho \leq 0$ (não há correlação positiva significativa).

$H_1 : \rho > 0$ (correlação positiva significativa).

Passo 2 (determinar a estatística de teste):

$$t_c = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} \text{ com } v = n - 2 \text{ graus de liberdade.}$$

Passo 3 (fixar o nível de significância): suponha $\alpha = 5\%$.

Passo 4 (calcular a estatística a partir da amostra):

Rejeitaremos a hipótese H_0 caso o valor t_c obtido a partir da amostra seja muito maior que $\rho = 0$ ou, ainda, quando t_c pertencer à região crítica $RC = \{T \in R \mid T \geq t\}$, em que t é obtido na tabela T. Observando a tabela na linha $v = 6 - 2 = 4$ e na coluna correspondente à probabilidade 5%, temos $t = 2,132$. Logo, $RC = \{T \in R \mid T \geq 2,132\}$.

Obtivemos $r \cong 0,960$ a partir de uma amostra de tamanho $n = 6$, logo, calculamos:

$$t_c = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{0,960}{\sqrt{\frac{1-(0,960)^2}{6-2}}} \cong 6,857 \in RC$$

Passo 5 (tomar uma decisão): como $t_c \in RC$, decidimos rejeitar H_0 , isto é, há indícios suficientes que nos permitem considerar a correlação entre X e Y positivamente significativa.



Lembre-se

Segundo Larson e Farber (2010, p. 403), “[...] um teste t pode ser usado se a correlação entre duas variáveis for significativa. A estatística

de teste é r e a estatística de teste padronizada $t_c = \frac{r}{\sigma_r} = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$

segue uma distribuição t com $n-2$ graus de liberdade”.



Faça você mesmo

Utilize os dados amostrados de seus colegas de classe na seção anterior, sobre a satisfação em relação às condições de trabalho e a satisfação em relação à remuneração, e teste a significância da correlação entre essas variáveis com um teste bilateral ao nível de confiança de 95%.

Faça valer a pena

1. Considere o conjunto de dados bivariados (X, Y) em que, por amostragem, coletou-se:

(5, 3), (14, 11), (15, 14), (5, 3), (9, 11), (13, 14), (7, 4)

Assinale a alternativa que contém o valor aproximado do coeficiente de correlação entre X e Y .

- a) 0,92365 c) 0,93265 e) 0,35629
b) 0,92634 d) 0,29356

2. Observe a amostra coletada para os dados bivariados (X, Y) a seguir.

X	1276	1445	1681	1917	1953	1584
Y	1108	1688	1494	2127	2108	1696

Assinale a alternativa que contém o valor aproximado da estatística de

teste $t_c = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$ considerando os dados anteriores.

- a) 3,9905 c) 3,8940 e) 0,8094
b) 0,8940 d) 0,9804

6. O coeficiente de correlação entre as variáveis X e Y , calculado a partir de uma amostra de tamanho $n = 20$, é $r = -0,85$. Construa a região crítica para o teste bilateral de significância para r com $\alpha = 5\%$.

7. Há indícios de que a correlação entre as variáveis X e Y seja significativa. Para comprovar essa suspeita, coletou-se a amostra a seguir.

X	44	28	76	49
Y	62	41	92	60

Utilize a abordagem valor-p para testar bilateralmente a significância da correlação entre essas variáveis com 98% de confiança.

Seção 4.3

Regressão linear

Diálogo aberto

Você aprendeu nas seções anteriores a mensurar a correlação linear entre duas variáveis e a testar a significância dessa correlação por meio de um teste estatístico de hipóteses. Vale lembrar que o índice que foi utilizado (o coeficiente de correlação) mede a correlação linear, ou seja, mede o quanto os pontos em um diagrama de dispersão se aproximam de uma reta. Ressaltamos isso porque não existe somente a correlação linear, mas sim uma grande variedade de associações entre duas variáveis, tais como a polinomial, a exponencial e a logarítmica.

Ao comprovarmos a significância da correlação linear entre duas variáveis, alguns questionamentos podem surgir: (1) há como estabelecer uma relação matemática, uma regra de associação entre uma variável X e uma variável Y ? (2) é possível realizar uma previsão pontual de Y a partir de um valor de X não amostrado?

Para darmos continuidade a essa investigação, considere a seguinte situação: imagine novamente que você é um funcionário da empresa M e que foi incumbido de descrever o perfil dos funcionários. A partir da Tabela 2.1, é possível estabelecer uma relação matemática entre a satisfação em relação à remuneração e a satisfação em relação às condições de trabalho? Um funcionário que avalie sua satisfação em relação à remuneração com a pontuação 9 avaliará com qual pontuação a satisfação em relação às condições de trabalho?

Veremos nesta seção um método para relacionar matematicamente duas variáveis correlacionadas linearmente. Vamos lá!

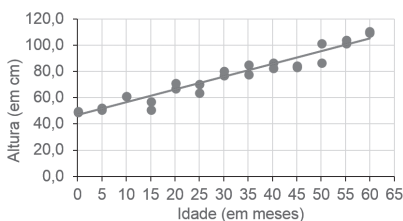
Não pode faltar

Quando testamos a significância da correlação linear entre duas variáveis X e Y , verificamos se os dados sustentam a hipótese de que a correlação entre elas é não nula, ou seja, de que as variações de Y são influenciadas pelas variações de X de modo linear. Sabendo dessa influência, algo natural é questionar se para um valor específico de X , não amostrado, é possível prever o valor correspondente de Y .

Para compreender melhor essa ideia, observe o diagrama da Figura 4.6, obtido a partir da Tabela 4.1. A linha que foi acrescentada é aquela que melhor se ajusta aos pontos e busca sinalizar uma tendência nos dados. Com base nisso, que valor você espera obter para y , dado que $x = 57$, isto é, qual seria a estatura de uma criança com idade de 57 meses?

Você pode verificar tanto na Figura 4.6 quanto na Tabela 4.1 que não há na amostra uma criança com 57 meses de idade.

Figura 4.6 | Linha de tendência para a idade e a altura das crianças de 0 a 5 anos



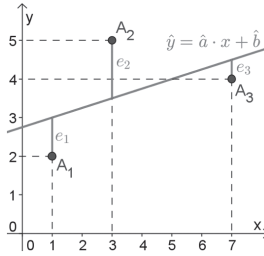
Fonte: Os autores (2015).

Logo, para realizar essa estimativa, você deve se basear nos demais valores amostrados e considerar a tendência indicada pela linha vermelha. Esperamos que você perceba que o valor esperado para y , nesse caso, gira em torno de 100. Assim, a previsão para a altura de crianças de 57 meses de idade é um valor próximo de 100 cm.

A linha reta representada na Figura 4.6, que é a reta de melhor ajuste, é denominada **reta de regressão**. O papel desempenhado por essa reta é o de representar geometricamente a associação entre as variáveis X e Y .

Você aprendeu na Unidade 1 que uma linha reta é descrita matematicamente por uma equação do tipo $y = a \cdot x + b$, em que a e b são números desconhecidos a serem determinados. Uma vez calculados os números a e b , também denominados **coeficientes**, podemos calcular o valor de y para qualquer valor de x dado. Vale observar que somente conseguiríamos calcular os valores exatos de a e b se tivéssemos os valores populacionais de (X, Y) . Como a nossa análise será feita com base em amostras, a equação $y = a \cdot x + b$ da reta de regressão é reescrita como $\hat{y} = \hat{a} \cdot x + \hat{b}$, em que \hat{y} , \hat{a} e \hat{b} são as estimativas para y , a e b .

Figura 4.7 | Retas de regressão para A_1 , A_2 e A_3



Fonte: Os autores (2015).

Para compreender como são calculados os valores de \hat{a} e \hat{b} a partir de uma amostra de dados bivariados (X, Y) , observe um exemplo de amostra na Tabela 4.3 e o diagrama de dispersão correspondente, na Figura 4.7.

Tabela 4.3 | Dados amostrados para as variáveis X e Y

X	1	3	7
Y	2	5	4

Fonte: Os autores (2015).

Você pode verificar na Figura 4.7 que foi adicionada a reta de equação $\hat{y} = \hat{a} \cdot x + \hat{b}$ e os erros cometidos ao aproximar os pontos A_1 , A_2 e A_3 por essa reta. Os erros e_1 , e_2 e e_3 são calculados por meio da diferença entre o valor de y amostrado (vide Tabela 4.3) e o valor correspondente de \hat{y} calculado por meio da reta de regressão, ou seja, $e_i = y_i - \hat{y}_i = y_i - (\hat{a} \cdot x_i + \hat{b})$.

Feito isso, como calculamos o erro total na aproximação pela reta de regressão? Instintivamente, poderíamos pensar em adicionar os erros e_i de cada aproximação, ou seja, $\sum e_i$. Entretanto, ocorre aqui algo semelhante ao que aconteceu no estudo das medidas de dispersão, em que sugerimos mensurar a dispersão dos dados efetuando $\sum (x_i - \bar{x})$. Você deve se lembrar de que essa soma é igual a zero. O que foi sugerido para driblar esse inconveniente é elevar ao quadrado cada desvio, ou seja, $\sum (x_i - \bar{x})^2$. Essa mesma estratégia pode ser utilizada para calcular o erro de aproximação da reta de regressão, e, assim, definimos a **soma dos quadrados dos erros** como:

$$SQ(\hat{a}, \hat{b}) = \sum e_i^2 = \sum [y_i - (\hat{a} \cdot x_i + \hat{b})]^2$$

A expressão $SQ(\hat{a}, \hat{b})$ é uma função que depende dos valores de \hat{a} e \hat{b} . Vamos calcular o valor de $SQ(\hat{a}, \hat{b})$ para alguns casos.

- Se $\hat{a} = 0$ e $\hat{b} = 1$, temos:

$$SQ(\hat{a} = 0, \hat{b} = 1) = [2 - (0 \cdot 1 + 1)]^2 + [5 - (0 \cdot 3 + 1)]^2 + [4 - (0 \cdot 7 + 1)]^2 = 26$$

- Se $\hat{a} = 1$ e $\hat{b} = 0$, temos:

$$SQ(\hat{a} = 1, \hat{b} = 0) = [2 - (1 \cdot 1 + 0)]^2 + [5 - (1 \cdot 3 + 0)]^2 + [4 - (1 \cdot 7 + 0)]^2 = 14$$

Observe que $SQ(\hat{a}, \hat{b})$ varia dependendo dos valores de \hat{a} e \hat{b} escolhidos. O desafio então é determinar valores específicos para \hat{a} e \hat{b} tais que $SQ(\hat{a}, \hat{b})$ seja a menor possível. A ferramenta utilizada para esse fim é denominada **método dos mínimos quadrados**.

Não entraremos em detalhes sobre o método dos mínimos quadrados, pois ele envolve recursos de Cálculo Diferencial, mas você pode se aprofundar nesse assunto acessando o *link*: http://repositorio.unesp.br/bitstream/handle/11449/101850/silva_mazm_dr_botfca.pdf?sequence=1 (acesso em: 20 jul. 2015). Iremos nos restringir a apresentar os resultados que podem ser obtidos por meio desse método.



Assimile

Os coeficientes \hat{a} e \hat{b} podem ser calculados pelas seguintes fórmulas:

$$\hat{a} = r \frac{Dp(Y)}{Dp(X)} \quad \text{e} \quad \hat{b} = \bar{y} - \hat{a} \cdot \bar{x},$$

em que $r = \rho(X, Y)$ é o coeficiente de correlação entre X e Y ; $Dp(Y)$ e $Dp(X)$ são, respectivamente, os desvios padrões amostrais; e \bar{y} e \bar{x} são, respectivamente, as médias dos valores de Y e X .

Vamos utilizar essas fórmulas para calcular os coeficientes de regressão da reta apresentada na Figura 4.7. Temos:

$$SQ(x) = (1^2 + 3^2 + 7^2) - \frac{(1+3+7)^2}{3} = 59 - \frac{11^2}{3} \cong 18,667;$$

$$SQ(y) = (2^2 + 5^2 + 4^2) - \frac{(2+5+4)^2}{3} = 45 - \frac{11^2}{3} \cong 4,667;$$

$$SQ(xy) = (1 \cdot 2 + 3 \cdot 5 + 7 \cdot 4) - \frac{(1+3+7)(2+5+4)}{3} = 45 - \frac{121}{3} \cong 4,667;$$

$$r = \frac{SQ(xy)}{\sqrt{SQ(x) \cdot SQ(y)}} = \frac{4,667}{\sqrt{18,667 \cdot 4,667}} \cong 0,5;$$

$$Dp(Y) \cong 1,528; \quad Dp(X) \cong 3,055; \quad \bar{y} \cong 3,667; \quad \bar{x} \cong 3,667$$

. Logo:

$$\hat{a} = r \frac{Dp(Y)}{Dp(X)} = 0,5 \frac{1,528}{3,055} \cong 0,25;$$

$$\hat{b} = \bar{y} - \hat{a} \cdot \bar{x} = 3,667 - 0,25 \cdot 3,667 \cong 2,75.$$

Portanto, a equação da reta de regressão da Figura 4.7 é $\hat{y} = 0,25 \cdot x + 2,75$.

Podemos fazer uma pequena verificação para esses valores, calculando $SQ(\hat{a}, \hat{b})$ e constatando se a soma dos quadrados dos erros é menor que os valores anteriormente calculados:

$$SQ(\hat{a} = 0,25; \hat{b} = 2,75) = [2 - (0,25 \cdot 1 + 2,75)]^2 + [5 - (0,25 \cdot 3 + 2,75)]^2 + [4 - (0,25 \cdot 7 + 2,75)]^2 = 1 + 2,25 + 0,25 = 3,5.$$

Observe que

$$SQ(\hat{a} = 0, 25; \hat{b} = 2, 75) < SQ(\hat{a} = 1, \hat{b} = 0) < SQ(\hat{a} = 0, \hat{b} = 1)$$

. Isso pode ser constatado para quaisquer outras escolhas de \hat{a} e \hat{b} , o que faz que os coeficientes de regressão sejam aqueles que fornecerem o menor erro, isto é, a reta de equação $\hat{y} = \hat{a} \cdot x + \hat{b}$ é a que melhor se ajusta aos pontos em um diagrama de dispersão.



Atenção

A reta de regressão pode ser obtida para quaisquer conjuntos de dados bivariados, havendo correlação linear ou não. O que se deve levar em consideração é o objetivo de calcular sua equação: prever um valor de y para um dado valor de x . Nesse contexto, só faz sentido calcularmos a equação da reta de regressão para os conjuntos de dados bivariados que possuem correlação linear significativa, pois, caso contrário, não conseguiremos realizar previsões concretas.

Com base nessa ideia, veja o exemplo a seguir.



Exemplificando

Obtenha a reta de regressão para os dados da Figura 4.6 e responda: qual seria a estatura de uma criança com idade de 57 meses?

Resolução:

Inicialmente, lembre-se de que na seção anterior testamos a correlação entre a idade (X) e a altura (Y) das crianças com até 5 anos de idade e constatamos que há significância. Logo, faz sentido obtermos a reta de regressão para esses dados. Recorde-se também de que já calculamos o coeficiente de correlação e obtivemos $r = 0,967$. Resta então calcular $Dp(Y)$, $Dp(X)$, \bar{y} e \bar{x} . Assim:

$$Dp(Y) \cong 20,0744 ; \quad Dp(X) \cong 19,8910 ; \quad \bar{y} = 75,8375 ; \\ \bar{x} = 30 .$$

Agora podemos calcular os coeficientes de regressão:

$$\hat{a} = r \frac{Dp(Y)}{Dp(X)} = 0,967 \frac{20,0744}{19,8910} \cong 0,9759;$$

$$\hat{b} = \bar{y} - \hat{a} \cdot \bar{x} = 75,8375 - 0,9759 \cdot 30 = 46,5605.$$

Portanto, $\hat{y} = 0,9759 \cdot x + 46,5605$ é a equação da reta de regressão que pode ser observada na Figura 4.6. Para prever a estatura de uma criança de 57 meses de idade, substituímos $x = 57$ na equação anterior. Temos:

$$\hat{y} = 0,9759 \cdot 57 + 46,5605 = 102,1868 \cong 102.$$

Concluimos então que a estimativa para a estatura de uma criança de 57 meses de idade é $\hat{y} \cong 102$ centímetros. Observe que esse valor é próximo do estimado visualmente (100 cm).

Existem fórmulas alternativas e equivalentes para calcular os coeficientes de regressão. São elas:

$$\hat{a} = \frac{n \sum xy - (\sum x)(\sum y)}{n \sum x^2 - (\sum x)^2} = \frac{SQ(xy)}{SQ(x)};$$

$$\hat{b} = \bar{y} - \hat{a} \cdot \bar{x} = \frac{\sum y}{n} - \hat{a} \frac{\sum x}{n}.$$



Pesquise mais

Veja mais detalhes sobre a regressão linear nos links: <<http://leg.ufpr.br/~paulojus/CE003/ce003/node9.html>>; <http://www.usp.br/fau/cursos/graduacao/arq_urbanismo/disciplinas/aut0516/Apostila_Regressao_Linear.pdf>; <http://www.pucrs.br/famat/rossana/psicologia/Aula18_Analise_regressao.pdf> (acesso em: 21 jul. 2015).

Sem medo de errar

Vamos retornar à situação-problema proposta no início desta seção: imagine novamente que você é um funcionário da empresa M e que foi incumbido de descrever o perfil dos funcionários. A partir da Tabela 2.1, é possível estabelecer uma relação matemática entre a satisfação em relação à remuneração e a satisfação em relação às condições de trabalho? Um funcionário que avalie sua satisfação em relação à remuneração com a pontuação 9 avaliará

com qual pontuação a satisfação em relação às condições de trabalho?

Você aprendeu nas seções anteriores a mensurar a correlação entre duas variáveis e a testar a significância dessa correlação. Durante esse aprendizado, foi visto que o coeficiente de correlação linear entre as variáveis G: satisfação em relação às condições de trabalho e H: satisfação em relação à remuneração é $r = 0,707$. Além disso, com um nível de confiança de 95%, foi atestada a significância dessa correlação. Logo, faz sentido determinarmos a equação da reta de regressão.

Da Seção 4.1, temos: $SQ(hg) = 57,65$; $SQ(h) = 46,55$. Logo,

$$\hat{a} = \frac{SQ(hg)}{SQ(h)} = \frac{57,65}{46,55} \cong 1,238. \text{ Segue da Tabela 2.1 que } \bar{h} = 5,15 \text{ e}$$

$$\bar{g} = 5,45. \text{ Assim, } \hat{b} = \bar{g} - \hat{a} \cdot \bar{h} = 5,45 - 1,238 \cdot 5,15 \cong -0,926.$$

Portanto, a equação da reta de regressão é $\hat{g} = 1,238 \cdot h - 0,926$. Para estimarmos qual pontuação em relação à condição de trabalho será atribuída por um funcionário que avaliar sua remuneração com a pontuação 9, substituímos $h = 9$ na equação anterior, ou seja:

$$\hat{g} = 1,238 \cdot 9 - 0,926 = 10,216 \cong 10.$$

O resultado foi arredondado, visto que a nota que deveria ser atribuída na pesquisa era um valor entre 0 e 10. Por fim, concluímos que um funcionário que atribua nota 9 a sua remuneração avaliará as condições de trabalho com a nota 10.

Avançando na prática

Pratique mais!

Instrução

Desafiamos você a praticar o que aprendeu transferindo seus conhecimentos para novas situações que pode encontrar no ambiente de trabalho. Realize as atividades e depois as compare com as de seus colegas.

1. Competência de fundamentos de área	Conhecer os conceitos matemáticos básicos e proporcionar o desenvolvimento do raciocínio lógico e quantitativo.														
2. Objetivos de aprendizagem	Obter a reta de regressão para dados bivariados com correlação significativa e realizar previsões para valores não amostrados.														
3. Conteúdos relacionados	Regressão linear.														
4. Descrição da situação-problema	A seguir, consta o valor gasto com propaganda e a quantidade vendida de um produto no mesmo mês.														
	<table border="1"> <tr> <td>Gastos com propaganda (× R\$ 1.000)</td> <td>10,0</td> <td>11,0</td> <td>12,2</td> <td>13,8</td> <td>14,4</td> <td>15,5</td> </tr> <tr> <td>Unidades vendidas (× 10.000)</td> <td>9,8</td> <td>9,7</td> <td>12,6</td> <td>14,4</td> <td>13,6</td> <td>16,2</td> </tr> </table>	Gastos com propaganda (× R\$ 1.000)	10,0	11,0	12,2	13,8	14,4	15,5	Unidades vendidas (× 10.000)	9,8	9,7	12,6	14,4	13,6	16,2
	Gastos com propaganda (× R\$ 1.000)	10,0	11,0	12,2	13,8	14,4	15,5								
Unidades vendidas (× 10.000)	9,8	9,7	12,6	14,4	13,6	16,2									
Obtenha a equação da reta de regressão para as variáveis "gastos com propaganda" e "unidades vendidas", e estime a quantidade de unidades vendidas para um gasto com propaganda igual a R\$ 13.000.															
5. Resolução da situação-problema	<p>Nas seções anteriores obtivemos $r \cong 0,960$ e concluímos que as variáveis X: gastos com propaganda e Y: unidades vendidas estão correlacionadas linearmente e positivamente, sendo este fato atestado em um teste de significância com 95% de confiança. Com base nisso, podemos agora obter a equação da reta de regressão. Temos $SQ(xy) = 26,168$ e $SQ(x) = 22,288$. Além disso, a partir da amostra podemos obter $\bar{x} \cong 12,817$ e $\bar{y} \cong 12,717$. Assim:</p> $\hat{a} = \frac{SQ(xy)}{SQ(x)} = \frac{26,168}{22,288} \cong 1,174;$ $\hat{b} = \bar{y} - \hat{a} \cdot \bar{x} = 12,717 - 1,174 \cdot 12,817 \cong -2,330.$ <p>A equação da reta de regressão correspondente é $\hat{y} = 1,174 \cdot x - 2,330$. Desse modo, a estimativa para a quantidade de unidades vendidas para um gasto com propaganda igual a R\$ 13.000 é:</p> $\hat{y} = 1,174 \cdot 13 - 2,330 = 12,932 \rightarrow 12.932 \text{ unidades.}$														



Lembre-se

A reta de melhor ajuste a um conjunto de pontos em um diagrama de dispersão é denominada **reta de regressão**.

Uma linha reta é descrita matematicamente por uma equação do tipo $y = a \cdot x + b$, em que a e b são números desconhecidos a serem determinados, também denominados **coeficientes**.

O **método dos mínimos quadrados** é uma ferramenta que busca determinar valores específicos para \hat{a} e \hat{b} tais que $SQ(\hat{a}, \hat{b})$, a **soma dos quadrados dos erros**, seja a menor possível.

Por mínimos quadrados, temos que os **coeficientes de regressão** são dados por $\hat{a} = r \cdot Dp(Y) / Dp(X)$ e $\hat{b} = \bar{y} - \hat{a} \cdot \bar{x}$.



Faça você mesmo

Acesse o link: http://nbcgib.uesc.br/lec/download/material_didatico/correlacao.pdf (acesso em: 21 jul. 2015) e resolva o problema proposto na página 6 desse material. Em seguida, determine a equação da reta de regressão que relaciona o índice DJIA com o S&P500. Por fim, realize uma previsão para o S&P500 quando o índice DJIA for 11.000.

Faça valer a pena

1. De determinada amostra de dados bivariados (X, Y) extraíram-se as seguintes informações:

$$r = 0,75 \quad Dp(Y) = 2 \quad Dp(X) = 3 \quad \bar{y} = 20 \quad \bar{x} = 32$$

Assinale a alternativa que contém a equação da reta de regressão para esse caso:

- a) $\hat{y} = 4,0 \cdot x + 0,5$ c) $\hat{y} = 0,5 \cdot x + 4,0$ e) $\hat{y} = 5,0 \cdot x + 4,0$
 b) $\hat{y} = 4,5 \cdot x + 4,0$ d) $\hat{y} = 0,5 \cdot x + 4,5$

da reta de regressão, assinale a alternativa que contém uma estimativa pontual para

y , dado $x = 60$.

- a) 34,1. c) 43,1. e) 43,4.
b) 31,4. d) 41,3.

6. Considere o conjunto de dados a seguir, cuja reta de melhor ajuste possui equação $\hat{y} = -2,3 \cdot x + 62,5$.

X	5	10	15	20
Y	50	40	30	15

Determine a soma dos quadrados dos erros $SQ(\hat{a}, \hat{b})$.

7. Determine a equação da reta de regressão para o conjunto de dados a seguir.

X	10	20	30	40
Y	94	82	68	57

Seção 4.4

Estudando resíduos

Diálogo aberto

Você aprendeu na seção anterior a obter a reta de regressão $\hat{y} = \hat{a} \cdot x + \hat{b}$ a partir de uma amostra de dados bivariados (X, Y) . Vimos que os coeficientes de regressão \hat{a} e \hat{b} são calculados por mínimos quadrados e são aqueles que minimizam a função $SQ(\hat{a}, \hat{b}) = \sum e_i^2$, a soma dos quadrados dos erros. Você também aprendeu que a equação da reta de regressão tem como objetivo gerar estimativas pontuais para valores de Y , dados os valores observados de X .

Lembre-se de que a regressão linear deve ser feita quando a correlação linear entre duas variáveis for significativa. Caso isso não ocorra, uma previsão \hat{y} feita a partir de um valor x pode ter grande imprecisão. Recorde-se também de que na Seção 3.2 construímos intervalos de confiança para a média amostral. Esse procedimento é muito comum na estatística, não somente para estimadores como a média, mas para todo estimador pontual, como é o caso do estimador \hat{y} de y . A construção de intervalos de confiança é feita com base na teoria de probabilidades e tem por objetivo estabelecer uma margem de erro para a estimativa.

Aproveitamos o momento para acrescentar que, apesar de a análise de significância ser feita com base em r (coeficiente de correlação), seu valor não auxilia a interpretar o quanto da variação de Y é devido a sua correlação com X e o quanto é devido ao acaso. Essa interpretação é possível por meio de um estudo dos resíduos de uma regressão linear, ou seja, dos erros ocorridos na geração de estimativas pelo processo de regressão. O estudo de resíduos também auxilia na construção de intervalos de confiança para os valores de regressão, e por esse motivo ele é o objeto de estudo desta seção.

Para compreender a importância do estudo dos resíduos, considere o seguinte problema: imagine novamente que você é

um funcionário da empresa M e que foi incumbido de descrever o perfil dos funcionários. Vimos nas seções anteriores que a partir da Tabela 2.1 é possível estabelecer uma relação matemática entre a satisfação em relação à remuneração e a satisfação em relação às condições de trabalho. Obtivemos a equação de regressão $\hat{g} = 1,238 \cdot h - 0,926$, em que H é a satisfação em relação à remuneração e G é a satisfação em relação às condições de trabalho. Estimamos que um funcionário que avalie sua satisfação em relação à remuneração com a pontuação 9 avaliará com a pontuação 10 a satisfação em relação às condições de trabalho. Neste ponto surgem alguns questionamentos: é possível estabelecer um intervalo de confiança para a estimativa $\hat{g}_0 = 10$? Quanto da variação de G é explicado pela variação de H e quanto é devido ao acaso e às características próprias de cada funcionário?

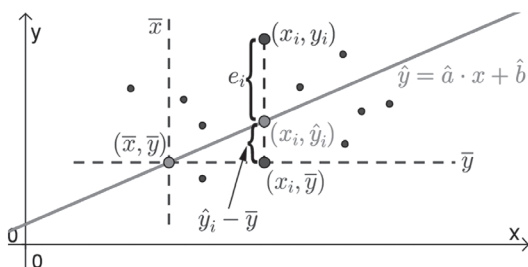
Não pode faltar

Resíduos

Quando realizamos uma regressão linear e obtemos os valores \hat{a} e \hat{b} , tais que a reta $\hat{y} = \hat{a} \cdot x + \hat{b}$ é aquela que melhor se ajusta ao conjunto de pontos correspondentes aos valores amostrados para as variáveis X e Y , sempre estamos sujeitos a erros. Em Estatística, tais erros são denominados **resíduos**.

Você aprendeu na seção anterior que a reta de regressão é determinada por meio da minimização de $SQ(\hat{a}, \hat{b}) = \sum e_i^2$, em que $e_i = y_i - \hat{y}_i$ é o erro (também denominado **desvio não explicado**) associado ao i -ésimo ponto no diagrama de dispersão. O erro e_i pode ser observado no diagrama da Figura 4.8.

Figura 4.8 | Resíduos em uma regressão linear



Fonte: Os autores (2015).

No diagrama foram representados os pontos amostrais obtidos para (X, Y) , a reta de regressão correspondente, o ponto de coordenadas (x_i, \hat{y}_i) , o ponto de coordenadas (x_i, \bar{y}) e o ponto de coordenadas (\bar{x}, \bar{y}) . Uma propriedade associada a toda regressão linear é que o ponto (\bar{x}, \bar{y}) pertence à reta de melhor ajuste. Vale observar que os pontos (\bar{x}, \bar{y}) , (x_i, \bar{y}) e (x_i, \hat{y}_i) foram adicionados ao diagrama e não necessariamente estão na amostra para (X, Y) . Para uma análise mais detalhada, definimos:



Assimile

O **desvio não explicado** e_i é a diferença entre o valor amostrado y_i e o valor previsto por regressão \hat{y}_i ou seja, $e_i = y_i - \hat{y}_i$.

O **desvio explicado** é a diferença entre o valor previsto por regressão \hat{y}_i e o valor médio \bar{y} , ou seja, $\hat{y}_i - \bar{y}$.

O **desvio total** é a diferença entre o valor amostrado y_i e o valor médio \bar{y} , ou seja, $y_i - \bar{y}$.



Lembre-se

Você já aprendeu o significado de desvio na Seção 2.4. O que se deve fazer agora é aprofundar este conceito e adaptá-lo ao contexto da análise de regressão.

As nomenclaturas utilizadas anteriormente são, de certa maneira, autoexplicativas. Contudo, vale observar que:

- O desvio não explicado se refere à diferença que pode ocorrer entre o valor previsto por regressão e o valor amostrado. Utilizamos essa terminologia porque a regressão por si só não explica a diferença ocorrida, de modo que a atribuímos à especificidade de cada ponto amostral e ao acaso.
- O desvio explicado é aquele devido à regressão e totalmente compreendido por meio dela.
- Por fim, o desvio total é a soma do desvio explicado com o não explicado, ou seja:

Desvio total = desvio explicado + desvio não explicado.

Com base nas definições anteriores, podemos também definir:



A **variação não explicada** é igual à soma dos quadrados dos desvios não explicados, ou seja, $\sum e_i^2 = \sum (y_i - \hat{y}_i)^2 = SQ(a,b)$.

Analogamente, a **variação explicada** é igual à soma dos quadrados dos desvios explicados, ou seja, $\sum (\hat{y}_i - \bar{y})^2$.

Por fim, a **variação total** é igual à soma dos quadrados dos desvios totais, ou seja, $\sum (y_i - \bar{y})^2$.

Não entraremos em maiores detalhes, mas é possível demonstrar que a variação total é igual à soma da variação explicada com a não explicada, isto é:

$$\begin{aligned} \text{Variação total} &= \text{variação explicada} + \text{variação não explicada.} \\ \sum (y_i - \bar{y})^2 &= \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \bar{y})^2. \end{aligned}$$

Com base nesses conceitos, podemos mensurar o quanto da variação total de Y pode ser explicado pela variação de X considerando a correlação existente entre essas variáveis.

Coefficiente de determinação (ou explicação)

O **coeficiente de determinação** é uma medida que tem por finalidade mensurar, em termos percentuais, o quanto da variação de uma variável Y é devido à variação de X , supondo que essas variáveis sejam correlacionadas. Esse coeficiente é calculado por meio da razão entre a variação explicada e a variação total, ou seja:

$$\text{coeficiente de determinação} = \frac{\text{variação explicada}}{\text{variação total}}.$$



Suponha duas variáveis X e Y correlacionadas linearmente, tais que $\sum (\hat{y}_i - \bar{y})^2 = 48$ e $\sum (y_i - \bar{y})^2 = 12$. Qual é o valor do coeficiente de determinação?

Resolução:

Para essas duas variáveis, a variação explicada é 48 e a variação não explicada é 12. Logo, a variação total é 60, já que corresponde à soma

das variações explicada e não explicada. Assim:

$$\text{coeficiente de determinação} = \frac{\text{variação explicada}}{\text{variação total}} = \frac{48}{60} = 0,8 = 80\%.$$

Portanto, 80% da variação de Y se deve à variação de X , e 20% (valor obtido efetuando $100\% - 80\%$) não se explica pela variação de X .

Existe uma relação estreita entre o coeficiente de correlação r e o coeficiente de determinação. Essa relação é expressa por:

$$\text{coeficiente de determinação} = r^2.$$

Por meio dessa relação, o coeficiente de determinação, ou explicação, pode ser calculado mais facilmente, simplesmente elevando ao quadrado o coeficiente de correlação.



Exemplificando

Duas variáveis X e Y estão negativamente correlacionadas de modo que $r = -0,9$. Quanto da variação de Y pode ser explicado por sua correlação e variação de X ?

Resolução:

Como $r = -0,9$, o coeficiente de explicação será:

$$r^2 = (-0,9)^2 = 0,81 = 81\%.$$

Logo, 81% da variação de Y se deve à variação de X . Temos ainda que 19% da variação de Y não se explica pela variação de X e se deve ao acaso.

Intervalos de previsão

Em Estatística, sempre que é realizada uma estimativa pontual, como é o caso da previsão para \hat{y} feita por meio da reta de regressão em que $\hat{y} = \hat{a} \cdot x + \hat{b}$, é natural pensarmos em construir um intervalo de confiança para a estimativa. Alguns autores também o denominam **intervalo de previsão**.

Segundo Larson e Farber (2010), dada uma equação de regressão linear $\hat{y} = \hat{a} \cdot x + \hat{b}$, para um valor específico x_i , o

intervalo de confiança para y_i é $\hat{y}_i - E < y_i < \hat{y}_i + E$ ou, ainda, $[\hat{y}_i - E, \hat{y}_i + E]$, em que E é a **margem de erro**.



Exemplificando

Dada a regressão linear $\hat{y} = 10,5 \cdot x + 4$, suponha que, ao nível de confiança de 95%, a margem de erro de previsão para \hat{y} seja $E = 2$. Determine o intervalo de confiança para o valor \hat{y}_0 correspondente a $x_0 = 15$.

Resolução:

A estimativa pontual para a variável Y , correspondente ao valor $x_0 = 15$, é calculada substituindo esse valor em $\hat{y} = 10,5 \cdot x + 4$. Logo:

$$\hat{y}_0 = 10,5 \cdot 15 + 4 = 161,5.$$

Assim, o intervalo de confiança para \hat{y}_0 será:

$$[\hat{y}_i - E, \hat{y}_i + E] = [161,5 - 2; 161,5 + 2] = [159,5; 163,5].$$

Podemos denotar um intervalo de confiança com probabilidade γ para \hat{y}_i como $IC(\hat{y}_i, \gamma) = [\hat{y}_i - E, \hat{y}_i + E]$. Para o exemplo anterior, temos $IC(\hat{y}_0 = 161,5; 95\%) = [159,5; 163,5]$.

Também de acordo com Larson e Farber (2010):



Assimile

Dada uma regressão linear $\hat{y} = \hat{a} \cdot x + \hat{b}$, a margem de erro E para uma estimativa \hat{y}_0 , calculada a partir de um valor x_0 , é dada por:

$$E = t_\gamma \cdot S_e \cdot \sqrt{1 + \frac{1}{n} + \frac{n(x_0 - \bar{x})^2}{n \sum x^2 - (\sum x)^2}},$$

em que t_γ é obtido a partir da tabela T, com $n - 2$ graus de liberdade,

de modo que $P(-t_\gamma < t < t_\gamma) = \gamma$. O valor S_e é denominado **erro padrão de estimativa** e calculado pela fórmula:

$$S_e = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{\sum y_i^2 - b \sum y_i - a \sum x_i y_i}{n-2}}.$$

Observe que $\sum (y_i - \hat{y}_i)^2$ é o que denominamos anteriormente de variação não explicada. Como $P(-t_\gamma < t < t_\gamma) = \gamma$, segue que a probabilidade de que \hat{y}_i pertença ao intervalo $[\hat{y}_i - E, \hat{y}_i + E]$ também é γ .

Para compreendermos melhor a construção de um intervalo de previsão, veja o exemplo a seguir.



Exemplificando

Considere a regressão linear $\hat{y} = 1,4x + 5,7$, obtida a partir do conjunto de dados a seguir.

i	1	2	3	4	5
X	5	10	15	20	25
Y	11,5	22,0	25,5	34,0	40,5

Determine um intervalo de previsão com 95% de confiança para \hat{y}_0 , dado $x_0 = 22$.

Resolução:

Primeiramente realizamos uma estimativa pontual por meio da regressão linear:

$$\hat{y}_0 = 1,4 \cdot 22 + 5,7 = 36,5.$$

Como a estimativa pontual para \hat{y}_0 é 36,5, o intervalo de previsão (ou de confiança) para \hat{y}_0 é $[36,5 - E; 36,5 + E]$, em que E é a margem de erro. Para calcularmos a margem de erro, precisamos determinar o erro padrão de estimativa S_e e obter $t_{\gamma=95\%}$. Temos:

$$\sum y_i^2 = 11,5^2 + 22,0^2 + 25,5^2 + 34,0^2 + 40,5^2 = 4062,75;$$

$$\sum y_i = 11,5 + 22,0 + 25,5 + 34,0 + 40,5 = 133,5;$$

$$\sum x_i y_i = 5 \cdot 11,5 + 10 \cdot 22,0 + 15 \cdot 25,5 + 20 \cdot 34,0 + 25 \cdot 40,5 = 2352,5$$

$$\sum x_i^2 = 5^2 + 10^2 + 15^2 + 20^2 + 25^2 = 1375;$$

$$\sum x_i = 5 + 10 + 15 + 20 + 25 = 75;$$

$$\bar{x} = \frac{\sum x_i}{n} = \frac{75}{5} = 15;$$

$$S_e = \sqrt{\frac{\sum y_i^2 - b \sum y_i - a \sum x_i y_i}{n-2}} = \sqrt{\frac{4062,75 - 5,7 \cdot 133,5 - 1,4 \cdot 2352,5}{5-2}} \cong 1,663.$$

Além disso, consultando a tabela T para $\nu = 5 - 2 = 3$ graus de liberdade e na coluna correspondente a 2,5% (obtido efetuando $\frac{1-\gamma}{2}$),

temos $t_{\gamma=95\%} = 3,182$. Logo:

$$E = t_{\gamma} \cdot S_e \cdot \sqrt{1 + \frac{1}{n} + \frac{n(x_0 - \bar{x})^2}{n \sum x^2 - (\sum x)^2}};$$

$$E = 3,182 \cdot 1,663 \cdot \sqrt{1 + \frac{1}{5} + \frac{5(22-15)^2}{5 \cdot 1375 - (75)^2}} = 5,291666 \cdot \sqrt{1 + \frac{1}{5} + \frac{245}{1250}};$$

$$E \cong 6,252.$$

Por fim, o intervalo de previsão para \hat{y}_0 com 95% de confiança é:

$$IC(\hat{y}_0 = 36,5; 95\%) = [36,5 - E; 36,5 + E] = [36,5 - 6,252; 36,5 + 6,252] = [30,248; 42,752]$$

Concluimos assim que há 95% de probabilidade de \hat{y}_0 , calculado a partir de $x_0 = 22$, pertencer ao intervalo $[30,248; 42,752]$.



Veja mais detalhes acerca dos intervalos de predição em: <<http://people.ufpr.br/~jomarc/regressao.pdf>>. Acesso em: 28 jul. 2015. (LARSON; FARBER, 2010; MORETTIN; BUSSAB, 2010).

Sem medo de errar

Agora que você já estudou os resíduos, vamos retomar a situação-problema proposta no início desta seção: é possível estabelecer um intervalo de confiança para a estimativa $\hat{g}_0 = 10$, obtida a partir de $h_0 = 9$? Quanto da variação de G é explicado pela variação de H e quanto é devido ao acaso e às características próprias de cada funcionário?

Você aprendeu que para calcular o quanto da variação de uma variável G é devido à variação de outra correlacionada H utilizamos o coeficiente de determinação, ou r^2 . Na Seção 4.1 o coeficiente de correlação dessas variáveis foi estimado em $r \cong 0,707$. Logo, $r^2 = 0,707^2 \cong 0,5 = 50\%$. Desse modo, apenas 50% da variação de G se deve à variação de H , e os outros 50% devem-se ao acaso.

Supondo um nível de confiança de $\gamma = 95\%$, para determinar um intervalo de predição para $\hat{g}_0 = 10$ precisamos calcular o erro padrão de estimativa S_e e a margem de erro E . Temos:

$$\sum g_i^2 = 8^2 + 5^2 + \dots + 10^2 + 9^2 = 737;$$

$$\sum g_i = 8 + 5 + \dots + 10 + 9 = 109;$$

$$\sum h_i g_i = 7 \cdot 8 + 4 \cdot 5 + \dots + 6 \cdot 10 + 8 \cdot 9 = 619;$$

$$\sum h_i^2 = 7^2 + 4^2 + \dots + 6^2 + 8^2 = 577;$$

$$\sum h_i = 7 + 4 + \dots + 6 + 8 = 103;$$

$$\bar{h} = \frac{\sum h_i}{n} = \frac{103}{20} = 5,15;$$

$$S_e = \sqrt{\frac{\sum g_i^2 - b \sum g_i - a \sum h_i g_i}{n-2}} = \sqrt{\frac{737 - (-0,926) \cdot 109 - 1,238 \cdot 619}{20-2}} \cong 1,995.$$

Além disso, consultando a tabela T para $\nu = 20 - 2 = 18$ graus de liberdade e na coluna correspondente a 2,5% (obtido efetuando $\frac{1-\gamma}{2}$), temos $t_{\gamma=95\%} = 2,101$. Logo:

$$E = t_\gamma \cdot S_e \cdot \sqrt{1 + \frac{1}{n} + \frac{n(h_0 - \bar{h})^2}{n \sum h^2 - (\sum h)^2}};$$

$$E = 2,101 \cdot 1,995 \cdot \sqrt{1 + \frac{1}{20} + \frac{20(9 - 5,15)^2}{20 \cdot 577 - (103)^2}} = 4,191495 \cdot \sqrt{1 + \frac{1}{20} + \frac{296,45}{931}};$$

$$E \cong 4,903.$$

Por fim, o intervalo de previsão para $\hat{g}_0 = 10$ com 95% de confiança é:

$$IC(\hat{g}_0 = 10; 95\%) = [10 - E; 10 + E] = [10 - 4,903; 10 + 4,903] = [5,097; 14,903].$$

Como é possível atribuir apenas uma pontuação inteira entre 0 e 10, temos 95% de confiança de que um funcionário que avalie sua remuneração com nota 9 irá atribuir uma nota de 5 a 10 para a satisfação em relação às condições trabalho.

Avançando na prática

Pratique mais!

Instrução

Desafiamos você a praticar o que aprendeu transferindo seus conhecimentos para novas situações que pode encontrar no ambiente de trabalho. Realize as atividades e depois as compare com as de seus colegas.

1. Competência de fundamentos de área	Conhecer os conceitos matemáticos básicos e proporcionar o desenvolvimento do raciocínio lógico e quantitativo.														
2. Objetivos de aprendizagem	Quantificar a variação de uma variável que se deve à correlação com outra e a variação que se deve ao acaso.														
3. Conteúdos relacionados	Regressão linear; coeficiente de determinação.														
4. Descrição da situação-problema	<p>No quadro a seguir, também apresentado nas seções anteriores, consta o valor gasto com propaganda e a quantidade vendida de um produto no mesmo mês.</p> <table border="1" style="margin-left: auto; margin-right: auto;"> <tr> <td>Gastos com propaganda (x R\$ 1.000)</td> <td>10,0</td> <td>11,0</td> <td>12,2</td> <td>13,8</td> <td>14,4</td> <td>15,5</td> </tr> <tr> <td>Unidades vendidas (x 10.000)</td> <td>9,8</td> <td>9,7</td> <td>12,6</td> <td>14,4</td> <td>13,6</td> <td>16,2</td> </tr> </table> <p>Quanto da variação da quantidade de unidades vendidas é explicado pela variação do gasto com propaganda e quanto é devido ao acaso?</p>	Gastos com propaganda (x R\$ 1.000)	10,0	11,0	12,2	13,8	14,4	15,5	Unidades vendidas (x 10.000)	9,8	9,7	12,6	14,4	13,6	16,2
Gastos com propaganda (x R\$ 1.000)	10,0	11,0	12,2	13,8	14,4	15,5									
Unidades vendidas (x 10.000)	9,8	9,7	12,6	14,4	13,6	16,2									
5. Resolução da situação-problema	<p>A medida de explicação da variação de uma variável em relação à variação de outra correlacionada é feita por meio do coeficiente de determinação, ou r^2. Lembre-se de que nas seções anteriores estimamos $r \cong 0,960$. Logo:</p> $r^2 = 0,960^2 = 0,9216 \cong 0,92 = 92\%$ <p>Desse modo, 92% da variação da quantidade de unidades vendidas pode ser explicada pela variação do gasto com propaganda, e os 8% restantes devem-se ao acaso.</p>														



Lembre-se

A **variação não explicada** é igual à soma dos quadrados dos desvios não explicados, ou seja, $\sum e_i^2 = \sum (y_i - \hat{y}_i)^2 = SQ(\hat{a}, \hat{b})$.

Analogamente, a **variação explicada** é igual à soma dos quadrados dos desvios explicados, ou seja, $\sum (\hat{y}_i - \bar{y})^2$.

Por fim, a **variação total** é igual à soma dos quadrados dos desvios totais, ou seja, $\sum (y_i - \bar{y})^2$.

A **variação total** é igual à soma da **variação explicada** com a **não explicada**.

O **coeficiente de determinação** é uma medida que tem por finalidade mensurar, em termos percentuais, o quanto da variação de uma variável **Y** é devido à variação de **X** , supondo que essas variáveis sejam correlacionadas.



Faça você mesmo

Na seção anterior propusemos que você acessasse o link: <http://nbcgib.uesc.br/lec/download/material_didatico/correlacao.pdf> (acesso em: 21 jul. 2015) e resolvesse o problema exposto na página 6 desse material. Em seguida, sugerimos que determinasse a equação da reta de regressão que relaciona o índice DJIA com o S&P500. Por fim, solicitamos que realizasse uma previsão para o S&P500 quando o índice DJIA for 11.000.

Agora, aproveitando o que você já desenvolveu e com 90% de confiança, construa um intervalo de previsão para a estimativa realizada.

Faça valer a pena

1. Duas variáveis X e Y estão correlacionadas linearmente de modo que os valores de Y são previstos a partir de X por regressão linear. Sabendo que $r = -0,75$, assinale a alternativa que contém o percentual da variação de Y não explicado pela variação de X .

a) 56,25%.

c) 43,75%.

e) 56,52%.

b) 52,65%.

d) 47,35%.

2. Considere o seguinte conjunto de dados.

X	1	2	3	4	5
Y	4	7	4	8	9

Assinale a alternativa que contém o valor aproximado do coeficiente de determinação para essas variáveis.

a) 75,5%.

c) 65,0%.

e) 71,4%.

b) 57,0%.

d) 51,2%.

3. Duas variáveis X e Y estão correlacionadas de modo que, quando os valores de X aumentam, os valores de Y diminuem. Sabendo que a variação não explicada pela correlação entre essas variáveis é igual a 19%, assinale a alternativa que contém o valor do coeficiente de correlação r :

- a) 0,81 c) 0,9 e) -0,19
b) -0,81 d) -0,9

4. Duas variáveis X e Y estão correlacionadas positivamente, sendo a equação da reta de regressão $\hat{y} = 1,33x + 4,35$. Essa equação foi estimada a partir de uma amostra de tamanho $n = 4$, sendo o erro padrão de estimativa $S_e = 9,27$. Sabendo que $\sum x^2 = 4600$ e $\sum x = 120$, assinale a alternativa que contém o valor da margem de erro de previsão E , a um nível de confiança de 98%, para a estimativa $\hat{y}_0 = 20,31$:

- a) 64. c) 78. e) 81.
b) 75. d) 80.

5. Duas variáveis X e Y estão correlacionadas de modo que Y pode ser estimado a partir de X por meio da equação $\hat{y} = 1,07x + 12,07$. Além disso, o erro padrão de estimativa obtido a partir de uma amostra de tamanho $n = 5$ é $S_e = 9,00$. Com essas informações, assinale a alternativa que contém o intervalo de confiança de 96% para \hat{y}_0 , sendo $x_0 = 10$ e dados $n(x_0 - \bar{x})^2 = 19220$ e $n\sum x^2 - (\sum x)^2 = 34400$.

- a) [18,79; 64,33] c) [-28,79; 54,33] e) [8,79; 54,33]
b) [-18,79; 64,33] d) [-8,79; 44,33]

6. Considere as variáveis (X, Y) correlacionadas e a amostra a seguir.

(100, 134), (150, 183), (200, 207), (250, 229), (300, 316)

Sabendo que a equação da reta de regressão para (X, Y) é $\hat{y} = 0,82x + 49,8$, construa uma tabela em que: na primeira coluna sejam listados os valores X_i ; na segunda coluna sejam listados os valores observados Y_i ; na terceira coluna sejam listados os valores estimados \hat{y}_i ; e na quarta e última coluna sejam listados os desvios não explicados $e_i = Y_i - \hat{y}_i$.

7. Duas variáveis (X, Y) estão correlacionadas, sendo $\hat{y} = 5x + 4$ a equação da reta de regressão correspondente. Sabendo que o desvio não explicado $e_0 = Y_0 - \hat{y}_0$ é 3, com $x_0 = 4$, determine o valor observado Y_0 .

Referências

ANDERSON, David R.; SWEENEY, Dennis J.; WILLIAMS, Thomas A. **Estatística aplicada à administração e economia**. 2. ed. São Paulo: Cengage Learning, 2011.

CRESPO, Antônio A. **Estatística fácil**. 17. ed. São Paulo: Saraiva, 2002.

FREUND, John E. **Estatística aplicada**: economia, administração e contabilidade. 11. ed. Porto Alegre: Bookman, 2006.

JOHNSON, Robert; KUBY, Patricia. **Estatística**. São Paulo: Cengage Learning, 2013.

LARSON, R.; FARBER, B. **Estatística aplicada**. 4. ed. São Paulo: Person Prentice Hall, 2010.

MEDEIROS, Valéria Z. (Coord.). **Métodos quantitativos com Excel**. São Paulo: Cengage Learning, 2008.

MORETTIN, Luiz G. **Estatística básica**: probabilidade e inferência. São Paulo: Pearson Prentice Hall, 2010.

MORETTIN, Luiz G.; BUSSAB, Wilton O. **Estatística básica**. São Paulo: Saraiva, 2010.

ISBN 978-65-8482-354-3



9 788584 823543 >