Capítulo

9

Introdução à Estatística Descritiva

Gilberto Pereira Sassi
gilberto.sassi@ufba.br

Objetivo do Capítulo

Este capítulo tem o objetivo de apresentar conceitos e técnicas para você realizar uma análise exploratória de dados usando gráficos e medidas resumo. Ao final da leitura deste capítulo, você deve ser capaz de:

- Entender os conceitos básicos de estatística.
- Representar graficamente o seu conjunto de dados.
- Resumir as informações em um conjunto de dados usando medidas de posição, medidas de dispersão e quantis.
- Estudar a associação entre duas colunas do seu conjunto de dados.

Era uma vez... *Maria, uma aluna de um programa de pós-graduação em Informática na Educação que decidiu analisar uma plataforma digital de ensino. Esta plataforma digital coleta várias informações sobre os alunos, como tempo para terminar uma atividade proposta, número de interações com outros alunos, notas nos testes semanais, idade do aluno, e outros. Antes de propor melhorias, Maria precisa analisar os padrões e comportamentos dos alunos dentro da plataforma digital. Para isso, Maria escolheu alguns alunos e coletou várias informações sobre eles e precisa descobrir padrões e comportamentos a partir dessa amostra de alunos. Será que podemos ajudar Maria?*

1 Introdução

A análise, interpretação e apresentação dos dados são etapas essenciais para qualquer indivíduo que deseja pesquisar na área de Informática na Educação e são exatamente essas preocupações da análise estatística descritiva e deste capítulo. Apresentaremos a você conceitos e métodos para extrair informações de sua base de dados. Repare que existe uma comunidade inteira pesquisando e sugerindo melhorias nos métodos estatísticos já existentes, então nosso foco será nas técnicas mais tradicionais.

Em estatística descritiva, estamos preocupados em representar os dados usando gráficos e diagramas, além do interesse em resumir em um (ou alguns) número todos os valores de uma coluna de sua base de dados. Neste capítulo, você vai aprender a representar graficamente e a resumir os dados. Na Seção 2, começamos introduzindo alguns conceitos básicos que serão usados em todos os capítulos. Na seção 3, mostramos como você pode representar graficamente as informações contidas em uma amostra. Em seguida, na seção 4, mostramos como você pode resumir os dados coletados usando medidas de posição, medidas de dispersão e quantis. Finalmente, na seção 5, você aprenderá como estudar a associação entre duas variáveis de sua amostra.

2 Conceitos Básicos

De um modo geral, podemos afirmar que existe duas maneiras para chegarmos a conclusões: usando inferência dedutiva e usando inferência indutiva.

A inferência dedutiva usa argumentos lógicos para chegar a conclusões a partir de premissas. Por exemplo: Premissa: “Todo ser humano nascido em solo brasileiro tem direito a cidadania Brasileira”; Maria nasceu em Salvador, então Maria tem direito a cidadania Brasileira. Esse tipo de inferência é muito usado em Filosofia e Matemática Abstrata e não abordaremos este assunto neste capítulo.

A inferência indutiva é um processo de generalização da parte para o todo. Ou seja, a partir de um número de casos suficientemente grande, fazemos conclusões sobre todos os casos possíveis. Por exemplo, na seção **Era uma vez...,** Maria pode escolher alguns alunos e coletar informações sobre estes alunos na plataforma digital, e, então, usar estatística para calcular medidas de resumo, desenhar gráficos e fazer afirmações para toda população.

Antes de apresentarmos técnicas de inferência indutiva, vamos estabelecer alguns nomes e conceitos que irão nos ajudar neste capítulo e em seu trabalho de pesquisa:

1. **População:** todos os indivíduos (ou elementos) alvo de um estudo ou pesquisa.
2. **Amostra:** parte da população.
3. **Parâmetro:** característica da população. Geralmente não é possível ou é muito caro (operacionalmente e/ou financeiramente) encontrar essa característica.
4. **Estimativa:** característica da amostra. Geralmente usamos uma estimativa para aproximar um parâmetro.

5. **Variável:** característica de um elemento/indivíduo da população. Geralmente usamos uma letra maiúscula do alfabeto latino para representar uma variável, e uma letra minúscula do alfabeto latino para representar o valor de uma variável para um indivíduo (ou elemento) da população. Por exemplo, podemos representar a variável "a idade dos alunos" por X e um valor de idade presente na amostra por $x = 23$ anos.

Variáveis podem ser classificadas em quatro categorias:

1. **Variável Qualitativa Nominal:** variável cujos valores possíveis são atributos não numéricos e estes atributos não tem hierarquia entre si. Por exemplo, uma variável "nacionalidade" com valores possíveis {Brasileiro, Estrangeiro} é uma variável qualitativa nominal, pois não existe motivo para supor a superioridade dos Brasileiros ou dos Estrangeiros.
2. **Variável Qualitativa Ordinal:** variável cujos valores possíveis são atributos não numéricos e estes atributos tem hierarquia entre si. Por exemplo, uma variável "satisfação com o atendimento" com valores possíveis {Completamente insatisfeito, insatisfeito, satisfeito, completamente satisfeito} é uma variável qualitativa ordinal, pois usuários "satisfeito" tiveram uma experiência superior aos usuários "insatisfeito".
3. **Variável Quantitativa Discreta:** variável cujos valores possíveis são números inteiros, geralmente provenientes de uma contagem. Por exemplo, a variável "Número de filhos" é uma variável quantitativa discreta.
4. **Variável Quantitativa Contínua:** variável cujo valor possível pode ser qualquer número. Por exemplo, a variável "nota em uma atividade" é uma variável quantitativa contínua.

Agora estamos prontos! Vamos começar aprendendo a representar os dados usando gráficos.

3 Métodos Gráficos

Após a coleta de dados, você tem em mãos uma base de dados em que cada linha é um indivíduo e cada coluna corresponde a uma variável. Dessa forma, cada célula vai conter o valor de uma variável para um indivíduo. A Figura 5 ilustra notas de cinco alunos em quatro testes realizados levantados por nossa personagem Maria. A primeira coisa que podemos fazer é contar. Os resultados dessa contagem podem ser organizados em uma tabela de distribuição de frequência. Vamos aprender a construir a tabela de distribuição de frequência por partes: primeiro para variáveis qualitativas (nominais ou ordinais), para variáveis quantitativas discretas e, finalmente, para variáveis quantitativas contínuas.

Suponha que X seja uma variável qualitativa (ordinal ou nominal) com valores possíveis $A_1, \dots, A_k$. Imagine que n_1 indivíduos tem valor de X igual a A_1, n_2 indivíduos tem valor de X igual a A_2, n_3 indivíduos tem valor de X igual a A_3, e assim por diante. Então ao final deste processo de contagem, obtemos a tabela 1.

Tabela 1: Tabela de distribuição de frequências para uma variável qualitativa.

X	Frequência	Frequência relativa	Porcentagem
A_1	n_1	$f_1 = \frac{n_1}{n}$	$100 \cdot f_1\%$
A_2	n_2	$f_2 = \frac{n_2}{n}$	$100 \cdot f_2\%$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
A_k	n_k	$f_k = \frac{n_k}{n}$	$100 \cdot f_k\%$
Total	$n = n_1 + n_2 + \cdots + n_k$	$1 = f_1 + f_2 \cdots + f_k$	100%

Podemos representar graficamente as informações da tabela 1 conforme a figura 1. Chamamos o diagrama da figura 1 de gráfico de barras. As alturas das barras rotuladas por $A_1, \ldots, A_k$ têm alturas iguais às frequências relativas: $f_1, \ldots, f_k$. Ou seja, em um gráfico de barras você precisa olhar a altura da barra. Ressaltamos que a largura da barra não importa em gráfico de barras. Além disso, no lugar das frequências relativas, $f_1, \ldots, f_k$, você poderia ter usado as frequências $n_1, \ldots, n_k$.

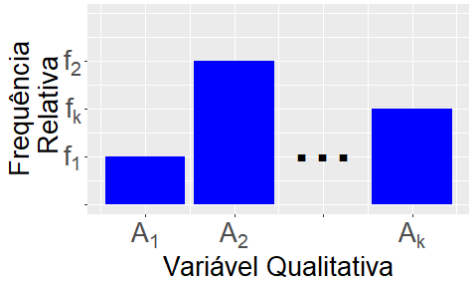


Figura 1: Gráfico de barras para variável qualitativa (ordinal ou nominal).

Note que o relevante não são os valores $n_1, n_2, \cdots, n_k$. O importante é quanto n_1 representa de n, quanto $n_2$representa de n, e assim por diante, e isso é representado pelas frequências relativas $f_1, f_2, \cdots, f_k$. Observe que as frequências relativas são números entre 0 e 1, e podemos facilitar a interpretação do leitor ao usar a porcentagem (multiplicando a frequência relativa por 100).

De maneira semelhante, poderíamos construir uma tabela de distribuição para a variável quantitativa discreta Y com valores possíveis $x_1, x_2, \cdots, x_k$, em que n_1 indivíduos tem valor de Y igual a x_1, em que n_2 indíviduos tem valor de Y igual a x_2, e assim por diante. Na tabela 2, mostramos a tabela de distribuição de frequência para a variável quantitativa discreta Y.

Tabela 2: Distribuição de frequências para uma variável quantitativa discreta Y.

Y	Frequência	Frequência relativa	Porcentagem
x_1	n_1	$f_1 = \frac{n_1}{n}$	$100 \cdot f_1\%$
x_2	n_2	$f_2 = \frac{n_2}{n}$	$100 \cdot f_2\%$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_k	n_k	$f_k = \frac{n_k}{n}$	$100 \cdot f_k\%$
Total	$n = n_1 + n_2 + \cdots + n_k$	$1 = f_1 + f_2 \cdots + f_k$	100%

Você pode representar uma tabela de distribuição de frequência usando um gráfico de barras de forma análoga ao que fizemos para variável qualitativa. Para cada valor $x_1, \ldots, x_k$ de Y desenhamos uma barra de altura $f_1, \ldots, f_k$. Na figura 2, ilustramos como ficaria o gráfico de barras de Y.

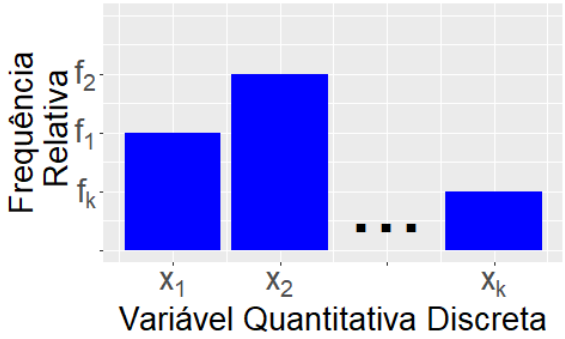


Figura 2: Gráfico de barras para uma variável quantitativa discreta.

Se Z é uma variável quantitativa contínua, criamos uma tabela de distribuição de frequência para as faixas de valores $(a_0, a_1]$, $(a_1, a_2]$, $\cdots$, $(a_{k-1}, a_k]$. Chamamos os valores $a_0, a_1, \cdots, a_k$ de *pontos de rotura,* em que a_0 é menor ou igual do que o menor valor de Z e a_k é o maior ou igual do que o maior valor de Z. Os símbolos "(" e ")" significam que as extremidades de uma faixa não devem ser consideradas na contagem, enquanto "[" e "]" significam que as extremidades de uma faixa de valores devem ser consideradas na contagem. Então, n_1é o número de indivíduos com valor de Z em $(a_0, a_1]$; n_2 é o número de indivíduos com valor de Z em $(a_1, a_2]$; n_3é o número de indivíduos com valor de Z em $(a_2, a_3]$; e assim por diante.

Tabela 3: Tabela de distribuição de frequência para uma variável quantitativa contínua.

Z	Frequência	Frequência relativa	Porcentagem
$(a_0, a_1]$	n_1	$f_1 = \frac{n_1}{n}$	$100 \cdot f_1\%$
$(a_1, a_2]$	n_2	$f_2 = \frac{n_2}{n}$	$100 \cdot f_2\%$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$(a_{k-1}, a_k]$	n_k	$f_k = \frac{n_k}{n}$	$100 \cdot f_k\%$
Total	$n = n_1 + n_2 + \cdots + n_k$	$1 = f_1 + f_2 \cdots + f_k$	100%

Para variáveis quantitativas contínuas, construímos um diagrama chamado histograma. O histograma é um gráfico de barras contíguas em que a área de cada barra é igual à frequência relativa. Como as áreas das barras são iguais às frequências relativas $f_1, f_2, f_3, \ldots, f_k$, as alturas das barras precisam ser iguais a $\frac{f_1}{a_2-a_1}, \frac{f_2}{a_3-a_2}, \frac{f_3}{a_4-a_3}, \ldots, \frac{f_k}{a_k-a_{k-1}}$. As razões $\frac{f_1}{a_2-a_1}, \frac{f_2}{a_3-a_2}, \frac{f_3}{a_4-a_3}, \ldots, \frac{f_k}{a_k-a_{k-1}}$ são denominadas de densidade de frequência. Na figura 3, mostramos o histograma para uma variável quantitativa contínua.

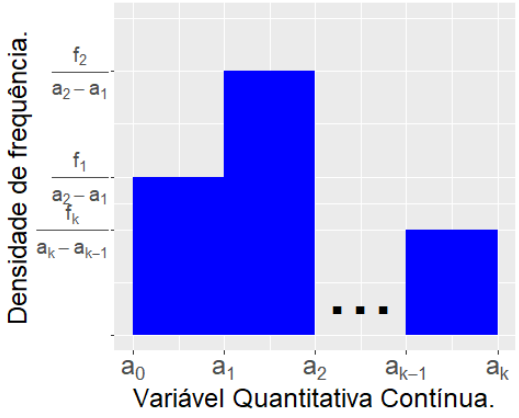


Figura 3: *Histograma para uma variável quantitativa contínua.*

4 Medidas Resumo

Além de construir gráficos, podemos resumir as informações de uma variável quantitativa com uma (algumas) medida(s) de resumo. Dificilmente você descobrirá informações úteis com o ato de olhar para todos os valores de uma variável em um banco de dados. Já aprendemos a representar visualmente os valores de uma variável usando gráficos, mas podemos ir além. Desejamos descobrir um (ou alguns) valor(es) que representa, da forma mais fidedigna possível, todos os valores de uma variável quantitativa. Para facilitar sua leitura, vamos dividir essa seção em três partes: medidas de posição, medidas de dispersão e quantis.

4.1 Medidas de Posição: Média, Moda e Mediana

Medida de posição é um valor representativo de uma variável quantitativa. Ou seja, se uma variável quantitativa X tem uma medida de posição com valor m_X, então quando você se deparar com um indivíduo (sem qualquer conhecimento prévio dele) você pode afirmar que o valor de X para tal indivíduo é m_X. Geralmente escolhemos como medida de posição quantidades frequentes na amostra ou quantidades que ocupam uma posição central entre os valores observados da variável quantitativa.

Existem três medidas de posição mais populares entre os pesquisadores na área de Informática na Educação: Moda, Média e Mediana. Geralmente, usamos a moda apenas para variáveis quantitativas discretas.

4.1.1 Moda

Seja X uma variável quantitativa discreta com valores observados $x_1, x_2, \dots, x_k$. A moda de X é o valor x_i que aparece mais vezes na amostra. Matematicamente, representamos a moda de X por $mo(X) = x_i$.

4.1.2 Média

Seja X uma variável quantitativa (discreta ou contínua) com valores observados $x_1, x_2, \dots, x_n$. Então, a média pode ser calculada por

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}.$$

Você pode interpretar a média como o centro de massa de uma barra em que pesos $x_1, x_2, \dots, x_n$ foram colocados em pontos igualmente espaçados, conforme ilustrado na figura 4.

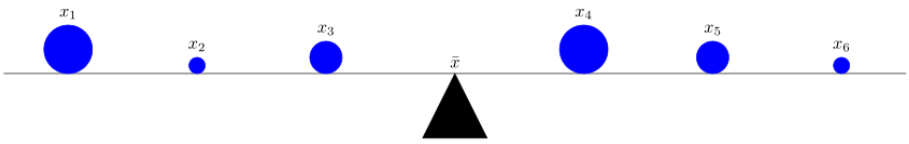


Figura 4: Interpretação da média. Na figura, representamos a média $\bar{x}$ da variável quantitativa X com valores observados $x_1, x_2, x_3, x_4, x_5, x_6$ representados por bolas sobre uma barra sem peso. A média $\bar{x}$ é o "ponto de equilíbrio" ou "centro da massa" dessa barra.

4.1.3 Mediana

Considere X uma variável quantitativa com valores observados $x_1, x_2, \dots, x_n$, então a mediana de X é um valor que divide a sequência ordenada de $x_1, x_2, \dots, x_n$ em duas partes iguais. Ou seja, a mediana é um valor $md(X)$ tal que 50% dos valores $x_1, x_2, \dots, x_n$ são menores ou iguais a $md(X)$ e 50% dos valores $x_1, x_2, \dots, x_n$ são maiores ou iguais a $md(X)$.

A primeira coisa que você precisa fazer para calcular a mediana é ordenar os valores do menor ao maior valor:

$$x_{(1)} \leq x_{(2)} \leq x_{(3)} \leq x_{(4)} \leq \cdots \leq x_{(n)},$$

em que $x_{(1)}$ é o menor valor entre $x_1, x_2, \ldots, x_n$; $x_{(2)}$ é o segundo menor valor entre $x_1, x_2, \ldots, x_n$; $x_{(3)}$ é o terceiro menor valor entre $x_1, x_2, \ldots, x_n$; $x_{(4)}$ é o quarto menor valor entre $x_1, x_2, \ldots, x_n$; e assim continua até $x_{(n)}$ (o último menor valor entre $x_1, x_2, \ldots, x_n$). Chamamos $x_{(1)}$, $x_{(3)}$, $x_{(4)}$, ... , $x_{(n)}$ de estatísticas de ordem.

Agora precisamos encontrar um valor $md(X)$ tal que:

1. 50% das estatísticas de ordem satisfaçam a desigualdade: $x_{(j)} \leq md(X)$;
2. 50% das estatísticas de ordem satisfaçam a desigualdade: $x_{(k)} \geq md(X)$.

Um valor que satisfaz as condições 1. e 2. é

$$md(X) = \begin{cases} x_{\left(\frac{n+1}{2}\right)}, & se\ n\ é\ ímpar, \\ \dfrac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2}, & se\ n\ é\ par. \end{cases}$$

4.2 Medidas de Dispersão: Variância, Desvio Padrão e Desvio Médio

Apesar da moda, média, e mediana serem úteis, estas medidas podem ser insuficientes para representar de forma fidedigna todos os valores de uma variável quantitativa. Para ilustrarmos, considere uma amostra de cinco alunos na plataforma digital de Maria que realizam uma bateria de testes em uma semana com as notas apresentadas na figura 5.

	A	B	C	D	E	F
1	Matrícula	teste A	teste B	teste c	teste D	
2	1	3	1	5	4	
3	2	4	3	5	5	
4	3	5	5	5	5	
5	4	6	7	5	6	
6	5	7	9	5	5	
7						

Figura 5: Notas nos testes A, B, C e D para a amostra de cinco alunos.

Para simplificar vamos representar a variável "as notas do teste A" por X, a variável "as notas do teste B" por Y, a variável "as notas do teste C" por Z, e "as notas do teste D" por W. Para as variáveis X, Y, Z e W, a média, a moda e a mediana são iguais a cinco. Se usarmos a moda ou a média ou a mediana para comparar o desempenho desses cinco alunos nos quatro testes, poderíamos chegar a conclusão equivocada que o desempenho foi semelhante. Contudo, no teste C todos os alunos tiraram cinco, ou seja, o desempenho foi homogêneo, enquanto no teste B teve aluno que tirou um e teve aluno que tirou nove, ou seja, o desempenho dos alunos foi heterogêneo. Na figura 6, usamos um diagrama para representar as variáveis X, Y, Z e W. Na figura 6, cada aluno é representado por uma bolinha.

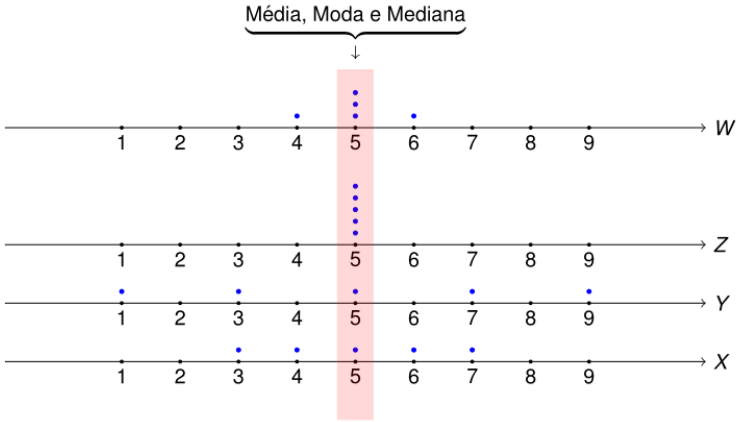


Figura 6: Distribuição das notas dos cinco alunos para o teste A (X), teste B (Y), teste C (Z) e o teste D (W). Note que a variável Z tem todos os valores iguais a cinco, enquanto a variável Y tem valores mais heterogêneos.

Se os valores (bolinhas no diagrama da figura 6) estão concentrados perto da média, então a variável é mais homogênea, e se os valores estão mais afastados da média, então a variável é mais heterogênea. A ideia das medidas de dispersão é calcular as distâncias entre os valores observados e a média: se as distâncias forem pequenas, então a variável é mais homogênea; se as distâncias forem grandes, então a variável é mais heterogênea. Para facilitar consideramos a média dos desvios $x_i - \bar{x}, i = 1, \dots, n$ (distâncias entre os valores observados e a média) e, com isso, obtemos três medidas de dispersão:

4.2.1 Desvio médio

Seja X uma variável quantitativa com valores observados $x_1, x_2, \dots, x_n$ com média $\bar{x}$, então você pode calcular o desvio médio por meio de

$$dm(x) = \frac{|x_1 - \bar{x}| + |x_2 - \bar{x}| + \cdots + |x_n - \bar{x}|}{n},$$

em que $|x_1 - \bar{x}| = \max\{(x_1 - \bar{x})\,;\, -(x_1 - \bar{x})\}$, ou seja, $|x_1 - \bar{x}|$ é o número sem o sinal. Chamamos $|x_1 - \bar{x}|$ de desvio absoluto, então o desvio médio é a média dos desvios absolutos.

4.2.2 Variância

Seja X uma variável quantitativa com valores observados $x_1, x_2, \dots, x_n$ com média $\bar{x}$, então você calcula a variância através de

$$var(x) = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n},$$

Note que $(x_1 - \bar{x})^2$ é a distância ao quadrado entre x_1 e $\bar{x}$ e chamamos $(x_1 - \bar{x})^2$ de desvio ao quadrado, então a variância é a média dos desvios ao quadrado.

4.2.3 Desvio Padrão

Suponha que X é uma variável quantitativa medida em cm (centímetros). Então, a unidade de $(x_1 - \bar{x})^2$ é cm^2 e a unidade da variância também vai ser em cm^2. Para manter a mesma unidade original dos dados, é comum considerar a raiz quadrada da variância

$$dp(x) = \sqrt{var(x)}.$$

Chamamos $dp(x)$ de Desvio Padrão.

Atenção para a seguinte interpretação: quanto menor o desvio padrão (ou variância ou desvio médio), mais homogênea a variável.

4.3 Quantis e diagrama de caixa (*boxplot*)

Você também pode resumir os dados usando o quantil de ordem *p*. Seja *X* uma variável quantitativa com valores observados $x_1, x_2, x_3, \dots, x_n$, o quantil q_p de ordem *p* é um valor tal que

- p100% dos valores de $x_1, x_2, x_3, \dots, x_n$ são menores ou igual a q_p;
- (1-p)100% dos valores de $x_1, x_2, x_3, \dots, x_n$ são maiores ou igual a q_p;

De forma análoga ao cálculo da mediana, a primeira coisa que você precisa fazer é encontrar as estatísticas de ordem $x_{(1)}$, $x_{(2)}$, $x_{(3)}$, $x_{(4)}$, ... , $x_{(n)}$. Lembre que $x_{(1)}$ é o menor valor de $x_1, x_2, x_3, \dots, x_n$; $x_{(2)}$ é o segundo menor valor de $x_1, x_2, x_3, \dots, x_n$; $x_{(3)}$ é o terceiro menor valor de $x_1, x_2, x_3, \dots, x_n$; e assim por diante. Então, se

- se $p < \frac{1}{n}$, então $q_p = x_{(1)}$;
- se existe um número inteiro $i = 1,2,3, \dots, n$ tal que $p = \frac{i}{n}$, então $q_p = x_{(i)}$;
- se $\frac{i}{n} < p < \frac{i+1}{n}$, então $q_p = \frac{x_{(i)} + x_{(i+1)}}{2}$;
- se $p > \frac{1}{n}$, então $q_p = x_{(n)}$.

Alguns quantis recebem um nome especial:

- se $p = 0{,}25$, chamamos $q_{0,25}$ de primeiro quartil e usamos a notação q_1;
- se $p = 0{,}5$, chamamos $q_{0,5}$ de segundo quartil ou mediana e usamos a notação q_2;
- se $p = 0{,}75$, chamamos $q_{0,75}$ de terceiro quartil e usamos a notação q_3.

Podemos representar o primeiro quartil, o segundo quartil e o terceiro quartil usando o diagrama de caixa (ou *boxplot* em inglês). Este diagrama de caixa é construído em formato de caixa conforme ilustração da figura 7.

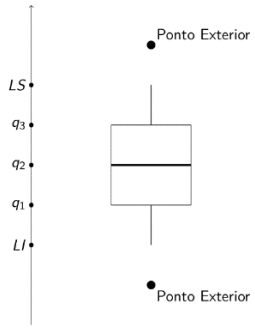


Figura 7: Diagrama de caixa (ou boxplot em inglês).

No diagrama de caixa representado na figura 7, calculamos *LS* e *LI* através de: $LS = q_3 + 1{,}5(q_3 - q_1)$ e $LI = q_1 - 1{,}5(q_3 - q_1)$. Se um valor da variável quantitativa é maior que *LS* ou é menor que *LI,* você classifica este valor como ponto exterior (*suspected outlier* em inglês) e um ponto exterior precisa de atenção do pesquisador. Ponto exterior pode ser um erro de digitação, ou de processamento, ou pode ser um valor possível mas raro.

No diagrama de caixa, se o valor do segundo quartil está mais próximo do primeiro quartil, significa que a variável quantitativa tem assimetria e os valores da variável tendem a ficar à direita. Nesse caso, dizemos que a variável quantitativa tem assimetria à direita. A Figura 8 ilustra essa ideia usando um histograma.

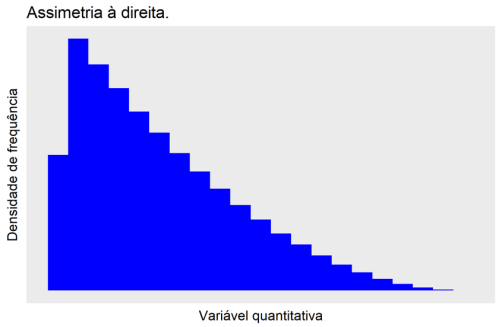


Figura 8: Histograma de uma variável quantitativa contínua com assimetria à direita.

Agora se no diagrama de caixa o valor do segundo quartil está mais próximo do terceiro quartil, significa que a variável quantitativa tem assimetria e os valores da variável tendem a ficar à esquerda. Neste caso, dizemos que a variável quantitativa tem assimetria à esquerda. A Figura 9 ilustra essa ideia usando um histograma.

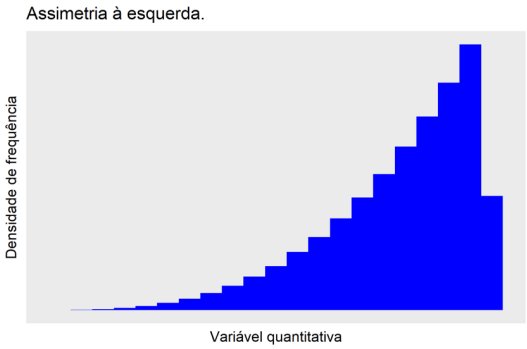


Figura 9: Histograma de uma variável quantitativa contínua com assimetria à esquerda.

Se o valor do segundo quartil está exatamente no meio entre o primeiro e terceiro quartil no diagrama de caixa, dizemos que a variável quantitativa é simétrica. A Figura 10 ilustra essa ideia.

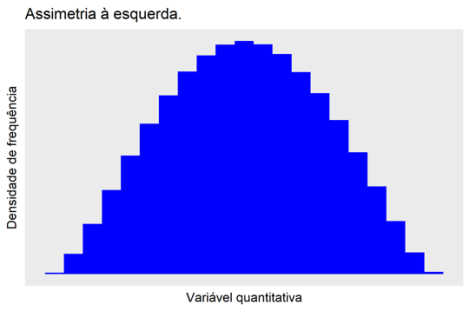


Figura 10: Histograma de uma variável quantitativa contínua com simetria.

Podemos transformar essa informação de simetria (ou assimetria) em uma medida que chamamos de coeficiente de assimetria de Bowley. A ideia é verificar se o segundo quartil está mais próximo do primeiro ou do terceiro quartil. Calculamos o coeficiente de assimetria de Bowley usando a seguinte equação:

$$B = \frac{(q_3 - q_2) - (q_2 - q_1)}{q_3 - q_1} = \frac{q_3 + q_1 - 2q_2}{q_3 - q_1}.$$

Note que

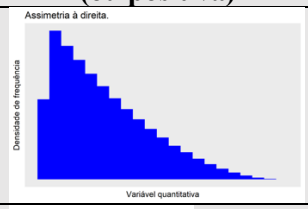
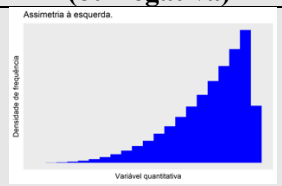
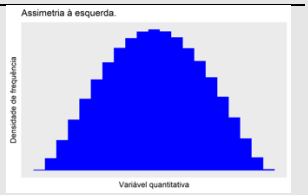
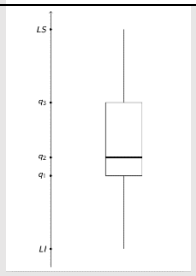
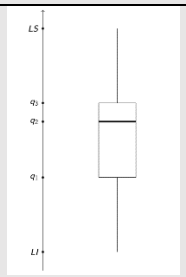
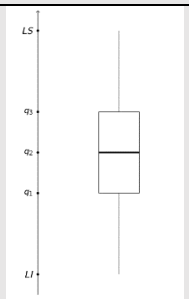
- B sempre está entre -1 e 1;
- Se o segundo quartil está mais próximo do primeiro quartil, então a variável tem assimetria a direita e $0 < B < 1$;
- Se o segundo quartil está mais próximo do terceiro quartil, então a variável tem assimetria a esquerda e $-1 < B < 0$;
- Se o segundo quartil está no meio do primeiro quartil e do terceiro quartil, então a variável é simétrica e $q_2 - q_1 \approx q_3 - q_2$. Ou seja, $B \approx 0$.
- Você pode usar, com parcimônia, a regra de ouro do quadro 1.

Quadro 1: Regra de ouro para o coeficiente de Bowley.

Interpretação	Coeficiente de Bowley	Interpretação	Coeficiente de Bowley
Forte assimetria à direita (positiva)	(0,9; 1]	Forte assimetria à esquerda (negativa)	[-1; -0,9)
Alta assimetria à direita (positiva)	(0,7; 0,9]	Alta assimetria à esquerda (negativa)	[-0,9; -0,7)
Moderada assimetria à direita (positiva)	(0,5; 0,7]	Moderada assimetria à esquerda (negativa)	[-0,7; -0,5)
Baixa assimetria à direita (positiva)	(0,3; 0,5]	Baixa assimetria à esquerda (negativa)	[-0,5; -0,3)
Simetria	[0; 0,3]	Simetria	[-0,3; 0]

No quadro 2, resumimos os tipos de assimetria (ou simetria) com o diagrama de caixa e o histograma.

Quadro 2: Diagrama de caixa, histograma e coeficiente de Bowley de acordo com o tipo assimetria.

	Assimetria à direita (ou positiva)	**Assimetria à esquerda (ou negativa)**	**Simetria**
Histograma			
Diagrama de caixa			
Coeficiente de Bowley	$0 < B < 1$	$-1 < B < 0$	$B \approx 0$

5 Associação entre Duas Variáveis

Nessa seção, você vai aprender a checar se duas variáveis estão associadas. Ou seja, queremos responder a seguinte pergunta: o conhecimento de uma variável X ajuda a entender ou descobrir o valor de uma variável Y? Vamos dividir essa seção em dois casos:

- X e Y são duas variáveis qualitativas;
- X e Y são duas variáveis quantitativas.

Decidimos focar nesses dois casos que provavelmente serão os que mais aparecerão em sua pesquisa. Caso você precise estudar a associação entre uma variável qualitativa e uma variável quantitativa ou entre duas variáveis qualitativas ordinais, apresentamos algumas referências ao final desse capítulo.

5.1 Associação entre Duas Variáveis Quantitativas

Nessa seção, você vai aprender a construir um gráfico e calcular uma medida numérica para mensurar a associação entre duas variáveis quantitativas. Primeiro apresentamos o gráfico de dispersão e, em seguida, você aprenderá a calcular o coeficiente de correlação linear de Pearson.

Sejam X e Y variáveis quantitativas com valores observados em uma amostra conforme a tabela 4.

Tabela 4: *Amostra de tamanho n para as variáveis X e Y.*

X	x_1	x_2	$\cdots$	x_n
Y	y_1	y_2	$\cdots$	y_n

Você pode representar os valores observados de X e Y usando o gráfico de dispersão. Em um gráfico de dispersão, representamos cada par $(x_1, y_1), (x_1, y_1), \ldots, (x_n, y_n)$ por um ponto no plano cartesiano, conforme os gráficos da figura 11. Três padrões podem estar presentes no gráfico de dispersão: associação positiva, negativa ou nula. No gráfico 11.**A** temos uma associação positiva e quanto maior o valor de X maior será o valor de Y. No gráfico 11.**B** temos uma associação negativa e quanto maior o valor de X, menor será o valor de Y. No gráfico 11.**C,** não temos associação entre as duas variáveis.

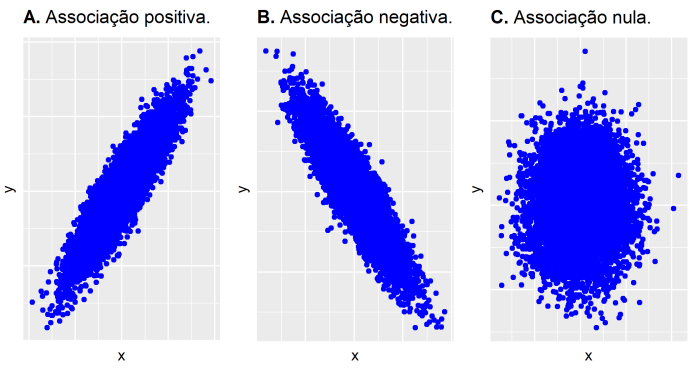


Figura 11: Gráfico de dispersão entre as variáveis quantitativas X e Y.

Além do gráfico de dispersão, você pode calcular o coeficiente de correlação de Pearson que representamos matematicamente por r. Note que

- o coeficiente de correlação de Pearson sempre está entre -1 e 1;
- se $r > 0$, então temos uma associação positiva;
- se $r < 0$, então temos uma associação negativa;
- se $r \cong 0$, então temos uma associação nula;
- você pode usar, com parcimônia, a regra de ouro da quadro 3.

Quadro 3: Regra de ouro para interpretação e uso do coeficiente de correlação linear de Pearson

Interpretação	Coeficiente de correlação linear de Pearson	Interpretação	Coeficiente de correlação linear de Pearson
Forte associação positiva	(0,9; 1]	Forte associação negativa	[-1; -0,9)
Alta associação positiva	(0,7; 0,9]	Alta associação positiva	[-0,9; -0,7)
Moderada associação positiva	(0,5; 0,7]	Moderada associação negativa	[-0,7; -0,5)
Baixa associação positiva	(0,3; 0,5]	Baixa associação negativa	[-0,5; -0,3)
Associação nula	[0; 0,3]	Associação nula	[-0,3; 0]

5.2 Associação entre Duas Variáveis Qualitativas

Antes de calcular, vamos explicar a você o que significa associação entre duas variáveis qualitativas. Mais especificamente, sejam X e Y duas variáveis qualitativas, então:

- X e Y estão associadas, se o conhecimento do valor de X para um indivíduo altera a plausibilidade dos valores de Y para este indivíduo da população;

- X e Y **não** estão associadas, se o conhecimento do valor de X para um indivíduo **não** altera a plausibilidade dos valores de Y para este indivíduo da população.

A primeira coisa que você deveria fazer ao estudar a associação entre duas variáveis qualitativas é contar, ou seja, construir uma tabela conjunta de distribuição de frequência como ilustrado na tabela 5. Na tabela 5, $n_{ij}, i,j = 1,2,3$ é o número de indivíduos com valor de X igual a A_i e com valor de Y igual a B_j; $n_{i.}, i = 1,2,3$ é o número elementos da amostra com valor de X igual a A_i; $n_{.j}, j = 1,2,3$ é o número elementos da amostra com valor de Y igual a B_j; e $n_{..}$ é o tamanho da amostra.

Tabela 5: *Tabela conjunta de distribuição de frequência para variável qualitativa Y com valores possíveis B_1, B_2, B_3 e para a variável qualitativa X com valores possíveis A_1, A_2, A_3.*

	Y			
X	B_1	B_2	B_3	Total
A_1	n_{11}	n_{12}	n_{13}	$n_{1.}$
A_2	n_{21}	n_{22}	n_{23}	$n_{2.}$
A_3	n_{31}	n_{32}	n_{33}	$n_{3.}$
Total	$n_{.1}$	$n_{.2}$	$n_{.3}$	$n_{..}$

Você pode facilitar ao leitor e calcular a frequência relativa em relação ao total da coluna (ou relação ao total da linha), conforme a tabela 6.

Tabela 6: *Tabela conjunta de distribuição frequência relativa ao total das colunas.*

	Y			
X	B_1	B_2	B_3	Total
A_1	$\frac{n_{11}}{n_{.1}}$	$\frac{n_{12}}{n_{.2}}$	$\frac{n_{13}}{n_{.3}}$	$\frac{n_{1.}}{n_{..}}$
A_2	$\frac{n_{21}}{n_{.1}}$	$\frac{n_{22}}{n_{.2}}$	$\frac{n_{23}}{n_{.3}}$	$\frac{n_{2.}}{n_{..}}$
A_3	$\frac{n_{31}}{n_{.1}}$	$\frac{n_{32}}{n_{.2}}$	$\frac{n_{33}}{n_{.3}}$	$\frac{n_{3.}}{n_{..}}$
Total	1	1	1	1

Se X e Y não são associadas, as colunas (ou linhas se você usar o total das linhas na tabela 6) devem ser iguais, ou seja:

- $\frac{n_{11}}{n_{.1}} = \frac{n_{12}}{n_{.2}} = \frac{n_{13}}{n_{.3}} = \frac{n_{1.}}{n_{..}}$;
- $\frac{n_{21}}{n_{.1}} = \frac{n_{22}}{n_{.2}} = \frac{n_{23}}{n_{.3}} = \frac{n_{2.}}{n_{..}}$;

- $\frac{n_{31}}{n_{.1}} = \frac{n_{32}}{n_{.2}} = \frac{n_{33}}{n_{.3}} = \frac{n_{3.}}{n_{..}}$;

Ou seja, as colunas (ou as linhas se você usar o total das linhas na tabela 6) vão ser todas iguais ou aproximadamente iguais. Se algum valor da linha (ou coluna) tiver um valor diferente, você já tem um sinal de que existe associação entre as duas variáveis qualitativas.

Podemos representar a tabela 6 usando um gráfico de barras conforme ilustrado abaixo na Figura 12. No gráfico 12.**B**, as variáveis qualitativas X e Y não são associadas e as barras são idênticas; e no gráfico 12.**A** as variáveis X e Y estão associadas e as barras são diferentes.

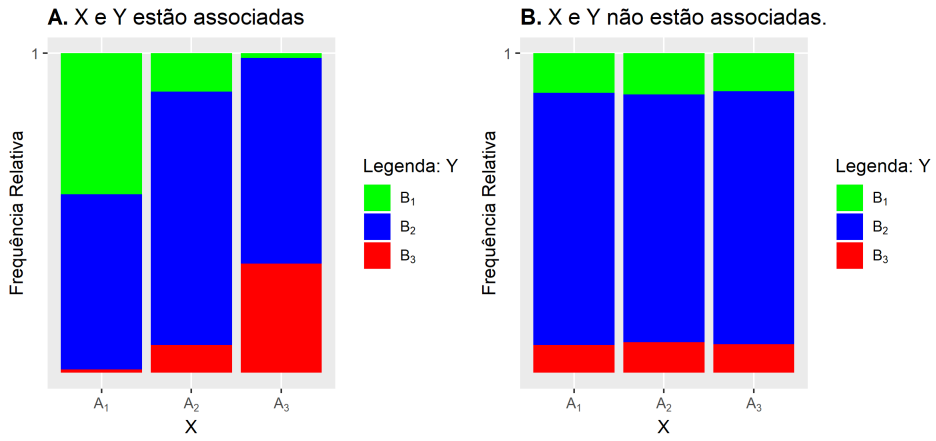


Figura 12: *Gráfico de barras para variáveis qualitativas.*

Você também pode avaliar a associação entre duas variáveis qualitativas usando uma medida numérica que em Estatística chamamos de coeficiente T de Tschuprow. O coeficiente T de Tschuprow satisfaz as seguintes propriedades:

- sempre está entre zero e um;
- quanto mais perto de um, maior a associação entre as duas variáveis qualitativas;
- quanto mais perto de zero, menor a associação a entre as duas variáveis qualitativas;
- você pode usar, com parcimônia, a regra de ouro do quadro 4.

Quadro 4: Regra de ouro para o coeficiente T de Tschuprow.

Interpretação	Coeficiente T de Tschuprow
Forte associação	(0,7; 1]
Moderada associação	(0,3; 0,7]
Sem associação	[0; 0,3]

6 Exemplo Ilustrativo

Nesta seção vamos usar o que aprendemos para ajudar Maria, a aluna de pós-graduação em informática na educação da seção **Era uma vez...** Maria decidiu acompanhar cem alunos da plataforma digital e coletou as seguintes variáveis para cada um dos alunos:

- `ID`: rótulo usado para identificar os alunos na plataforma digital;
- `Nota`: nota em matemática;
- `tempo`: tempo (em minutos) que o aluno ficou logado na plataforma digital na semana;
- `genero`: gênero declarado pelo aluno;
- `localizacao`: variável qualitativa com dois valores possíveis – Capital e Interior. Capital indica que o aluno mora em uma capital ou região metropolitana, e Interior indica que o aluno não mora em uma capital ou região metropolitana.

Estas variáveis foram armazenadas em arquivo excel denominado `"data_plataforma.xlsx"`. Vamos ajudar Maria em duas tarefas:

(1) Resumir, descrever e analisar as variáveis `Nota`, `tempo` e `genero`;
(2) Estudar a associação entre `tempo` e `Nota` e a associação entre `localizacao` e `genero`.

Quadro 5: Maria vai usar os seguintes pacotes do R.

Pacote	Descrição do pacote
`tidyverse`	Oferece algumas ferramentas que podem simplificar e economizar tempo no R.
`readxl`	Importa arquivos excel no R.
`DescTools`	Calcula o coeficiente de T de Tschuprow e várias medidas descritivas.

Vamos começar carregando os pacotes e arquivo `"data_plataforma.xlsx"`.

```
# Carregando os pacotes
library(tidyverse)
library(DescTools)
library(readxl)

# importando o arquivo no R
dados_maria <- read_xlsx("data_plataforma.xlsx", sheet = "dados",
                         col_names = TRUE)

# as cinco primeiras observações
head(dados_maria, n = 5)
```

```
## # A tibble: 5 x 5
##      ID  Nota tempo genero    localizacao
##   <dbl> <dbl> <dbl> <chr>     <chr>
## 1 71573  8.72 106.  Feminino  Capital
## 2 88855  8.68  93.5 Feminino  Interior
## 3 52826  5.39  62.8 Feminino  Interior
## 4 14692  7.79  88.6 Feminino  Interior
## 5 28539  8.55  87.4 Masculino Capital
```

Vamos primeiro fazer uma análise descritiva das variáveis: `Nota, tempo` e `genero`. Vamos começar com a variável `genero`. A sua primeira tarefa é construir a tabela de distribuição de frequência, conforme código abaixo.

```
# Tabela de distribuição de frequência
dados_maria %>% group_by(genero) %>%
  summarise(frequencia = n()) %>%
  mutate(frequencia_relativa = frequencia / sum(frequencia),
         porcentagem = 100 * frequencia_relativa)
```

```
## # A tibble: 3 x 4
##   genero    frequencia frequencia_relativa porcentagem
##   <chr>          <int>               <dbl>       <dbl>
## 1 Feminino          49                0.49          49
## 2 Masculino         47                0.47          47
## 3 Outro              4                0.04           4
```

Podemos representar a tabela de distribuição de frequência usando um gráfico de barras, conforme figura 13. No gráfico da figura 13, notamos que existe uma minoria com sexo “outro” e um equilíbrio entre o número de alunos do sexo “masculino” e do sexo “feminino”.

```
# Gráfico de barras: Gênero
ggplot(dados_maria) +
  geom_bar(aes(x = genero, y = ..prop.., group = 1),
           fill = "blue") +
  xlab("Gênero") + ylab("Frequência relativa") +
  ylim(c(0,1))
```

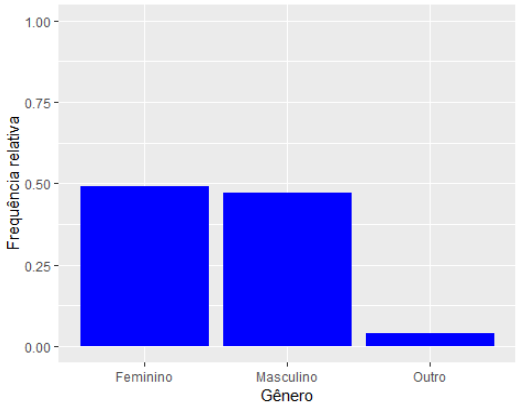


Figura 13: Gráfico de barras para variável `genero`.

Agora vamos estudar a associação entre `localizacao` e `genero`. Primeiramente, vamos construir a tabela conjunta de distribuição de frequência, e em seguida construir a tabela conjunta de distribuição de frequência relativa ao total das colunas, conforme o código abaixo.

```
# tabela conjunta de distribuição de frequências
dados_maria %>% group_by(localizacao, genero) %>%
  summarise(frequencia = n()) %>%
  spread(key = genero, value = frequencia)
```

```
## # A tibble: 2 x 4
## # Groups:   localizacao [2]
##   localizacao Feminino Masculino Outro
##   <chr>          <int>     <int> <int>
## 1 Capital           39        32     3
## 2 Interior          10        15     1
```

```
# tabela conjunta de distribuição de frequências relativa ao total das
# colunas
dados_maria %>% group_by(localizacao, genero) %>%
  summarise(frequencia = n()) %>%
  spread(key = genero, value = frequencia) %>% ungroup() %>%
  mutate( Feminino = 100 * Feminino / sum(Feminino),
          Masculino = 100 * Masculino / sum(Masculino),
          Outro = 100 * Outro / sum(Outro) )
```

```
## # A tibble: 2 x 4
##   localizacao Feminino Masculino Outro
##   <chr>          <dbl>     <dbl> <dbl>
## 1 Capital         79.6      68.1    75
## 2 Interior        20.4      31.9    25
```

Podemos representar essas duas tabelas conjuntas de distribuição de frequência usando um gráfico de barras, conforme código abaixo. O trecho de código abaixo produz o gráfico da figura 14. Note que as barras do gráfico da figura 14 são semelhantes e isso indica a ausência de associação entre Localização e Gênero.

```
# Gráfico de barras entre Localização e Gênero
# Associação: gráfico de barras
ggplot(dados_maria) +
  geom_bar(aes(x=genero, fill=localizacao),
           position = "fill") +
  xlab("Gênero") + ylab("Frequência relativa")
```

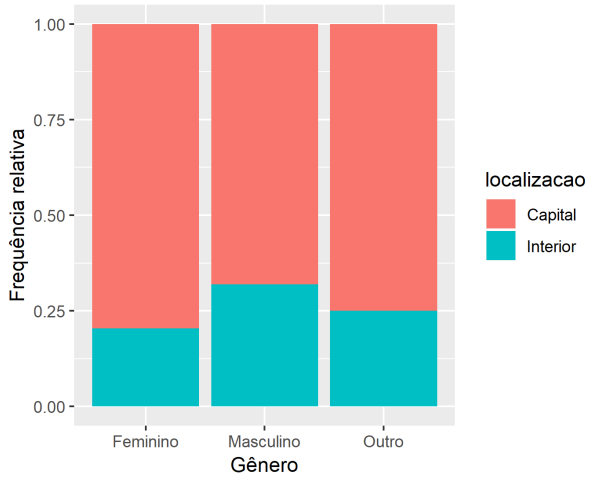


Figura 14: Associação entre `localizacao` e `genero`.

Você pode calcular o coeficiente T de Tschuprow, que transforma a informação visual do gráfico da figura 14 em uma medida ente 0 e 1: quanto mais perto de zero, menor a associação e quanto mais próximo de 1, mais forte a associação. No R, vamos usar a função `TschuprowT` do pacote `DescTools`. O coeficiente de contingência é $C = 0{,}11$, um valor próximo de 0, e temos mais um indício da não associação entre `localizacao` e `genero`.

```
# Coeficiente T de Tschuprow
TschuprowT(dados_maria$genero,
           dados_maria$localizacao)
```

```
## [1] 0.1081155
```

Vamos analisar agora a variável `Nota` e começamos com a tabela de distribuição de frequência com o código a seguir. Notamos que as faixas de notas (7,8], (8,9] e (9,10] são as mais frequentes entre os alunos.

```
# tabela de distribuição de frequência
dados_maria %>% group_by(Nota= cut(Nota, breaks = seq(from=0, to=10,
by=1)) ) %>%
  summarise(frequencia = n()) %>%
  mutate(frequencia_relativa = frequencia / sum(frequencia),
            porcentagem = frequencia_relativa * 100)
```

```
## # A tibble: 9 x 4
##   Nota   frequencia frequencia_relativa porcentagem
##   <fct>       <int>               <dbl>       <dbl>
## 1 (0,1]           2                0.02           2
## 2 (2,3]           1                0.01           1
## 3 (3,4]           2                0.02           2
## 4 (4,5]           2                0.02           2
## 5 (5,6]           3                0.03           3
## 6 (6,7]           8                0.08           8
## 7 (7,8]          17                0.17          17
```

```
## 8 (8,9]          25                0.25          25
## 9 (9,10]         40                0.4           40
```

Podemos ir além e resumir a variável `Nota` usando as medidas de posição, medidas de dispersão e quantis usando o código abaixo.

```
# Medidas resumo para variável nota
dados_maria %>%
  summarise(media = mean(Nota),
            mediana = median(Nota),
            desvio_padrao = (Nota - media)^2 %>% mean() %>% sqrt(),
            desvio_medio = (Nota - media) %>% abs() %>% mean(),
            Q1 = quantile(Nota, probs = 0.25),
            Q3 = quantile(Nota, probs = 0.75))
```

```
## # A tibble: 1 x 6
##   media mediana desvio_padrao desvio_medio    Q1    Q3
##   <dbl>   <dbl>         <dbl>        <dbl> <dbl> <dbl>
## 1  8.12    8.72          1.87         1.31  7.68  9.34
```

De forma análoga, podemos construir a tabela de distribuição de frequência e calcular medidas de resumo para a variável `tempo` usando o código abaixo.

```
# tabela de distribuição de frequência
dados_maria %>%
  group_by(tempo = cut(tempo, breaks = seq(from=30,to=160, by = 10)))
%>%
  summarise(frequencia = n()) %>%
  mutate(frequencia_relativa = frequencia / sum(frequencia),
         porcentagem = 100 * frequencia_relativa)
```

```
## # A tibble: 13 x 4
##    tempo     frequencia frequencia_relativa porcentagem
##    <fct>          <int>               <dbl>       <dbl>
##  1 (30,40]            1                0.01           1
##  2 (40,50]            1                0.01           1
##  3 (50,60]            1                0.01           1
##  4 (60,70]            2                0.02           2
##  5 (70,80]           13                0.13          13
##  6 (80,90]           15                0.15          15
##  7 (90,100]          18                0.18          18
##  8 (100,110]         16                0.16          16
##  9 (110,120]         11                0.11          11
## 10 (120,130]         14                0.14          14.
## 11 (130,140]          4                0.04           4
## 12 (140,150]          3                0.03           3
## 13 (150,160]          1                0.01           1
```

```
# medidas resumo
dados_maria %>%
  summarise(media = mean(tempo), mediana = median(tempo),
            desvio_padrao = (tempo - media)^2 %>% mean() %>% sqrt(),
            desvio_medio = (tempo - media) %>% abs() %>% mean(),
            Q1 = quantile(tempo, probs=0.25),
            Q3 = quantile(tempo, probs=0.75))
```

```
## # A tibble: 1 x 6
##   media mediana desvio_padrao desvio_medio    Q1    Q3
##   <dbl>   <dbl>         <dbl>        <dbl> <dbl> <dbl>
## 1 100.0    98.6          22.1         18.1  84.7  116.
```

Pela tabela de distribuição de frequência, notamos que a maioria dos alunos gasta entre 70 e 130 minutos logados na plataforma digital e isso já é uma informação adicional que dificilmente você obteria nos valores individuais do arquivo `"data_plataforma.xlsx"`. As medidas de resumo são úteis para entender variáveis quantitativas. A nota média dos estudantes foi 8,12 e tempo médio logado foi 100 minutos.

Quando e onde puder, use gráficos! Eles ajudam o leitor a entender o que está acontecendo com as variáveis. Para as variáveis quantitativas contínuas `Nota` e `tempo` vamos construir o Diagrama de Caixa e o Histograma. No R, o Diagrama de Caixa pode ser construído usando a função `geom_boxplot` e o histograma pode ser construído usando a função `geom_histogram`. As funções `geom_boxplot` e `geom_histogram` são funções que estão inclusas no pacote `tidyverse`. Vamos apresentar o código para calcular o histograma e o diagrama de caixa para a variável `tempo`, e o mesmo código, com as devidas alterações, pode ser usada para a variável `Nota`.

```
# histograma: tempo
m <- (1 + dados_maria %>% nrow() %>% log2()) %>% ceiling()
ggplot(dados_maria)+
  geom_histogram(aes(x=tempo, y=..density..), bins = m,
                 fill = "blue")+
  xlab("Tempo") + ylab("Densidade de frequência")

# diagrama de caixa: tempo
ggplot(dados_maria) +
  geom_boxplot(aes(x="",y=tempo)) +
  xlab("") + ylab("Tempo")
```

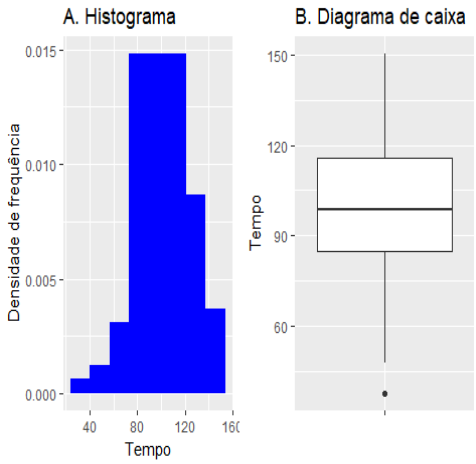


Figura 15: No gráfico à esquerda temos o histograma e no gráfico à direita temos o diagrama de caixa (ou *boxplot* em inglês) para a variável `tempo`.

No histograma e no diagrama de caixa da figura 15, notamos uma simetria em torno da média 100 minutos. Podemos calcular o coeficiente de Bowley para avaliar a simetria da variável `tempo` conforme o código abaixo. O coeficiente de Bowley é $B = 0{,}10$ e podemos interpretar que a variável é simétrica.

```
# Coeficiente de Bowley: tempo
quartis <- 1:3 %>% map_dbl(function(i) dados_maria$tempo %>%
                             quantile(probs = i/4 ))
(B <- (quartis[3] - 2 * quartis[2] + quartis[1]) /
  (quartis[3] - quartis[1]))
```

```
## [1] 0.1033201
```

Você também pode construir o diagrama de caixa e o histograma para a variável quantitativa contínua `Nota`, como mostrado na figura 16. Você pode notar que a mediana está ligeiramente mais perto do terceiro quartil no diagrama de caixa, e os valores tendem a ficar acumulados a esquerda no histograma.

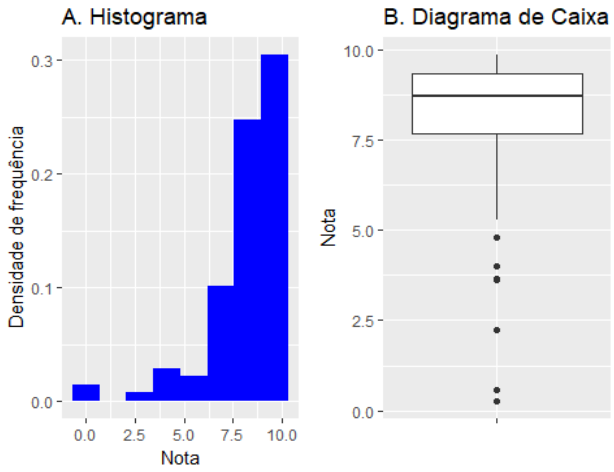


Figura 16: No gráfico da esquerda, mostramos o histograma para a variável `Nota`. No gráfico da direita, mostramos o diagrama de caixa da variável `Nota`.

Adicionalmente, podemos calcular o coeficiente de Bowley para avaliar a assimetria da variável Nota, conforme o código abaixo. O coeficiente de Bowley é $B = -0{,}25$, e a variável Nota é aproximadamente simétrica.

```
# Coeficiente de Bowley: Nota
quartis <- 1:3 %>% map_dbl(function(i) dados_maria$Nota %>%
                           quantile(probs = i/4 ))
(B <- (quartis[3] - 2 * quartis[2] + quartis[1]) /
  (quartis[3] - quartis[1]))
```

```
## [1] -0.2541353
```

Depois de estudar cada variável quantitativa individualmente, está na hora de estudar a associação entre as variáveis `Nota` e `tempo`. Primeiro construímos o gráfico de dispersão, mostrado na figura 17. Observamos uma tendência: os alunos que gastam mais tempo na plataforma digital têm notas maiores. Ou seja, as variáveis estão positivamente associadas.

```
# gráfico de dispersão
ggplot(dados_maria)+
  geom_point(aes(x=Nota, y = tempo), color = "blue")
```

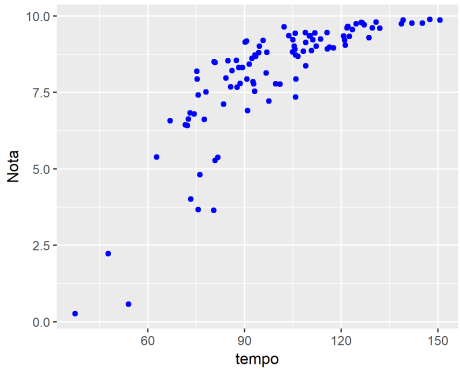


Figura 17: Gráfico de dispersão entre as variáveis quantitativas `tempo` e `Nota`.

Você pode calcular uma medida inspirada no diagrama de dispersão chamada de coeficiente de correlação linear de Pearson r. O valor do coeficiente de correlação de Pearson é $r = 0{,}79$ e temos uma alta associação positiva entre `Nota` e `tempo`. Podemos usar a função `cor` para calcular o coeficiente de correlação linear de Pearson.

```
# coeficiente de correlação linear entre Nota e tempo
cor(dados_maria$tempo, dados_maria$Nota)
```

```
## [1] 0.7949051
```

Resumindo, chegamos às seguintes conclusões:

(1) Existe uma igual proporção de homens e mulheres na plataforma, com uma mino ria de alunos se declarando “Outro” (pessoas que não se identificam como Home m nem como mulher);

(2) Não existe associação entre `localizacao` e genero;

(3) Obtemos as medidas resumo para as variáveis `tempo` e `Nota` apresentadas na tabe la 7;

Tabela 7: Medidas resumo para as variáveis `Nota` e `tempo`.

Variável	Média	Mediana	Desvio Padrão	Desvio Médio	Q1	Q3
Tempo	99,99	98,57	22,06	18,06	84,70	115,64
Nota	8,12	8,72	1,87	1,31	7,68	9,34

(4) As variáveis `tempo` e `Nota` são aproximadamente simétricas;

(5) As variáveis `tempo` e `Nota` estão positivamente associadas.

7 Resumo

Neste capítulo, estudamos estatística descritiva. Começando com a observação de que existe duas formas básicas de inferência: dedutiva e indutiva, e que estatística tem foco na inferência indutiva. Aprendemos conceitos básicos de estatística como população, amostra, parâmetro, estimativa e variável. Em seguida, mostramos como representar graficamente (gráfico de barras e histograma) os valores observados de uma variável com o objetivo de visualizar informações e padrões. Fomos além e aprendemos a encontrar valores que representam todos os valores observados usando medidas resumo: medidas de posição (média, moda e mediana), medidas de dispersão (variância, desvio padrão e desvio médio) e quantis. O próximo passo foi estudar a associações entre duas variáveis usando gráficos e medidas de associação como gráfico de dispersão, coeficiente de correlação linear de Pearson e o coeficiente T de Tschuprow. Finalmente, na última seção deste capítulo, ilustramos os conceitos e métodos apresentados com um exemplo no R.

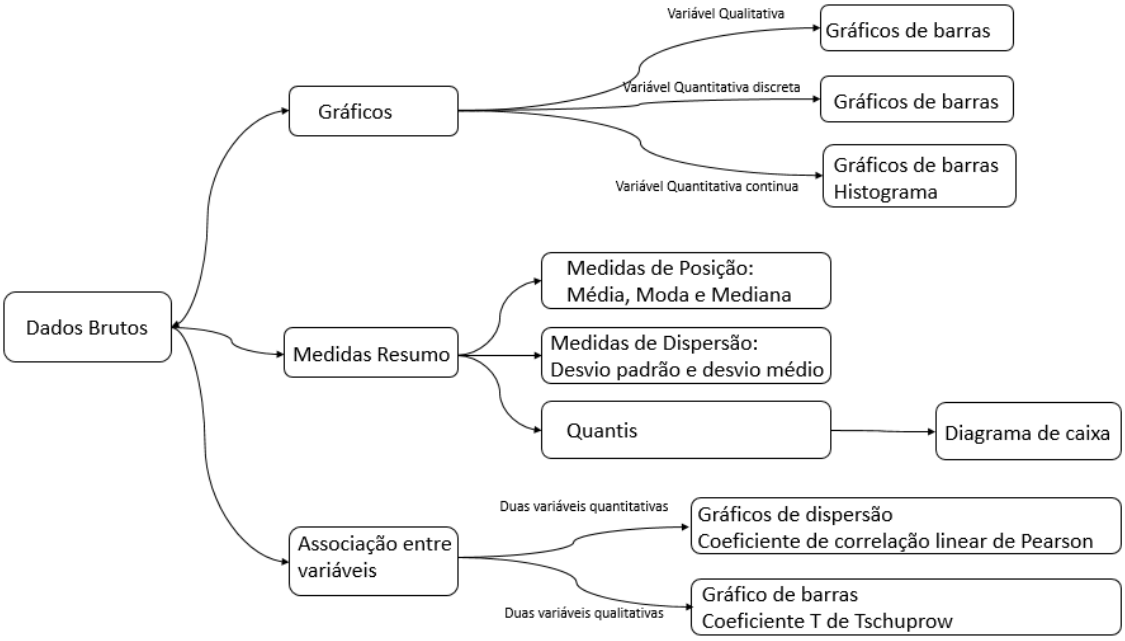


Figura 18: Mapa mental para estatística descritiva.

8 Leituras Recomendadas

- **Estatística Básica** (BUSSAB; MORETTIN, 2014). Este livro é uma referência bastante usada para cursos introdutórios de Estatística Descritiva e Inferencial. O livro cobre os tópicos deste capítulo com mais detalhes, além de apresentar outros tópicos de estatística, incluindo estatística inferencial, regressão linear e probabilidade. Os autores mostram como estudar a associação entre uma variável qualitativa e uma variável quantitativa. Além disso, os autores disponibilizam os conjuntos de dados usados no livro no endereço eletrônico: https://www.ime.usp.br/~pam/EstBas.html.

- **Estatística Aplicada às Ciências Sociais** (BARBETTA, 2008). Este livro é uma referência bastante usada para cursos introdutórios de Estatística Descritiva e Inferencial para ciências humanas. Existe um grande foco em aplicação e interpretação.
- **Understanding Robust and Exploratory Data Analysis** (HOAGLIN; MOSTELLIER; TUKEY, 2000). Este é um texto clássico para entender estatística descritiva. Os autores apresentaram as medidas resumo e os métodos gráficos que vimos nesse capítulo.
- **R for Data Science** (WICKHAM; GROLEMUND, 2016). Este livro apresenta desde estatística descritiva até modelagem usando o R. O autor diz explicitamente que está preocupado em difundir o uso de R para análise de dados. O livro tem uma leitura fluída e existe a opção de você ler o livro gratuitamente no endereço eletrônico: http://r4ds.had.co.nz/ .

9 Artigos Exemplos

- **Reduced GUI for an interactive geometry software: Does it affect students' performance?** (BORGES et al., 2015). Este artigo foi publicado na revista *Computers in Human Behaviour* e usa medidas resumo e diagrama de caixa.
- **What do students do on-line? Modeling students' interactions to improve their learning experience** (PAIVA et al., 2016). Este artigo analisa a interação de estudantes com um ambiente de aprendizado on-line (MeuTutor). Os autores construíram histogramas e gráficos de barras.

10 Checklist

- Identificar o tipo de variável de cada coluna de sua base de dados: qualitativa ordinal, qualitativa nominal, quantitativa discreta e quantitativa contínua.
- Construir gráfico de barras para variáveis qualitativas.
- Construir histogramas para variáveis quantitativas.
- Calcular quantis e construir o diagrama de caixa para as variáveis quantitativas, e analisar a assimetria das variáveis quantitativas.
- Estude a associação entre duas variáveis qualitativas usando gráficos de barras e coeficiente T de Tschuprow.
- Estude a associação entre duas variáveis quantitativas usando gráficos de dispersão e coeficiente de correlação linear de Pearson.

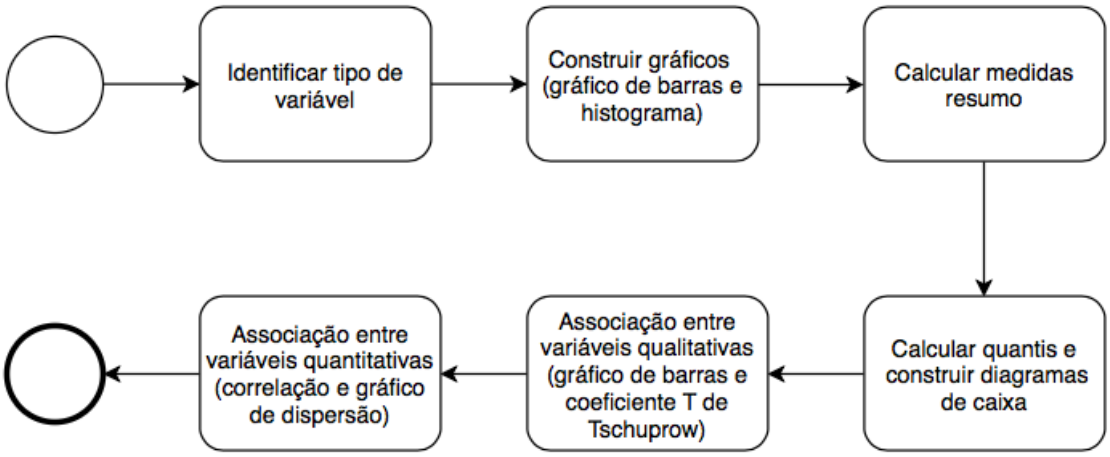


Figura 19: *Fluxograma para análise descritiva.*

11 Exercícios

1) Um professor de estatística coletou as notas finais e a idade de turma com 15 alunos ao final do semestre corrente. Ele salvou as informações no arquivo `"notas_finais.xlsx"`.

a. Calcule a média, mediana, desvio médio e desvio para as variáveis quantitativas `"notas"` e `"idade"`;
b. Construa o histograma para as variáveis quantitativas `"notas"` e `"idade"`. Interprete os resultados.
c. Construa o gráfico de barras para as variáveis `"genero"` e `"localização"`, e calcule o coeficiente T de Tschuprow. Você acha que estas variáveis estão associadas?
d. Construa o gráfico de dispersão entre `"notas"` e `"idade"`, e calcule o coeficiente de correlação linear entre `"notas"` e `"idade"`. Você acha que estas variáveis estão associadas?

2) Considere o conjunto de dados `"nlschools"` do pacote MASS do R. Esse conjunto de dados tem as variáveis quantitativas QI e a nota em linguagem para 2287 crianças holandesas.

a. Calcule a média, mediana, desvio médio e desvio para as variáveis quantitativas `"lang"` e `"IQ"`;
b. Construa o histograma para as variáveis quantitativas `"lang"` e `"IQ"`. Interprete os resultados.
c. Construa o gráfico de dispersão entre `"lang"` e `"IQ"`, e calcule o coeficiente de correlação linear entre `"lang"` e `"IQ"`. Você acha que estas variáveis estão associadas?

3) Considere o conjunto de dados `"Caschool"` do pacote Ecdat do R. Esse conjunto contém informações de 420 escolas na Califórnia, incluindo a nota média de testes em leitura e matemática em 1999.

a. Calcule a média, mediana, desvio médio e desvio para as variáveis: `"computer"` (número de computadores na escola), `"readscr"` (nota média em leitura) e `"mathscr"` (nota média em matemática).
b. Construa o histograma para a variável `"computer"`. Você acha que esta variável é simétrica? Justifique a sua resposta.
c. Você acha que as `"computer"` e `"mathscr"` estão associadas? E as variáveis `"mathscr"` e `"readscr"`? Justifique a sua resposta.

12 Referências

BARBETTA, Pedro A. **Estatística aplicada às ciências sociais**. Editora UFSC. **Florianópolis**, 2008.

BORGES, Simone S. et al. Reduced GUI for an interactive geometry software: Does it affect students' performance? **Computers in Human Behavior**, v. 54, p. 124-133, 2016.

BUSSAB, Wilton O.; MORETTIN, Pedro A. **Estatística Básica.** Editora Saraiva. **São Paulo**, 2005.

CASELLA, George; BERGER, Roger L. **Statistical inference**. Pacific Grove, CA: Duxbury, 2002.

COSTNER, Herbert L. Criteria for measures of association. **American Sociological Review**, p. 341-353, 1965.

KIM, Tae-Hwan; WHITE, Halbert. On more robust estimation of skewness and kurtosis. **Finance Research Letters**, v. 1, n. 1, p. 56-73, 2004.

MOOD, Alexander McFarlane; GRAYBILL, Franklin A.; BOES, Duane C. **Introduction to the Theory of Statistics**. McGraw-Hill Kogakusha, 1974.

PAIVA, Ranilson et al. What do students do on-line? Modeling students' interactions to improve their learning experience. **Computers in Human Behavior**, v. 64, p. 769-781, 2016.

ROYSTON, Patrick. Which measures of skewness and kurtosis are best? **Statistics in Medicine**, v. 11, n. 3, p. 333-343, 1992.

TAYLOR, Richard. Interpretation of the correlation coefficient: a basic review. **Journal of Diagnostic Medical Sonography**, v. 6, n. 1, p. 35-39, 1990.

TUKEY, John Wilder; MOSTELLER, Frederick; HOAGLIN, David Caster (Ed.). **Understanding robust and exploratory data Analysis**. Wiley, 1983.

WICKHAM, Hadley; GROLEMUND, Garrett. **R for data science: import, tidy, transform, visualize, and model data**. O'Reilly Media, Inc., 2016.

Sobre o Autor

Gilberto Pereira Sassi

http://lattes.cnpq.br/7008457711842107

Doutor em Estatística e mestre em Ciência da Computação e Matemática Computacional e graduado em Matemática pela Universidade de São Paulo, Brasil. Atualmente, Gilberto é Professor Adjunto na Universidade Federal da Bahia, e sua pesquisa está focada em Análise de Dados Funcionais.