

1

Medidas de posição e dispersão

O QUÊ?

Medidas de posição: média, mediana, moda e quartis. Medidas de dispersão: desvio médio, variância, desvio padrão, amplitude amostral, distância entre quartis e coeficiente de variação. Construção do desenho-esquemático (boxplot).

POR QUÊ?

As medidas resumo (posição e dispersão) correspondem a uma síntese do conjunto de dados observados e ao passo preliminar para fazer uma inferência estatística, ou seja, a partir das informações obtidas na amostra, expandir nossas conclusões para a população. Como as distribuições podem apresentar formas variadas é importante conhecer diferentes tipos de medidas resumo, tanto de posição como de dispersão, para usar medidas apropriadas em cada caso.

Projeto: LIVRO ABERTO DE MATEMÁTICA



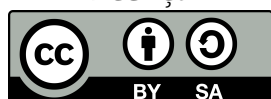
Cadastre-se como colaborador no site do projeto: umlivroaberto.org

Versão digital do capítulo:

https://www.umlivroaberto.org/BookCloud/Volume_1/master/view/PE104.html

Título:	Medidas de Posição e Dispersão
Ano/ Versão:	2020 / versão 1.2 de 17 de novembro de 2020
Editora	Instituto Nacional de Matemática Pura e Aplicada (IMPA-OS)
Realização:	Olimpíada Brasileira de Matemática das Escolas Públicas (OBMEP)
Produção:	Associação Livro Aberto
Coordenação:	Fabio Simas e Augusto Teixeira (livroaberto@impa.br)
Autores:	Flávia Landim (coordenadora da equipe - UFRJ), Nei Rocha (UFRJ), Vanessa Matos (SEduc Angra dos Reis e Mesquita).
Revisão:	Cydara Ripoll Alexandre Silva
Design:	Andreza Moreira (Tangentes Design)
Ilustrações:	—
Gráficos:	Beatriz Cabral e Tarso Caldas (Licenciandos da UNIRIO)
Capa:	Foto de Clement Chen, no Unsplash https://unsplash.com/photos/EvoIUzuW89s

Licença:



Desenvolvido por



Patrocínio:



Medidas de Posição e Dispersão

Neste capítulo serão trabalhadas as seguintes habilidades da BNCC para o Ensino Médio:

EM13MAT202

Planejar e executar pesquisa amostral usando dados coletados ou de diferentes fontes sobre questões relevantes atuais, incluindo ou não, apoio de recursos tecnológicos, e comunicar os resultados por meio de relatório contendo gráficos e interpretação das medidas de tendência central e das de dispersão.

EM13MAT316

Resolver e elaborar problemas, em diferentes contextos, que envolvem cálculo e interpretação das medidas de tendência central (média, moda, mediana) e das de dispersão (amplitude, variância e desvio padrão).

EM13MAT408

Construir e interpretar tabelas e gráficos de frequências, com base em dados obtidos em pesquisas por amostras estatísticas, incluindo ou não o uso de softwares que inter-relacionem estatística, geometria e álgebra.

EM13MAT409

Interpretar e comparar conjuntos de dados estatísticos por meio de diferentes diagramas e gráficos, como o histograma, o de caixa (boxplot), o de ramos e folhas, reconhecendo os mais eficientes para sua análise.

EM13MAT510

Investigar conjuntos de dados relativos ao comportamento de duas variáveis numéricas, usando tecnologias da informação, e, se apropriado, levar em conta a variação e utilizar uma reta para descrever a relação observada.

Objetivo Geral do Capítulo

Compreender como algumas medidas, chamadas medidas resumo, podem revelar informações sobre a distribuição dos dados. Para alcançar este objetivo geral, este capítulo inclui:

a) definições de medidas de posição (média, mediana, moda e quartis);

b) definições de medidas de dispersão (desvio médio, variância e desvio padrão, amplitude (amostral), distância entre quartis, coeficiente de variação);

c) sinalizações para a importância das medidas de dispersão para avaliar a representatividade da média na distribuição;

d) revisão dos conceitos apresentados no capítulo **A Natureza da Estatística**: parâmetro e estatística, definindo média populacional e média amostral, variância populacional e variância amostral.

e) apresentação de representação gráfica para variáveis quantitativas conhecida como *boxplot* (gráfico caixa).

Observação: Embora muitos textos denominem as medidas de posição (média, mediana e moda) como medidas de tendência central, preferimos denominá-las medidas de posição, pois, por exemplo, na presença de forte assimetria, a média não será um valor central, pois será influenciada pelos valores extremos da distribuição.

Por que estudar o assunto?

As medidas resumo (posição e dispersão) correspondem a uma síntese do conjunto de dados observados e ao passo preliminar para fazer uma inferência estatística, ou seja, a partir das informações obtidas na amostra, expandir nossas conclusões para a população. Como as distribuições podem apresentar formas variadas é importante conhecer diferentes tipos de medidas resumo, tanto de posição como de dispersão, para usar medidas apropriadas em cada caso.

Desafios do Capítulo

a) Entender as especificidades das variadas medidas resumo e reconhecer suas limitações em cada contexto.

b) A partir de algumas medidas resumo, ser capaz de ter uma noção quanto à forma da distribuição dos dados.

Conteúdos abordados

a) Medidas de posição: média, mediana, moda e quartis.

b) Medidas de dispersão: amplitude, distância entre quartis, desvio médio, variância, desvio padrão e coeficiente de variação.

c) Parâmetro versus Estimador (estimativa) (Média populacional versus média amostral, Variância populacional versus variância amostral, etc.).

d) Construção do *boxplot* (gráfico caixa).

Pré-requisitos

EF07MA29

Compreender, em contextos significativos, o significado de média estatística como indicador da tendência de uma pesquisa, calcular seu valor e relacioná-lo, intuitivamente, com a amplitude do conjunto de dados.

EF08MA22

Obter os valores de medidas de tendência central de uma pesquisa estatística (média, moda e mediana) com a compreensão de seus significados e relacioná-los com a dispersão de dados, indicada pela amplitude.

EM11MT03

Realizar pesquisas, considerando: o planejamento, a discussão (se será censitária ou por amostra), a seleção de amostras, a elaboração e aplicação de instrumentos de coleta, a organização e representação dos dados (incluindo agrupamentos de dados em classe), a construção de gráficos apropriados (incluindo o histograma), a interpretação e a análise crítica apresentadas em relatórios descritivos.

Desdobramentos imediatos

EM15MT06

Analisar criticamente os métodos de amostragem em relatórios de pesquisas divulgadas pela mídia e as afirmativas feitas para toda a população baseadas em uma amostra.

Abordagem do Capítulo

Pretende-se ao longo do capítulo, além de apresentar as definições das variadas medidas, enfatizar suas propriedades. Por exemplo, no caso da média, pretende-se explorar as seguintes propriedades:

- a) a média de um conjunto de números é um valor que pertence ao intervalo delimitado pelos valores extremos deste conjunto (mínimo e máximo);
- b) a soma dos desvios da média é sempre zero para qualquer conjunto de números;
- c) a média é influenciada por todos os valores no conjunto;
- d) a média pode ser um valor que não pertence ao conjunto analisado;
- e) a média é um valor representativo do conjunto analisado.

Após definir a média e explorar algumas de suas propriedades, situações em que ela pode ser inadequada como valor representativo de uma distribuição (distribuições com forte assimetria) são apresentadas, motivando assim a definição da mediana, uma medida que é pouco afetada por valores extremos.

Uma vez definidas média e mediana, uma situação em que ambas podem ser consideradas inadequadas como valores representativos de uma distribuição (distribuições bimodais simétricas) é apresentada, motivando assim a definição da moda.

Em adição às medidas de posição, medidas de dispersão complementam a descrição de uma distribuição. A motivação para a necessidade de definir medidas de dispersão será realizada com base em uma atividade em que dois conjuntos de dados apresentam média, mediana e moda coincidentes, mas suas distribuições são diferentes.

Após a definição das medidas de dispersão, uma atividade, na qual dois conjuntos de dados apresentam a mesma variância, é proposta. A finalidade dessa atividade será avaliar a magnitude da variância em relação ao conjunto e definir o coeficiente de variação amostral, usado para avaliar essa magnitude.

Na seção seguinte, será proposta uma atividade para construir uma representação gráfica alternativa ao histograma conhecida como boxplot. O esquema dos cinco números, conjunto formado pelas medidas: mínimo, quartis e máximo é usado na construção do *boxplot*. Os

boxplots são gráficos simples e muito usados na comparação de diferentes conjuntos de dados. Na construção do boxplot, é apresentado um critério de classificação de um valor do conjunto como valor atípico em relação aos demais valores do conjunto de dados. Na conclusão desta seção, apresenta-se uma regra empírica para a determinação de frequência de observações nos intervalos centrados na média de comprimentos dados por dois desvios padrões e quatro desvios padrões para uma situação nas quais não existem valores atípicos ou estes são raros e, a forma da distribuição não é muito assimétrica.

Na seção **Para saber mais** são apresentadas fórmulas para o cálculo de algumas medidas trabalhadas no capítulo para dados agrupados e algumas demonstrações de resultados trabalhados no capítulo.

Diferencial do Capítulo

De acordo com Russel e Mokros (1991), citados em Batanero e Borovnik (2016), a compreensão da ideia de "valor representativo" implica em três competências diferentes:

- a) selecionar o melhor valor representativo para um dado conjunto de dados;
- b) construir um conjunto de dados tendo um determinado valor representativo, por exemplo, a moda;
- c) compreender o efeito que uma mudança em parte dos dados tem sobre os possíveis valores representativos.

Pretende-se explorar estas três competências nas atividades e exercícios do capítulo.

A fórmula de cálculo da variância é apresentada de forma detalhada para que o aluno compreenda o significado desta medida. No entanto, dada a sua complexidade, evitaremos seu uso direto, propondo o uso da tecnologia para obtê-la.

A definição de quartis e a construção do boxplot são propostas inovadoras em relação ao conteúdo usual de Estatística nos livros didáticos do Ensino Médio. Os conceitos relativamente simples de quartis aliados à grande utilidade do *boxplot* na comparação de grupos diferentes, reforçam a pertinência em tratá-los no Ensino Médio.

Dificuldades típicas dos estudantes (distratores)

Com base no texto de Batanero e Borovnik (2016), apesar da maior parte dos métodos de análise exploratória de

dados envolverem apenas cálculos e interpretações de medidas estatísticas simples, bem como, construções de gráficos e suas respectivas leituras, pesquisadores sugerem que os estudantes apresentam problemas na compreensão de conceitos, e em relacionar os mesmos ao contexto de forma significativa. Uma razão para isso é que, em geral, os professores focam sobre a aplicação de métodos em vez da interpretação de resultados em um dado contexto. Neste texto, um resumo de resultados de pesquisas realizadas neste tema é apresentado. A seguir, algumas destas dificuldades são destacadas.

- a) Cálculo de médias combinadas a partir das médias de diferentes grupos: desprezam-se os tamanhos dos diferentes grupos, calculando uma média simples das médias dos grupos.
- b) Cálculo de média para dados agrupados: ignoram-se as frequências, considerando apenas os pontos médios dos intervalos, somando-os e dividindo pelo número de intervalos, ou simplesmente, considerando apenas o valor da variável, quando a variável é quantitativa discreta.
- c) Compreensão das medidas de posição: média, mediana e moda.
- d) Compreensão das medidas de dispersão, em particular, da variância e do desvio padrão.
- e) Interpretação dos valores obtidos no contexto considerado.

Os dois primeiros itens estendem-se para o cálculo da variância e do desvio padrão.

Os distratores serão explorados nas atividades e nos exercícios.

Exemplos

Princípios norteadores dos exemplos selecionados:

- a) propor problemas cuja solução requer dados a serem coletados pelos alunos ou que de alguma forma estão disponíveis para consulta.
- b) contextualizar sempre os problemas propostos, pois o contexto é fundamental nas investigações e interpretações.

Estratégia pedagógica

Usar um processo reflexivo baseado no pensamento estatístico.

- a) Fórmulas e algoritmos para obter as medidas resumo, embora importantes neste capítulo, não serão valorizados.
- b) Dar importância à compreensão dos conceitos e à interpretação dos resultados.
- c) As atividades deverão estar sempre bem caracterizadas a um problema a ser resolvido em um contexto específico.
- d) Neste capítulo, o uso de recursos tecnológicos para a realização dos cálculos de medidas resumo é fundamental. Recomenda-se o GeoGebra e planilhas de cálculo.

O uso de calculadoras é fortemente recomendável. No entanto, cabe alertar o estudante quanto à notação adotada no Brasil para o separador decimal: a vírgula, e que é adotada neste livro. No entanto, a notação utilizada, em geral, nas calculadoras e na maioria dos programas e aplicativos é o ponto decimal. Por exemplo, o GeoGebra usa o ponto como separador de casas decimais e, em geral as planilhas eletrônicas estão formatadas para números, usando a vírgula como separador decimal. Assim, copiar e colar os dados de um aplicativo para outro pode acarretar em grandes variações. Cabe também ressaltar que é muito importante treinar o estudante a usar corretamente as calculadoras: não é raro, apesar de permitir o uso das mesmas, ocorrerem erros pelo manuseio incorreto da calculadora. O estudante deve ser alertado para a ordem de hierarquia das operações, muito trabalhada nos anos iniciais do segundo segmento do Ensino Fundamental, mas que é aparentemente esquecida quando vão utilizar calculadoras.

Cabe reforçar também que apesar do capítulo poder à primeira vista parecer pesado pelo excesso de definições e fórmulas, é possível verificar que as atividades são relativamente simples e visam muito mais à interpretação das medidas do que o cálculo das mesmas. Sempre serão fornecidas informações para facilitar cálculos quando estes forem solicitados. Nas avaliações, a não ser que seja permitido o uso de calculadora, recomenda-se fortemente não pedir para calcular variâncias e desvios padrões, estes deverão ser, em geral dados, e perguntas envolvendo o significado dos mesmos devem ser feitas.

Enfim o que deve ser valorizado no capítulo é conhecer o significado das medidas aqui apresentadas. O cálculo

das mesmas pode ser feito, usando-se aplicativos e, as fórmulas, se por acaso forem necessárias para resolver algum problema, deverão ser sempre fornecidas. Principalmente, em se tratando de medidas de dispersão.

Estrutura do Capítulo

Explorando 1 - Medidas de Posição

Nesta seção serão trabalhadas duas atividades. A primeira propõe duas transformações simples nos dados de um conjunto e procura avaliar o efeito destas transformações na distribuição dos dados. A segunda foca especificamente no cálculo de medidas de posição tais como média, mediana e moda, que já devem ser conhecidas do Ensino Fundamental. Também proporemos a divisão do conjunto de dados em quatro intervalos de classes de frequências iguais a $1/4$ para definir os três quartis de uma distribuição.

- a) Atividade: Distribuição de notas para perceber o efeito de transformações simples (multiplicação e/ou adição de um valor) no dado na posição e escala(forma) da distribuição, comparando histogramas.
- b) Atividade: Apresentação de diferentes conjuntos de dados sobre tempos para completar uma "maratona" que apresentam diferentes tipos de assimetria.

Organizando as ideias 1 - Medidas de posição

Definições de média; mediana; moda e quartis.

Praticando o assunto 1

Atividades explorando conceitos e propriedades apresentados no organizando as ideias 1.

Explorando 2 - Medidas de dispersão

Proposição de uma atividade envolvendo dois conjuntos de dados reais, todos com medidas de posição iguais, mas apresentando diferenças em suas distribuições caracterizando a necessidade da medida de dispersão.

Organizando as ideias 2 - Medidas de dispersão

Definições de amplitude; distância entre quartis; desvio-médio; variância; desvio padrão e coeficiente de variação.

Nesta seção também serão retomados os conceitos de parâmetro e estimador, tratados no capítulo A Natu-

reza da Estatística apresentando a definição de variância populacional e amostral, desvio-padrão populacional e amostral e, média populacional e amostral.

Proposição de uma atividade apresentando dois conjuntos de dados com a mesma variância, mas com medidas de posição diferentes para motivar a definição de coeficiente de variação.

Praticando o assunto 2

Atividades que usam os conceitos e propriedades apresentados no organizando 2 e que buscam dar significado às medidas de dispersão definidas.

Explorando 3

Proposição de atividade de construção de representação de dados usando Mínimo, Q_1 , Mediana, Q_3 e Máximo.

Organizando as ideias 3

- a) Definição do boxplot (gráfico caixa) representação gráfica para variáveis quantitativas alternativa ao histograma.
- b) Descrição do critério de classificação de um valor como valor atípico do conjunto de dados adotado na construção do boxplot.
- c) Apresentação de regra empírica para avaliar a frequência de dados nos intervalos $\bar{x} \pm s$ e $\bar{x} \pm 2 \cdot s$.

Praticando o assunto 3

Proposição de atividades de comparação de grupos, usando o boxplot.

Para saber mais

Nesta seção serão apresentadas

- a) fórmulas para o cálculo de medidas apresentadas no capítulo para dados agrupados;
- b) demonstrações de propriedades trabalhadas nas seções organizando as ideias;

Material Suplementar

Um applet do GeoGebra foi disponibilizado com manual de instruções nesta seção. Nele será possível gerar conjuntos de dados para os quais serão fornecidas as medidas resumo do conjunto bem como o histograma e o boxplot. Neste applet também será possível entrar com o seu próprio conjunto de dados para obter os gráficos e as medidas resumo.

Exercícios

Nesta seção são propostos exercícios do ENEM, Vestibulares entre outros, abordando os conteúdos desse capítulo. Nos exercícios serão tratados os distratores.

EXPLORANDO MEDIDAS DE POSIÇÃO

No capítulo **A Natureza Estatística** trabalhamos com representações gráficas de conjuntos de dados com a finalidade de obter informações sobre estruturas da sua distribuição como estratégia para resumir os dados. No exemplo dos registros de tempo deste capítulo, os 64 dados, no quadro a seguir

Tabela 1.2: Registros de tempo de atividade do capítulo **A Natureza Estatística**

A	B	C	D	E	F	G	H
3,03	4,37	5,04	5,73	4,03	5,37	6,04	6,74
3,38	4,46	5,11	5,84	4,38	5,46	6,11	6,84
3,60	4,55	5,19	5,95	4,60	5,55	6,19	6,96
3,78	4,63	5,29	6,08	4,78	5,64	6,29	7,08
3,92	4,71	5,36	6,23	4,92	5,72	6,36	7,23
4,04	4,79	5,45	6,41	5,04	5,79	6,45	7,40
4,16	4,87	5,54	6,62	5,16	5,87	6,54	7,63
4,27	4,95	5,64	6,97	5,26	5,95	6,64	7,97

foram organizados em 10 intervalos de classe, como mostra a tabela a seguir

Tabela 1.3: Registros de tempo agrupados em intervalos de classe

Intervalo de classe	Número de observações
[3,0 ; 3,5 [2
[3,5 ; 4,0 [3
[4,0 ; 4,5 [7
[4,5 ; 5,0 [9
[5,0 ; 5,5 [11
[5,5 ; 6,0 [11
[6,0 ; 6,5 [9
[6,5 ; 7,0 [7
[7,0 ; 7,5 [3
[7,5 ; 8,0 [2

que, por sua vez, foi usada para construir um gráfico, o histograma a seguir.

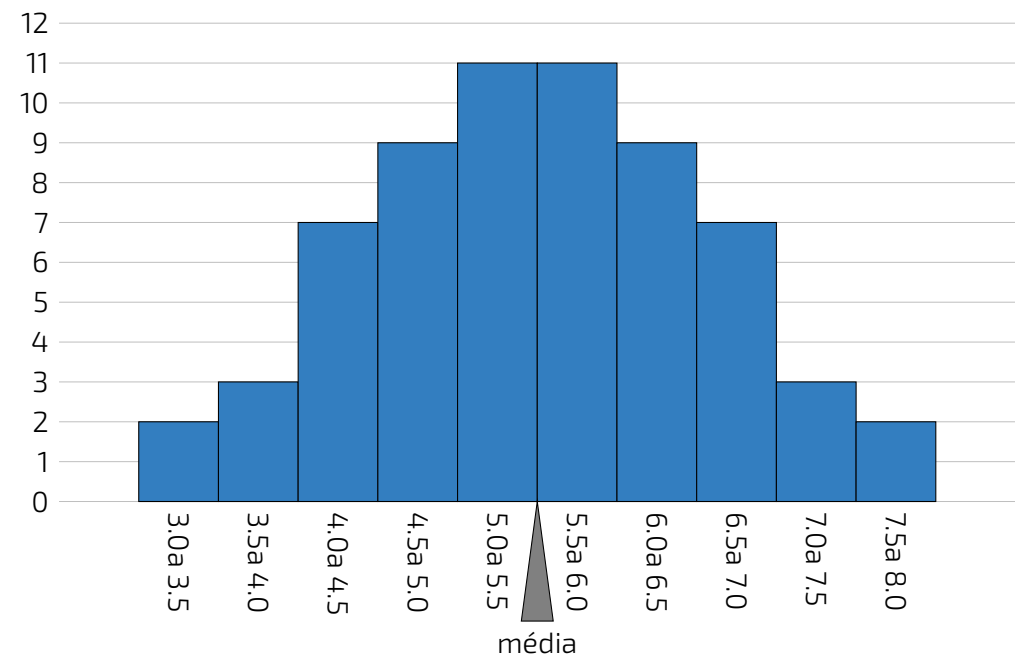


Figura 1.4: Histograma dos registros de tempo

Observe que os 64 registros de tempo foram resumidos numa representação gráfica que revela o comportamento destes dados: registros de tempo entre 3,0 e 8,0, estrutura simétrica em torno da média dos registros de tempo que é 5,5.

O capítulo **Medidas de Posição e Dispersão** tem como objetivo responder, entre outras, as seguintes perguntas sobre um conjunto de dados quantitativos.

- a) É possível encontrar valor(es) para resumir as observações? Qual(is) seria(m) este(s) valor(es)? Como encontrá-lo(s)?
- b) Como medir se os dados estão "próximos" ou "afastados" uns dos outros?
- c) Como você classifica a forma do gráfico construído para representar os dados?
- d) Existe algum valor muito diferente dos demais? Como identificá-lo?

Ao longo deste capítulo veremos como resumir a informação dos dados, usando apenas algumas medidas que caracterizam a distribuição em vez de usar toda a coleção de dados para descrevê-la. Por esta razão, tais medidas são chamadas medidas resumo.

Como as distribuições podem apresentar formas variadas é importante conhecer diferentes tipos de medidas resumo, tanto de posição como de dispersão, para usar medidas apropriadas em cada caso.

Notas de Artes

Atividade 1

Ao final de um trimestre, um professor de Artes registrou as seguintes notas de seus 35 alunos, listadas no quadro a seguir, em ordem crescente.

0,8	20	2,0	2,5	2,5	3,5	4,5
5,0	5,4	5,5	5,5	5,5	6,0	6,0
6,0	6,0	6,3	6,5	6,8	6,8	7,0
7,0	7,0	7,0	7,3	7,3	7,5	7,5
7,5	7,5	7,8	8,0	8,0	8,0	8,0

Este professor verificou que a média da turma foi aproximadamente 5,93 (soma das notas $S = 207,5$). Como a participação da turma foi muito boa ao longo do trimestre, o professor resolveu dar uma bonificação na nota de cada aluno desta turma, pensando em duas possibilidades:

- acrescentar um ponto para cada aluno da turma;
- aumentar em 20% a nota de cada aluno da turma.

A [tabela 1.4](#) contém os intervalos de classe considerados na construção do histograma das notas sem bonificação, ilustrado na [figura 1.5](#)

Tabela 1.4: Distribuição de frequências das notas antes de bonificação

Intervalo	Frequência absoluta
[0,2[1
[2,4[5
[4,6[6
[6,8]	23

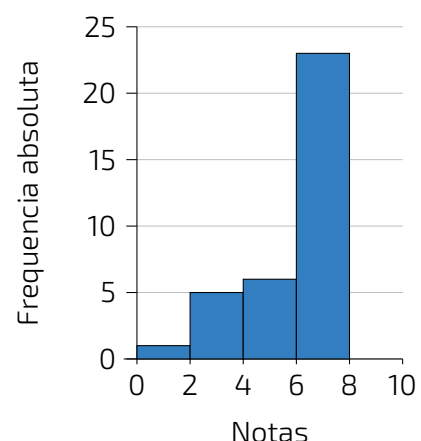


Figura 1.5: Histograma das notas de Artes sem bonificação

Os dois histogramas a seguir, na [figura 1.6](#) correspondem às notas, após usar cada uma das possibilidades consideradas pelo professor, mantendo quatro intervalos de classe, conforme as [tabelas 1.5 e 1.6](#).

Objetivos Específicos

Notas de arte

Estudar o efeito de uma transformação simples numa distribuição de dados: adição (posição) ou multiplicação (escala).

Sugestões e discussões

Notas de arte

Esta atividade tem como objetivo principal levar o aluno a perceber efeitos que certas transformações simples nos dados (adição e multiplicação) acarretam em uma distribuição de frequências e, consequentemente, levá-lo a avaliar possíveis mudanças nas medidas de posição e dispersão que serão tratadas neste capítulo. Como ela é uma atividade introdutória, essas propriedades não serão totalmente exploradas na atividade, mas ao longo da capítulo ela será retomada. Os dados desta atividade podem ser obtidos neste [link](#), e sugere-se o uso do GeoGebra ou uma planilha para realizar as transformações indicadas, embora não seja necessário para a realização da atividade. No item **e)** não há uma resposta certa, mas ele deverá ser explorado futuramente com o objetivo de avaliar os efeitos em uma distribuição quando somamos um valor constante a todos os dados e quando multiplicamos um valor constante a todos os dados.

Solução: Notas de arte

- a) Analisando o primeiro histograma apresentado com o original, percebe-se que o primeiro apresenta uma pequena alteração com intervalos de classe mais largos, ou seja de comprimento 2,4 (os comprimentos originais dos intervalos são iguais a 2). Já, o segundo, mantém intervalos de classe com mesmo comprimento aos do original, apresentando um deslocamento dos intervalos em uma unidade para à direita.
- b) Com o acréscimo de 1 ponto a cada nota, a nota maior que é 8,0 passa a ser 9,0; já com o aumento de 20% sobre a nota de cada um, a nota maior passa a ser 9,6. Portanto, analisando os dois histogramas dados, conclui-se que o primeiro corresponde ao aumento de 20% na nota de cada um e, o segundo, ao acréscimo de 1 ponto na nota de cada um.
- c) Observe que se todos os alunos tiverem o acréscimo de 1 ponto, a soma total das notas será acrescida de 35 pontos (pois são 35 alunos). Ao dividir o total por 35, percebe-se que a nova média será alterada exatamente pelo acréscimo de 1 ponto, passando a ser 6,93. Ou seja, a nova média é dada por $\frac{207,5 + 35}{35} \approx 5,93 + 1$. Já no caso do aumento de 20% sobre a nota de cada aluno, teremos que a nova soma total de notas será dada pela soma original acrescida de 20% tal que a média será dada por $\frac{S + 0,2 \cdot S}{35} = \frac{1,2 \cdot S}{35} = 1,2 \times \underbrace{\frac{S}{35}}_{\approx 5,9 \text{ média original}} = 1,2 \times 5,93 \approx 7,12$, em que $S = 207,5$.
- d) Não há uma resposta certa para este item. Se cada aluno olhar o seu ponto de vista particular, para alguns será melhor ganhar um ponto e para outros será melhor ter um aumento de 20% sobre a nota. Mais especificamente, para quem tiver obtido nota 5,0 será indiferente; para quem tiver obtido nota inferior a 5,0 será melhor ganhar um ponto e, para os restantes, será melhor o acréscimo de 20% sobre a nota.

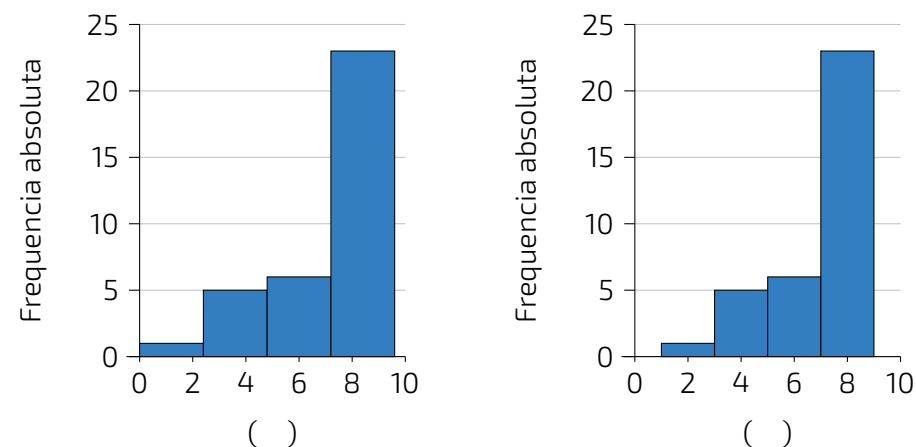


Figura 1.6: Histogramas das notas de Artes com bonificação

Tabela 1.5: Distribuição de frequências das notas após acréscimo de 1 ponto a cada nota

Intervalo	Frequência absoluta
[1; 3[1
[3; 5[5
[5; 7[6
[7; 9]	23

Tabela 1.6: Distribuição de frequências das notas após aumento de 20% sobre a nota

Intervalo	Frequência absoluta
[0; 2,4[1
[2,4; 4,8[5
[4,8; 7,2[6
[7,2; 9,6]	23

- a) Compare os histogramas das notas com bonificação com o histograma original. O que mudou em cada um deles em relação ao original?
- b) Considerando os [Histogramas das notas de Artes com bonificação](#), identifique qual deles corresponde ao acréscimo de 1,0 ponto, assinalando (a) e qual deles corresponde ao aumento de 20% das notas originais, assinalando (b).
- c) Dada a informação inicial de que a média da turma foi 5,93, de quanto será a média se o professor acrescentar um ponto a cada aluno? E se ele aumentar em 20% a nota de cada aluno?
- d) Se você fosse um aluno desta turma, que possibilidade de bonificação você escolheria? Para que notas é melhor cada uma das estratégias?

A maratona

Atividade 2

A maratona é uma prova de atletismo que consiste em correr uma distância de 42,195km. Pelas suas características, este tipo de prova é realizada nas ruas de uma grande cidade ou na estrada. As principais cidades do mundo realizam um destes eventos anualmente, recebendo milhares de atletas profissionais e amadores que encaram o desafio e almejam finalizar a corrida ou melhorar o próprio tempo do passado.

Uma das mais famosas é a Maratona da Cidade de Nova Iorque, nos Estados Unidos. Veja na figura [Corredores participando da Maratona de Nova York, Wikipedia](#) realização de uma maratona em Nova Iorque. Com mais de 50.000 participantes cada ano, é um dos principais eventos do atletismo mundial, junto com as maratonas de Chicago, Londres, Boston, Berlim e Tóquio.



Figura 1.7: Corredores participando da Maratona de Nova York, [Wikipedia](#)

Os resultados do evento são divididos nas categorias de homens e mulheres, além disso, no evento participam cadeirantes e pessoas usando triciclos de mão (*handcycle*), categorias cujos resultados são premiados e publicados separadamente. Qual das categorias você acha que terá os melhores resultados na maratona? Em quanto tempo você acha que uma pessoa percorre os 42,195 km? O que você acha ser mais rápido: correr em cadeira de rodas ou em triciclo de mão?



A seguir analisaremos os tempos de corrida das 100 melhores atletas na categoria de Mulhe-

Objetivos Específicos

A maratona

- Identificar a posição da média e dos quartis no gráfico da distribuição.
- Apresentar representação gráfica alternativa ao histograma: o *box-plot* sem valores discrepantes.

Sugestões e discussões

A maratona

Nesta atividade serão apresentados conjuntos diferentes de dados envolvendo tempos para completar maratonas. Os dados estão disponíveis neste [link](#). Serão fornecidos os totais para que o cálculo das médias envolva apenas uma divisão e possa ser feito com uma calculadora simples. Pretende-se levar o aluno a perceber que na presença de forte assimetria (histograma alongado à direita ou à esquerda), a média pode não ser a medida mais adequada para representar o conjunto e, com isso, motivar a definição de mediana.

É importante discutir as perguntas na caixa Para refletir em sala de aula com o intuito de que os estudantes percebam a necessidade de tratar previamente dados de determinada natureza antes de usá-los numericamente, como é o caso do tempo considerado em unidades distintas (hora:minuto:segundo).

Na sequência se inclui a tabela com a respectiva conversão para minutos em números decimais de modo a simplificar os cálculos na atividade, mas deve-se deduzir com os estudantes como calcular a conversão.

Expressão utilizada para calcular o resultado em minutos decimais (minutos_{10}):

$$\text{minutos}_{10} = \text{Horas} \cdot 60 + \text{Minutos} + \frac{\text{Segundos}}{60}$$

É importante comentar com os estudantes a diferença observada entre a média e a mediana e que esta se deve a uma forte assimetria na distribuição dos dados, representada pela forma de um histograma alongado para à esquerda com frequências pequenas, tornando a média inferior à mediana.

Sugere-se como atividade interdisciplinar, a realização de corridas com os estudantes na Educação Física, medindo os tempos totais, calculando a velocidade média do percurso e comparando com a velocidade média do primeiro e centésimo lugares da maratona. Recomenda-se explicitar o vínculo com a Física para o cálculo e a interpretação da velocidade média, assim como a colocação de questões críticas que facilitem a interpretação dos resultados, por exemplo, é possível manter essa velocidade média durante quase 3 horas (tempo médio da maratona de mulheres)?

res da Maratona de Nova York do ano 2017, dados disponíveis no [site oficial da competição](#).

Observe no quadro a seguir, que os tempos já estão ordenados do menor para o maior e que para identificar o tempo da quadragésima sétima chegada, basta tomar a interseção da linha 7 com a coluna +40 para obter o tempo 2 : 55 : 36, ou seja, duas horas 55 minutos e 36 segundos.

Tabela 1.7: 100 melhores tempos de finalização da Maratona de Nova Iorque 2017 para mulheres (hora:minuto:segundo)

	+0	+10	+20	+30	+40	+50	+60	+70	+80	+90
1	2:26:53	2:32:01	2:42:52	2:49:44	2:53:59	2:56:58	2:58:35	2:59:36	3:01:24	3:03:43
2	2:27:54	2:32:09	2:44:26	2:49:59	2:54:42	2:57:05	2:58:36	2:59:41	3:01:26	3:03:46
3	2:28:08	2:33:18	2:44:48	2:50:04	2:54:52	2:57:10	2:58:50	2:59:43	3:01:28	3:04:02
4	2:29:36	2:34:10	2:45:20	2:50:05	2:55:04	2:57:40	2:58:52	2:59:46	3:01:44	3:04:04
5	2:29:39	2:34:23	2:45:52	2:51:11	2:55:25	2:57:49	2:58:56	2:59:51	3:02:09	3:04:17
6	2:29:39	2:36:38	2:46:45	2:53:01	2:55:34	2:57:49	2:59:01	2:59:56	3:02:15	3:04:26
7	2:29:41	2:37:22	2:47:04	2:53:02	2:55:36	2:57:50	2:59:03	3:00:02	3:02:39	3:04:42
8	2:29:56	2:37:33	2:47:30	2:53:02	2:55:39	2:58:08	2:59:10	3:00:05	3:02:41	3:04:49
9	2:31:21	2:39:01	2:47:35	2:53:19	2:56:47	2:58:23	2:59:16	3:00:49	3:02:56	3:04:58
10	2:31:44	2:40:09	2:49:37	2:53:38	2:56:57	2:58:26	2:59:23	3:01:18	3:03:32	3:05:09

PARA REFLETIR

- Como você calcularia a média de valores em horas, minutos e segundos, como os da tabela?
- Como você construiria um histograma com estes dados? Como você definiria os limites dos intervalos? (Consulte a atividade [Construção do Histograma](#) do capítulo [A Natureza Estatística](#) em caso de dúvida.)
- Qual o maior tempo em que uma corredora deveria completar a maratona para ficar entre as 25 primeiras? E entre as 50 primeiras?

Para calcular a média destes dados é conveniente reduzi-los a uma única unidade de medida, pois, caso contrário, seria necessário calcular três médias e, ainda fazer conversões apropriadas para obter a resposta em hora:minuto:segundo. Convertendo todos os tempos para minutos, obtemos o seguinte quadro de tempos arredondados para duas casas decimais.

	+0	+10	+20	+30	+40	+50	+60	+70	+80	+90
1	146,88	152,02	162,87	169,73	173,98	176,97	178,58	179,60	181,40	183,72
2	147,90	152,15	164,43	169,98	174,70	177,08	178,60	179,68	181,43	183,77
3	148,13	153,30	164,80	170,07	174,87	177,17	178,83	179,72	181,47	184,03
4	149,60	154,17	165,33	170,08	175,07	177,67	178,87	179,77	181,73	184,07
5	149,65	154,38	165,87	171,18	175,42	177,82	178,93	179,85	182,15	184,28
6	149,65	156,63	166,75	173,02	175,57	177,82	179,02	179,93	182,25	184,43
7	149,68	157,37	167,07	173,03	175,60	177,83	179,05	180,03	182,65	184,70
8	149,93	157,55	167,50	173,03	175,65	178,13	179,17	180,08	182,68	184,82
9	151,35	159,02	167,58	173,32	176,78	178,38	179,27	180,82	182,93	184,97
10	151,73	160,15	169,62	173,63	176,95	178,43	179,38	181,30	183,53	185,15

- a) Construa um histograma dos dados convertidos para horas, completando a [tabela 1.8](#), que indica os intervalos de classe (fechados no limite inferior e abertos no limite superior).

Intervalo	Frequência
[146,0; 150,0[
[150,0; 154,0[
[154,0; 158,0[
[158,0; 162,0[
[162,0; 166,0[
[166,0; 170,0[
[170,0; 174,0[
[174,0; 178,0[
[178,0; 182,0[
[182,0; 186,0[

Tabela 1.8: Intervalos de classes

- b) Calcule o tempo médio dos 100 melhores tempos das corredoras, sabendo que a soma dos tempos é 17.191,66 minutos. Localize o valor encontrado no eixo horizontal do histograma. Em que posição ficaria uma corredora cujo tempo no qual completou a maratona é igual ao tempo médio calculado neste item?
- c) Trace linhas verticais no histograma de modo a separar as classificações em 4 grupos: uma linha vertical que identifica o 25° lugar, separando os 25 primeiros colocados dos demais; outra, que identifica a 50ª classificação e, por fim, uma que marca o 75° tempo na classificação geral.

As marcações dos tempos das 25ª, 50ª e 75ª posições neste conjunto de 100 observações são chamadas de quartis da distribuição, este conceito será formalizado adiante.

Solução: A maratona

- a) A tabela com as frequências por intervalo e o histograma ficam de seguinte forma:

Intervalo	Frequência
[146,0; 150,0[8
[150,0; 154,0[5
[154,0; 158,0[5
[158,0; 162,0[2
[162,0; 166,0[5
[166,0; 170,0[7
[170,0; 174,0[9
[174,0; 178,0[16
[178,0; 182,0[27
[182,0; 186,0[16

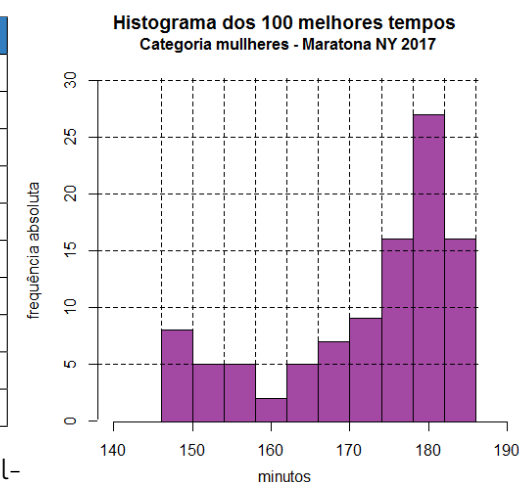


Tabela 1.1: Guia para o cálculo de frequências do histograma

Figura 1.1: Histograma dos tempos da categoria de mulheres na Maratona de NY

- b) O tempo médio das primeiras 100 corredoras é de aproximadamente 171,92 minutos. Uma corredora com esse tempo teria ficado na entre a 35ª e 36ª posição.
- c) Para ficar entre os primeiros 25 lugares, uma corredora teria que terminar a corrida em até 165,87 minutos. Já para ficar nas primeiras 50, precisaria terminar o percurso em 176,95 minutos ou menos. Finalmente, para ficar entre as primeiras 75, seu tempo teria que ser menor ou igual a 179,85 minutos.

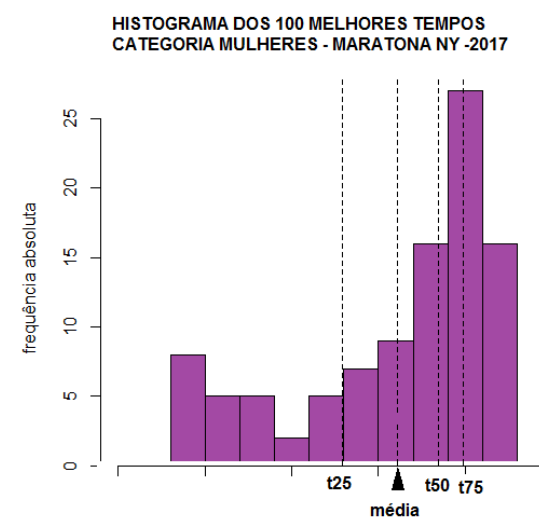


Figura 1.2: Histograma dos tempos da categoria de mulheres na Maratona de NY mostrando os quartis, a mediana e a média

Solução: A maratona

d) Os comprimentos dos intervalos são dados por:

Intervalo	Comprimento
1° a 25°	18,99
25° a 50°	11,08
50° a 75°	2,90
75° a 100°	5,30

Lembre-se que se o histograma for construído considerando os intervalos acima, deve-se trabalhar com a escala de densidade de frequência (absoluta ou relativa: razão da frequência pelo comprimento do intervalo), pois os comprimentos dos intervalos são diferentes. No histograma construído nesta atividade, usou-se a escala da frequência absoluta, pois os intervalos considerados têm comprimentos iguais a 4.

- e) Não coincide com nenhuma delas (25°, 50°, e 75°)
- f) Tem-se que o tempo médio foi 171,92 minutos e o tempo da posição 50 foi 176,95 minutos e, portanto, são diferentes. Adiante vamos trabalhar a razão desta diferença neste conjunto. Isto se deve à forma da distribuição dos tempos de chegada ilustrada pelo histograma. Observando o histograma com as marcações, verifique que a média está em um intervalo de frequência não muito alta (9 tempos, com 32 tempos nos intervalos anteriores e 59 tempos nos intervalos posteriores), enquanto a o tempo da 50ª posição, mais à direita estão em um intervalo de frequência mais alta (16 tempos, com 41 tempos nos intervalos anteriores e 43 tempos nos intervalos posteriores). Neste caso, o tempo da 50ª posição representa melhor o centro desse conjunto.
- g) Percebe-se uma estrutura com assimetria à esquerda, isto é, frequências baixas (menores ou iguais a 8) nos tempos iniciais (7 primeiros intervalos) e grande concentração (frequências altas - maiores ou iguais a 16) nos três intervalos finais. Esse tipo de estrutura resulta na média inferior ao tempo localizado na posição do meio (50 nesse exemplo).
- h) A figura 1.3 construída é chamada boxplot sem a sinalização de valores discrepantes (vamos chamar de boxplot simplificado). A construção do boxplot com sinalização de valores discrepantes será trabalhada na seção de medidas de dispersão. Veja como fica o boxplot simplificado, adotando-se orientação horizontal.

Nota 1

- d) Calcule os comprimentos dos intervalos de tempo determinados pela proposta de divisão no item d) e compare-os.

Intervalo	Comprimento
1° a 25°	
25° a 50°	
50° a 75°	
75° a 100°	

Observa-se que os comprimentos dos intervalos são bem diferentes, sendo maiores no início e mais estreitos no final.

Observe que apesar das diferenças de comprimento desses intervalos, cada um deles corresponde a cerca de 25 tempos.

- e) O valor obtido para o tempo médio coincide com alguma das outras marcas feitas no histograma?
- f) Observe que o tempo médio dos 100 melhores tempos para mulheres e o tempo da corredora que chegou em 50º. lugar são diferentes. Qual deles você escolheria como medida resumo destes dados? Por quê?
- g) Que características da distribuição dos 100 melhores tempos para mulheres podem ser destacadas, analisando-se o histograma construído?

(Construção de boxplot simplificado) Considerando os valores mínimo (146,88 min.) e máximo (185,15 min.), trace uma reta (vertical ou horizontal), incluindo esse intervalo de variação. Por exemplo, de 146 a 186.

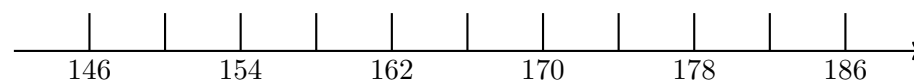


Figura 1.8: Sugestão de escala para a construção do boxplot dos 100 melhores tempos (em minutos) para a categoria mulheres

Em seguida, trace um retângulo, cujas bases estão nas posições referentes aos 25º e 75º tempos de chegada, cortando o retângulo por um segmento na posição referente ao 50º tempo de chegada. Para terminar a construção, trace um segmento, partindo do ponto médio da parte inferior até o valor mínimo e repita para o ponto médio da parte superior até o valor máximo.

A figura obtida é conhecida como boxplot (gráfico-caixa) do 100 melhores tempos sem a sinalização de valores discrepantes.

O boxplot é uma representação gráfica de dados quantitativos alternativa ao histograma. Esse gráfico é útil na comparação do comportamento de uma variável, considerando diferentes grupos, por exemplo, 100 melhores tempos de chegada entre homens e mulheres.

ORGANIZANDO MEDIDAS DE POSIÇÃO

Medidas de Posição, como o próprio termo indica, visam a resumir um conjunto de dados em geral numa única medida em algum lugar geométrico entre os extremos observados do conjunto (mínimo e máximo). Veja na [figura 1.9](#), as marcações da média e da mediana das notas de Artes sem bonificação.

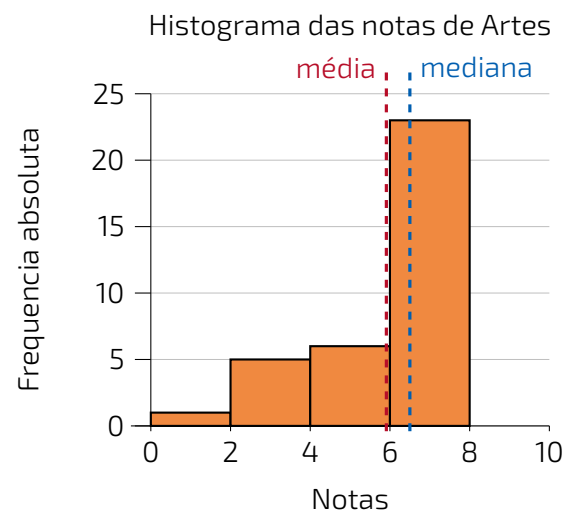


Figura 1.9: Média e mediana assinaladas no Histograma de das notas de Artes

Só é possível obter medidas como a média e a mediana, se nossas observações são de natureza quantitativa, pois, como vimos no capítulo [A Natureza Estatística](#), as variáveis qualitativas estão no domínio da frequência apenas, ou seja, só podemos contar quantas observações ocorrem em cada categoria da variável qualitativa, mas não podemos operar matematicamente com as categorias em si. Por exemplo, na atividade Prática de Atividades Físicas deste capítulo, trabalhamos com a variável modalidade do esporte praticado. As modalidades correspondem à "Futebol", "Caminhada", "Fitness", etc. Observe que são respostas não numéricas e, por isso, não podemos calcular uma média e não existe uma relação de ordem natural das respostas. Apenas podemos ordenar as respostas pela frequência na qual elas ocorreram.

As principais medidas de posição usadas na Estatística são a média, a mediana, a moda e os quartis da distribuição. Lembre-se que essas medidas buscam de alguma forma resumir a informação do conjunto.

Para definir as medidas a serem estudadas neste capítulo vamos adotar a notação descrita no exemplo a seguir.

EXEMPLO 1 Idade de pessoas que tomaram a vacina da febre amarela

Suponha que na primeira segunda-feira do mês de março de 2018, um Posto de Saúde tenha registrado as idades (em anos completos) das seis primeiras pessoas que chegaram para tomar a vacina da febre amarela e, os registros, obtidos foram $\{55, 22, 30, 14, 25, 40\}$, nessa ordem.

Neste exemplo dizemos que o número de observações, denotado por n , é 6 e que as observações são dadas por $x_1 = 55$, $x_2 = 22$, $x_3 = 30$, $x_4 = 14$, $x_5 = 25$ e $x_6 = 40$.

De um modo geral, sejam x_1, x_2, \dots, x_n , os n valores observados de uma variável quantitativa tal que

x_1 é o primeiro valor observado; x_2 é o segundo valor observado; e, assim por diante, tal que x_n é o último valor observado.

Os valores observados não ocorrem necessariamente de forma ordenada do menor para o maior. Neste exemplo, das idades das três primeiras pessoas que chegaram para tomar a vacina no Posto de Saúde foram $x_1 = 55$, $x_2 = 22$ e $x_3 = 30$ de modo que $x_1 > x_2$ e $x_2 < x_3$.

Para definir a mediana, será útil usar uma notação para representar os dados ordenados.

Sejam $x_{(1)}$ o menor valor do conjunto $\{x_1, x_2, \dots, x_n\}$; $x_{(2)}$, o segundo menor valor do conjunto $\{x_1, x_2, \dots, x_n\}$; e assim sucessivamente até $x_{(n)}$, o maior valor do conjunto $\{x_1, x_2, \dots, x_n\}$.

Desse modo, $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ são os valores ordenados do conjunto $\{x_1, x_2, \dots, x_n\}$.

No exemplo das idades das seis primeiras pessoas que chegaram para tomar a vacina no Posto de Saúde, os registros obtidos foram 55, 22, 30, 14, 25, 40 tal que

$$x_1 = 55, x_2 = 22, x_3 = 30, x_4 = 14, x_5 = 25 \text{ e } x_6 = 40$$

e

$$x_{(1)} = 14, x_{(2)} = 22, x_{(3)} = 25, x_{(4)} = 30, x_{(5)} = 40 \text{ e } x_{(6)} = 55$$

A letra maiúscula sigma (Σ) é usada para denotar somatório, simplificando algumas fórmulas. Por exemplo,

$$\sum_{i=1}^n x_i = x_1 + x_2 + \dots + x_n$$

$$\sum_{i=1}^n x_i^2 = x_1^2 + x_2^2 + \dots + x_n^2$$

Observe que neste exemplo, das idades das seis primeiras pessoas que chegaram para tomar a vacina no Posto de Saúde,

$$\begin{aligned} \sum_{i=1}^6 x_i &= x_1 + x_2 + x_3 + x_4 + x_5 + x_6 = \\ &55 + 22 + 30 + 14 + 25 + 40 = 186 \end{aligned}$$

e

$$\begin{aligned} \sum_{i=1}^6 x_i^2 &= x_1^2 + x_2^2 + x_3^2 + x_4^2 + x_5^2 + x_6^2 = \\ &55^2 + 22^2 + 30^2 + 14^2 + 25^2 + 40^2 = 6.830 \end{aligned}$$

Média

A definição de média de um conjunto de dados quantitativos já é conhecida desde o Ensino Fundamental e, consiste na soma dos valores do conjunto dividida pelo número de observações. No exemplo das idades das seis primeiras pessoas que chegaram para tomar a vacina no Posto de Saúde, a soma das idades é 186 tal que a média será dada por $\frac{186}{6} = 31$ anos.

De modo mais geral, considere um conjunto contendo n valores de uma variável quantitativa representado por $\{x_1, x_2, \dots, x_n\}$. A média deste conjunto, denotada por \bar{x} , é definida por

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} = \frac{x_1 + x_2 + \dots + x_n}{n}$$

Observe que a média pode substituir todas as observações sem alterar a soma dos valores, isto é,

$$x_1 + x_2 + \dots + x_n = \bar{x} + \bar{x} + \dots + \bar{x} = n \cdot \bar{x}$$

forneendo a expressão que define a média, denotada por \bar{x} .

Esta é justamente a ideia por trás da definição de qualquer média: uma medida que de alguma forma representa o conjunto de dados, segundo uma formulação, e se situa entre os extremos das observações. É claro que, em geral, haverá valores diferentes no conjunto e, neste caso, a média será um valor pertencente ao intervalo de variação dos valores neste conjunto e não necessariamente, um valor que tenha sido observado.

No exemplo das idades das seis primeiras pessoas que chegaram para tomar a vacina no Posto de Saúde a média é 31 anos, porém não se observou uma idade igual a 31 anos.

Você já calculou a média dos dados das duas primeiras atividades, a saber, [Notas de Arte](#) e [A Maratona](#). Identifique nos histogramas correspondentes a posição em que estas médias ficaram.

Média para dados agrupados

Quando os dados disponíveis estão agrupados em intervalos de classe, não é possível calcular a soma total exata dos dados. Neste caso, usamos uma aproximação para o cálculo da média como mostra o exemplo a seguir.

Suponha que um coordenador tenha tido acesso apenas ao [Histograma das notas de Artes sem bonificação](#), sem conhecer as notas separadamente. Como este coordenador poderia calcular a média da turma, considerando as notas antes da bonificação?

Temos a seguinte distribuição de frequências das notas antes da bonificação:

Tabela 1.9: Distribuição de freqüências das notas antes da bonificação

Intervalo	Frequência absoluta	Ponto médio do intervalo
[0, 2[1	1,0
[2, 4[5	3,0
[4, 6[6	5,0
[6, 8]	23	7,0

Apenas sabemos que, por exemplo, entre 2 e 4 existem cinco notas, mas não conhecemos o valor exato de cada uma destas cinco notas. Portanto, a soma exata destas cinco notas não é

conhecida. A estratégia é tomar o ponto médio desta classe $\left(\frac{2+4}{2}\right) = 3$ como a nota representativa das cinco observações, pois espera-se que os erros cometidos para mais e para menos sejam compensados na classe. Desse modo estimamos a soma das notas neste intervalo como $3 + 3 + 3 + 3 + 3 = 5 \cdot 3 = 15$.

Esse procedimento é adotado para todas as classes a fim de obter uma estimativa da soma total dos dados, a saber,

$$1 \cdot 1 + 5 \cdot 3 + 6 \cdot 5 + 23 \cdot 7 = 207$$

Logo, a média correspondente a este agrupamento, a ser considerada pelo coordenador é estimada por

$$\text{média} = \bar{x} = \frac{1 \times 1 + 5 \times 3 + 6 \times 5 + 23 \times 7}{35} = \frac{207}{35} \approx 5,91$$

Observe que este agrupamento resultou numa soma 207, muito próxima da soma exata dada por 207,5. Por esta razão dizemos que o agrupamento não incorreu em grande perda de informação para efeito de calcular a soma dos dados: em vez de usar as 35 notas, foi possível com quatro intervalos de classe avaliar de forma precisa a soma original dos dados. Consequentemente, a média estimada por este agrupamento (5,91) não se diferencia muito da média considerando os dados brutos (dados não agrupados) dada por (5,93).

Na seção [Para saber mais](#) apresenta-se notação e fórmula para o cálculo da média numa situação genérica de dados agrupados.

Interpretação da média como ponto de equilíbrio no histograma

Observe o [Histograma das notas de Artes sem bonificação](#), em que as notas dispostas ao longo do eixo horizontal. Suponha que o histograma seja mais do que uma representação da distribuição de frequências, que seja um objeto. Assim, cada ponto que compõe as notas teria massa e poderia ser associado a um peso. Por exemplo, a nota 1 corresponderia a 1kg, a nota 5 a 5kg e a nota 6,3 a 6,3 Kg. esse caso, podemos perguntar onde se encontrará o ponto de equilíbrio (ou centro de massa) do histograma que representa a distribuição de frequências dos dados. É natural pensar na média como o ponto de equilíbrio, como mostra o histograma na [figura 1.10](#), com destaque para a média. Veja adiante a seção sobre desvios da média para reforçar esta noção de ponto de equilíbrio.

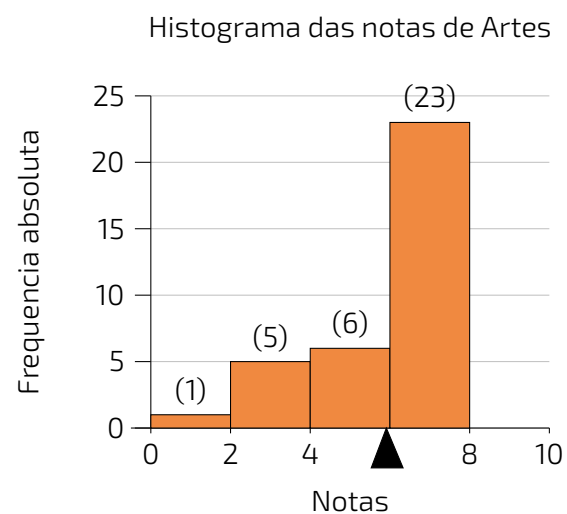


Figura 1.10: Histograma com destaque para a média como ponto de equilíbrio

Se fossemos tentar equilibrar o histograma num ponto acima da média, considerando esta interpretação, o mesmo penderia para à esquerda, conforme ilustra a [figura 1.11](#).

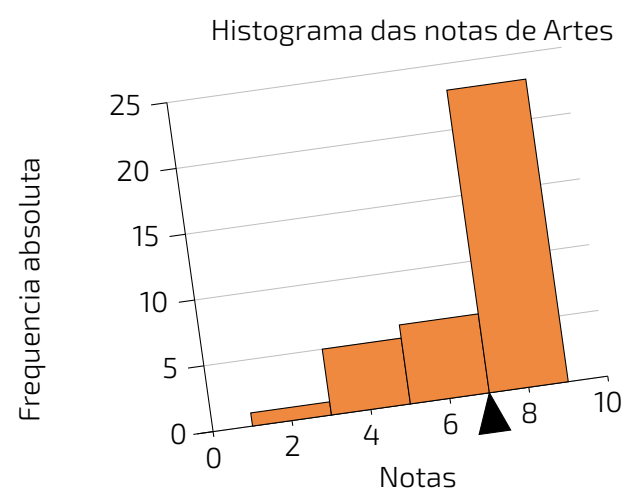


Figura 1.11: Histograma inclinado para à esquerda

Se fossemos tentar equilibrar o histograma num ponto abaixo da média, considerando esta interpretação, o mesmo penderia para à direita, conforme ilustra a [figura 1.12](#).

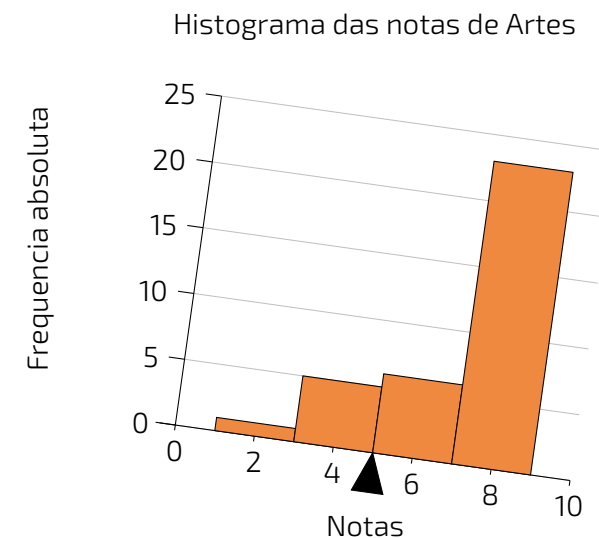


Figura 1.12: Histograma inclinado para à direita

Cuidado com esta interpretação: o ponto de equilíbrio corresponde à posição para a qual a soma dos valores, interpretada como peso, é a mesma à esquerda e à direita dela. Esta posição, correspondendo à posição da média, não é necessariamente a posição na qual a área total do histograma é dividida em duas metades (mediana). É claro que, se a forma do histograma for simétrica, estas duas posições serão coincidentes. Veja na [figura 1.13](#), um histograma em uma situação de simetria na qual a média e a mediana coincidem.

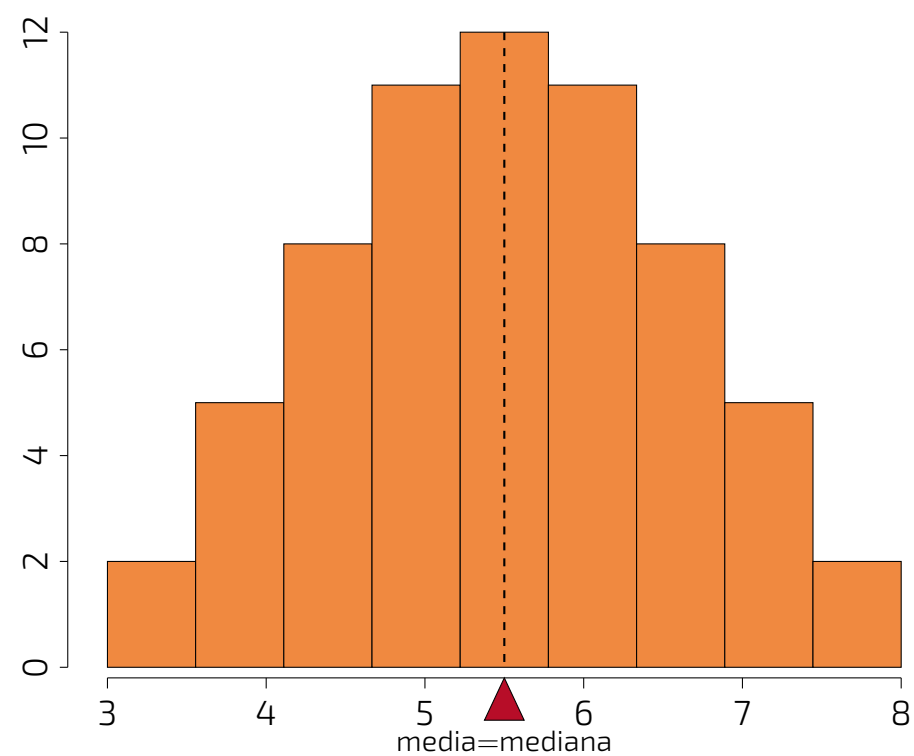


Figura 1.13: Histograma dos registros de tempo de atividade do Capítulo [A Natureza Estatística](#)

EXEMPLO 2 O cartão de crédito de supermercado

Numa tarde, 10 clientes interessados em obter um cartão de crédito oferecido por uma rede de supermercados informaram a uma atendente seus salários (em salários mínimos): $\{1, 1, 2, 3, 4, 5, 5, 6, 9, 10\}$.

A média destes dados é, então, $\bar{x} = \frac{46}{10} = 4,6$, que representa bem este conjunto, pois nele existem cinco valores acima da média e cinco valores abaixo da média e, estes valores não estão muito afastados do valor da média, conforme ilustrado no Diagrama de Pontos na figura 1.14.

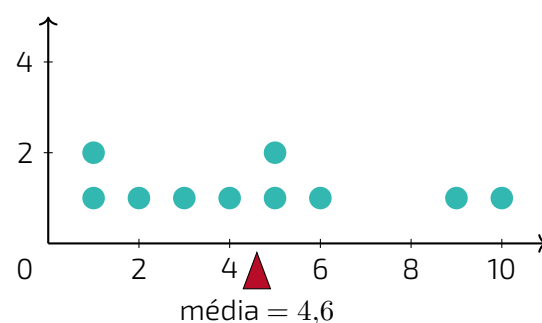


Figura 1.14: Diagrama de pontos do conjunto $\{1, 1, 2, 3, 4, 5, 5, 6, 9, 10\}$ com destaque para a média do conjunto

Suponha uma pequena variação do conjunto de dez salários na qual o salário igual a 10 salários mínimos foi substituído por um igual a 100 salários mínimos. Assim, os dados agora são: $\{1, 1, 2, 3, 4, 5, 5, 6, 9, 100\}$.

Há apenas uma diferença entre os dois conjuntos no valor máximo: no primeiro é 10 e no segundo é 100. O que esta única diferença nos dois conjuntos acarreta na média?

Com os dados do segundo conjunto, a média é dada por $\frac{1 + 2 + \dots + 100}{10} = \frac{136}{10} = 13,6$, valor maior do que a maioria dos dados observados no conjunto, a saber, apenas uma observação é bem superior a 13,6. Observe, que para representar o diagrama de pontos destes dados (figura 1.15), usou-se um recurso de quebra do eixo dos dados devido ao valor atípico 100, em relação aos demais valores.

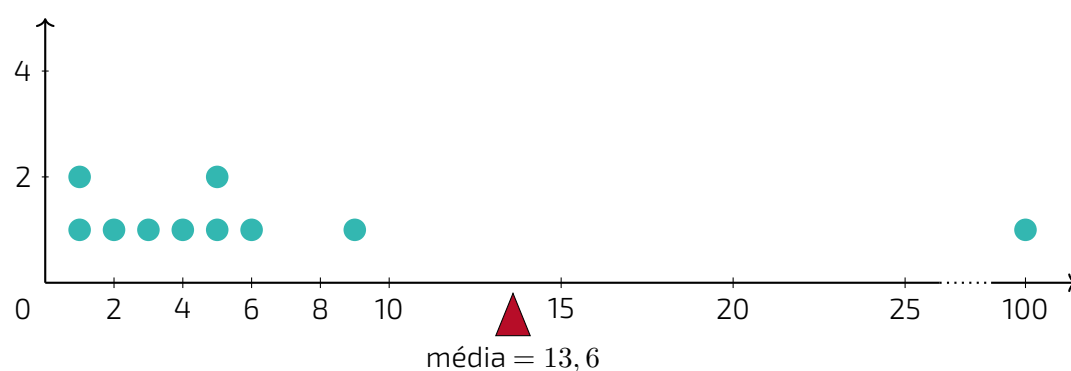


Figura 1.15: Diagrama de pontos do conjunto $\{1, 1, 2, 3, 4, 5, 5, 6, 9, 100\}$ com destaque para a média do conjunto e quebra do eixo devido ao valor atípico

Este exemplo simples mostra que na presença de dados atipicamente altos, deve-se tomar cuidado em escolher a média como medida de posição das observações coletadas. Uma medida pouco afetada para valores atípicos, conhecida como medida robusta, deverá ser considerada em situações deste tipo. A mediana, que trataremos a seguir, é considerada uma medida robusta.

Do exemplo *cartão de crédito de supermercado* podemos concluir que deve-se ter cautela em resumir os dados com a média quando sua distribuição, representada pelo histograma, apresenta forma muito assimétrica, como mostram as figuras 1.16 e 1.16.

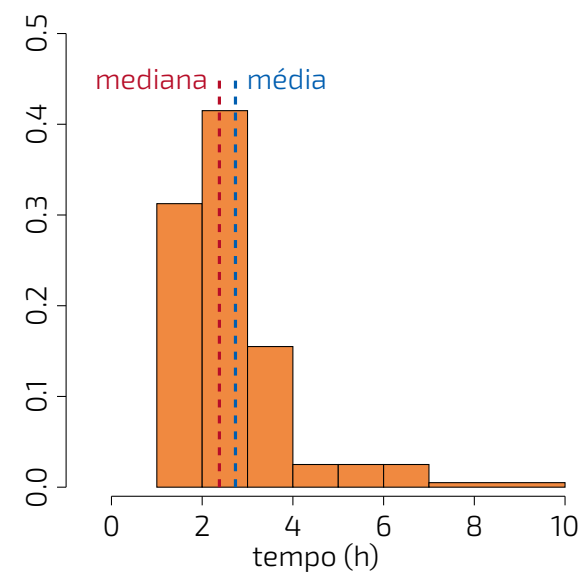


Figura 1.16: Histograma da distribuição dos tempos de chegada (em horas) na categoria triciclo de mão revelando assimetria à direita (mediana < média)

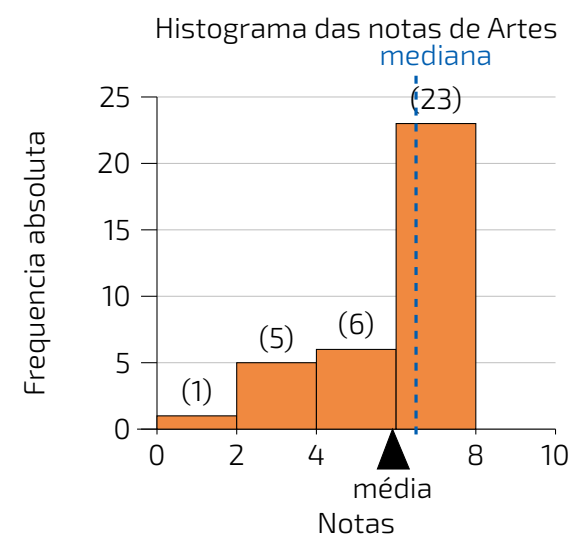


Figura 1.17: Histograma das notas de Artes sem bonificação (distribuição com assimetria à esquerda)

Alguns textos usam os termos assimetria positiva para indicar assimetria à direita e assimetria negativa para indicar assimetria à esquerda.

Mediana

A mediana de um conjunto de valores numéricos é definida como o valor que ocupa a posição central dos dados ordenados.

Se o conjunto de dados tem uma quantidade ímpar de elementos então, considerando os dados ordenados, a mediana ocupará a posição central. Por exemplo, se o conjunto de dados tiver $n = 9$ elementos, a posição central será a quinta. Nesse caso, haverá, ordenadamente, quatro elementos anteriores e quatro posteriores à mediana.

EXEMPLO 3 Idades de crianças atendidas em Posto de Saúde

Considere o seguinte registro das idades de crianças atendidas (na ordem de atendimento) em um ambulatório pediátrico de um Posto de Saúde na primeira segunda-feira do mês de março no turno da manhã: $\{4, 6, 9, 3, 2, 3, 7, 8, 7\}$.

Temos ao todo 9 observações cujos valores ordenados são:

$$2 \leq 3 \leq 3 \leq 4 \leq \overbrace{6}^{\text{valor da quinta posição}} \leq 7 \leq 7 \leq 8 \leq 9$$

mediana

Se o conjunto de dados tem uma quantidade par de elementos não será possível identificar “um” elemento central. Nesse caso, para a determinação da mediana serão considerados os dois elementos centrais da sequência ordenada. A mediana é dada pela média aritmética desses elementos. Por exemplo, se o conjunto de dados tiver 10 elementos, então as posições centrais são a 5ª. e a 6ª. A mediana será a média dos elementos que ocupam essas posições na sequência ordenada.

EXEMPLO 4 Cartão de crédito de supermercado (2)

Considere o conjunto de salários de 10 clientes interessados em obter um cartão de crédito oferecido por uma rede de supermercados e que informaram à atendente seus salários (em salários mínimos):

$$\{1, 1, 2, 3, \overbrace{4}^{5^{\text{a. posição}}}, \underbrace{5}_{6^{\text{a. posição}}}, 5, 6, 9, 100\}$$

Observe que os valores já estão ordenados e que o salário da 5ª. posição é 4 e, o da 6ª., é 5. Logo, a mediana dos salários será dada por

$$\frac{4 + 5}{2} = 4,5$$

Lembre que a média destes dados resultou em 13,6. Este exemplo ilustra a propriedade de que a mediana é pouco afetada na presença de valores atipicamente grandes (ou pequenos). Já a média não possui esta propriedade, sendo muito afetada na presença de valores atípicos.

De maneira geral, se $x_{(1)}, x_{(2)}, \dots, x_{(n)}$ são os valores ordenados do conjunto de dados, a mediana será dada por

$$\text{Mediana} = \begin{cases} x_{(\frac{n+1}{2})}, & \text{se } n \text{ for ímpar} \\ \frac{1}{2}[x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}], & \text{se } n \text{ for par,} \end{cases}$$

EXEMPLO 5 Mediana do conjunto de dados das atividades 1 e 2

Na atividade [Notas de Arte](#) na qual tem-se $n = 35$ notas. Como 35 é ímpar, usando a definição anterior, podemos concluir que a mediana das notas será a nota na 18ª. posição

$$\left(\frac{35+1}{2} = 18\right), \text{ a saber, mediana} = x_{(18)} = 6,5$$

Na atividade [A Maratona](#) dispõe-se dos $n = 100$ melhores tempo de chegada entre as mulheres. Como 100 é um número par, usando a definição anterior, podemos concluir que a mediana dos 100 melhores tempos será dada pela média dos tempos na 50ª e na 51ª. chegada, a saber,

$$\text{mediana} = \frac{x_{(50)} + x_{(51)}}{2} = \frac{2,949 + 2,949}{2} = 176,96 \text{ minutos}$$

Mediana para dados agrupados

Voltando à atividade [Notas de Arte](#), suponha novamente que o coordenador tenha tido acesso apenas ao [Histograma das notas de Artes sem bonificação](#), sem conhecê-las separadamente. Como ele poderia calcular a mediana da turma, considerando as notas antes da bonificação? Sabemos que a posição da mediana deve ser a posição central depois de ter as notas ordenadas. Na tabela de frequências observe que os intervalos já estão ordenados, mas apenas conhecemos a quantidade de notas que ocorreram em cada intervalo e não as notas individualmente. No entanto, é fácil, a partir da tabela, identificar em que intervalo estará a mediana, bastando para isso encontrar o intervalo que compreende a nota da posição 18. Aqui, vamos introduzir o conceito de frequência absoluta acumulada de um intervalo de classe que corresponde à soma da frequência absoluta do intervalo mais a soma acumulada das frequências absolutas de todos os intervalos anteriores. Veja a [tabela 1.10](#), incluindo as frequências absolutas acumuladas.

Tabela 1.10: Notas de artes agrupadas e frequência absoluta acumulada

Intervalo	Frequência absoluta	Ponto médio do intervalo	Freq. absoluta acumulada
$[0, 2[$	1	1,0	1
$[2, 4[$	5	3,0	$1 + 5 = 6$
$[4, 6[$	6	5,0	$6 + 6 = 12$
$[6, 8[$	23	7,0	$12 + 23 = 35$

Observe que a nota da posição 18 está no último intervalo, pois até o intervalo anterior, $]4, 6]$, acumularam-se apenas 12 das 35 notas.

Uma forma de estimar a mediana no caso em que não conhecemos as notas separadamente é tomar o ponto médio do intervalo de classe que compreende o valor da posição central. Neste

caso, teríamos como uma aproximação para a nota mediana o valor 7,0, que corresponde ao ponto médio do intervalo de classe que contém a mediana ([6, 8]). Comparando este valor com o valor da mediana obtido, usando-se as 35 notas individuais, percebe-se que o erro da aproximação é de apenas 0,5 ponto já que sabemos que a nota da posição 18 é 6,5. Novamente, observe que o agrupamento das notas não incorreu em grande perda de informação para efeito da avaliação da mediana do conjunto de notas.

Resumindo, quando dispomos dos dados apenas na forma agrupada, para obter uma aproximação da mediana, deve-se identificar o intervalo de classe que compreende o valor da posição central e, então, calcular o ponto médio desta classe como valor aproximado da mediana.

Existem outras formas de avaliar a mediana quando os dados estão agrupados e uma delas foi proposta no [exercício 17](#) do capítulo [A Natureza Estatística](#).

Escolha entre a média e a mediana como valor mais adequado para resumir a informação do conjunto de dados

Vimos que a média é uma medida muito afetada na presença de valores atípicos (muito afastados da maioria dos dados) e de distribuições fortemente assimétricas (caracterizadas por histogramas alongados para à direita ou para à esquerda). A mediana, por sua vez, é pouco afetada para valores atípicos na distribuição, e por isso é dita ser uma medida robusta.

Por exemplo, vamos voltar ao exemplo sobre as informações de salário entre os interessados para obter um cartão de crédito de uma rede de supermercados. Lembre-se que trabalhamos com dois conjuntos de dados, a saber, $C_1 = \{1, 1, 2, 3, 4, 5, 5, 6, 9, 10\}$ e $C_2 = \{1, 1, 2, 3, 4, 5, 5, 6, 9, 100\}$.

A média dos dados do conjunto C_1 é $\bar{x} = \frac{46}{10} = 4,6$ e, a mediana $= \frac{x_{(5)} + x_{(6)}}{2} = \frac{4 + 5}{2} = 4,5$.

Tanto a média, como a mediana do conjunto C_1 são valores que o representam bem: observe que os demais valores no conjunto C_1 não estão muito afastados dos valores da média e da mediana e, de forma equilibrada, alguns estão abaixo deles e outros, acima deles.

Por outro lado, a **média** dos dados do conjunto C_2 é $\frac{136}{10} = 13,6$, enquanto que a **mediana** é dada por $\frac{x_{(5)} + x_{(6)}}{2} = \frac{4 + 5}{2} = 4,5$. Este último exemplo ilustra como a média é fortemente influenciada pela presença do valor atípico 100, enquanto a mediana não. Na presença do valor atípico (100), a média é muito afetada, mudando de 4,6 para 13,6, enquanto que a mediana não foi afetada, mantendo-se igual a 4,5. Observe que apenas um valor no conjunto C_2 está acima da média.

Em distribuições aproximadamente simétricas (veja a atividade [Histograma dos registros de tempo de atividade do Capítulo A Natureza Estatística](#)) temos que a média e a mediana são valores próximos um do outro, esta é uma das razões que levam muitas pessoas a confundir estas duas medidas, achando que elas representam a mesma posição na distribuição dos dados qualquer que seja a situação. Mas, vimos que em distribuições com assimetria à direita, veja, por exemplo a figura [figura 1.16](#), a média é maior do que a mediana e, em distribuições com assimetria à esquerda, veja por exemplo a figura [figura 1.17](#), a média é menor do que a mediana.

Moda

A moda é a observação mais frequente de um conjunto de dados.

Caso não haja observação mais frequente, ou seja, todos os valores aparecem apenas uma única vez no conjunto de dados, a distribuição é dita amodal. Um conjunto é dito unimodal se houver apenas uma moda; bimodal se houver duas modas; ou multimodal se houver três ou mais modas no conjunto de dados coletados.

Vejamos exemplos das diversas situações possíveis. Considere os conjuntos de notas da prova de Matemática dos alunos de quatro turmas diferentes dadas pela tabela a seguir.

Tabela 1.11: Exemplos de diversas possibilidades quanto à moda

Turma	Notas	Moda	Distribuição
I	2; 4; 6; 7; 8; 9; 10	Não existe	Amodal
II	2; 4; 5; 5; 8; 9; 10	5	Unimodal
III	2; 4; 5; 5; 8; 9; 9; 10	5 e 9	Bimodal
IV	2; 2; 4; 5; 5; 8; 9; 9; 10	2; 5 e 9	Multimodal

O conceito de moda é adequado para conjuntos de dados qualitativos ou quantitativos discretos, pois quando os dados são quantitativos contínuos, potencialmente todas as observações são distintas entre si tal que raramente existirá um valor mais frequente e, mesmo quando um valor se repetir, não necessariamente é por que ele corresponderá a uma moda. Neste último caso, o que fazemos é, agrupar os dados em intervalos de classe para identificar um intervalo de classe modal ou intervalos de classe modais, isto é, o(s) intervalo(s) de classe com maior frequência. Uma vez identificado(s) o(s) intervalo(s) de classe modal(ais), uma estimativa para a(s) moda(s) é dada pelo ponto médio do intervalo de classe modal correspondente.

A pergunta que surge naturalmente agora é: Quando a moda será preferível à média ou à mediana?

Se o histograma da distribuição é aproximadamente simétrico, e há uma única moda, então as três medidas-resumo (média, mediana e moda) serão valores aproximadamente iguais. Nesse caso, em geral, preferiremos usar a média como medida de posição, pois ela possui propriedades relevantes para a inferência estatística. Veja uma ilustração desse caso na [figura 1.18](#).

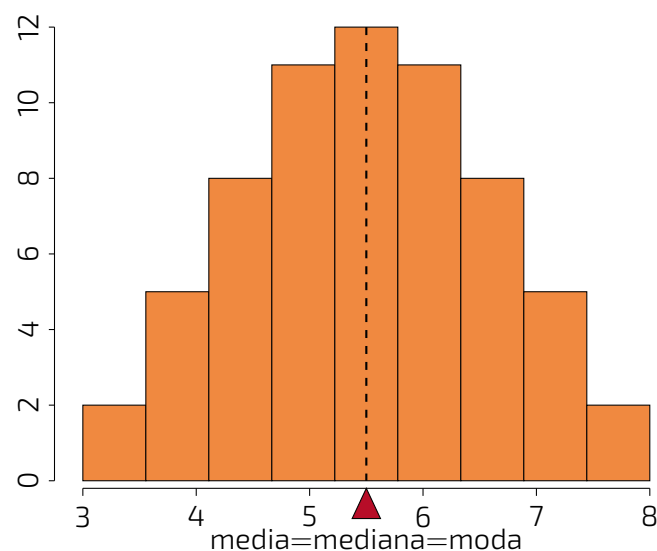


Figura 1.18: Histograma simétrico: distribuição unimodal (Dados: Registros de tempo de atividade do capítulo [A Natureza Estatística](#))

Se, no entanto, a distribuição apresenta forte assimetria com a presença valores atípicos e unimodal, então preferiremos, em geral, tomar a mediana como medida resumo. Veja uma ilustração desse caso na [figura 1.19](#).

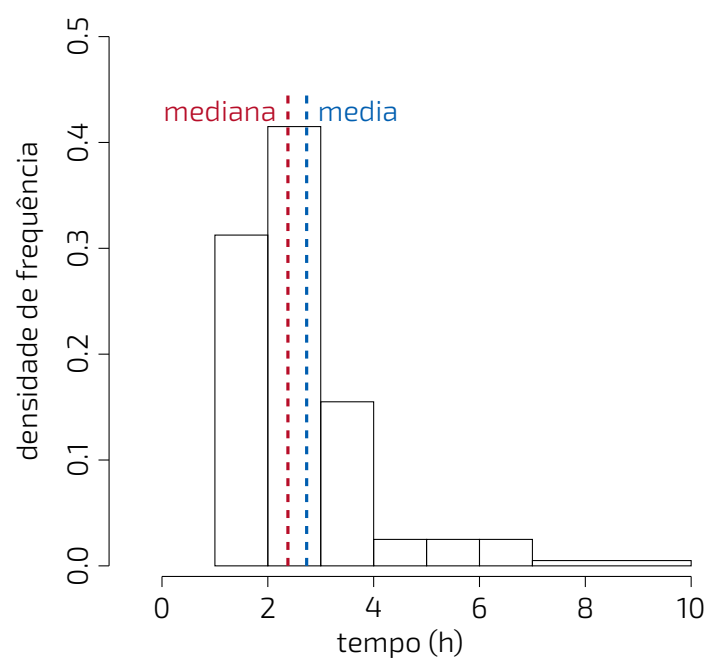


Figura 1.19: Histograma de distribuição com assimetria à direita (Tempos de chegada para a categoria Triciclo de mão na maratona de Nova Iorque/2017).

Se, por outro lado, o histograma da distribuição é do tipo simétrico e bimodal como na representação esquemática na [figura 1.20](#), então nem a média, nem a mediana serão indicadas como medidas de representação dos dados, pois observe na figura, que elas estarão situadas bem no centro onde há pouca incidência de valores. Assim, neste caso, as duas modas serão mais

úteis para descrever de forma resumida este conjunto de dados.

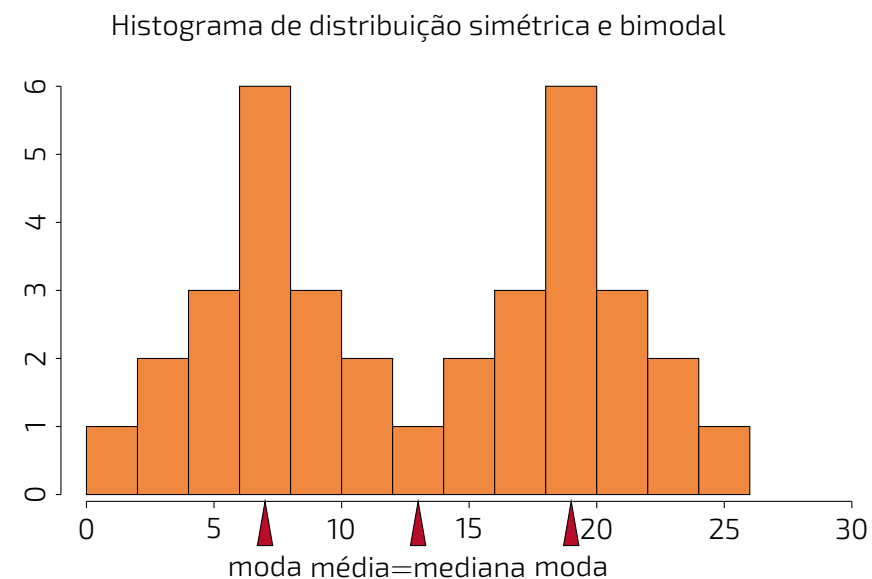


Figura 1.20: Histograma de distribuição simétrica e bimodal

Quartis

Os quartis são os três valores que dividem a distribuição em quatro partes de frequências iguais.

- a) Primeiro quartil (Q_1) é o valor da distribuição para o qual a frequência relativa de valores abaixo dele é igual 25% do número de observações do conjunto de dados e, consequentemente, acima dele, é 75% do número de observações do conjunto de dados.
- b) O segundo quartil (Q_2) é equivalente à mediana, é o valor da distribuição para o qual a frequência relativa de valores abaixo dele é 50% do número de observações do conjunto de dados e, consequentemente, acima dele, é 50% do número de observações do conjunto de dados.
- c) Terceiro quartil (Q_3) é o valor da distribuição para o qual a frequência relativa de valores abaixo dele é igual 75% do número de observações do conjunto de dados e, consequentemente, acima dele, é 25% do número de observações do conjunto de dados.

EXEMPLO 6 Quartis da distribuição dos dados: mulheres na maratona

Na atividade [A Maratona](#) foram identificados os tempos das posições 25ª, 50ª e 75ª. A saber, $t_{25} = 165,87$ minutos, $t_{50} = 176,96$ minutos e $t_{75} = 179,85$ minutos.

Como o número total de observações nesse conjunto de dados é $n = 100$, podemos tomar como o primeiro quartil, o tempo da 25ª. posição, pois $(\frac{1}{4} \cdot 100 = 25)$. Assim, $Q_1 = 165,87$.

O mesmo raciocínio se for usado para o segundo quartil (mediana) indicará o tempo da 50ª. posição que é 176,95 minutos. No entanto, vamos usar preferencialmente a regra apresentada para obter a mediana quando temos um número par de observações. Lembre-se que já fizemos isso, calculando a média dos tempos das posições 50 e 51, obtendo 176,96 minutos,

bem próximo ao tempo da posição 50.

Para o terceiro quartil, podemos tomar o o valor da 75ª. posição, pois $((\frac{3}{4} \cdot 100 = 75))$. Assim, $Q_3 = 179,85$ min.

Finalmente, observe que com essas informações dispõe-se de um agrupamento dos dados em quatro intervalos de comprimentos desiguais, a saber,

- a) [mínimo, Q_1 [,
- b) [Q_1 , mediana[,
- c) [mediana, Q_3 [e
- d) [Q_3 , máximo[,

porém todos eles com frequências relativas iguais a $\frac{1}{4} = 0,25$.

Já vimos como determinar mediana (ou segundo quartil) de um conjunto de n dados. Um método simples para obter os demais quartis, Q_1 e Q_3 , é considerar dois novos conjuntos de dados, o primeiro, consistindo da primeira metade dos valores ordenados e, o segundo, consistindo da segunda metade. Depois, tome como primeiro quartil a mediana da primeira metade do conjunto e, como terceiro quartil, a mediana da segunda metade do conjunto de dados.

O esquema dos cinco números e o boxplot

O esquema dos cinco números é composto pelas seguintes medidas:

- a) valor mínimo,
- b) os três quartis e
- c) valor máximo.

Com essas cinco informações podemos construir o boxplot (gráfico-caixa) de um conjunto de dados quantitativos, conforme foi visto no último item da atividade [A Maratona](#) na seção anterior. O boxplot é uma representação gráfica de dados quantitativos alternativa ao histograma. Veja na [figura 1.21](#) uma representação simultânea dos dois gráficos para os 100 melhores tempos de chegada na categoria mulheres.

HISTOGRAMA DOS CEM MELHORES TEMPOS - MULHERES - MARATONA - NY - 2017

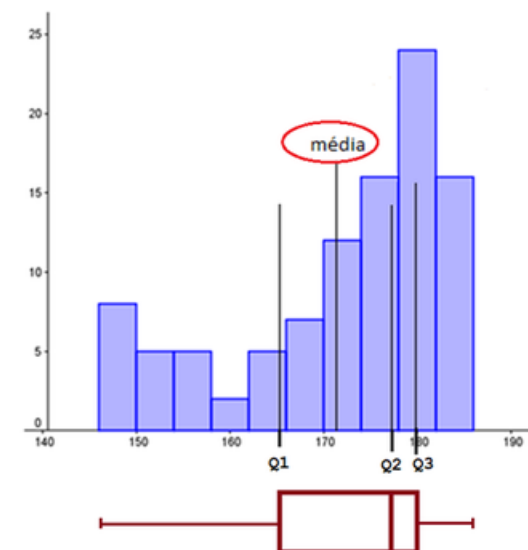


Figura 1.21: Histograma e boxplot (sem sinalização de valores discrepantes) para os 100 melhores tempos na categoria mulheres

O boxplot é um gráfico muito utilizado em comparações do comportamento de uma variável quantitativa para diferentes categorias, por exemplo, podemos construir boxplots para as categorias mulheres, homens, "cadeira de rodas" e "triciclo de mão" para identificar diferenças ou semelhanças entre elas.

Veremos adiante como sinalizar valores discrepantes no boxplot. Valores discrepantes em um conjunto de dados são valores que destoam em relação aos demais valores no conjunto. No exemplo cartão de crédito no supermercado (2), o valor 100 é um valor discrepante em relação aos demais valores do conjunto que variam entre 0 e 10.

Ramo-e-folhas

O ramo-e-folhas é uma representação gráfica simples para dados quantitativos. Como cada valor observado é considerado nessa representação, sua construção para conjuntos de dados com mais de 30 observações é recomendada apenas com o auxílio de tecnologia.

Na construção do ramo-e-folhas sem o auxílio de tecnologia precisamos primeiro conhecer os valores mínimo e máximo para definirmos os ramos e as folhas que são unidades e sub-unidades dos valores observados, respectivamente. Para facilitar, veja o exemplo a seguir.

EXEMPLO 7 Valores consumidos na cantina de uma escola entre 7h e 9h

Suponha que 20 pessoas consumiram na cantina de uma escola em um dia no período entre 7 e 9 horas da manhã e que os valores em reais foram registrados: 38, 20, 15, 8, 1, 32, 29, 18, 13, 10, 6, 19, 22, 25, 3, 13, 21, 30, 12 e 25, nessa ordem.

Observe que os valores registrados variaram entre 1 real e 38 reais. Nesse caso, podemos pensar nos ramos como as dezenas 0, 1, 2 e 3 e as folhas como as respectivas unidades dessas dezenas. Assim temos a seguinte representação de ramo-e-folhas para esses dados

0		0 1 6 8
1		0 2 3 3 5 8 9
2		0 1 2 5 5 9
3		0 2 8

Na terceira linha do ramo-e-folhas desse exemplo temos o ramo da dezena 20. Assim, o número 0 representado logo a seguir, à direita da barra vertical, corresponde a uma observação de valor 20, o número 1 em seguida, corresponde a uma observação de valor 21, e assim por diante, de modo que os 20 valores observados ficam organizados de maneira rápida e simples.

Observe que o ramo-e-folhas parece um histograma simplificado com as posições trocadas: classes de valores na escala vertical e frequências na escala horizontal (4 observações no primeiro ramo, 7 observações no segundo ramo, 6 observações no terceiro ramo e 3 observações no quarto ramo). As classes correspondem aos intervalos abertos à direita $[0, 10[$, $[10, 20[$, $[20, 30[$ e $[30, 40[$.

O ramo-e-folhas pode ser útil na identificação das medidas do esquema dos 5 números. Tem-se mínimo e máximo 1 e 38, respectivamente. O segundo quartil, como temos uma quantidade par de valores é a média dos valores nas posições centrais 10 e 11 que são facilmente identificadas no ramo-e-folhas como as observações 18 e 19 reais. Assim, o segundo quartil (mediana) é dado por R\$ 18,50. Para identificarmos os primeiro e terceiro quartis, podemos dividir o conjunto em duas metades de tamanho 10. Assim o primeiro quartil é dado pela média dos valores nas posições 5 e 6 e, o terceiro quartil, pela média dos valores nas posições 15 e 16.

$$Q_1 = \frac{10 + 12}{2} = 11 \text{ e } Q_3 = \frac{25 + 25}{2}$$

Resumindo, os cinco números são dados por R\$ 1,00, R\$ 11,00, R\$ 18,50, R\$ 25,00 e R\$ 38,00. Veja na [figura 1.22](#) o boxplot desses dados, usando a escala dos valores na orientação vertical.

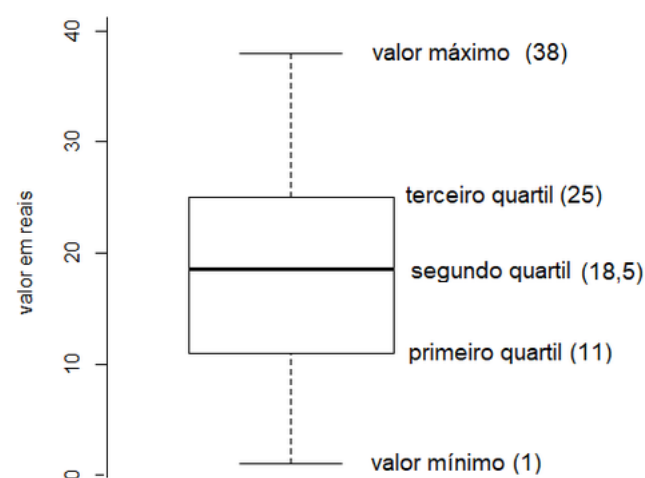


Figura 1.22: Boxplot dos valores consumidos na cantina

Objetivos Específicos

Categoria homens na maratona

Usar medidas de posição para a comparação das distribuições de uma mesma variável em dois grupos diferentes.

Sugestões e discussões

Categoria homens na maratona

Nesta atividade serão comparados os dados dos 100 melhores tempos na maratona de Nova Iorque/2017 para as categorias homens e mulheres. A tabela com os 100 melhores tempos em minutos para a categoria homens é fornecida. A comparação será feita com base nas medidas de posição média, quartis, mínimo e máximo, que são fáceis de serem determinadas apesar da quantidade de dados ser 100. No caso da média, a soma dos 100 tempos é informada.

Solução: Categoria homens na maratona

Veja a [tabela 1.12](#) com as medidas resumo para as categorias homens e mulheres, correspondendo aos 100 melhores tempos na maratona de Nova Iorque - 2017.

	Mulheres	Homens
Mínimo	146,88	130,88
Máximo	185,15	158,33
Média	171,92	150,70
Mediana	176,96	153,00
Q_1	165,87	148,32
Q_3	179,85	156,62

Tabela 1.12: Tabela das medidas resumo para as categorias mulheres e homens - Maratona de Nova Iorque/2017

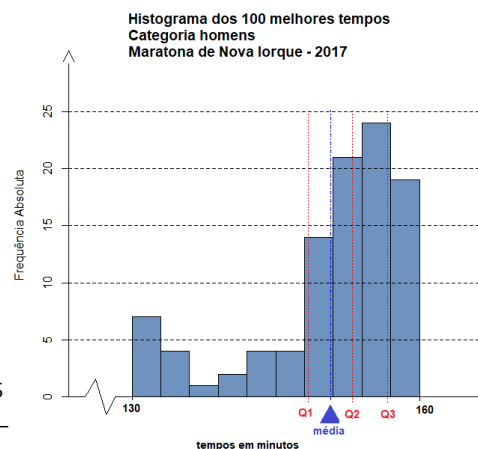


Figura 1.23: Histograma dos resultados da categoria de Homens da Maratona da Cidade de Nova York do ano 2017, com média, mediana, Q_1 e Q_3 indicados

PRATICANDO

MEDIDAS DE POSIÇÃO

Categoria homens na maratona

Atividade 3

Considere os dados da categoria Homens da Maratona da Cidade de Nova Iorque do ano 2017 apresentados na [tabela 1.13](#), já convertidos para minutos.

Tabela 1.13: 100 melhores tempos de finalização da Maratona de Nova Iorque 2017 para homens

	+0	+10	+20	+30	+40	+50	+60	+70	+80	+90
1	130,88	135,48	147,42	150,00	151,55	153,08	154,38	156,10	156,95	157,85
2	130,93	138,65	147,68	150,08	151,65	153,13	154,50	156,38	157,25	157,85
3	131,53	140,48	148,12	150,10	151,78	153,25	154,63	156,45	157,25	157,88
4	131,87	141,50	148,28	150,43	151,85	153,30	154,65	156,62	157,30	158,03
5	132,02	142,60	148,32	150,47	151,87	153,42	155,27	156,62	157,38	158,08
6	132,65	142,77	148,43	150,85	151,98	153,73	155,27	156,72	157,52	158,12
7	132,80	143,67	148,70	151,05	152,50	153,75	155,45	156,75	157,58	158,13
8	133,35	143,85	149,20	151,20	152,77	154,05	155,50	156,77	157,63	158,18
9	133,97	145,58	149,68	151,40	152,88	154,25	155,68	156,80	157,68	158,33
10	134,95	147,18	149,78	151,43	152,92	154,37	155,80	156,82	157,77	158,33

Veja na [figura 1.27](#) um histograma destes dados, considerando-se 10 intervalos de classe, a saber, $[130, 133[$, $[133, 136[$, $[136, 139[$, $[139, 142[$, $[142, 145[$, $[145, 148[$, $[148, 151[$, $[151, 154[$, $[154, 157[$ e $[157, 160[$. As frequências absolutas de cada intervalo de classe estão destacadas no topo dos retângulos do histograma.

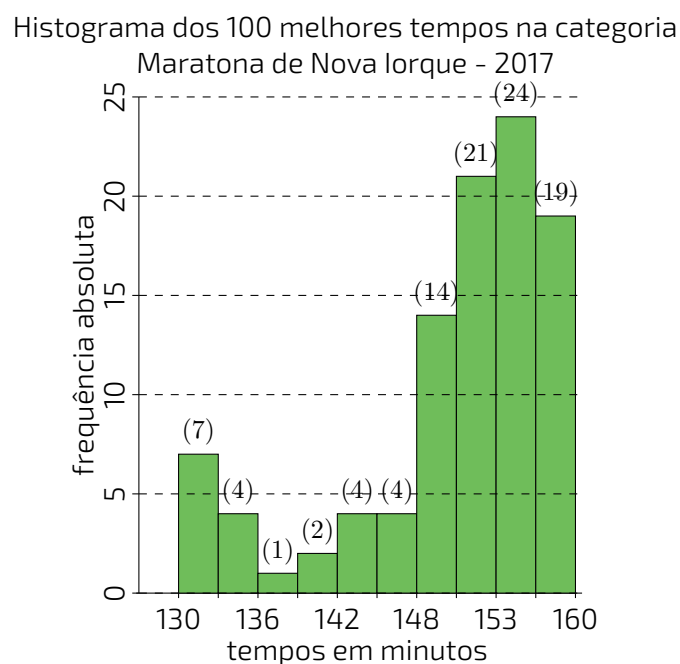


Figura 1.27: Histograma dos resultados da categoria de Homens da Maratona da Cidade de Nova Iorque do ano 2017

- Calcule a média dos 100 melhores tempos na categoria homens, sabendo que a soma dos tempos é dada por 15.069,70 horas.
- Calcule a mediana dos 100 melhores tempos na categoria homens.
- Identifique o intervalo de classe modal dos 100 melhores tempos na categoria homens.
- Determine o primeiro e o terceiro quartis dos 100 melhores tempos na categoria homens.
- Localize no histograma a média e os quartis.
- Compare com os resultados obtidos para a categoria homens com os obtidos para a categoria mulheres na atividade [A Maratona](#) completando a [tabela 1.14](#).

Tabela 1.14: Tabela de medidas-resumo para Mulheres e Homens - Maratona de Nova Iorque/2017

	Mulheres	Homens
Mínimo		
Máximo		
Média		
Mediana		
Q_1		
Q_3		

Solução: Categoria homens na maratona

Comparando as duas distribuições (homens e mulheres) é possível perceber que ambas têm a mesma forma com assimetria à esquerda, o que pode ser visualizado pelos histogramas. No entanto, percebe-se que os homens são mais rápidos (todas as medidas da [tabela 1.12](#) são menores para os homens). Além disso, podemos notar que há mais dispersão entre as mulheres, calculando-se a amplitude amostral. Entre as mulheres, a amplitude amostral é $185,15 - 146,88 = 38,27$ minutos, enquanto que entre os homens é de $158,33 - 130,88 = 27,45$ minutos. Veja na [figura 1.24](#), dois histogramas correspondentes às categorias homens e mulheres, construídos na mesma escala, ilustrando os comentários da comparação das duas categorias.

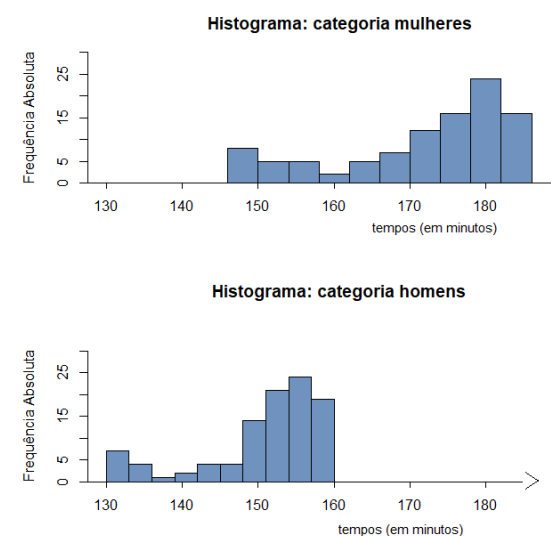


Figura 1.24: Histogramas dos 100 melhores tempos das categorias homens e mulheres

Observe como a comparação entre as duas categorias (homens e mulheres) é mais simples se usarmos a representação dos dados com o [boxplot](#).

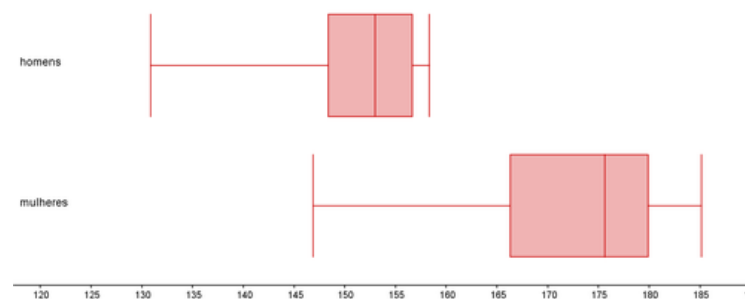


Figura 1.25: Boxplots dos 100 melhores tempos das categorias homens e mulheres (sem sinalização de valores discrepantes)

Objetivos Específicos

Comparação das diferentes categorias na maratona

- Comparar distribuições de uma mesma variável para grupos distintos a partir dos histogramas.
- Perceber a necessidade de usar a mesma escala nos eixos do histograma, para tornar os mesmos comparáveis.

Sugestões e discussões

Comparação das diferentes categorias na maratona

Esta atividade introduz os elementos necessários para a comparação de dois histogramas, a saber: mesmas escalas nos eixos e colunas de frequências relativas.

Os histogramas são apresentados com uma série de perguntas de discussão que podem motivar a formulação do conceito de dispersão de forma intuitiva, que será trabalhado na seguinte seção. Além de mostrar como apenas as medidas de posição não dizem suficiente sobre uma distribuição.

As perguntas não têm respostas fechadas, têm o intuito de gerar uma discussão sobre os assuntos já colocados.

Solução: Comparação das diferentes categorias na maratona

- As escalas horizontal e vertical são coincidentes em ambos os histogramas para permitir a comparação. Isto foi discutido no capítulo **A Natureza da Estatística**.
- Observando-se os histogramas o maior tempo está na categoria de Triciclo de mão.
- A simples visualização pode remeter a estimativa de 4 horas para estas categorias, valor bem maior do que as médias para homens e mulheres (ativ-maratona-categoria-homens). No entanto, este valor excede muito as médias reais, apresentadas. Isto mostra que estimar o centro de equilíbrio de uma distribuição a partir do histograma não é trivial.
- Ambos os histogramas apresentam assimetria à direita: grande concentração de dados à esquerda e forma alongada para a direita.

- Construa, usando a mesma escala, os boxplots (sem sinalização de valores discrepantes) para as categorias homens e mulheres e destaque as diferenças encontradas nesse gráfico.

PARA REFLETIR

- O que seria necessário considerar para poder comparar o histograma da categoria homens com o da categoria mulheres?
- Como poderiam ser utilizados os quartis para comparar duas distribuições de dados? Pense em alguma forma de comparar esse dados de forma visual e descreva-a.

Comparação das diferentes categorias na maratona

Atividade 4

Observe os histogramas da [figura 1.28](#) referentes aos tempos de chegada das categorias “cadeira de rodas” e “triciclo de mão” na Maratona de Nova Iorque em 2017. Nesse caso, os tempos de chegada foram convertidos para horas e referem-se ao total de participantes que completaram a prova: 51 na categoria cadeira de rodas e 69 na categoria triciclo de mão.

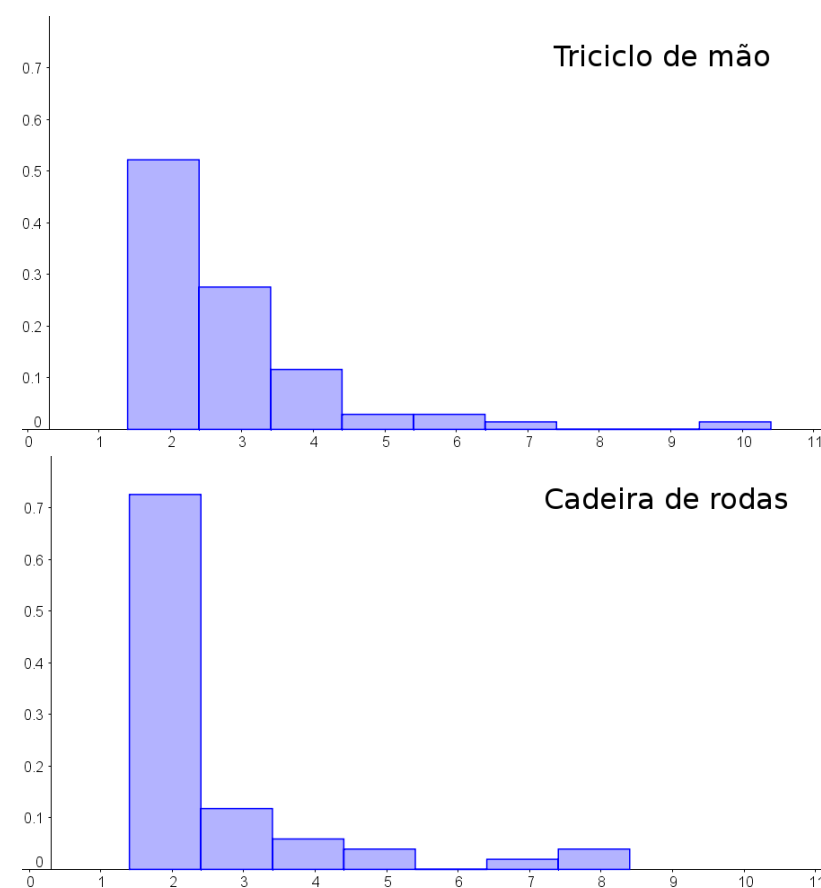


Figura 1.28: Histogramas comparativos das quatro modalidades da maratona de Nova Iorque 2017

- a) Compare as escalas utilizadas na construção destes histogramas, tanto no eixo horizontal, como no eixo vertical. O que você observou?
- b) Em qual categoria se encontra o atleta que completou a maratona no maior tempo?
- c) Você consegue estimar o tempo médio destas categorias observando os histogramas? Você acha que os tempos médios dessas duas categorias serão muito diferentes dos tempos médios das categorias de homens e de mulheres (atividade [Categoria Homens na Maratona](#))?
- d) Observe o quadro a seguir e marque as médias (em horas) nos histogramas. Comente sobre a posição da média em cada caso e sobre a simetria ou assimetria de cada distribuição de dados.

Tabela 1.15: Média das quatro categorias da maratona de Nova Iorque 2017

Categoria	Cadeira de rodas	Triciclo de mão
Média	2,59	2,73

- e) Observe que as médias não são muito diferentes. Se você conhecesse apenas a média, seria capaz de perceber a forma destes histogramas? Por quê?
- f) Comparando os dois histogramas, qual distribuição apresenta maior dispersão? Por quê?
- g) A partir dos esquemas de cinco números para os tempos de chegada nas categorias cadeira de rodas e triciclo de mão, informados a seguir, construa os respectivos boxplots sem sinalização de valores discrepantes, usando a mesma escala de valores para as duas categorias.

Categoria	Mínimo	Q_1	Q_2	Q_3	Máximo
Cadeira de rodas	1,62	1,8	2,09	2,68	7,81
Triciclo de mão	1,48	1,78	2,38	3,09	9,52

Solução: Comparação das diferentes categorias na maratona

- e) Não é possível, a informação apenas da média é insuficiente para caracterizar a forma do histograma. Outras informações são necessárias para isto.
- f) O triciclo de mão apresenta maior dispersão, considerando a amplitude.
- g) Na figura a seguir veja uma construção dos boxplots sem sinalização dos valores discrepantes.

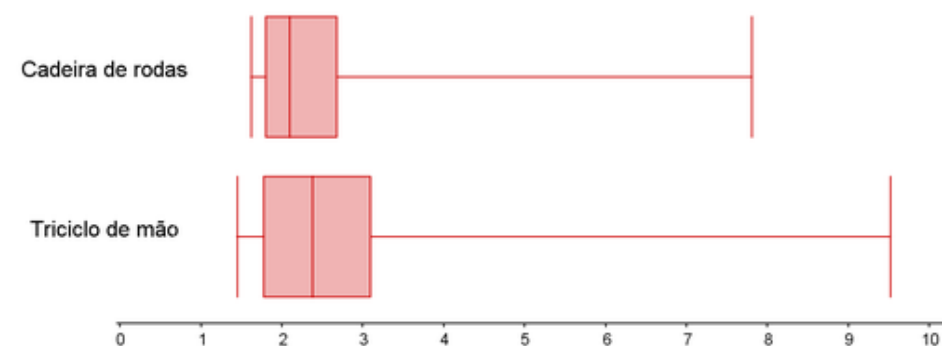


Figura 1.26: Boxplots sem sinalização de valores discrepantes para os tempos de chegada em horas das categorias "cadeira de rodas" e "triciclo de mão"

Estratégia de investimento**Atividade 5**

Para investir na bolsa de valores compramos ações de empresas por intermédio de uma corretora a um certo preço e depois de um período de tempo vendemos estas ações na expectativa de que seus preços tenham aumentado. No entanto, também podemos perder com o investimento, caso o preço da ação diminua no período de investimento. Uma ação é a menor parte do capital de uma empresa. Veja na figura a seguir um esquema simplificado do investimento na bolsa de valores.

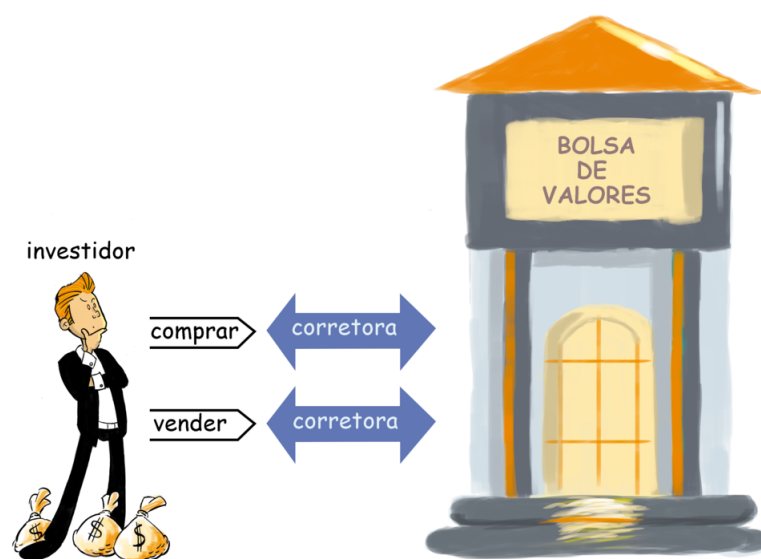


Figura 1.29: Esquema simplificado de investimento na bolsa de valores

Suponha que você tenha a oportunidade de investir um capital, comprando ações de uma de duas Companhias *A* ou *B* e para escolher uma das duas, disponha de duas amostras de preços do valor destas ações (em reais) registrados no fechamento da bolsa de valores em dez sextas-feiras consecutivas. Veja na figura e na tabela a seguir a cotação das ações ao longo das últimas 10 semanas.

Objetivos Específicos**Estratégia de investimento**

Definir medidas que caracterizam a dispersão de um conjunto de dados.

Sugestões e discussões**Estratégia de investimento**

Nessa atividade são apresentados dois conjuntos de dados temporais que apresentam mesma média, mesma mediana e mesma moda e, no entanto, seus gráficos de linha são distintos. O objetivo principal é mostrar que as medidas de posição podem ser insuficientes para caracterizar a distribuição dos dados, levando à necessidade de usar medidas de dispersão. Lembre que é esperado, do Ensino Fundamental, que os estudantes já tenham a noção de amplitude amostral, uma medida bruta de dispersão, pois só leva em conta o mínimo e o máximo observados.

Esta atividade pode ser vinculada às disciplinas de História ou Geografia, por exemplo, no estudo do período da Crise Econômica de 1929 ou outros temas relacionados com o PIB e crescimento econômico.

Solução: Estratégia de investimento

- a) A escolha pode ser tanto pela *A* como pela *B*, mas deve vir acompanhada de uma justificativa. Por exemplo, "eu escolheria a companhia *A* porque os preços oscilam menos", "escolheria a companhia *B* porque os preços oscilam mais", "escolheria a companhia *B* porque foi a que apresentou maior valor de cotação entre os dias observados" etc.
- b) Dado que são 10 observações em cada um dos conjuntos e que as somas das 10, resultam em 615, segue que a média das cotações na companhia *A* é R\$ 61,50, que também é a média das cotações na companhia *B*.
- c) Para obter as medianas é necessário antes ordenar os valores. Na tabela a seguir os valores das cotações foram ordenados para cada companhia.

A	56	56	57	58	61	63	63	67	67	67
B	33	43	48	52	57	67	67	77	82	90

Como são 10 observações em cada conjunto e 10 é um número par, temos que a mediana será dada pela média das duas posições centrais, a saber, posições 5 e 6: $\text{Mediana} = \frac{x_{(5)} + x_{(6)}}{2}$.

Na companhia *A* teremos $\text{Mediana} = \frac{61 + 63}{2} = 62$ reais e, na companhia *B*, $\text{Mediana} = \frac{57 + 67}{2} = 62$ reais.

- d) Na companhia *A* o valor mais frequente foi 67, ocorrendo 3 vezes. Na companhia *B*, o valor mais frequente foi 67, ocorrendo duas vezes. Logo, tanto em *A* como em *B* o valor da moda foi 67 reais.
- e) Não, pois tais medidas são idênticas nas duas companhias.
- f) Analisando os gráficos de linha da figura 57, percebe-se que as cotações da companhia *B* variam mais do que as da companhia *A* e, portanto, como menor risco envolve menos variação, escolheria a companhia *A*. Observe que as amplitudes (diferença entre o maior e menor valores) observadas nas companhias *A* e *B* são $67 - 56 = 11$ e $90 - 33 = 57$, respectivamente, confirmando que na companhia *A* a variação das cotações é menor.

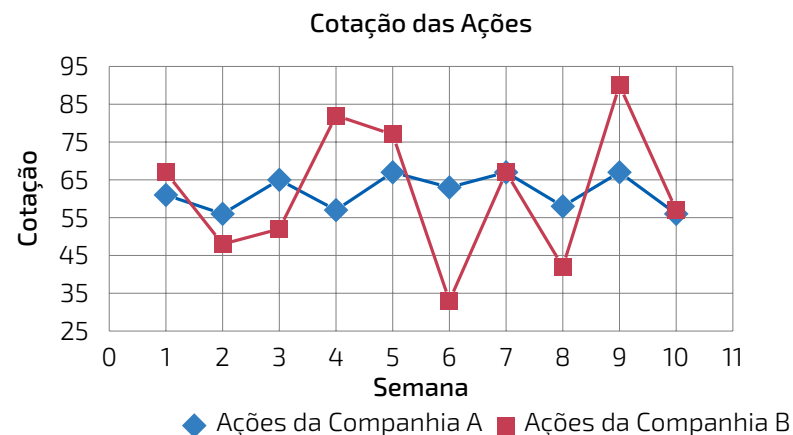


Figura 1.30: Gráficos de linha da cotação das ações

Semana	1	2	3	4	5	6	7	8	9	10	Total
A	61	56	63	57	67	63	67	58	67	56	615
B	67	48	52	82	77	33	67	42	90	57	615

- a) Observando o gráfico, qual das duas companhias você escolheria para investir? Por quê?
- b) Obtenha as médias das cotações das ações das companhias *A* e *B* nas semanas observadas e compare-as.
- c) Obtenha as medianas das cotações das ações das companhias *A* e *B* nas semanas observadas e compare-as, lembrando que os dados da tabela estão apresentados na ordem temporal.
- d) Obtenha as modas das cotações das ações das companhias *A* e *B* nas semanas observadas e compare-as.
- e) Analisando apenas as medidas de posição obtidas em a), b) e c), pode-se dizer que as duas companhias diferem uma da outra? Por quê?
- f) Um investimento que apresenta grandes ganhos e perdas pode ser chamado de alto risco, já investimentos cujos valores flutuam pouco são considerados de baixo risco. Se você é um investidor da bolsa de valores avesso ao risco, isto é, você gostaria de escolher o investimento com menores flutuações, em qual das companhias você investiria o seu dinheiro? Por quê?

ORGANIZANDO MEDIDAS DE DISPERSÃO

Pela atividade anterior, você deve ter notado que usar apenas medidas de posição para caracterizar uma distribuição não é suficiente. Nos dois conjuntos analisados, vimos que ambos apresentaram média, mediana e moda iguais. No entanto, vimos que um deles apresenta maiores variações de valores do que o outro. A ideia por trás de variação é a noção de dispersão.

Enquanto as medidas de posição procuram resumir o conjunto de dados em alguns valores situados entre dados coletados, as medidas de dispersão buscam avaliar quão dispersos são os dados coletados. Isso é de fundamental importância, pois podemos ter dois conjuntos de dados com as mesmas medidas de posição, como na [Estratégia de Investimento](#), mas com dispersões diferentes, fazendo com que os valores qualitativos dessas medidas de posição sejam também diferentes.

Há uma piada irônica que conta que o Estatístico é o profissional que diz que uma pessoa, ao se sentar numa cadeira com duas placas de metal, uma aquecida a 100°C e outra resfriada a -40°C , estará em média confortável, pois temperatura média é de 30°C . Na verdade, um Estatístico jamais diria isso, pois ele não toma decisões apenas por uma medida de posição, mas leva em conta também a dispersão dos dados em torno de uma medida de posição. Uma cadeira com duas placas de metal, uma aquecida a 35°C e outra a 25°C , também tem temperatura média de 30°C , mas há menos dispersão da temperatura nessa cadeira que na outra. Assim, embora quantitativamente iguais, os dois valores de 30°C não são qualitativamente equivalentes. Há, portanto, que se avaliar a dispersão dos dados coletados, a fim de poder obter conclusões adequadas.

Nesta seção serão apresentadas medidas que buscam caracterizar a dispersão dos dados em um conjunto.

Amplitude amostral e distância entre quartis

Entre as medidas de dispersão mais simples, define-se a amplitude amostral (R) como a diferença entre o maior valor e menor valor observados. Usando a notação apresentada anteriormente, dado um conjunto com n observações, temos

$$\text{Amplitude amostral} = R = \overbrace{x_{(n)}}^{\text{maior valor do conjunto}} - \underbrace{x_{(1)}}_{\text{menor valor do conjunto}}$$

Uma desvantagem desta medida é que ela considera apenas os dois extremos do conjunto. Ainda é possível que dois conjuntos, tendo mesmas média, moda e mediana, apresentem a mesma amplitude e, no entanto, eles tenham comportamentos diferentes.

EXEMPLO 8 Notas de Matemática

Considere os seguintes conjuntos de notas de Matemática de duas turmas de reforço, cada uma com 10 alunos.

Notas da turma A = $\{1, 1, 1, 5, 5, 5, 5, 9, 9, 9\}$ e Notas da turma B = $\{1, 3, 3, 5, 5, 5, 5, 7, 7, 9\}$

Observe que os dois conjuntos apresentam a mesma média 5 (a soma das notas é 50 em ambos), a mesma moda 5 (5 é o valor mais frequente em ambos), a mesma mediana 5 (5 é

o valor da posição central em ambos) e a mesma amplitude amostral 8.

No entanto, comparando os diagramas de pontos correspondentes a cada um deles, ilustrados na [figura 1.31](#), é possível perceber diferenças quanto à dispersão das notas em torno da média 5,0 nos dois conjuntos.

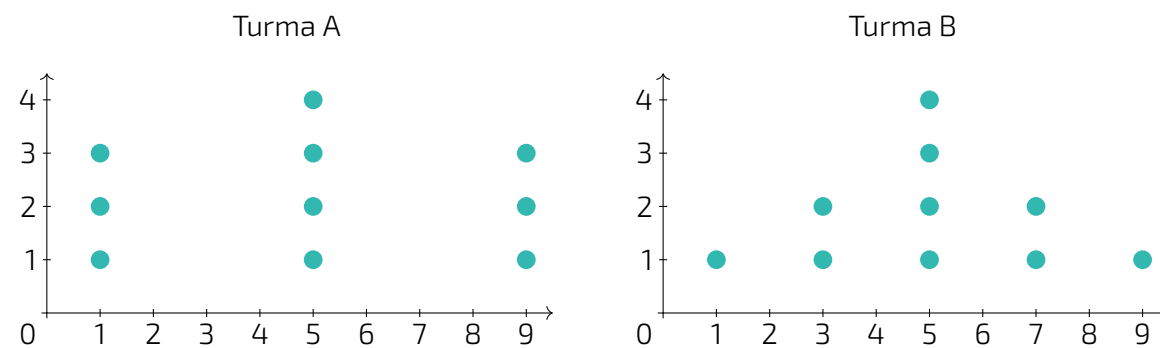


Figura 1.31: Diagramas de pontos das notas nas turmas A e B

Na turma A as notas são mais dispersas em torno da média comparando com a turma B. Mas essa diferença não é captada pela amplitude amostral.

O exemplo *Notas de Matemática* ilustra a necessidade de uma medida um pouco mais refinada, mas ainda sem considerar todos os valores no conjunto, é a distância entre quartis (DQ), definida como a diferença entre o terceiro e primeiro quartis da distribuição. Usando a notação apresentada anteriormente,

$$DQ = Q_3 - Q_1$$

EXEMPLO 9 Notas de Matemática 2

Vamos retomar os dados do exemplo Notas de Matemática no qual tem-se dois conjuntos de notas, cada conjunto tem 10 observações.

Notas da turma A = {1, 1, 1, 5, 5, 5, 5, 9, 9, 9} e Notas da turma B = {1, 3, 3, 5, 5, 5, 5, 7, 7, 9}

Podemos dividir os dois conjuntos em duas metades com cinco observações cada e tomar as medianas desses subconjuntos para identificar os primeiro e terceiros quartis das notas.

$$\text{Notas da turma A} = \overbrace{\{1, 1, 1, 5, 5\}}^{\text{primeira metade}}, \underbrace{\{5, 5, 9, 9\}}_{\text{segunda metade}}$$

Deste modo, temos para a turma A, $Q_1 = 1$ (mediana da primeira metade) e $Q_3 = 9$ (mediana da segunda metade) tal que $DQ = 9 - 1 = 8$ e, para a turma B, usando o mesmo raciocínio, $DQ = 7 - 3 = 4$, indicando que na turma B, considerando a distância entre quartis, temos menor dispersão, comparada à turma A, observação que pode ser verificada nos diagramas de pontos da [figura 1.31](#).

De fato, a distância entre quartis (DQ) também apresenta a desvantagem de somente considerar o primeiro e terceiro quartis, não considerando todas as observações do conjunto. A seguir, serão definidas medidas de dispersão que levam em conta todas as observações realizadas.

Boxplot com sinalização de valores discrepantes

Já vimos como construir o boxplot de um conjunto de dados quantitativos sem a sinalização de valores discrepantes, conhecendo o esquema dos cinco números do conjunto de dados. Agora veremos como é feita a construção do boxplot com essa sinalização.

A distância entre quartis ($DQ = Q_3 - Q_1$) é a medida de dispersão utilizada na classificação de valores da distribuição como valores discrepantes, isto é, valores que destoam dos demais no conjunto de dados.

O critério adotado para classificar um valor como discrepante na construção do boxplot é descrito a seguir.

Defina

$$\text{Cerca inferior} = Q_1 - 1,5 \cdot DQ \text{ e Cerca superior} = Q_3 + 1,5 \cdot DQ$$

Qualquer observação do conjunto de dados que for **menor do que a cerca inferior ou maior do que a cerca superior** é classificada como valor discrepante e assinalada no boxplot com um asterisco ou algum outro caracter, de acordo com o eixo na escala dos dados.

Na finalização da construção do boxplot, traçam-se segmentos paralelos ao eixo considerado (vertical ou horizontal) partindo dos pontos médios das bases do retângulo e terminando nos maior e menor valores não discrepantes que foram observados.

EXEMPLO 10 Construção do boxplot da atividade 3

O esquema dos cinco números para a categoria homens dos 100 melhores tempos de chegada (em minutos) é dado por: mínimo= 130,88; $Q_1 = 148,37$; $Q_2 = 152,99$; $Q_3 = 156,66$ e máximo= 158,33.

Portanto, $DQ = Q_3 - Q_1 = 156,66 - 148,37 = 8,29$, tal que

- a) $1,5 \cdot DQ = 12,435$
- b) Cerca inferior= $148,37 - 12,435 = 135,935$
- c) Cerca superior= $156,66 + 12,435 = 169,095$

Como o máximo é 158,33, tem-se que não há valores discrepantes para o lado superior da distribuição. Mas, a cerca inferior é 135,935 minutos, enquanto que o valor mínimo é 130,88 minutos. Nesse caso, será necessário voltar aos dados individuais para poder destacar os valores discrepantes para o lado inferior da distribuição.

Consultando os valores ordenados tem-se que os 12 primeiros tempos de chegada são dados por 130,88; 130,93; 131,53; 131,86; 132,01; 132,65; 132,8; 133,35; 133,96; 134,95; 135,48 e 138,65. Logo, concluímos que os 11 primeiros tempos de chegada na categoria homens são valores discrepantes. Veja na [figura 1.32](#) o boxplot com a sinalização de tais valores.

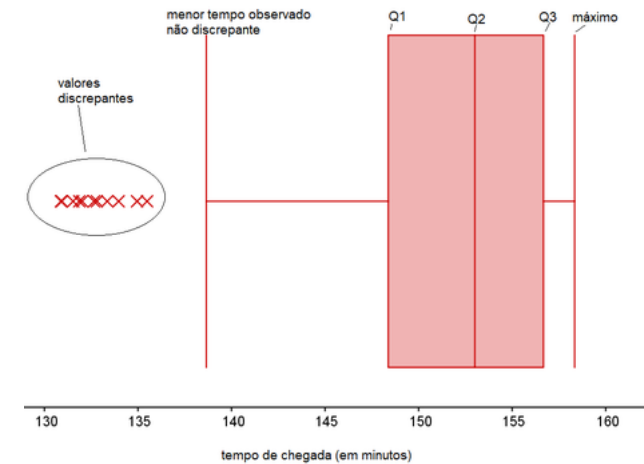


Figura 1.32: Boxplot dos 100 primeiros tempos de chegada (em horas) na maratona de Nova York (categoria homens)

O retângulo do boxplot corresponde aos 50% valores centrais da distribuição, ou seja, metade dos dados estão no intervalo delimitado pela caixa (retângulo) e, a outra metade, está nos dois intervalos delimitados fora da caixa, sendo 25% acima e 25% abaixo da caixa.

As medidas do esquema dos cinco números nos permitem avaliar o grau de assimetria da distribuição. Por exemplo, se

- a) $\text{Mediana} - Q_1 \approx Q_3 - \text{mediana}$
- b) $Q_1 - x_{(1)} \approx x_{(n)} - Q_3$
- c) $\text{Mediana} - x_{(1)} \approx x_{(n)} - \text{mediana}$

podemos concluir que a distribuição é aproximadamente simétrica, porém se alguns destes pares de intervalos apresentarem comprimentos muito diferentes, isso indica que a distribuição apresenta algum tipo de assimetria. Os pares de intervalos citados são

- a) $[Q_1, \text{mediana}]$ e $[\text{mediana}, Q_3]$;
- b) $[\text{mínimo}, Q_1]$ e $[Q_3, \text{máximo}]$;
- c) $[\text{mínimo}, \text{mediana}]$ e $[\text{mediana}, \text{máximo}]$;

Afinal, para que servem os quartis da distribuição?

- a) identificar valores discrepantes da distribuição (se houver), também conhecidos como valores atípicos ou *outliers*;
- b) avaliar o grau de assimetria da distribuição empírica do conjunto de dados e
- c) construir um gráfico alternativo ao histograma para representar dados quantitativos conhecido como boxplot.

EXEMPLO 11 Construção do bloxplot da atividade 2

O esquema dos cinco números para a categoria mulheres dos 100 melhores tempos de chegada (em minutos) é dado por: mínimo=146,88; $Q_1 = 166,31$; $Q_2 = 175,62$; $Q_3 = 179,89$ e máximo=185,15.

Portanto, $DQ = Q_3 - Q_1 = 179,89 - 166,31 = 13,58$, tal que

- a) $1,5 \cdot DQ = 20,37$
- b) Cerca inferior= $166,31 - 20,37 = 145,94$
- c) Cerca superior= $179,89 + 20,37 = 200,265$

Como o máximo é 185,15, tem-se que não há valores discrepantes para o lado superior da distribuição. O mesmo vale para o lado inferior da distribuição, pois o mínimo é 146,88 e a cerca inferior é 145,94. Logo, na distribuição dos 100 melhores tempos de chegada na categoria mulheres não existem valores discrepantes.

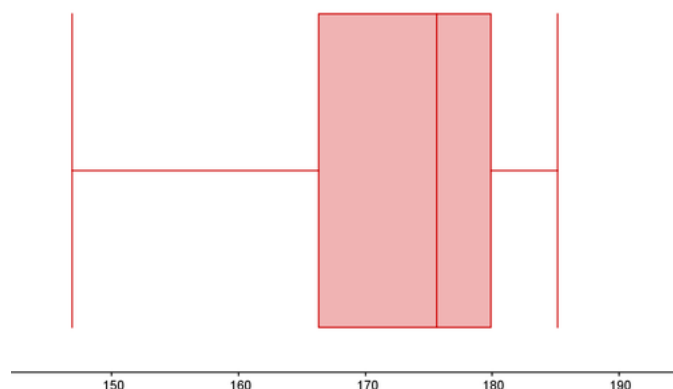


Figura 1.33: Boxplot dos 100 primeiros tempos de chegada (em horas) na maratona de Nova York (categoria mulheres)

Analisando o boxplot da [figura 1.33](#) podemos observar que a distribuição dos 100 melhores tempos na categoria mulheres apresenta assimetria à esquerda, pois

- a) $Q_2 - Q_1 > Q_3 - Q_2$;
- b) $Q_1 - \text{valor mínimo} \gg \text{valor máximo} - Q_3$;
- c) $Q_2 - \text{valor mínimo} \gg \text{valor máximo} - Q_2$ em que o símbolo \gg é usado para representar "bem maior do que".

Podemos concluir o mesmo na categoria homens em que a assimetria à esquerda é mais aparente devido à presença de valores discrepantes à esquerda (lado inferior da distribuição).

Desvios da Média

Considerando o conjunto $\{x_1, x_2, \dots, x_n\}$ com n observações, seja \bar{x} a média deste conjunto. Define-se como um desvio da média, a diferença entre uma observação e a média, a saber,

$$d_i = x_i - \bar{x}, \quad i = 1, 2, \dots, n$$

Na atividade [Estratégia de Investimento](#) os desvios da média, para cada uma das Companhias estão registrados na tabela a seguir.

Semana	Companhia A	Companhia B
1	-0,5	5,5
2	-5,5	-13,5
3	1,5	-9,5
4	-4,5	20,5
5	5,5	15,5
6	1,5	-28,5
7	5,5	5,5
8	-3,5	-19,5
9	5,5	28,5
10	-5,5	-4,5
Soma	0	0

Poderíamos pensar em usar os desvios da média para definir uma medida de dispersão dos dados em relação à média do conjunto, no entanto, a não ser que todos os valores sejam iguais, teremos valores acima da média e valores abaixo da média de tal modo que os desvios da média poderão apresentar sinais positivos ou negativos. Vimos que a média pode ser interpretada como o centro de massa (ponto de equilíbrio) dos dados e, esta propriedade pode ser descrita da seguinte forma: a soma dos desvios da média de qualquer conjunto de dados é sempre nula.

Com os dados da atividade [Estratégia de Investimento](#) você pôde comprovar esta propriedade. Veja na [figura 1.34](#) a ilustração dos desvios da média das duas companhias na qual a linha pontilhada representa a cotação média da companhia e os segmentos em vermelho indicam o tamanho do desvio da média.

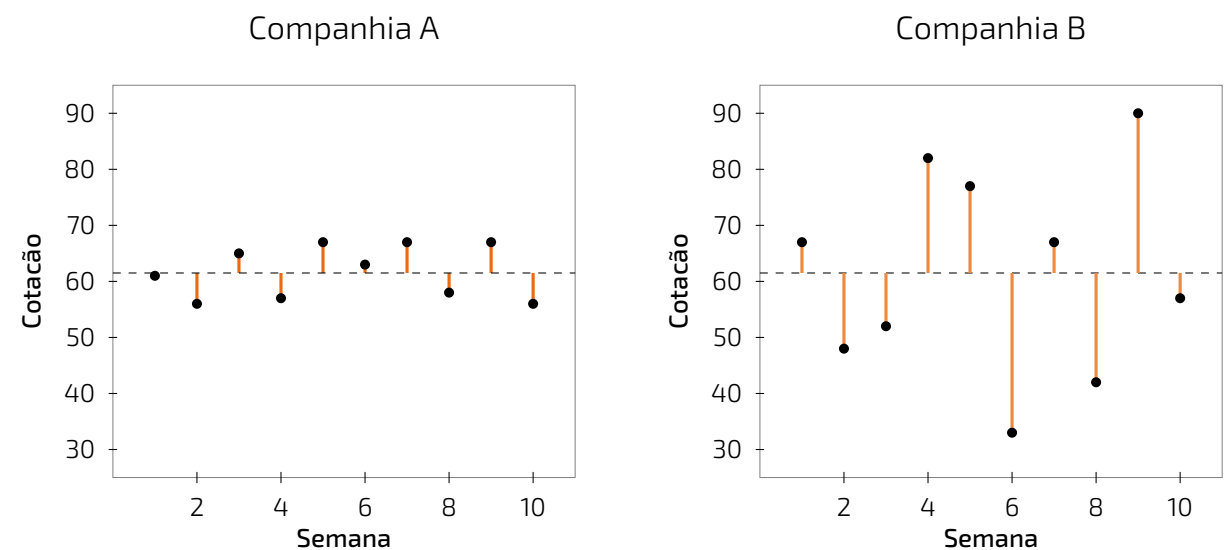


Figura 1.34: Desvios da média das cotações nas companhias A e B

A [figura 1.34](#) reforça a conclusão anterior, da atividade [Estratégia de Investimento](#), de que as cotações da companhia A variam bem menos em torno da média do que as cotações da companhia B.

Em símbolos, a propriedade de que a soma dos desvios da média é sempre nula, pode ser traduzida em

$$\sum_{i=1}^n d_i = \sum_{i=1}^n (x_i - \bar{x}) = 0,$$

qualquer que seja o conjunto $\{x_1, x_2, \dots, x_n\}$

Portanto, não será possível usar a soma dos desvios da média como medida de dispersão de um conjunto de dados, pois ela sempre resultará em zero. Isso se deve ao fato de que a soma em valor absoluto dos desvios de sinal negativo é sempre igual a soma dos desvios de sinal positivo, uma consequência da propriedade da média como centro de massa.

Na Companhia A a soma dos desvios negativos é $-19,5$ e, dos desvios positivos, $19,5$. Na Companhia B a soma dos desvios negativos é $-75,5$ e, dos desvios positivos, $75,5$.

Uma forma de contornar esta situação, de modo a usar os desvios da média para definir uma medida de dispersão, é eliminar o sinal negativo dos desvios da média de tal forma que a soma nula destes desvios transformados ocorra apenas quando todos os dados são iguais, ou seja, quando qualquer medida de dispersão bem definida deve ser nula.

Veja na seção [Para saber mais](#) a demonstração da propriedade de que a soma dos desvios da média é sempre nula.

Desvio Médio Absoluto

Considerando os desvios da média em valor absoluto ($|x_i - \bar{x}|$) observe que todos serão não-negativos tal que a soma dos desvios da média em valor absoluto ($\sum_{i=1}^n |x_i - \bar{x}|$) será nula apenas quando todos os valores do conjunto forem iguais.

Com base na observação anterior, pode-se definir uma medida de dispersão dos dados, considerando todas as observações, chamada desvio médio absoluto (DM) que é definida como a média dos desvios da média tomados em valor absoluto.

Na [tabela 8.16](#) são apresentados os desvios da média em valor absoluto das cotações nas companhias A e B e, a respectiva soma.

Tabela 1.16: Desvios da média em valores absolutos para as companhias A e B

Semana	A	B
1	0,5	5,5
2	5,5	13,5
3	1,5	9,5
4	4,5	20,5
5	5,5	15,5
6	1,5	28,5
7	5,5	5,5
8	3,5	19,5
9	5,5	28,5
10	5,5	4,5
Soma	39,0	151,0

Logo, concluímos que o desvio médio absoluto na companhia A é $DM = \frac{39}{10} = 3,9$ reais e, na companhia B, $DM = \frac{151}{10} = 15,1$ reais, indicando que, de fato, a dispersão em torno da média na companhia B é cerca de 4 vezes maior do que na companhia A com relação ao desvio médio ($15,1/3,9 \approx 3,89$).

De maneira geral, o desvio médio absoluto do conjunto de dados $\{x_1, x_2, \dots, x_n\}$ é

$$DM = \frac{1}{n} \cdot \sum_{i=1}^n |x_i - \bar{x}| = \frac{|x_1 - \bar{x}| + |x_2 - \bar{x}| + \dots + |x_n - \bar{x}|}{n}$$

Variância e Desvio Padrão

Uma outra forma de eliminar o sinal negativo dos desvios da média é elevar ao quadrado cada um deles, tornando-os não-negativos. A variância é definida como uma média dos desvios da média elevados ao quadrado.

$$\text{variância} = \frac{1}{n} \cdot \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n}$$

Na [tabela 1.17](#) são apresentados os desvios da média elevados ao quadrado das cotações nas companhias A e B e, a respectiva soma.

Tabela 1.17: Desvios da média elevados ao quadrado para as companhias A e B

Semana	A	B
1	0,25	30,25
2	30,25	182,25
3	2,25	90,25
4	20,25	420,25
5	30,25	240,25
6	2,25	812,25
7	30,25	30,25
8	12,25	380,25
9	30,25	812,25
10	30,25	20,25
Soma	188,5	3018,5

Logo, concluímos que a variância na companhia A é $\frac{188,5}{10} = 18,85 \text{ reais}^2$ e, na companhia B, $\frac{3018,5}{10} = 301,85 \text{ reais}^2$, indicando que a dispersão em torno da média na companhia B é cerca de 16 vezes maior do que na companhia A com relação à variância ($301,85/18,85 \approx 16$).

Quando lidamos com grande quantidade de dados, calcular a variância usando a definição apresentada será uma tarefa maçante, pois após calcular a média de muitos dados, teremos que calcular cada desvio da média, elevá-los ao quadrado e, finalmente, somá-los. Para conjuntos de dados com mais de 10 elementos será, em geral, muito trabalhoso calcular a variância desta forma. Um modo mais simples para calcular a variância é apresentado a seguir. Pode-se mostrar que o numerador da fórmula da variância é dado por

$$\sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - n \cdot \bar{x}^2$$

Assim, basta conhecer a soma simples ($\sum_{i=1}^n x_i$), para determinar a média \bar{x} , e a soma de quadrados ($\sum_{i=1}^n x_i^2$) para calcular a variância.

A demonstração desta igualdade está na Seção [Para saber mais](#).

Na atividade [Estratégia de Investimento](#), podemos verificar que na companhia A, $\bar{x} = 61,5$ e $\sum_{i=1}^{10} x_i^2 = 38.011$ tal que a variância em A pode ser calculada por

$$\text{variância} = \frac{1}{10} \cdot (38.011 - 10 \cdot 61,5^2) = 18,85 \text{ reais}^2$$

e, na companhia B,

$\bar{x} = 61,5$ e $\sum_{i=1}^{10} x_i^2 = 40.841$ tal que a variância em B pode ser calculada por

$$\text{variância} = \frac{1}{10} \cdot (40.841 - 10 \cdot 61,5^2) = 301,85 \text{ reais}^2$$

Vimos que o desvio médio absoluto da companhia B foi aproximadamente 4 vezes maior do que o da companhia A. Na comparação de variâncias, a variância da companhia B foi cerca de 16 vezes maior do que a da companhia A. Este grande aumento deve-se ao fato de que consideramos os desvios da média elevados ao quadrado no cálculo da variância. Observe que a unidade de medida na variância é o quadrado da unidade de medida das observações. Para retornar à escala de medida das observações, basta extrair a raiz quadrada da variância, levando a definição de desvio padrão, uma medida de dispersão em torno da média, na mesma unidade das observações.

$$\text{desvio padrão} = \sqrt{\text{variância}}$$

No exemplo das cotações, podemos verificar que na companhia A,

$$\text{desvio padrão} = \sqrt{18,85} \approx 4,34 \text{ reais}$$

e, na companhia B,

$$\text{desvio padrão} = \sqrt{301,85} \approx 17,37 \text{ reais}$$

Verifique que o desvio padrão da companhia B é aproximadamente 4 vezes maior do que o da companhia A.

Por que o desvio padrão é preferível ao desvio médio?

Você deve estar se perguntando por que se utiliza o desvio padrão na Estatística em detrimento do desvio médio, cujo cálculo é bem mais simples. A resposta é um tanto complexa para o nível em que estamos, mas ela está associada à necessidade na Estatística de se minimizar estruturas de maneira simples. O desvio médio faz uso da função modular $f(x) = |x|$, que não possui boas propriedades matemáticas para a minimização, por possuir na sua forma uma mudança abrupta em torno de $x = 0$, enquanto que a variância faz uso da função quadrática $f(x) = x^2$, representando parábolas de vértice suave e cujas propriedades analíticas são bem conhecidas. Veja a [figura 1.35](#).

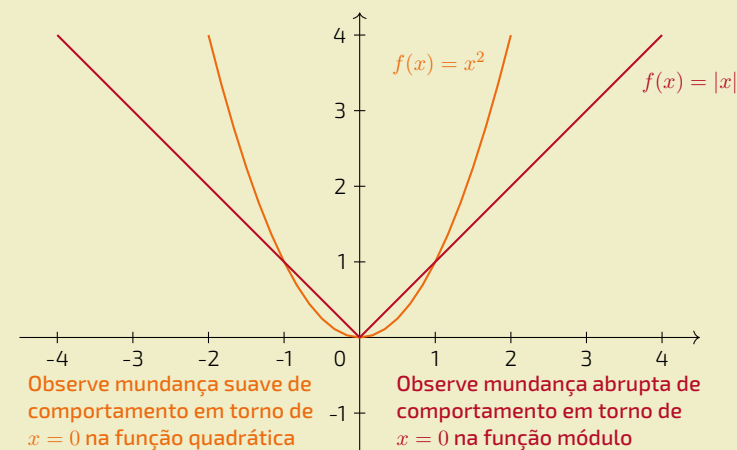


Figura 1.35: Funções modular e quadrática com destaque para o comportamento em torno de $x = 0$.

Muitos problemas de estimação de posição de astros na Física são resolvidos por funções quadráticas por esse motivo, um legado deixado pelo matemático alemão Carl Friedrich Gauss (1777 - 1855) no chamado [Método dos Mínimos Quadrados](#).



Figura 1.36: Carl Friedrich Gauss

Variância populacional e amostral, desvio padrão populacional e amostral

No capítulo **A Natureza Estatística** foram apresentados os conceitos parâmetro e estimador. Parâmetro é uma característica numérica da população, em geral desconhecida; enquanto estimador é uma função dos dados da amostra (subconjunto da população), usada para estimar o parâmetro. Em geral, usam-se letras gregas para denotar parâmetros.

Se dispomos de uma amostra da população, de fato, calculamos a média amostral e a variância amostral (funções dos dados da amostra) e usamos estes resultados como estimativas da média populacional e da variância populacional. Como já foi comentado anteriormente, a média amostral tem boas propriedades como estimador da média populacional. No entanto, é possível mostrar que a variância calculada pela fórmula apresentada no início deste capítulo é um estimador que tende a produzir valores menores do que o valor da variância da população. Dizemos que é um estimador viesado por essa razão.

Para contornar este defeito do estimador, usamos o denominador $n - 1$ no lugar de n . Observe que com isto os valores produzidos serão um pouco maiores, pois o denominador é um pouco menor.

Assim, as expressões que deverão ser usadas quando o conjunto de dados sob estudo é uma amostra da população são dadas por

$$\begin{aligned}\text{variância amostral} &= s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \\ \text{desvio padrão amostral} &= \sqrt{s^2} = s\end{aligned}$$

Na maioria das vezes trabalhamos com amostras. Assim, neste capítulo, salvo menção em contrário, estaremos sempre calculando a variância amostral (s^2) e o desvio padrão amostral (s), mesmo que o termo "amostral" esteja omitido.

Se você estiver trabalhando com uma amostra e usar o denominador n para calcular a variância, isso implicará que você escolheu um estimador viesado, pois tende a produzir estimativas que são menores do que o verdadeiro valor da variância. Observe que se você estiver trabalhando com amostras muito grandes, essa diferença não será importante, pois haverá pouca diferença entre dividir por n ou por $n - 1$.

Expressões que deverão ser consideradas quando o conjunto de dados sob estudo refere-se à população com N elementos:

$$\begin{aligned}\text{variância populacional} &= \sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2 \\ \text{desvio padrão populacional} &= \sqrt{\sigma^2} = \sigma\end{aligned}$$

em que μ representa a média populacional.

EXEMPLO 12 Variância populacional versus variância amostral

Para ilustrar a discussão anterior suponha uma população com apenas 4 elementos, caracterizada pelo conjunto de valores 1, 2, 3, 4.

Nesse caso, temos todas as informações sobre a população. Observe que

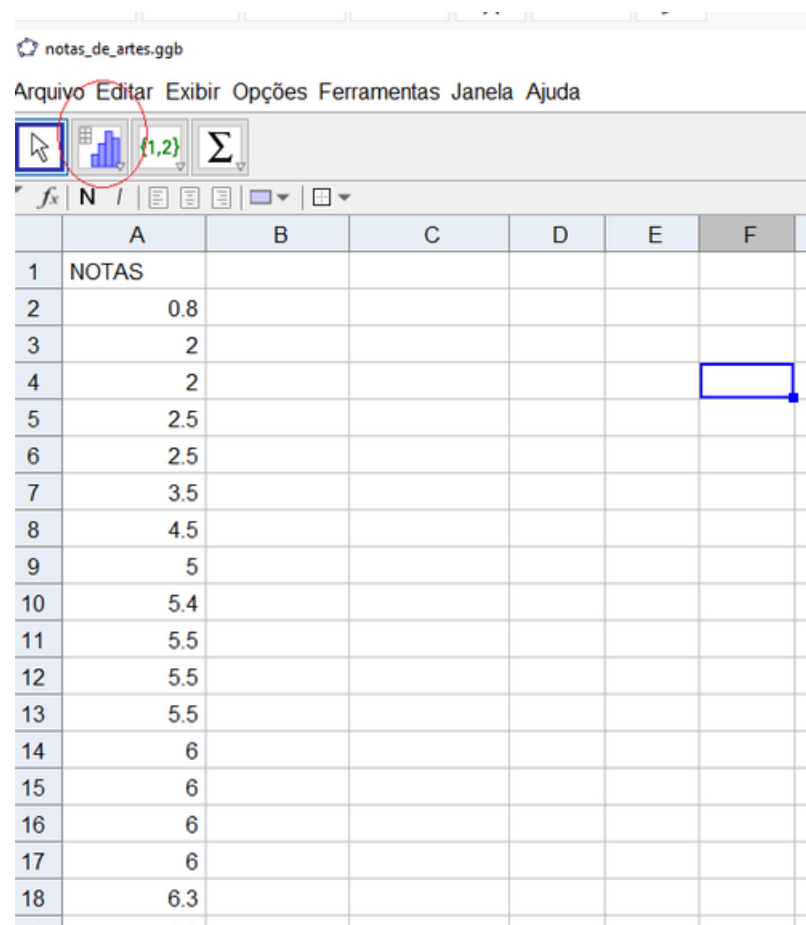
- a) o tamanho da população é $N = 4$, a média da populacional é dada por $\mu = \frac{1+2+3+4}{4} = 2,5$ e a variância populacional é dada por $\sigma^2 = \frac{(1-2,5)^2 + (2-2,5)^2 + (3-2,5)^2 + (4-2,5)^2}{4} = 1,25$.

Suponha que iremos trabalhar com amostras de tamanho $n = 2$ dessa população. Usando amostragem aleatória simples (sorteios com reposição dos elementos da população), observe que existem ao todo 16 amostras possíveis. Para cada amostra possível, vamos calcular a média da amostra e a variância da amostra, considerando os denominadores $n - 1$ e n . Veja os resultados na tabela 8.19.

Amostra	Valores	Média amostral	$s^2(n-1)$	$\sigma^2(n)$
1	1 e 1	1	0	0
2	1 e 2	1,5	0,5	0,25
3	1 e 3	2	2	1
4	1 e 4	2,5	4,5	2,25
5	2 e 1	1,5	0,5	0,25
6	2 e 2	2	0	0
7	2 e 3	2,5	0,5	0,25
8	2 e 4	3	2	1
9	3 e 1	2	2	1
10	3 e 2	2,5	0,5	0,25
11	3 e 3	3	0	0
12	3 e 4	3,5	0,5	0,25
13	4 e 1	2,5	4,5	2,25
14	4 e 2	3	2	1
15	4 e 3	3,5	0,5	0,25
16	4 e 4	4	0	1
	soma	40	20	11




EXEMPLO 13 Medidas de posição e dispersão usando o GeoGebra

Podemos usar o GeoGebra para calcular várias das medidas trabalhadas nesse capítulo, bem como para construir boxplots e histogramas. Considere o conjunto das Notas de Artes sem bonificação da primeira atividade do capítulo. São 35 notas que podem ser inseridas na planilha do GeoGebra, conforme [figura 1.37](#). Lembre-se que o GeoGebra adota o ponto como separador decimal.



notas_de_artes.ggb

Arquivo Editar Exibir Opções Ferramentas Janela Ajuda

   (1,2)

	A	B	C	D	E	F
1	NOTAS					
2	0.8					
3	2					
4	2					
5	2.5					
6	2.5					
7	3.5					
8	4.5					
9	5					
10	5.4					
11	5.5					
12	5.5					
13	5.5					
14	6					
15	6					
16	6					
17	6					
18	6.3					

Figura 1.37: Tela de planilha do GeoGebra com as notas de artes, destacando o botão com imagem de um histograma

Para calcular as medidas ou mesmo construir gráficos desse conjunto de dados, basta selecionar a coluna com os dados e clicar no botão cuja imagem é um histograma. Veja a [figura 1.38](#).

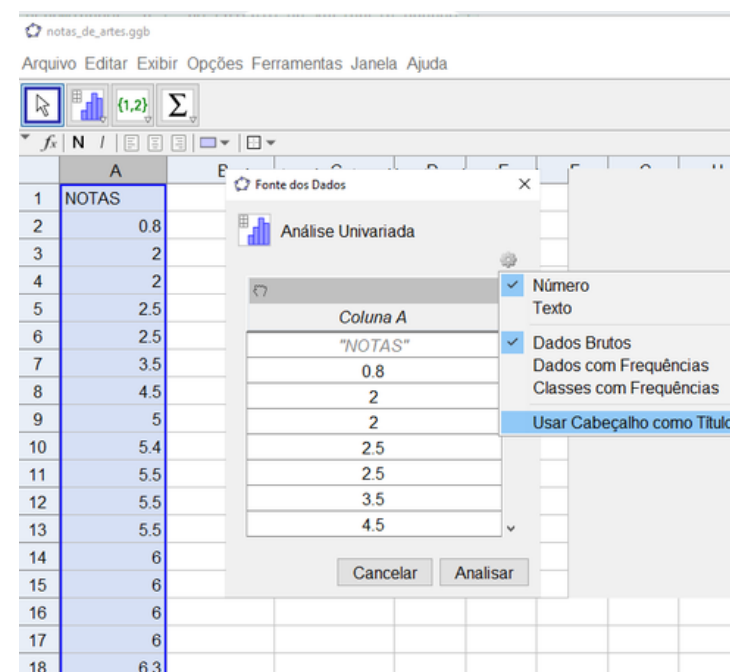


Figura 1.38: Tela de planilha do GeoGebra com as notas de artes, destacando a opção análise univariada

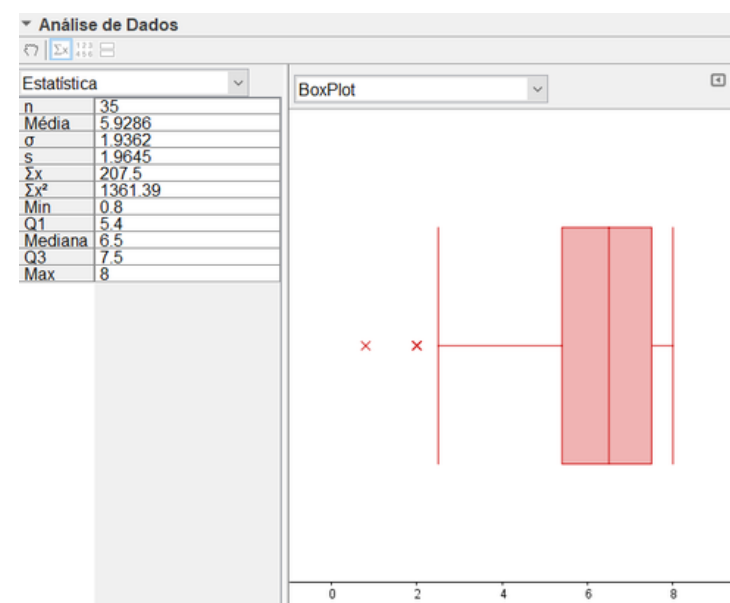


Figura 1.39: Medidas resumo e boxplot do conjunto de notas de Artes usando o GeoGebra

A [figura 1.39](#) mostra a saída do GeoGebra da análise univariada das notas de artes com a construção do respectivo boxplot.

a) n – número de observações consideradas,

- b) média – média aritmética,
- c) σ – desvio padrão obtido a partir do cálculo da variância considerando o denominador n , recomendado quando o conjunto representa a população e não uma amostra da população,
- d) s – desvio padrão amostral calculado a partir da variância amostral usando denominador $n - 1$,
- e) $\sum x$ – soma simples dos valores de um conjunto,
- f) $\sum x^2$ – soma de quadrados dos valores do conjunto,
- g) Min, Q_1 , Mediana Q_3 e Max: medidas do esquema dos cinco números, a saber, mínimo, primeiro, segundo e terceiro quartis e máximo.

Inflação anual

Atividade 6

A seguir são apresentados dados sobre as inflações anuais em dois países. Antes de trabalhar com os dados, vamos tentar explicar o que é inflação. De uma maneira bem simples, pode-se dizer que a inflação é o aumento contínuo nos preços de produtos e serviços. Esse aumento costuma ser avaliado de forma mensal, gerando os índices de inflação, que refletem a variação nos preços.

A inflação pode ser medida de várias formas. O índice oficial de inflação no Brasil é o IPCA (Índice de Preços ao Consumidor Amplo), que mede a variação mensal de preços de produtos considerando o consumo de famílias com renda mensal entre 1 e 40 salários mínimos. O IBGE (Instituto Brasileiro de Geografia e Estatística) é o órgão responsável pela medição e divulgação do IPCA. Veja neste [link](#), um vídeo produzido pelo IBGE, explicando o IPCA.

Foram observadas as inflações anuais de dois países A e B para os anos de 2011 a 2015, conforme tabela a seguir.

Tabela 1.18: Inflação anual

País	2011	2012	2013	2014	2015	Soma
A	2,00%	1,80%	2,10%	2,20%	1,90%	10,00%
B	0,01%	-0,19%	-0,09%	0,21%	0,11%	0,05%

- a) Calcule as médias das inflações anuais dos dois países. Há diferenças entre elas?
- b) Calcule as variâncias das inflações anuais dos dois países, sabendo que para o país A , $\sum_{i=1}^5 x_i^2 = 20,1 (\% ^2)$ e para o país B , $\sum_{i=1}^5 x_i^2 = 0,1005 (\% ^2)$. Há diferença entre elas?

Objetivos Específicos

Inflação anual

- Comparar diferentes conjuntos de dados que apresentam a mesma variância, mas suas médias são diferentes.
- Perceber a necessidade de definir uma medida que avalie a magnitude da variância (desvio padrão) em relação à média.

Sugestões e discussões

Inflação anual

Nesta atividade são apresentados dois conjuntos de dados cujas variâncias são iguais, mas cujas médias são distintas. Pretende-se na discussão, levar à definição de coeficiente de variação, uma medida útil para avaliar a magnitude da variância. Como o dado observado é a inflação anual de um país, a atividade começa com um pequeno texto introdutório sobre inflação.

Solução: Inflação anual

- a) No país A , a inflação média anual, considerando estes 5 anos, é $\bar{x} = \frac{10}{5} = 2,00\%$. No país B , a inflação média anual, considerando estes 5 anos, é $\bar{x} = 0,055 = 0,01\%$. Logo, as inflações anuais médias dos dois países são bem diferentes.
- b) Usando a fórmula simplificada para o cálculo da variância, temos, para o país A , $s^2 = \frac{1}{5-1} (20,1 - 5 \cdot 2^2) = 0,025 (\% ^2)$. Para o país B , temos $s^2 = \frac{1}{5-1} (0,1005 - 5 \cdot 0,012) = 0,025 (\% ^2)$. Logo, as variâncias destes dois conjuntos de inflações anuais são iguais e, consequentemente, os desvios padrões também são iguais.
- c) Verifique que os cinco desvios da média produzidos pelos dados dos dois países são idênticos, levando à mesma variância (mesmo desvio padrão). No entanto, a média no país A (2%) é bem maior do que no país B , indicando uma variação relativa à média menos forte do que no país B . A seguir, será definido o coeficiente de variação, que avalia essa propriedade de dispersão relativa à média. Observe que o desvio padrão para os dois países é $\sqrt{0,025} \approx 0,16\%$ de modo que no país A o desvio padrão corresponde a 8% da média ($\frac{0,16}{2} = 0,08$), enquanto que no país B , corresponde a 1.600% da média ($\frac{0,16}{0,01} = 16$), ou seja, a flutuação em torno da média é muito mais forte no país B .

- c) Qual dos países apresenta maior variação inflacionária quando comparada à média inflacionária?

Coeficiente de variação

Nem sempre uma variância pequena (e consequentemente desvio-padrão pequeno) significa pouca dispersão. Tampouco uma variância grande é sempre indicador de alta dispersão. Esses valores podem ser altos ou baixos devido à magnitude (ordem de grandeza) dos dados observados. Se medimos observações em microscópio, por exemplo, teremos inevitavelmente valor numericamente baixo de variância, podendo no entanto haver alta dispersão dos dados no nível microscópico. Da mesma maneira, ao medir os produtos internos brutos brasileiros em dólares em vários anos teremos valores observados de alta magnitude, gerando variância numericamente grande, mas não necessariamente indicando alta dispersão.

Na atividade [Inflação anual](#), estudamos dois conjuntos de dados que apresentam médias diferentes, mas variâncias iguais. Podemos dizer que o impacto da variância em relação à média é o mesmo para os dois conjuntos? Comparando o valor do desvio padrão de cerca de 0,16% à média do país A de 2%, vemos que ele é pequeno em relação à média. Comparando o valor do desvio padrão 0,16% em relação à média do país B de 0,01%, vemos que ele é muito grande em relação à média. Neste caso dizemos que no país A os dados apresentam variação relativa em torno da média pequena. Já, no país B, os dados apresentam variação relativa em torno da média grande.

O coeficiente de variação é uma medida usada para calcular a variação relativa dos dados de um conjunto em torno da média: quanto maior seu valor, maior é a variação relativa em torno da média.

Coeficiente de variação é a razão entre o desvio padrão e a média. Em geral, ele é descrito em termos percentuais.

O coeficiente de variação amostral, em termos percentuais, é calculado por

$$CVA = \frac{s}{\bar{x}} \cdot 100\%$$

em que s é o desvio padrão amostral e \bar{x} é a média amostral.

Esta expressão é usada quando dispomos de uma amostra da população. Se, dispomos dos dados da população, então o coeficiente de variação populacional é dado por

$$CVP = \frac{\sigma}{\mu} \cdot 100\%$$

em que σ é o desvio padrão populacional e μ é a média populacional.

Observe que o coeficiente de variação só é definido para conjuntos cuja média é diferente de zero.

Uma regra empírica para avaliar frequências de valores em intervalos em torno da média que pode ser útil, é obtida a partir das propriedades de um modelo teórico conhecido como densidade normal de probabilidades. Entre várias propriedades desta densidade, destaca-se que ela é simétrica e unimodal tal que média, mediana e moda são iguais. Veja na [figura 1.40](#) 8.39 uma ilustração da densidade normal com média μ e desvio padrão σ , também conhecida como a curva em forma de sino.

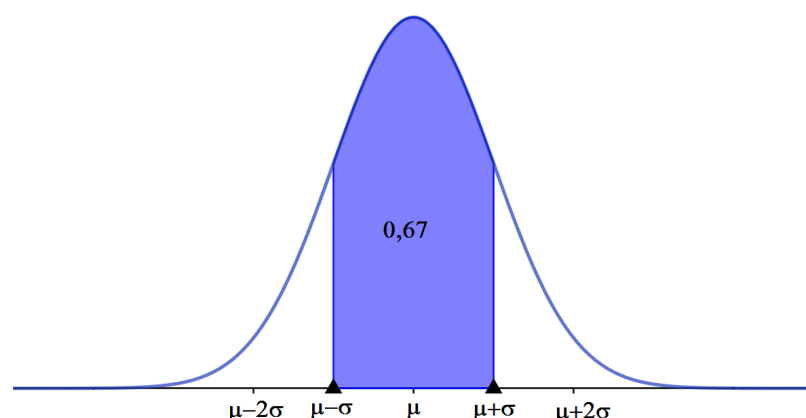


Figura 1.40: Densidade Normal com região colorida no intervalo entre $\mu - \sigma$ e $\mu + \sigma$, cuja área corresponde a aproximadamente 0,67 da área total igual a 1.

A regra empírica estabelece que em distribuições aproximadamente simétricas para as quais são raros ou mesmo não existem valores discrepantes,

- a) a frequência relativa de valores no intervalo $[\bar{x} - s; \bar{x} + s]$ é aproximadamente 67%,
- b) a frequência relativa de valores no intervalo $[\bar{x} - 2 \cdot s; \bar{x} + 2 \cdot s]$ é aproximadamente 95%.

No caso dos dados da atividade [A Maratona](#) vimos que não existem valores discrepantes na categoria mulheres, mas a distribuição apresenta assimetria à esquerda. Ainda assim, contando frequência de observações que nos intervalos definidos por $[\bar{x} - s; \bar{x} + s]$ e $[\bar{x} - 2 \cdot s; \bar{x} + 2 \cdot s]$, obtém-se 69% e 93%, respectivamente. Observe que estes valores estão próximos dos valores estipulados pela regra empírica, mesmo com este conjunto apresentando assimetria à esquerda.

PRATICANDO

MEDIDAS DE DISPERSÃO

Comparação de conjuntos de dados

Atividade 7

Para realizar esta atividade será necessário coletar dois conjuntos de dados da mesma natureza, correspondentes a grupos distintos, os quais queremos comparar. Por exemplo:

- alturas de homens e mulheres;
- alturas de alunos de 1º e de 9º ano do Ensino Fundamental;
- notas de disciplinas distintas;
- notas de turmas distintas na mesma disciplina;
- medições de produtos naturais: comprimento das folhas de vegetais (alface, rúcula, etc) comprados em lojas distintas, altura de árvores ou plantas similares locais da cidade distintos;

Objetivos Específicos

Comparação de conjuntos de dados

Comparar diferentes distribuições de uma mesma variável quando separada por grupos.

Sugestões e discussões

Comparação de conjuntos de dados

Nesta atividade serão coletados dados de uma mesma variável que possa ser separada em grupos, com o intuito de comparar as suas medidas de posição e dispersão. Sugerem-se algumas opções, dependendo do tamanho da turma e do contexto escolar, podem até ser escolhidas variáveis distintas para grupos pequenos de alunos, por exemplo, um grupo trabalha com as médias de Matemática, outro grupo trabalha com alturas, etc.

Uma vez coletados os dados, serão calculadas suas medidas de posição e dispersão e comparadas, tentando orientar os estudantes a comentar as observações e não apenas fazer os cálculos. Para a realização dos cálculos deve ser usado suporte tecnológico: calculadoras, aplicativos, etc.

O intuito é dar uma perspectiva para os estudantes da forma em que a estatística é utilizada na ciência para responder perguntas como:

- Uma determinada espécie vegetal cresce melhor perto de uma fonte de água ou longe da mesma? Na sombra de uma árvore ou recebendo luz direta do sol?
- As meninas são mais altas que os meninos numa certa idade? Acontece o mesmo em todas as idades?

De forma ideal, pode ser formulada primeiro a pergunta, e depois coletados os dados, apelando a informações encontradas num artigo científico ou numa publicação de jornal, com o intuito de tentar contrastar uma afirmação dada num texto com dados coletados diretamente.

1

CAPÍTULO

MEDIDAS DE POSIÇÃO E DISPERSÃO

entre outros que podem ser escolhidos dependendo da região e dos recursos disponíveis na escola.

No seu caderno ou em uma planilha eletrônica, registre os dados coletados, como indicado no modelo de tabela a seguir, lembrando que quanto mais dados você coletar com os critérios definidos, os resultados do experimento terão maior chance de refletir a realidade.

Variável: altura em cm	
Turma A	Turma B
155	165
168	159
⋮	⋮

Para calcular as medidas de posição e dispersão, utilize de forma cuidadosa as fórmulas apresentadas. De forma alternativa, você pode digitar os dados no [Aplicativo de medidas de posição e dispersão do Livro Aberto](#) e obter as medidas resumo dos dados.

Tabela 1.19: Registre os seus resultados

Nome da categoria	Grupo A	Grupo B
Mínimo ($x_{(1)}$)		
Máximo ($x_{(n)}$)		
Média		
Q_1		
Mediana		
Q_3		
Amplitude amostral (R)		
Dist. entre quartis (DQ)		
Desvio médio absoluto (DM)		
Variância amostral (s^2)		
Desvio padrão amostral (s)		

Sugere-se a construção dos histogramas para comparar os dois grupos. Você pode usar o GeoGebra para esta construção.

- a) Discuta as suas observações com a turma. Lembre-se de interpretar as medidas de dispersão e não apenas as de posição, que informação adicional oferecem?
- b) Analisando os dois conjuntos de dados obtidos, que medida de posição você julga mais adequada para resumir a informação do conjunto? Por quê?
- c) Os resultados que você obteve parecem refletir a realidade? Existe algum resultado científico que suporte estas observações? Consulte professores de outras áreas sobre suas conclusões.

Aproximação para o valor do desvio padrão amostral

Atividade 8

Nos conjuntos de dados, quando não há valores atípicos (valores muito altos ou muito baixos em relação à maior parte dos valores no conjunto), a maior parte dos valores se situará no intervalo centrado na média distando 2 desvios padrões à esquerda e à direita da média. A partir desta suposição, pode-se obter uma fórmula para estimar o valor do desvio padrão amostral s .

$$\begin{cases} \text{Max} = x_{(n)} \approx \bar{x} + 2 \cdot s \\ \text{Min} = x_{(1)} \approx \bar{x} - 2 \cdot s \end{cases}$$

Tomando a diferença das primeiras expressões apresentadas, obtemos

$$R = \text{Max} - \text{Min} \approx 4 \cdot s$$

tal que

$$s \approx \frac{R}{4}$$

- a) Use esta fórmula para estimar o valor do desvio padrão amostral dos dados da atividade [Notas de Arte](#) e compare o valor obtido com o desvio padrão amostral s . Use os dados na tabela a seguir, produzidos pelo GeoGebra.

Estatística	
n	35
Média	5,9286
σ	1,9362
s	1,9645
Σx	207,5
Σx^2	1361,39
Min	0,8
Q_1	5,4
Mediana	6,5
Q_3	7,5
Max	8

Tabela 1.20: Estatísticas resumo das Notas de Artes

- b) Idem para estimar o valor do desvio padrão amostral dos dados da atividade [A Maratona](#) e compare o valor obtido com o desvio padrão amostral s . Use os dados na figura a seguir, produzidos pelo GeoGebra.

Objetivos Específicos

Aproximação para o valor do desvio padrão amostral

- Calcular uma aproximação grosseira do desvio padrão amostral em função da amplitude amostral.
- Comparar os resultados obtidos pela fórmula de aproximação com os valores exatos do desvio padrão amostral.
- Avaliar o valor obtido do desvio padrão, comparando-o com a aproximação.

Sugestões e discussões

Aproximação para o valor do desvio padrão amostral

Nesta atividade pretende-se apresentar interpretações para o desvio padrão, evitando que ele torne-se apenas uma medida a mais sem muito sentido para o aluno. Além disso, esta atividade pode ser útil para o aluno avaliar se ele calculou corretamente um desvio padrão. É muito comum, mesmo informando-se somatórios e permitindo-se o uso de calculadoras, a produção de resultados incorretos para a variância e, conseqüentemente, para o desvio padrão. Uma ferramenta útil pode ser comparar o valor obtido do desvio padrão com a razão $\frac{R}{4}$. Se a diferença for grande (mais de 50% do valor obtido de s) recomenda-se verificar novamente o cálculo de s .

Solução: Aproximação para o valor do desvio padrão amostral

- a) Da [tabela 1.20](#) vemos que $s \approx 1,96$ e que $R = 8 - 0,8 = 7,2$. Pela fórmula apresentada temos $s \approx \frac{7,2}{4} = 1,8$. Comparando o valor aproximado de $s(1,8)$ com o valor calculado de $s(1,96)$ vemos que a aproximação é um pouco menor do que o valor de s . O erro percentual cometido por esta aproximação corresponde a 8% do valor de s , pois $\frac{|1,8 - 1,96|}{1,96} \approx 0,08$.

Solução: Aproximação para o valor do desvio padrão amostral

b) Da [tabela 1.21](#), para a categoria homens, vemos que $s \approx 7,70$ minutos e que $R = 158,33 - 130,88 = 27,45$. Pela fórmula apresentada temos $s \approx \frac{27,45}{4} \approx 6,86$ minutos. Comparando o valor aproximado de $s(6,86)$ com o valor calculado de $s(7,70)$ vemos que a aproximação é um pouco menor do que o valor de s . O erro percentual cometido por esta aproximação corresponde a cerca de 11% do valor de s , pois $\frac{|6,86-7,70|}{7,70} \approx 0,11$.

Da [tabela 1.21](#), para a categoria mulheres, vemos que $s \approx 11,13$ minutos e que $R = 185,15 - 146,88 = 38,27$. Pela fórmula apresentada temos $s \approx 38,274 = 9,57$ minutos. Comparando o valor aproximado de $s(9,57)$ com o valor calculado de $s(11,13)$ vemos que a aproximação é um pouco menor do que o valor de s . O erro percentual cometido por esta aproximação corresponde a cerca de 14% do valor de s , pois $\frac{|9,57-11,13|}{11,13} \approx 0,14$.

c) Da [tabela 1.22](#) vemos que, para a companhia A, $s \approx 4,5765$ e que $R = 67 - 56 = 11$. Pela fórmula apresentada temos $s \approx \frac{11}{4} = 2,75$. Comparando o valor aproximado de $s(2,75)$ com o valor calculado de $s(4,5765)$ vemos que a aproximação é menor do que o valor de s . O erro percentual cometido por esta aproximação corresponde a 40% do valor de s , pois $\frac{|2,75-4,5765|}{4,5765} \approx 0,4$.

Da [tabela 1.22](#) vemos que, para a companhia B, $s \approx 17,3738$ e que $R = 90 - 33 = 57$. Pela fórmula apresentada temos $s \approx \frac{57}{4} = 14,25$. Comparando o valor aproximado de $s(14,25)$ com o valor calculado de $s(17,3738)$ vemos que a aproximação é menor do que o valor de s . O erro percentual cometido por esta aproximação corresponde a 18% do valor de s , pois $\frac{|14,25-17,3738|}{17,3738} \approx 0,18$.

Objetivos Específicos**Frequência de valores no intervalo centrado na média**

Calcular a frequência relativa de dados que caem no intervalo centrado na média mais ou menos dois desvios padrões.

Sugestões e discussões**Frequência de valores no intervalo centrado na média**

Esta atividade será útil no final da próxima seção que trata da construção do boxplot e seus resultados serão retomados adiante. Além disso, será útil na verificação da afirmação feita na atividade anterior de que quando não há valores atípicos, a grande maioria dos dados situa-se entre a média mais ou menos dois desvios padrões.

Nota 2

	Homens	Mulheres
n	100	100
Média	150,6942	171,9166
σ	7,6617	11,075
s	7,7003	11,1308
Min	130,88	146,88
Q_1	148,37	166,31
Mediana	152,995	175,625
Q_3	156,66	158,33
Max	158,33	185,15

Tabela 1.21: Estatísticas resumo dos 100 melhores tempos para homens e mulheres - Maratona de Nova Iorque/2017

c) Idem para estimar o valor de desvio padrão amostral dos dados da atividade [Estratégia de Investimento](#). Use os dados na figura a seguir, produzidos pelo GeoGebra.

Companhia A		Companhia A	
n	10	n	10
Média	61.5	Média	61.5
σ	4.3417	σ	17.3738
s	4.5765	s	18.3136
Σx	615	Σx	615
Σx^2	28011	Σx^2	40841
Min	56	Min	33
Q_1	57	Q_1	48
Mediana	62	Mediana	62
Q_3	67	Q_3	77
Max	67	Max	90

Tabela 1.22: Estatísticas resumo das cotações das ação nas Companhias A e B.

Frequência de valores no intervalo centrado na média**Atividade 9**

Para os conjuntos de dados considerados na atividade [Aproximação para o Valor do Desvio Padrão Amostral](#), calcule a frequência absoluta de dados que estão no intervalo $[\bar{x} - 2 \cdot s, \bar{x} + 2 \cdot s]$ e comente sobre os resultados obtidos.

Comparação das bonificações nas Notas de Artes

Atividade 10

Vamos retornar à atividade [Notas de Arte](#) e às duas possibilidades de bonificação das notas: acrescentar um ponto a todos os alunos ou aumentar em 20% a nota de cada aluno. Suponha, que o professor deseja que o resultado geral de sua turma apresente o menor coeficiente de variação. Partindo deste ponto de vista, qual das duas possibilidades é mais interessante para o professor adotar?

Para facilitar, use as informações a seguir.

Tabela 1.23: Dados sobre as somas simples e somas de quadrados das notas antes da bonificação (antes), após serem acrescidas de um ponto (1 pt) e após serem aumentadas em 20% (20%)

$n = 35$	Antes	1 pt	20%
$\sum x$	207,5	242,5	249,0
$\sum x^2$	1361,39	1811,39	1960,402

Modalidades da maratona de Nova Iorque 2017

Atividade 11

Na [figura 1.41](#) e na [tabela 1.24](#) a seguir apresentam-se

- a) os boxplots dos 100 melhores tempos para na maratona de Nova Iorque no ano de 2017 para as categorias homens e mulheres e os boxplots dos concluintes nas categorias cadeira de rodas (51 ao todo) e triciclo de mão (69 ao todo) e
- b) as medidas resumo calculadas pelo GeoGebra para as quatro categorias.

Para construir os quatro gráficos na mesma escala, todos os tempos foram convertidos para horas.

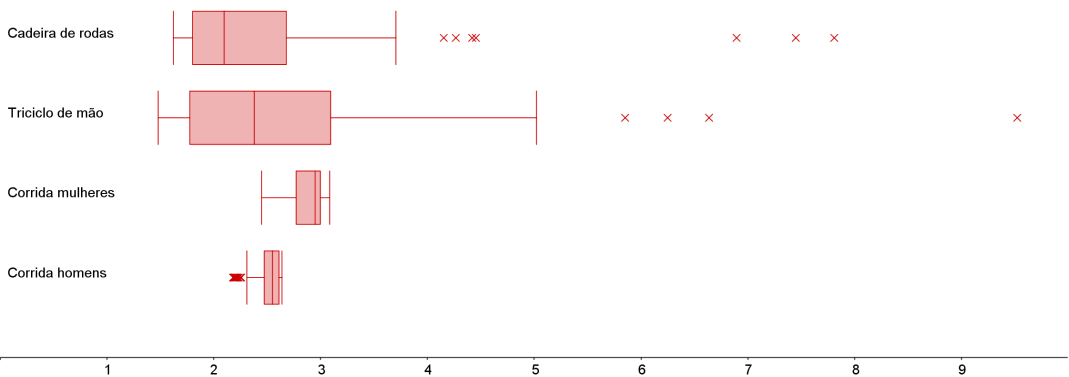


Figura 1.41: Boxplots para os 100 melhores tempos das categorias homens e mulheres e dos melhores tempos das categorias cadeira de rodas e triciclo de mão da maratona de Nova Iorque/2017

Objetivos Específicos

Comparação das bonificações nas Notas de Artes

Avaliar o efeito no coeficiente de variação de um conjunto de dados quando realizamos transformações de adição de uma constante e de multiplicação por uma constante.

Sugestões e discussões

Comparação das bonificações nas Notas de Artes

Nesta atividade pretende-se retornar ao item e) da [Notas de Arte](#) quando foi perguntado ao estudante o que ele achava melhor: ganhar um ponto ou um acréscimo de 20% em sua nota. A ideia será propor a mesma pergunta de um ponto de vista do professor, que prefere que a distribuição das notas apresente o menor coeficiente de variação.

Solução: Comparação das bonificações nas Notas de Artes

O professor deverá escolher o aumento de um ponto para cada estudante, pois esta bonificação acarretará num coeficiente de variação menor, implicando em maior homogeneidade da turma em relação à média, conforme os cálculos a seguir.

Considerando o acréscimo de um ponto a todos os alunos temos que a média passa a ser $\bar{x} = \frac{242,5}{35} \approx 6,93$. A variância, calculada por s^2 é dada por $\frac{1811,39 - 35 \cdot 6,93^2}{35 - 1} \approx 3,84$ e, o desvio padrão, $s \approx 1,96$. Assim, o coeficiente de variação da turma, resultante desta bonificação será dado por $CV = \frac{1,96}{6,93} \cdot 100 \approx 28\%$.

Considerando um aumento de 20% para cada nota temos que a média passa a ser $\bar{x} = \frac{249,0}{35} \approx 7,11$. A variância, calculada por s^2 é dada por $\frac{1960,402 - 35 \cdot 7,11^2}{35 - 1} \approx 5,56$ e, o desvio padrão, $s \approx 2,36$. Assim, o coeficiente de variação da turma, resultante desta bonificação será dado por $CV = \frac{2,36}{7,11} \cdot 100 \approx 33\%$.

Objetivos Específicos

Modalidades da maratona de Nova Iorque 2017

Comparar diferentes conjuntos de dados, considerando a mesma variável.

Sugestões e discussões

Modalidades da maratona de Nova Iorque 2017

Nesta atividade retomaremos as quatro categoriais da maratona de Nova Iorque para usar o boxplot como esquema gráfico para auxiliar na comparação dos resultados para as diferentes categorias, a saber, homens, mulheres, cadeira de rodas e triciclo de mão. Os dados estão disponíveis neste [link](#).

Nota 3

	n	Média	σ	s	Min	Q_1	Mediana	Q_3	Max
Corrida homens	100	2.5116	0.1277	0.1383	2.1814	2.4729	2.55	2.6111	2.6389
Corrida mulheres	100	2.8698	0.1858	1.867	2.4481	2.7718	2.9493	2.9982	3.0858
Triciclo de mão	69	2.7338	1.3679	1.3779	1.48	1.7764	3.3797	3.0946	9.4206
Cadeira de rodas	51	2.5855	1.4069	1.4209	1.6225	1.8025	2.0978	2.6794	7.8081

Tabela 1.24: Medidas resumo para as quatro categorias da maratona de Nova Iorque/2017

- Qual das modalidades apresentou maior dispersão?
- Qual(ais) modalidade(s) apresentaram valores atípicos?
- Como você avalia, em relação à simetria, cada uma das distribuições?
- Faça uma análise comparativa das distribuições das modalidades homens e mulheres, usando a [figura 1.42](#).

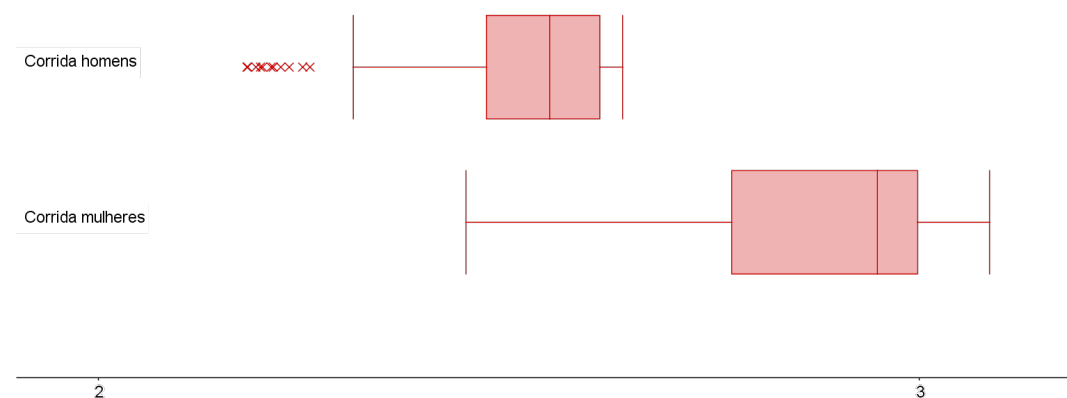


Figura 1.42: Boxplot dos 100 melhores tempos para homens e mulheres na maratona de Nova Iorque/2017

- Faça uma análise comparativa das distribuições das modalidades cadeira de rodas e triciclo de mão.

PARA SABER + CÁLCULOS PARA DADOS AGRUPADOS

Média

Considere um conjunto de n dados agrupados em c intervalos de classe.

Sejam $\tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_c$ os pontos médios dos c intervalos de classe e, n_1, n_2, \dots, n_c , as frequências absolutas dos c intervalos de classe, respectivamente. Lembre que o ponto médio de um intervalo de classe corresponde à média aritmética dos extremos do intervalo. Neste caso a média é calculada por

$$\text{média} = \bar{x} = \frac{n_1 \cdot \tilde{x}_1 + n_2 \cdot \tilde{x}_2 + \dots + n_c \cdot \tilde{x}_c}{\underbrace{n_1 + n_2 + \dots + n_c}_{=n}} = \frac{1}{n} \cdot \sum_{i=1}^c n_i \cdot \tilde{x}_i$$

Denotando por $f_i = \frac{n_i}{n}$ a frequência relativa do i -ésimo intervalo classe, temos

$$\text{média} = \bar{x} = f_1 \cdot \tilde{x}_1 + f_2 \cdot \tilde{x}_2 + \dots + f_c \cdot \tilde{x}_c = \sum_{i=1}^c f_i \cdot \tilde{x}_i$$

Quando os dados estão agrupados em intervalos de classe, a média é calculada como uma média ponderada dos pontos médios das classes em que os pesos são dados pelas frequências absolutas (ou relativas) das classes.

Mediana

Para obter uma aproximação da mediana quando os dados estão agrupados, deve-se primeiro determinar as frequências acumuladas (absoluta ou relativa) associadas a cada intervalo. Se as frequências forem absolutas, deve-se identificar em qual intervalo encontra-se a observação na posição central ($\frac{n+1}{2}$ se n for ímpar), ou as duas posições centrais ($\frac{n}{2}$ e $\frac{n}{2} + 1$) se n for par. Depois, como foi sugerido anteriormente, tome como mediana o ponto médio do intervalo de classe que compreende a(s) posição(ões) central(is).

Variância e desvio padrão amostrais

$$s^2 = \frac{1}{n-1} \sum_{i=1}^c n_i (\tilde{x}_i - \bar{x})^2 = \frac{1}{n-1} \left(\sum_{i=1}^c n_i \tilde{x}_i^2 - n \bar{x}^2 \right)$$

em que \bar{x} é a média amostral. Se conhecemos apenas as frequências relativas do conjunto de dados, também podemos calcular a variância amostral por $s^2 = \sum_{i=1}^c f_i (\tilde{x}_i - \bar{x})^2 = \sum_{i=1}^c f_i \tilde{x}_i^2 - \bar{x}^2$.

O desvio padrão amostral é, então, calculado por $s = \sqrt{s^2}$.

Objetivos Específicos

Medidas para dados agrupados

Determinar a média, mediana e variância, a partir de um histograma.

Sugestões e discussões

Medidas para dados agrupados

Esta atividade pretende mostrar a utilidade das fórmulas apresentadas nesta seção para obter medidas de posição e dispersão, quando não se conhecem os dados separadamente.

Nota 4

Medidas para dados agrupados

Atividade 12

Os resultados obtidos na prova de seleção para vagas de estágio numa empresa estão representados no histograma a seguir.

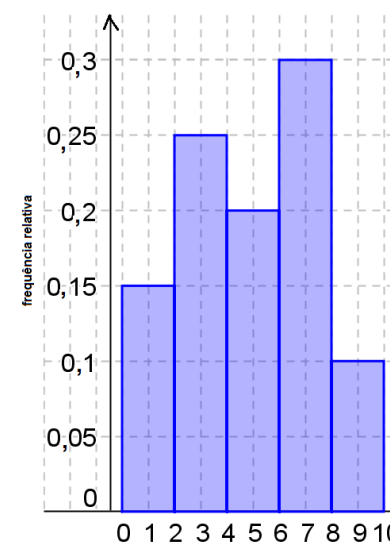


Figura 1.43: Histograma das notas na prova de seleção para vagas de estágio

- Com base neste histograma, calcule a média, a variância, a mediana, a moda, o primeiro quartil e o terceiro quartil.
- Usando a informação do histograma, faça um esboço do boxplot destes dados.

Um método para a determinação dos quartis

Existem métodos diferentes para determinar os quartis de um conjunto $\{x_1, x_2, \dots, x_n\}$ de n observações. Um método simples será descrito a seguir.

Tome Q_1 como o valor correspondente à posição $\frac{n+1}{4}$ depois de ordenar os dados.

Tome Q_2 como a mediana do conjunto de dados, calculada pelo método apresentado para o cálculo da mediana.

Tome Q_3 como o valor correspondente à posição $\frac{3n+1}{4}$ depois de ordenar os dados.

Se os resultados de $\frac{n+1}{4}$ e $\frac{3n+1}{4}$ não forem números inteiros, arredonde-os para o inteiro mais próximo. Se a parte decimal do resultado destas operações for 0,5; calcule a média dos dois valores nas posições correspondentes. Por exemplo, suponha $n = 21$ tal que $(21 + 1)/4 = 5,5$. Assim, neste caso, para obter o primeiro quartil, calcule a média dos valores nas posições 5 e 6.

Vamos voltar aos dados da atividade [Notas de Arte](#). Como $n = 35$, para o primeiro quartil tomaremos o valor da posição $\frac{35+1}{4} = 9$, a saber, $Q_1 = 5$, já vimos que a mediana é 6,5 e, para o terceiro quartil tomaremos o valor da posição $\frac{3 \cdot 35 + 1}{4} = 26,5$. Como 26,5 é equidistante das posições 26 e 27, tomaremos o terceiro quartil como a média dos dois valores nestas duas posições, a saber, $Q_3 = \frac{7,3 + 7,5}{2} = 7,4$. Logo, podemos dizer que na turma cerca de 25% das notas foram menores do que 5 e cerca de 25% das notas foram maiores do que 7,4.

Soma dos desvios da média

Considerando o conjunto $\{x_1, x_2, \dots, x_n\}$ com n observações, seja \bar{x} a média deste conjunto. Define-se como um desvio da média, a diferença entre uma observação e a média, a saber,

$$d_i = x_i - \bar{x}, \quad i = 1, 2, \dots, n$$

Uma propriedade dos desvios da média é dada por

$$\sum_{i=1}^n d_i = \sum_{i=1}^n (x_i - \bar{x}) = 0,$$

qualquer que seja o conjunto $\{x_1, x_2, \dots, x_n\}$.

Demonstração:

$$\sum_{i=1}^n (x_i - \bar{x}) = (x_1 - \bar{x}) + (x_2 - \bar{x}) + \dots + (x_n - \bar{x}) = \underbrace{(x_1 + x_2 + \dots + x_n)}_{=n \cdot \bar{x}} - n \cdot \bar{x} = 0$$

lembrando que $\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n}$.

Veja um exemplo na seção sub-desvios da média.

Fórmula para o cálculo da variância amostral

Vimos que a variância amostral do conjunto de dados $\{x_1, x_2, \dots, x_n\}$ é definida por

$$s^2 = \frac{1}{n-1} \cdot \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n-1}$$

De fato, é possível mostrar que

$$s^2 = \frac{1}{n-1} \cdot \left(\sum_{i=1}^n x_i^2 - n \cdot \bar{x}^2 \right)$$

Demonstração: Expandindo a soma no numerador da fórmula da variância é possível concluir que

$$\sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - n \cdot \bar{x}^2$$

Lembre que $(x_i - \bar{x})^2 = x_i^2 - 2 \cdot \bar{x} \cdot x_i + \bar{x}^2$. Assim,

$$\begin{aligned} \sum_{i=1}^n (x_i - \bar{x})^2 &= \sum_{i=1}^n (x_i^2 - 2 \cdot \bar{x} \cdot x_i + \bar{x}^2) = \\ &= (x_1^2 - 2 \cdot \bar{x} \cdot x_1 + \bar{x}^2) + (x_2^2 - 2 \cdot \bar{x} \cdot x_2 + \bar{x}^2) + \dots + (x_n^2 - 2 \cdot \bar{x} \cdot x_n + \bar{x}^2) \end{aligned}$$

Como a soma é finita, podemos reunir os termos semelhantes, obtendo

$$\begin{aligned}
 \sum_{i=1}^n (x_i - \bar{x})^2 &= \\
 (x_1^2 + x_2^2 + \cdots + x_n^2) &\underbrace{- 2 \cdot \bar{x} \cdot (x_1 + x_2 + \cdots + x_n)}_{=-2 \cdot n \cdot \bar{x}^2} \underbrace{+ n \cdot \bar{x}^2}_{=n \cdot \bar{x}^2} = \\
 \sum_{i=1}^n x_i^2 - n \cdot \bar{x}^2 &=
 \end{aligned}$$

Vamos voltar aos dados da atividade [Notas de Arte](#). Temos $n = 35$, $\sum_{i=1}^{35} x_i = 207,5$ e $\sum_{i=1}^{35} x_i^2 = 1361,39$ tal que $\bar{x} = \frac{207,5}{35} \approx 5,93$ e

$$s^2 = \frac{1}{34} (1361,39 - 35 \cdot 5,93^2) \approx 3,8417$$

tal que o desvio padrão amostral é, aproximadamente, 1,96.

Análises exploratórias bivariadas

Em geral, nos estudos de investigação estatística, observam-se mais de uma variável para cada unidade. Quando faz sentido, investigam-se possíveis relações entre as variáveis observadas e o primeiro passo nessa investigação é a visualização dos dados observados em gráficos. Apresentaremos aqui sugestões de representações gráficas adequadas para pares de variáveis sob investigação.

Os gráficos adequados para investigar possíveis associações entre duas variáveis dependem da natureza das variáveis sob estudo que podem ser ambas quantitativas, ambas qualitativas ou uma qualitativa e outra qualitativa. Lembre-se que classificamos uma variável em quantitativa se ela assume valores numéricos e qualitativa se ela assume respostas não numéricas tal como profissão: nível médio, professor, engenheiro, advogado etc.

Vamos considerar as três situações distintas: (1) duas variáveis quantitativas, (2) duas variáveis qualitativas (categóricas) e (3) uma variável qualitativa e uma variável quantitativa, conforme os exemplos a seguir.

EXEMPLO 14 Consumos de luz e de gás

Um morador da cidade do Rio de Janeiro registrou seus consumos de gás (em metros cúbicos) e de luz (quilowatt-hora) ao longo dos meses de maio de 2019 até junho de 2020. Os dados são apresentados na tabela a seguir.

mês/ano	gás	luz
mai/19	34	1031
jun/19	44	538
jul/19	47	835
ago/19	43	739
set/19	39	850
out/19	31	1303
nov/19	31	1050
dez/19	24	951
jan/20	17	1123
fev/20	20	1202
mar/20	25	1250
abr/20	24	1040
mai/20	28	1020
jun/20	29	920

Para visualizar o comportamento das duas variáveis simultaneamente podemos contruir o chamado diagrama de dispersão que nada mais é do que o plano cartesiano cujos 14 pontos assinalados correspondem aos pares ordenados dados por (consumo de gás, consumo de luz) referentes aos meses de maio de 2019 até junho de 2020. Veja uma ilustração do diagrama de dispersão desses dados na [figura 1.44](#).

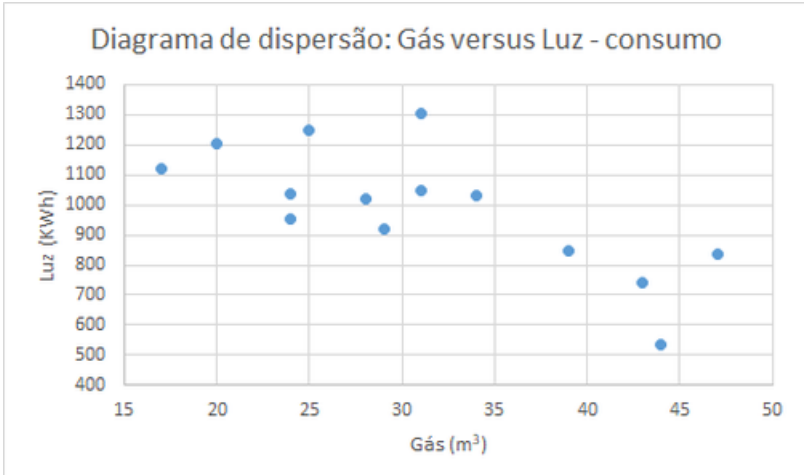
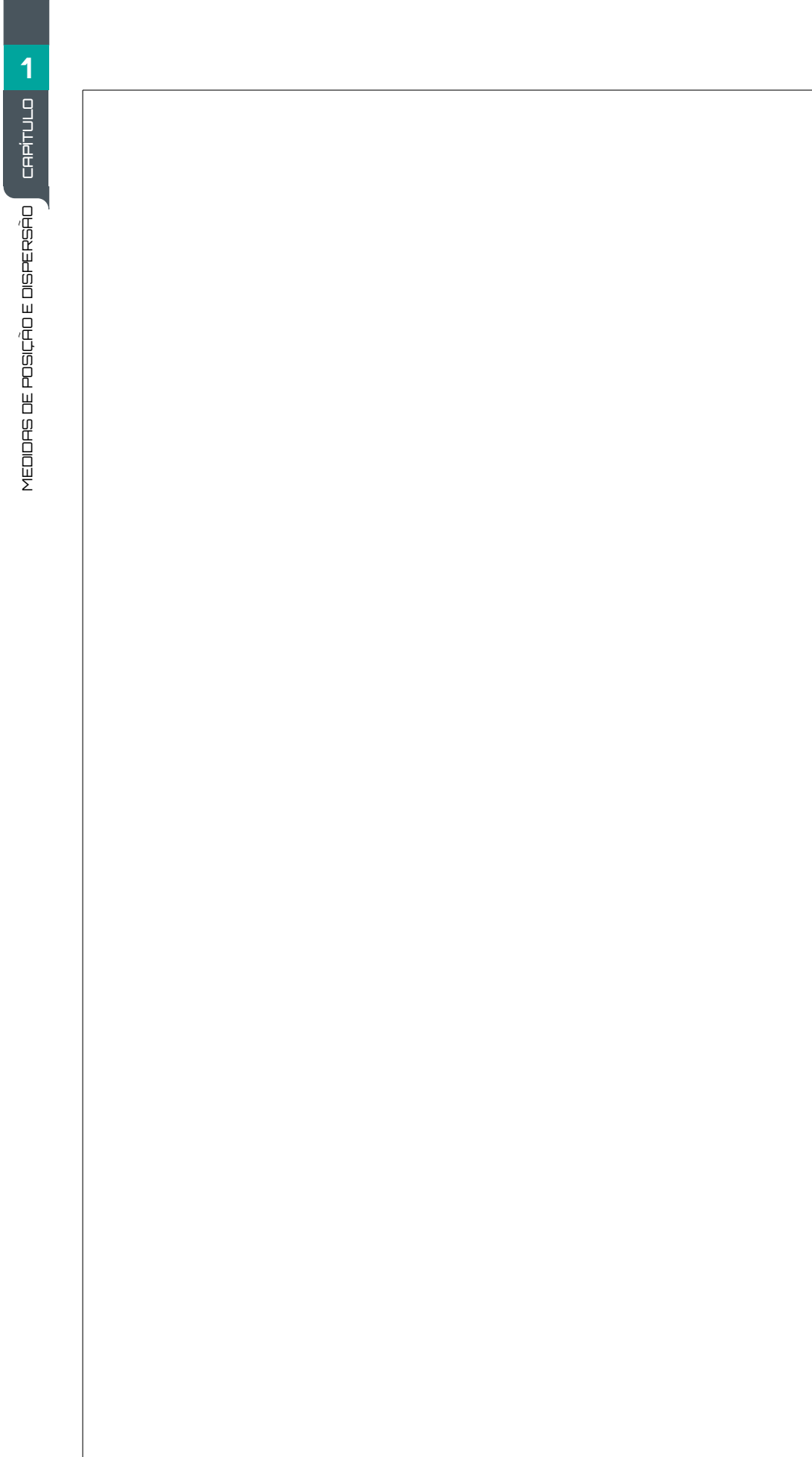


Figura 1.44: Diagrama de dispersão do consumo gás versus o consumo de energia

É possível perceber, a partir do diagrama de dispersão da [figura 1.44](#), uma certa tendência: a medida que o consumo de gás aumenta, o consumo de energia parece diminuir. Essa observação pode ser explicada para uma residência cujo sistema de aquecimento de água é a gás e que tem aparelhos de ar condicionado. Nesse caso, o uso de gás diminui em épocas de temperaturas mais altas devido ao menor número de banhos quentes, enquanto o



consumo de energia aumenta devido ao uso mais intenso do ar condicionado. E, em épocas de temperaturas mais baixas, aumenta o consumo de gás, enquanto diminui o consumo de energia.

Observe que nesse exemplo é importante perceber que não podemos afirmar que exista uma relação de causa e efeito entre essas duas variáveis (consumo de gás e consumo de energia). De fato, a variável que afeta principalmente esses consumos é a temperatura. No entanto, outras informações sobre o local analisado também podem ajudar a explicar o comportamento dessas variáveis.

Nesse exemplo, apresentamos o diagrama de dispersão, usado para representar o comportamento conjunto de duas variáveis quantitativas.

EXEMPLO 15 Mortes por COVID-19 em Nova York (2020)

Os dados que serão trabalhados nesse exemplo são preliminares e sujeitos a variações na medida em que novos casos sejam investigados. Eles referem-se a casos de óbitos por COVID-19 de pessoas residentes na cidade de Nova York coletados até 14 de abril de 2020. Um total de 5.235 registros de óbitos foi analisado e observaram-se a faixa de idades e o fato da pessoa ter ou não comorbidades, ou seja, ser ou não do grupo de risco (idosos, diabéticos, hipertensos etc.) Os dados foram resumidos na tabela

Faixa etária	Com comorbidades	Sem comorbidades	Total
18 a 44 anos	244	25	269
45 a 64 anos	1343	59	1402
65 a 74 anos	1272	26	1298
75 ou mais	2289	27	2316
Total	5148	137	5285

Tabela 1.25: Distribuição conjunta dos óbito por COVID-19 em função da faixa etária e do fato de pertencer ou não ao grupo de risco. Fonte: [NYC Health](#). Acesso em junho de 2020.

Observe que nessa investigação, a unidade de observação é um óbito de pessoa residente na cidade de Nova York. As variáveis observadas para cada unidade são faixa etária (qualitativa, apesar da idade ser numérica aqui elas foram agrupadas em classes) e o fato de ter ou não comorbidades (também qualitativa).

Olhando rapidamente a tabela é imediato perceber que é muito raro o óbito ocorrer para pessoas sem comorbidades, pois de um total de 5285, apenas 137 foram nesse grupo, correspondendo a apenas 2,62% do total. É possível também perceber, que pessoas da faixa etária 18 a 44 anos tem menos chances de vir a óbito, pois apenas 269 pessoas dessa faixa vieram a óbito de um total de 5.285, correspondendo a 5,14% do total.

A questão que surge para esse caso, duas variáveis qualitativas, é como representá-los em um gráfico para visualizarmos o comportamento conjunto dessas duas variáveis.

Um gráfico adequado nesse caso é gráfico de barras múltiplas representando perfis-linha ou coluna da tabela de distribuição conjunta que costuma ser chamada de tabela de dupla-entrada (ou tabela de contingência).

Por exemplo, vamos construir os perfis-coluna dessa tabela, ou seja, as distribuições percentuais de óbitos por faixa etária condicionadas aos grupos com e sem comorbidades. Veja os perfis obtidos na tabela a seguir, incluindo o perfil com todos os casos observados.

Faixa etária	Com comorbidades	Sem comorbidades	Total
18 a 44 anos	4,7	18,2	5,1
45 a 64 anos	26,1	43,1	26,5
65 a 74 anos	24,7	19,0	24,6
75 ou mais	44,5	19,7	43,8
Total	100	100	100

Tabela 1.26: Distribuição conjunta dos óbito por COVID-19 em função da faixa etária e do fato de pertencer ou não ao grupo de risco. Fonte: [NYC Health](#). Acesso em junho de 2020.

Veja na [figura 1.45](#) 8.49 um gráfico de barras múltiplas, ilustrando os dados da [tabela 1.26](#).

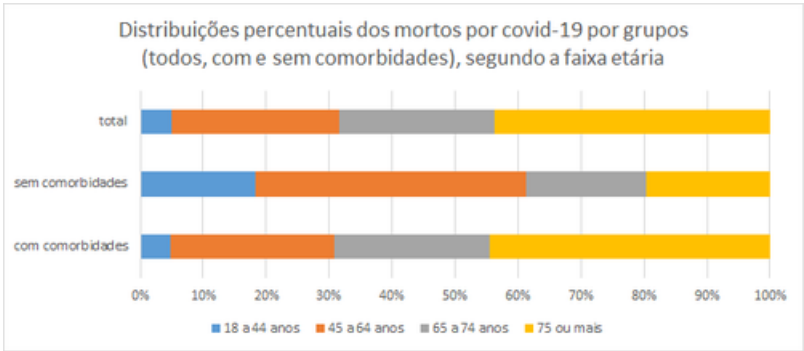


Figura 1.45: Distribuições percentuais dos óbitos por faixa-etária segundo o grupo (todos, sem comorbidades e com comorbidades)

Podemos facilmente perceber que o grupo das pessoas com comorbidades apresenta comportamento diferente em relação aos casos por faixa etária quando comparado ao grupo sem comorbidades. No primeiro caso, 44,5% dos óbitos ocorreram na faixa etária de 75 anos ou mais, enquanto que no segundo caso, o total de óbitos nessa faixa etária foi de 19,7%.

As diferenças observadas nos dois perfis (com comorbidades e sem comorbidades) nos levam a concluir que deve haver alguma associação entre essas duas características. No caso do grupo sem comorbidades 43,1% dos óbitos ocorreram na faixa etária 45 a 64 anos, enquanto que no outro grupo esse percentual foi de 26,1%.

Nesse exemplo, apresentamos o gráfico de barras múltiplas usado para representar o comportamento conjunto de duas variáveis qualitativas.

EXEMPLO 16 Maratona de Nova York (2017)

Na [primeira atividade](#) desse capítulo trabalhamos com os dados da maratona de Nova York (2017), analisando os 100 melhores tempos de chegada na categoria homens e os 100 melhores tempos de chegada na categoria mulheres. Observe que nesse caso, temos duas variáveis observadas a categoria do maratonista (homem ou mulher) e o tempo de chegada. Assim tem-se uma variável qualitativa e uma variável quantitativa. Já vimos como representar dados dessa natureza para fazer comparações. Lembre-se que construímos boxplots dos 100 melhores tempos de chegada para cada categoria. Na [figura 1.46](#) reproduzimos novamente os boxplots.

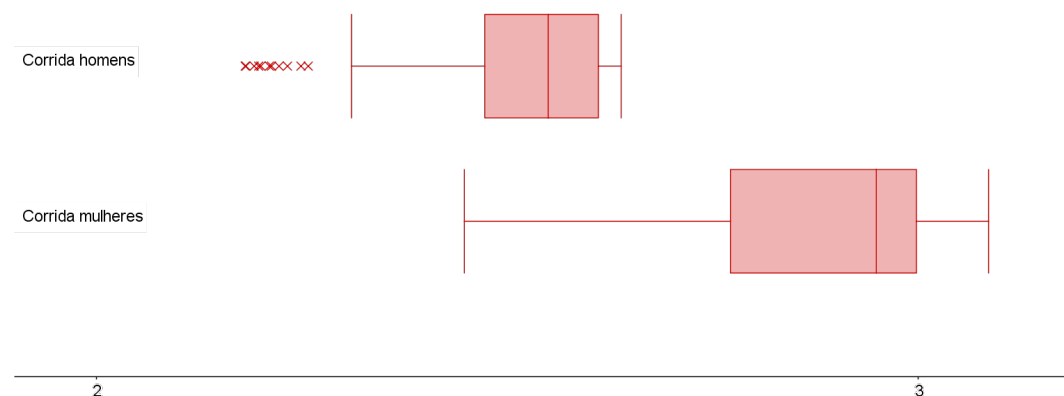


Figura 1.46: Boxplots dos 100 melhores tempos de chegada (em horas) na maratona de Nova York (2017) nas categorias homens e mulheres

Resumindo, como primeiro passo em análises exploratórias de dados bivariados devemos entender a natureza das duas variáveis. Se ambas as variáveis são quantitativas, podemos começar a investigação, construindo o diagrama de dispersão (veja o exemplo consumo de gás e luz). Se ambas as variáveis são qualitativas, podemos começar, construindo o gráfico de barras múltiplas, considerando os perfis-linha ou os perfis-coluna (veja o exemplo Mortes por COVID-19 em Nova York (2020)). Finalmente, se uma das variáveis é qualitativa e a outra é quantitativa, podemos começar a investigação, construindo os boxplots da variável quantitativa para cada categoria de resposta da variável qualitativa (veja o exemplo Maratona de Nova York (2017)).

Em um problema real, em geral, mais de duas variáveis são observadas e muitas vezes queremos entender como essas variáveis se relacionam. Podemos pensar nas análises bivariadas como um primeiro passo para uma análise mais geral.

EXERCÍCIOS

- 1 Numa Escola de Ensino Médio os estudantes precisam fazer um exame no final do ano, se a média dos bimestres for inferior a 7. Um estudante de Ensino Médio desta Escola gostaria de saber em quantas disciplinas ele pode ser aprovado sem fazer exame final. No final do segundo bimestre ele obteve as notas registradas no quadro a seguir. Indique quanto deverão somar, no mínimo, as duas notas dos dois últimos bimestres para evitar o exame final, conclua quando isto ainda é possível.

Disciplina	1º	2º	Soma mínima das notas	Exemplo de notas possíveis
Língua portuguesa	7	4		
Física	5	4		
Matemática	8	8		
História	3	4		
Geografia	5	5		
Filosofia	7	9		
Educação Física	9	8		
Inglês	7	5		
Química	3	7		
Biologia	8	6		

- 2 Suponha que o aluno do exercício anterior conseguiu evitar o exame final das disciplinas de Língua Portuguesa, Física e Biologia, por ter obtido média sete. As notas deste aluno dos quatro bimestres estão indicadas no quadro a seguir.

Disciplina	1º	2º	3º	4º
Língua portuguesa	7	4	8	9
Física	5	4	9	10
Biologia	8	6	7	7

- a) Complete o quadro seguir com os desvios da média de cada disciplina.

Disciplina	Desvios da média				Soma
Língua portuguesa					
Física					
Biologia					

- b) Em qual das disciplinas foi maior o desvio padrão das notas? E o menor?
- c) Você acha que a mediana das notas seria um bom critério para a aprovação? Apresente exemplos para os quais a mediana das notas é 7 e a média é:
- inferior a 7;
 - igual a 7;
 - superior a 7.

Solução: Exercícios

- 1 A tabela fica da seguinte forma:

Disciplina	1º	2º	Soma mínima das notas nos dois últimos bimestres	Exemplo de notas possíveis	
				3º	4º
Língua portuguesa	7	4	17	8	9
Física	5	4	19	9	10
Matemática	8	8	12	6	6
História	3	4	21	Exame Final	
Geografia	5	5	18	9	9
Filosofia	7	9	12	6	6
Educação Física	9	8	11	5	6
Inglês	7	5	16	8	8
Química	3	7	18	8	10
Biologia	8	6	14	7	7

- 2 O estudante deve obter as respostas dos seguintes quadros:

Disciplina	Desvios da média				Soma	Desvio padrão (s)
Língua portuguesa	0	-3	1	2	0	2,16
Física	-2	-3	2	3	0	2,94
Biologia	1	-1	0	0	0	0,82

Notas				Média	Mediana
1	7	7	7	5,5	7
5	6	8	9	7	7
7	7	7	10	7,75	7

Solução: Exercícios

3 A primeira afirmação é verdadeira e, na segunda, os dados são insuficientes para uma conclusão.

- a) Como nem todos os recrutas têm a mesma altura, se nenhum deles medisse mais de 1,81m, a média seria menor do que 1,81m. Logo, pelo menos um recruta tem altura maior do que 1,81m. Analogamente, se nenhum recruta medisse menos de 1,81m, a média seria maior do que 1,81m. Logo, ao menos um recruta mede menos de 1,81 m.
- b) Por exemplo, pode-se ter no grupo 51 recrutas com 1,81 m, exatamente um com 1,80 m e exatamente um com 1,82 m, o que tornaria a sentença falsa. No entanto, também pode-se ter 49 recrutas com 1,81 m, dois com 1,80 m e dois com 1,82 m, o que tornaria a sentença verdadeira. Portanto, os dados são insuficientes para uma conclusão.

4 Com esta transformação:

- a) A média do novo conjunto será dada pela média inicial acrescida da constante a , pois

$$\begin{aligned}\bar{y} &= \frac{y_1 + y_2 + \cdots + y_n}{n} = \frac{x_1 + a + x_2 + a + \cdots + x_n + a}{n} \\ &= \frac{\sum_{i=1}^n x_i + n \cdot a}{n} = \frac{\sum_{i=1}^n x_i}{n} + \frac{n \cdot a}{n} = \bar{x} + a\end{aligned}$$

- b) Podemos verificar que a soma dos desvios da média tomados ao quadrado é a mesma nos dois conjuntos, pois

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n \left[\underbrace{(x_i + a)}_{=y_i} - \overbrace{(\bar{x} + a)}^{\bar{y}} \right]^2 = \sum_{i=1}^n (x_i - \bar{x})^2.$$

Portanto, a variância do novo conjunto, denotada por s_y^2 será igual à variância do conjunto inicial, a saber, $s_y^2 = s^2$ e, assim, o desvio padrão do novo conjunto será igual ao desvio padrão do conjunto inicial, $s_y = s$.

- c) Com base nas respostas anteriores, o coeficiente de variação do novo conjunto será dado por

$$CV_y = \frac{s_y}{\bar{y}} \cdot 100 = \frac{s}{\bar{x} + a} \cdot 100.$$

Logo, se $a > 0$, o coeficiente de variação do novo conjunto será menor do que o coeficiente de variação do conjunto inicial.

Porém, se $a < 0$, o coeficiente de variação do novo conjunto será maior do que o coeficiente de variação do conjunto inicial.

Nota 5

3 (UFRJ - 2005 - adaptado) A altura média de um grupo de 53 recrutas é 1,81m. Sabe-se também que nem todos os recrutas do grupo têm a mesma altura. Diga se cada uma das afirmações a seguir é verdadeira, falsa ou se os dados são insuficientes para uma conclusão. Em cada caso, justifique a sua resposta.

- a) "Há, no grupo em questão, pelo menos um recruta que mede mais de 1,81m e pelo menos um que mede menos de 1,81m."
- b) "Há, no grupo em questão, mais de um recruta que mede mais de 1,81m e mais de um que mede menos de 1,81m."

4 Seja $\{x_1, x_2, \dots, x_n\}$ uma amostra de tamanho n de uma população, em que a média amostral é dada por \bar{x} , o desvio padrão amostral é dado por s e o coeficiente de variação amostral é dado por $CV = \frac{s}{\bar{x}} \cdot 100\%$.

Defina um novo conjunto de dados $\{y_1, y_2, \dots, y_n\}$ em que

$$y_i = x_i + a, \quad i = 1, 2, \dots, n$$

e a é um número real fixado, ou seja, o novo conjunto compreende todos os elementos do conjunto inicial acrescidos de uma constante a . Na atividade [Notas de Arte](#) essa transformação será realizada sobre o conjunto das notas, se o professor acrescentar 1,0 ponto às notas dos alunos da turma.

- a) Em função da média do conjunto inicial, \bar{x} , determine a média do novo conjunto.
- b) Em função do desvio padrão do conjunto inicial, s , determine o desvio padrão do novo conjunto.
- c) Compare o coeficiente de variação do novo conjunto com o do conjunto inicial. São iguais? Por quê?

5 Seja $\{x_1, x_2, \dots, x_n\}$ uma amostra de tamanho n de uma população, em que a média amostral é dada por \bar{x} , o desvio padrão amostral é dado por s e o coeficiente de variação amostral é dado por $CV = \frac{s}{\bar{x}} \cdot 100\%$. Defina um novo conjunto de dados $\{y_1, y_2, \dots, y_n\}$ em que $y_i = c \cdot x_i, i = 1, 2, \dots, n$ e c é um número real fixado e $c > 0$, ou seja, o novo conjunto compreende todos os elementos do conjunto inicial multiplicados por uma constante $c > 0$. Na atividade [Notas de Arte](#) essa transformação será realizada sobre o conjunto das notas, se o professor aumentar em 20% a nota de cada aluno, isto é, multiplicar cada nota pelo fator 1,2.

- a) Em função da média do conjunto inicial, \bar{x} , determine a média do novo conjunto.
- b) Em função do desvio padrão do conjunto inicial, s , determine o desvio padrão do novo conjunto.
- c) Compare o coeficiente de variação do novo conjunto com o do conjunto inicial. Houve alguma alteração? Por quê?

- 6 (ENEM 2015) Em uma seletiva para a final dos 100 metros livres de natação, numa olimpíada, os atletas, em suas respectivas raias, obtiveram os tempos no quadro a seguir. Escolha a opção que indica o valor da mediana dos tempos apresentados.

- a) 20,70s.
b) 20,77s.
c) 20,80s.
d) 20,85s.
e) 20,90s.

Tabela 1.27: Tempos em segundos

Raia	1	2	3	4	5	6	7	8
Tempo(s)	20,90	20,90	20,50	20,80	20,60	20,60	20,90	20,96

- 7 (ENEM 2016-adaptado) Em uma cidade, o número de casos de dengue confirmados aumentou consideravelmente nos últimos dias. A prefeitura resolveu desenvolver uma ação, contratando funcionários para ajudar no combate à doença, os quais orientarão os moradores a eliminarem criadouros do mosquito *Aedes aegypti*, transmissor da dengue. A tabela a seguir apresenta o número atual de casos confirmados, por região da cidade.

A prefeitura optou pela seguinte quantidade de funcionários a serem contratados: (I) 10 funcionários para cada região da cidade cujo número de casos seja maior que a média dos casos confirmados e (II) 7 funcionários para cada região da cidade cujo número de casos seja menor ou igual à média dos casos confirmados. Quantos funcionários a prefeitura deverá contratar para efetivar a ação?

- a) 59 b) 65 c) 68 d) 71 e) 80

Tabela 1.28: Número atual de casos por região da cidade

Região	Casos confirmados
Oeste	237
Centro	262
Norte	158
Sul	159
Noroeste	160
Leste	278
Centro-Oeste	300
Centro-Sul	278
Soma	1.832

Solução: Exercícios

- 8 O primeiro passo é colocar os tempos do quadro apresentado em ordem crescente, a saber, $20,50 < 20,60 < 20,80 < 20,90 \leq 20,90 < 20,96$. Como o número de observações é par ($n = 8$), segue que a mediana é dada por $\frac{x_{(4)} + x_{(5)}}{2} = \frac{20,80 + 20,90}{2} = 20,85$. A resposta correta encontra-se na opção **d**.
- 9 A média do número de casos confirmados é dada por $\frac{1.832}{8} = 229$. Logo, o número de regiões da cidade cujo número de casos confirmados é maior do que 299 é 5, e o número de regiões da cidade cujo número de casos confirmados é menor do que 299 é 3. Assim, o número de funcionários que devem ser contratados pela prefeitura é $5 \cdot 10 + 3 \cdot 7 = 71$. A resposta correta encontra-se na opção **d**.

Solução: Exercícios**6**

- a) Concordo, pois podemos perceber que os comprimentos dos intervalos à direita são maiores:

$$Q_3 - \text{mediana} = 13 - 6,5 = 6,5 > \text{mediana} - Q_1 = 6,5 - 6 = 0,5$$

$$\text{Max} - Q_3 = 13 - 9 = 4 > Q_1 - \text{Min} = 6 - 4 = 2 \text{ e}$$

$$\text{Max} - \text{mediana} = 13 - 6,5 = 6,5 > \text{mediana} - \text{Min} = 6,5 - 4 = 2,5.$$

- b) Concordo, pois há assimetria à direita.
 c) Concordo: este gráfico não nos revela a existência de um intervalo de maior frequência, pois os quatro intervalos nele considerados têm frequências relativas iguais a 0,25
 d) Concordo, considerando a aproximação dada por $\frac{R}{4} = \frac{13-4}{4} = \frac{9}{4} = 2,25$
 e) Concordo, pois $DQ = Q_3 - Q_1 = 9 - 6 = 3$ gols
 f) Concordo: o gráfico não apresenta pontos destacados. Também podemos verificar que a cerca inferior é dada por

$$Q_1 - 1,5 \cdot DQ = 6 - 1,5 \cdot 3 = 1,5$$

e a cerca superior é dada por

$$Q_3 + 1,5 \cdot DQ = 9 + 4,5 = 13,5.$$

Como o valor mínimo é 4 e o máximo é 13, conclui-se que não existem valores atípicos

- g) Concordo, pois o *boxplot* agrupou os dados em quatro intervalos de frequências relativas dadas por 0,25, a saber, $[4; 6]$, $[6; 6,5]$, $[6,5; 9]$ e $[9; 13]$. Os valores dentro dos parênteses na expressão indicada correspondem aos pontos médios de cada um destes intervalos. Ou seja, esta média foi calculada com base na fórmula

$$\sum_{i=1}^4 f_i \cdot \bar{x}_i$$

Nota 6**8**

O *boxplot* a seguir representa a distribuição do número de gols da artilharia nas Copas do Mundo desde a Copa de 1930 até a Copa de 2006. Vamos chamar este número de **recorde**. Observe que só é considerado o **recorde**, sem levar em conta se houve mais de um artilheiro na Copa. Desse modo, nestas 18 Copas do Mundo, a figura leva em consideração os 18 **recordes** observados.

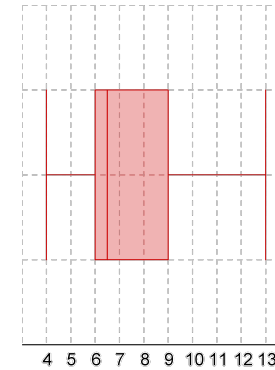


Figura 1.50: Boxplot dos **recordes** das Copas do Mundo de 1936 a 2006.

Com base neste gráfico, as seguintes afirmações foram feitas a cerca da distribuição dos **recordes** nestas Copas do Mundo.

- a) A distribuição apresenta assimetria à direita.
 b) A média dos **recordes** é maior do que a mediana dos **recordes**.
 c) O boxplot não nos permite avaliar a existência de moda.
 d) Uma aproximação grosseira para o valor do desvio padrão dos **recordes** nestas Copas é dada por 2,25 gols.
 e) A distância entre quartis desta distribuição é 3 gols.
 f) Esta distribuição não apresentou valores atípicos.
 g) Uma aproximação para o valor da média dos **recordes** pode ser calculada por $0,25 \cdot (5 + 6,25 + 7,75 + 11) = 7,5$ gols.

Responda se concorda ou não com cada uma destas afirmações, justificando cada resposta.

9

Na questão anterior foram consideradas 18 Copas do Mundo. Sabe-se que a soma exata dos **recordes** destas Copas é dada por $\sum_{i=1}^{18} x_i = 132$ e que a soma dos quadrados dos **re-**

cordes é dada por $\sum_{i=1}^{18} x_i^2 = 1060$.

- a) Com base nestas informações, calcule a média e o desvio padrão dos **recordes** e compare com as aproximações obtidas no exercício anterior.
 b) Consultando os **recordes** referentes às Copas de 2010 e 2014, verificou-se que eles foram 5 e 6, respectivamente. Determine a média e o desvio padrão dos **recordes**, considerando as 20 Copas do Mundo até 2014.

- 10 (ENEM-2010) O quadro seguinte mostra o desempenho de um time de futebol no último campeonato. A coluna da esquerda mostra o número de gols marcados e a coluna da direita informa em quantos jogos o time marcou aquele número de gols.

Tabela 1.29: Desempenho de um time

Gols marcados	Quantidade de partidas
0	5
1	3
2	4
3	3
4	2
5	2
7	1

Se X , Y e Z são, respectivamente, a média, a mediana e a moda desta distribuição, então:

- a) $X = Y < Z$ b) $Z < X = Y$ c) $Y < Z < X$ d) $Z < X < Y$ e) $Z < Y < X$

- 11 Um professor de Matemática suspeita que seus alunos do turno da tarde são mais fracos do que os seus alunos do turno da manhã. Para verificar sua suspeita, logo no início do ano letivo ele aplicou um teste básico de questões envolvendo conteúdos básicos e esperados para o nível a ser iniciado em duas amostras, uma de alunos do turno da manhã e outra de alunos do turno da tarde. A seguir, estão os resultados para as duas amostras.

Tabela 1.30: Notas de uma amostra de alunos do turno da manhã

7,4	7,3	6,2	6,3	4,1
5,7	10,0	6,2	4,9	6,0
8,7	6,5	3,0	5,8	7,0
8,0	8,0	4,9	7,4	6,8
6,7	7,6	6,1	6,2	8,5
7,4	4,4	8,1	5,8	6,6
4,2	5,3	4,9	8,1	6,8
6,8	4,4	5,4	7,1	6,1
5,3	5,2	5,7	9,9	8,3

Tabela 1.31: Notas de uma amostra de alunos do turno da tarde

5,1	4,7	5,7	4,7	5,0
4,2	4,9	6,0	4,4	4,4
6,0	4,9	5,6	6,2	6,6
6,2	4,7	6,0	4,6	3,6
5,4	5,2	5,6	5,5	5,2
5,8	4,5	5,0	3,8	4,6
4,1	4,7	4,2	6,8	5,6
5,3	4,5	4,7	5,1	5,2

Usando todas as ferramentas estudadas neste capítulo, ajude este professor, fazendo um relatório detalhado e comparativo sobre os dois turnos. Se preferir, você poderá baixar estes dados no [link](#), mas lembre-se que como eles estão registrados no GeoGebra, a vírgula foi trocada por ponto.

Solução: Exercícios

- 10 É fácil ver que a moda é zero tal que $Z = 0$. Somando o número de partidas jogadas vemos foram consideradas 20 partidas. Assim, o valor da mediana é o valor que ocupa as posições centrais 10 e 11. Da tabela, calculando as frequências acumuladas, vemos que até 1 gol acumularam-se 8 partidas e até 2 gols, acumularam-se 12 partidas. Assim, podemos concluir que nas posições 10 e 11 o número de gols foi 2, tal que a mediana $= Y = 2$. A média é dada por

$$\bar{x} = X = \frac{5 \cdot 0 + 3 \cdot 1 + 4 \cdot 2 + 3 \cdot 3 + 2 \cdot 4 + 2 \cdot 5 + 1 \cdot 7}{20} = \frac{45}{20} = 2,25.$$

Logo, tem-se $Z < Y < X$ e a resposta correta encontra-se na opção e)

- 11 As figuras a seguir ilustram os respectivos histogramas (ambos na mesma escala e usando a frequência absoluta no eixo vertical) e boxplots das notas para os alunos da manhã e da tarde.

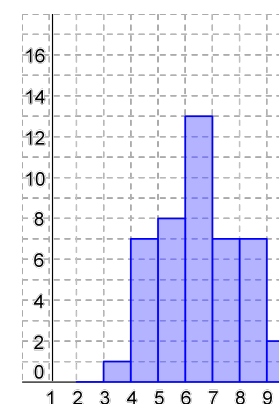


Figura 1.47: Histograma das notas dos alunos do turno da manhã

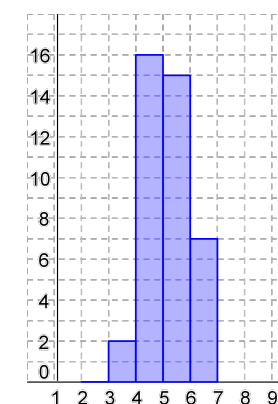


Figura 1.48: Histograma das notas dos alunos do turno da tarde

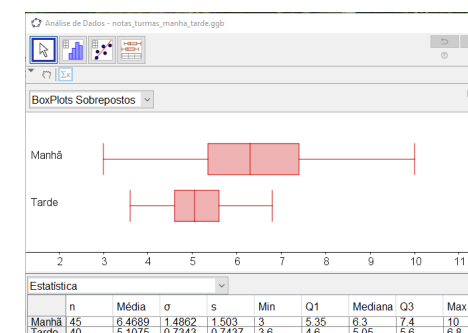


Figura 1.49: Boxplots das notas dos alunos do turno da manhã e do turno da tarde e quadro das medidas resumo gerados pelo GeoGebra

Nota 7

Solução: Exercícios

12 a) $I = \frac{3 \cdot (5,93 - 6,5)}{1,945} \approx -0,87$, indicando alguma assimetria à esquerda.

b) Na categoria cadeira de rodas, temos

$$I = \frac{3 \cdot (2,5855 - 2,0978)}{1,4209} \approx 1,03,$$

indicando alguma assimetria à direita. Na categoria mulheres temos

$$I = \frac{3 \cdot (2,8698 - 2,9493)}{0,1867} \approx -1,28,$$

indicando forte assimetria à esquerda. Na categoria homens temos

$$I = \frac{3 \cdot (1,5116 - 1,55)}{0,1283} \approx -0,9,$$

indicando assimetria à esquerda.

c) No turno da manhã temos

$$I = \frac{3 \cdot (6,4689 - 6,3)}{1,503} \approx 0,3,$$

e no turno da tarde

$$I = \frac{3 \cdot (6,1075 - 5,05)}{0,7437} \approx 0,2/$$

Pela análise dos boxplots destas duas distribuições, avaliamos que ambas eram aproximadamente simétricas. Valores de I entre $-0,3$ e $0,3$ podem indicar dados aproximadamente simétricos.

Nota 8**Solução: Exercícios**

14 a)

Medida	X	Y	Z
Média	10	12	12
s	1,487	2,81	0,725
Min	7	7	11
Q_1	9	10	11,75
Mediana	10	12	12
Q_3	11	14	12,25
Max	13	18	13

12 Quando comparou-se a média com a mediana falou-se em grau de assimetria da distribuição (**Organizando as ideias: medidas de posição**). Na seção **Para saber mais** falou-se novamente em grau de assimetria. A assimetria pode ser medida pelo **índice de assimetria de Pearson**

$$I = \frac{3 \cdot (\bar{x} - \text{mediana})}{s}$$

Se $I \approx 0$, os dados são considerados aproximadamente simétricos. Um valor de I negativo, indica assimetria à esquerda e, um valor de I positivo, assimetria à direita.

Se $I \geq 1,00$ ou $I \leq -1,00$, os dados podem ser considerados fortemente assimétricos à direita ou à esquerda, respectivamente. Calcule o índice de assimetria de Pearson, para os dados de

- Notas de Arte;
- Atividade: modalidades da maratona de Nova Iorque 2017;
- Exercício 10.

13 Em provas aplicadas em grande escala é comum divulgar as notas transformadas da seguinte forma

$$y_i = 500 + 100 \cdot \frac{(x_i - \bar{x})}{s}, \quad i = 1, 2, \dots, n$$

em que x_i é a nota obtida pelo i -ésimo candidato, $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$, s é o desvio padrão amostral das notas do conjunto $\{x_1, x_2, \dots, x_n\}$ e y_i é a nota transformada do i -ésimo candidato. Considere as afirmações a seguir.

- A média das notas transformadas é 500.
- O desvio padrão das notas transformadas é 100.
- Se a distribuição de notas é aproximadamente simétrica e com poucas notas atípicas, cerca de 67% dos candidatos obtiveram notas transformadas entre 400 e 600.
- Se a distribuição de notas é aproximadamente simétrica e com poucas notas atípicas, cerca de 95% dos candidatos obtiveram notas transformadas entre 300 e 700.

Responda se concorda ou não com cada uma destas afirmações, justificando cada resposta.

14 (Dados trabalhados na Atividade "Comparação de Medicamentos" no Capítulo **A Natureza Estatística**)

Deseja-se comparar três medicamentos, X, Y e Z, no tratamento da dor de cabeça. Para isso 60 pacientes com perfis similares foram separados aleatoriamente em três grupos de 20 cada. Para cada grupo, será ministrado um dos medicamentos e observado o tempo de cura da dor de cabeça (em minutos). No quadro a seguir estão dispostos os dados obtidos.

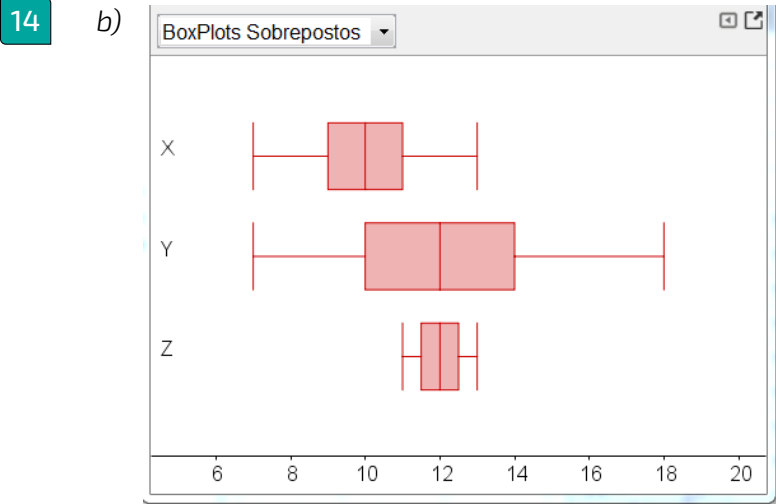
Dados ordenados	X	Y	Z
1	7	7	11
2	8	8	11
3	8	9	11
4	9	9	11
5	9	10	11
6	9	10	12
7	9	11	12
8	10	11	12
9	10	11	12
10	10	12	12
11	10	12	12
12	10	12	12
13	10	13	12
14	11	13	12
15	11	14	12
16	11	14	13
17	11	15	13
18	12	15	13
19	12	16	13
20	13	18	13
Soma simples	200	240	240
Soma de quadrados	2042	3030	2890

a) Complete o quadro a seguir.

Medida	X	Y	Z
Média			
Moda			
S			
Min			
Q_1			
Mediana			
Q_3			
Max			

- a) Construa os boxplots para os três conjuntos de dados.
- b) Como você avalia a forma das distribuições quanto à assimetria? Por quê?
- c) Com base nas informações obtidas, que medicamento você escolheria? Por quê?

Solução: Exercícios



c) Observando-se nos boxplots os comprimentos dos seguintes pares de intervalos:

- Q_1 e mínimo e máximo e Q_3 ;
- Mediana e Q_1 e Q_3 e mediana;
- Mediana e mínimo e Máximo e mediana.

conclui-se que são aproximadamente iguais, concluindo-se que as distribuições, para os três medicamentos, são simétricas. Observe que os índices de assimetria são iguais a zero em cada uma, pois tem-se média=mediana, nas três distribuições.

d) Como todas as distribuições são simétricas, podemos usar a regra empírica de frequência entre a média mais ou menos dois desvios padrões. Para o medicamento X , com cerca de 95% de chance a dor de cabeça será curada entre 8,513min e 11,487min. Para o medicamento Y , com cerca de 95% de chance a dor de cabeça será curada entre 9,19min e 14,81min. Para o medicamento Z , com cerca de 95% de chance a dor de cabeça será curada entre 11,275min e 12,725min. Observe que apesar do intervalo de 95% de chance para o tempo de cura no medicamento em Z ser mais estreito do que o mesmo intervalo para o medicamento X , o intervalo para o medicamento X apresenta valores menores: de fato, cerca de 75% dos valores em X são menores do que o valor mínimo em Z . Por esta razão, neste exemplo, a melhor escolha parece ser o medicamento X .

Material Suplementar

Como material de suporte para este capítulo foi desenhado um aplicativo interativo de Geogebra para a visualização de medidas de posição e dispersão de uma distribuição, que pode ser encontrado [aqui](#). O aplicativo pode ser usado diretamente no explorador de internet de sua preferência ou baixado e usado em computadores e celulares com [Geogebra](#) instalado.

O aplicativo gera dados de forma aleatória, mas você pode inserir seus próprios dados na primeira coluna da planilha e verá o histograma correspondente na área gráfica, escolhendo a quantidade de partições do intervalo que você deseja.

O aplicativo permite visualizar, além do histograma, as medidas de posição da distribuição além das medidas de dispersão, mostrando: mínimo, máximo, média, mediana, Q_1 , Q_3 , variância e desvio padrão amostrais e populacionais.

Finalmente, é possível construir o boxplot na mesma área gráfica para que o estudante se familiarize visualmente com a relação entre o histograma e o boxplot.

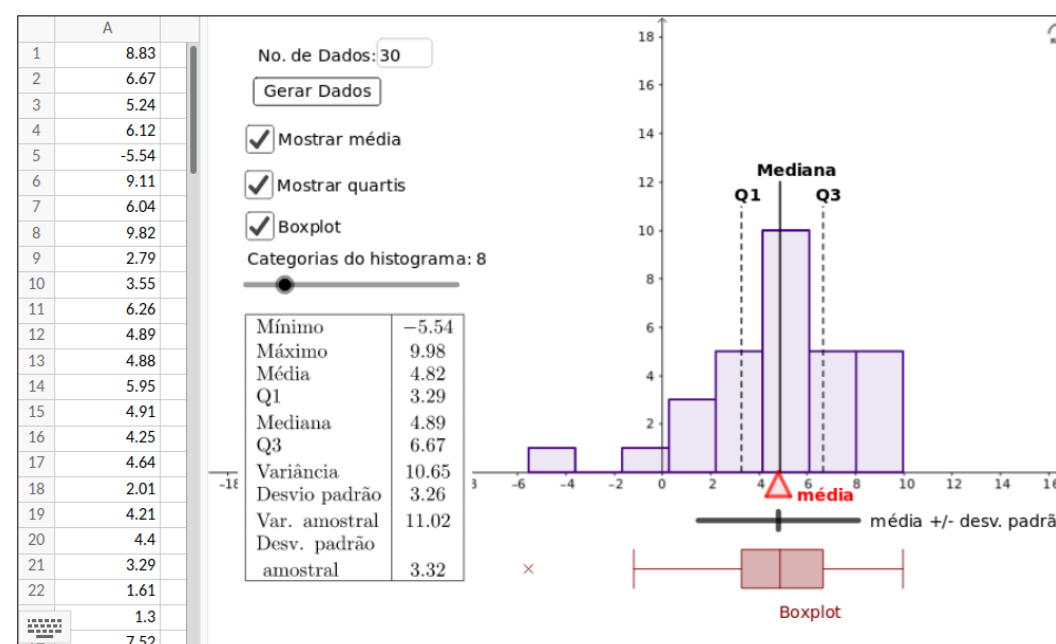


Figura 1.51: Aplicativo interativo em Geogebra para a visualização de medidas de posição e dispersão de uma distribuição

Notas

1

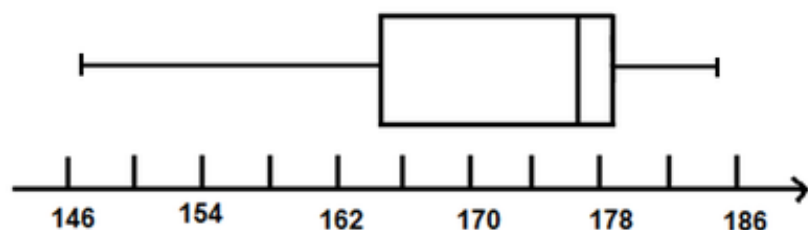


Figura 1.3: Boxplot (sem sinalização de valores discrepantes) dos 100 melhores tempos para a categoria mulheres

2 Solução: Frequência de valores no intervalo centrado na média

No caso dos dados da atividade [Notas de Arte](#) temos $\bar{x} = 5,93$ e $s = 1,96$ tal que os limites deste intervalo são, respectivamente, 2,01 e 9,85. Portanto, das 35 notas podemos ver que 32 observações caem dentro destes limites, ou equivalentemente, cerca de 91% das observações.

No caso dos dados da atividade [A Maratona](#), categoria mulheres, temos $\bar{x} = 171,91$ e $s = 11,13$ tal que os limites deste intervalo são, respectivamente, 149,65 e 194,17. Portanto, dos 100 tempos podemos ver que 96 caem dentro destes limites, ou equivalentemente, 96% dos tempos.

No caso dos dados da atividade [Categoria homens na maratona](#), temos $\bar{x} = 150,69$ e $s = 7,70$ tal que os limites deste intervalo são, respectivamente, 135,29 e 166,09. Portanto, dos 100 tempos podemos ver que 90 caem dentro destes limites, ou equivalentemente, 90% dos tempos.

No caso dos dados da atividade [Estratégia de Investimento](#), para a companhia A, temos $\bar{x} = 61,5$ e $s = 4,5765$ tal que os limites deste intervalo são, aproximadamente, 52,3 e 70,7. Portanto, das 10 cotações podemos ver que todas caem dentro destes limites, ou equivalentemente, 100% das cotações.

No caso dos dados da atividade [Estratégia de Investimento](#), para a companhia B, temos $\bar{x} = 61,5$ e $s = 18,3136$ tal que os limites deste intervalo são, aproximadamente, 24,9 e 98,1. Portanto, das 10 cotações podemos ver que todas caem dentro destes limites, ou equivalentemente, 100% das cotações.

Observe que para os cinco conjuntos considerados nessa atividade, de fato, a maior parte dos dados (90% ou mais) situa-se entre os limites de uma média mais ou menos 2 desvios padrões.

3 Solução: Modalidades da maratona de Nova Iorque 2017

- Considerando a amplitude amostral é fácil perceber que a maior dispersão ocorre na categoria triciclo de mão. O mesmo vale se considerarmos a distância entre quartis. Pela [tabela](#) podemos ver que esta resposta também valerá se considerarmos o desvio padrão.
- Pela [figura 1.41m](#) podemos ver que a única categoria que não apresentou valores atípicos foi a categoria das mulheres, pois não há pontos destacados no boxplot correspondente às mulheres.
- Considerando as categorias "cadeira de rodas" e "triciclo de mão", vemos que

$$\begin{aligned} Q_1 - \text{Min} &<< \text{Max} - Q_3; \\ \text{Mediana} - Q_1 &< Q_3 - \text{Mediana e} \\ \text{Mediana} - \text{Min} &<< \text{Max} - \text{mediana}, \end{aligned}$$

em que o símbolo $<<$ é usado para indicar "bem menor do que".

Logo, conclui-se que nestas categorias tem-se assimetria à direita acentuada. Observe, que nestes dois casos tem-se que a mediana é menor do que a média. Reveja os histogramas construídos na [ativ-comparacao-de-diferentes-grupos](#). Considerando as categorias "homens" e "mulheres", vemos que

$$\begin{aligned} Q_1 - \text{Min} &>> \text{Max} - Q_3; \\ \text{Mediana} - Q_1 &> Q_3 - \text{Mediana e} \\ \text{Mediana} - \text{Min} &>> \text{Max} - \text{mediana}, \end{aligned}$$

em que o símbolo $>>$ é usado para indicar "bem maior do que". Logo, conclui-se que nestas categorias tem-se assimetria à esquerda acentuada. Observe, que nestes dois casos tem-se que a mediana é maior do que a média. Reveja os histogramas construídos na atividade [Comparação de conjuntos de dados](#).

- Podemos perceber que ambas as categorias apresentam distribuições com assimetria à esquerda, mas na categoria mulheres não há valores atípicos. Também podemos perceber que a dispersão na categoria mulheres é maior do que na categoria homens, considerando a amplitude, a distância entre quartis e também o desvio padrão. Por esta razão, a categoria mulheres não apresentou valores atípicos. Já para a categoria homens, por ter apresentado menos dispersão, apresentou vários valores atípicos pequenos, que certamente, devem se referir aos

tempos dos atletas profissionais. Reveja os histogramas construídos na [Comparação de conjuntos de dados](#).

- e) Considerando as categorias "cadeira de rodas" e "triciclo de mão" vemos que na primeira, 51 completaram a maratona e, na segunda, 69 completaram a maratona. Quanto à amplitude, vemos que ela foi maior na categoria "triciclo de mão", valendo o mesmo para a distância entre quartis e para o desvio padrão. Possivelmente, esta diferença nas dispersões das duas categorias esteja sendo acarretada pelo maior valor atípico da categoria "triciclo de mão", a saber, 9,5206h. Já foi observado que ambas as categorias apresentam distribuições com assimetria à direita de modo que a mediana é menor do que a média. Reveja os histogramas construídos na [Comparação de conjuntos de dados](#).

4 Solução: Medidas para dados agrupados

- a) A média pode ser calculada por

$$\begin{aligned}\bar{x} &\approx 0,15 \cdot 1 + 0,25 \cdot 3 + 0,20 \cdot 5 + 0,3 \cdot 7 + 0,1 \cdot 9 \\ &= 0,15 + 0,75 + 1 + 2,1 + 0,9 = 4,9.\end{aligned}$$

Para calcular a variância, primeiro obtemos uma aproximação para a soma de quadrados das notas, dada por

$$\begin{aligned}0,15 \cdot 12 + 0,25 \cdot 32 + 0,20 \cdot 52 + 0,3 \cdot 72 + 0,1 \cdot 92 = \\ 0,15 + 2,25 + 5 + 14,7 + 8,1 = 30,2,\end{aligned}$$

assim, $s^2 \approx 30,24,92 = 6,19$.

A classe modal corresponde ao intervalo delimitado por 6 e 8, uma aproximação para o valor modal é considerar o ponto médio da classe modal. Neste caso, temos que 7 é uma aproximação para o valor da moda nesta distribuição.

Não podemos identificar quem é o valor central ou valores centrais, pois não foi dada a informação do número de candidatos que fizeram a prova. Mas isso não é problema, pois a mediana divide a distribuição em dois intervalos de frequências iguais a 50%. Logo, precisamos identificar em que intervalo, cairá a mediana e, como apresentado na Organizando as ideias: medidas de posição tomar o ponto médio desta classe como aproximação para o valor da mediana. Observe na figura que a frequência do primeiro intervalo é 0,15; a frequência acumulada, considerando os dois primeiros intervalos é $0,15 + 0,25 = 0,40$ ainda é menor do que 0,5. Considerando os três primeiros intervalos, a frequência acumulada é $0,4 + 0,2 = 0,6$. Logo, a mediana está no intervalo delimitado por 4 e 6, de modo que tomamos o ponto médio deste intervalo como uma aproximação para o valor da mediana, a saber, 5.

O mesmo raciocínio utilizado para obter a mediana, pode ser usado para obter aproximações do primeiro e terceiro quartis. Em vez de 50% na frequência acumulada, deveremos encontrar 25% e 75%, respectivamente. Como a frequência

do primeiro intervalo é 0,15 e a frequência acumulada, considerando os dois primeiros intervalos é $0,15 + 0,25 = 0,40$, seque que o primeiro quartil deve estar no segundo intervalo delimitado por 2 e 4. Logo, tomamos o ponto médio deste intervalo como uma aproximação para o primeiro quartil, a saber, 3. Até o terceiro intervalo a frequência acumulada é 0,6, considerando o quarto intervalo, a frequência acumulada é 0,9. Logo, como o terceiro quartil está no quarto intervalo, tomamos o ponto médio 7 com aproximação para o terceiro quartil.

- b) Com base no histograma temos o seguinte esquema dos cinco números $Min = 0$, $Q_1 = 3$, Mediana = 5, $Q_3 = 7$, $Max = 10$. $DQ = 7 - 3 = 4$. Cerca inferior = $3 - 6 = -3$, cerca superior = $7 + 6 = 13$. Logo, não existem valores discrepantes. A figura a seguir ilustra um boxplot para este esquema dos cinco números.

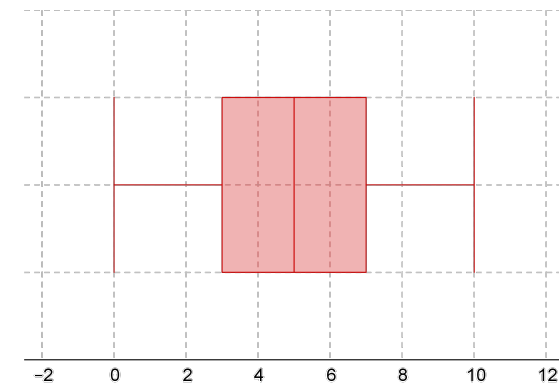


Figura 1.4: Boxplot dos resultados dos candidatos na prova de seleção

5 Solução: Exercícios

5 $y_i = c \cdot x_i, i = 1, 2, \dots, n.$

- a) A média do novo conjunto será dada pela média inicial multiplaca pela constante c , pois

$$\bar{y} = \frac{y_1 + y_2 + \dots + y_n}{n} = \frac{c \cdot x_1 + c \cdot x_2 + \dots + c \cdot x_n}{n} = \frac{c}{n} \cdot \sum_{i=1}^n x_i = c \cdot \bar{x}$$

- b) Podemos verificar que a soma dos desvios da média tomados ao quadrado será dada pela soma original dos desvios da média elevados ao quadrado multiplicada por c^2 , pois

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (c \cdot x_i - c \cdot \bar{x})^2 = c^2 \cdot \sum_{i=1}^n (x_i - \bar{x})^2.$$

Portanto, a variância do novo conjunto, denotada por s_y^2 será igual à variância do conjunto inicial multiplicada por c^2 , a saber, $s_y^2 = c^2 \cdot s^2$ e, assim, o desvio padrão do novo conjunto será igual ao desvio padrão do conjunto inicial multiplado por c , $s_y = c \cdot s$. Lembre que estamos considerando $c > 0$.

- c) Com base nas respostas anteriores, o coeficiente de variação do novo conjunto será dado por

$$CV_y = \frac{c \cdot s_y}{c \cdot y} \cdot 100 = \frac{s}{x} \cdot 100.$$

Logo, o coeficiente de variação do novo conjunto será igual ao coeficiente de variação do conjunto inicial.

6 Solução: Exercícios

- 9 a) A média é dada por $\bar{x} = \frac{132}{8} \approx 7,33$, e o desvio padrão amostral é dado por

$$\sqrt{\frac{1}{18-1}(1060 - 18 \cdot 7,33^2)} \approx 2,33.$$

Comparando a média aproximada de 7,5 com a média 7,33, conclui-se que o erro de aproximação é bem pequeno, representando apenas cerca de 2,3% da média obtida com a soma exata $\frac{|7,5-7,33|}{7,33} \approx 0,023$. Comparando o desvio padrão aproximado de 2,25 com o desvio padrão 2,33, conclui-se que o erro de aproximação é bem pequeno, representando apenas cerca de 3,4% o desvio padrão obtido com as sobas exatas $\frac{|2,25-2,33|}{2,33} \approx 0,034$.

- b) Nesse caso temos 20 dados, e

$$\sum_{i=1}^{20} = 1 = 132 + 5 + 6 = 143$$

tal que a média dos **recordes** nas 20 Copas do Mundo até 2014 é $\bar{x} = \frac{143}{20} = 7,15$. Para o cálculo do desvio padrão temos que considerar a soma de quadrados dos 20 **recordes**, a saber, considerando as 20 Copas, temos

$$\sum_{i=1}^{20} x_i^2 = 1060 + 5^2 + 6^2 = 1060 + 25 + 36 = 1121.$$

Assim, o desvio padrão amostral é dado por

$$\sqrt{\frac{1}{20-1}(1121 - 20 \cdot 7,15^2)} \approx 2,28.$$

7 Solução: Exercícios

11 Comparando os dois turnos pode-se perceber que

- ambas as distribuições de notas são aproximadamente simétricas (observe que em ambas as distribuições os valores de média e mediana são próximos, a saber, 6,47 e 6,3 no turno da manhã e 5,11 e 5,05 no turno da tarde);
- a dispersão das notas no turno da tarde é inferior à dispersão das notas do turno da manhã, mas o boxplot revela que o "centro" no turno da tarde, caracterizado pelo retângulo no boxplot está mais para à esquerda em relação ao retângulo do boxplot para os alunos do turno da manhã, indicando inferioridade de notas (50% das notas centrais no turno da tarde estão entre 4,6 e 5,6; enquanto que 50% das notas centrais do turno da manhã estão entre 5,35 e 7,4);
- a distribuição das notas no turno da manhã é mais homogênea em relação à média do que a distribuição das notas do turno da tarde, observação que pode ser comprovada pelo cálculo do coeficiente de variação amostral de ambos os turnos, a saber, $CVA_{\text{manhã}} = \frac{1,503}{6,4686} \cdot 100 = 23,2\%$ e $CV_{\text{tarde}} = \frac{0,7437}{5,1075} \cdot 100 = 14,6\%$;
- a frequência de notas em torno da média mais ou menos um desvio padrão no turno da manhã é 62,2% e, no turno da tarde, 70% (estes valores estão perto do valor estipulado pela regra empírica de 67%);
- a frequência de notas em torno da média mais ou menos dois desvios padrões no turno da manhã é 93,3% no turno da tarde, 92,5% (estes valores estão perto do valor estipulado pela regra empírica de 95%);
- utilizando a aproximação grosseira para o cálculo do desvio padrão amostral, obtém-se $\frac{10-3}{4} = 1,75$ para o turno da manhã (um erro relativo de 16,4%) em relação ao valor de s calculado para o turno da manhã e, $\frac{6,8-3,6}{4} = 0,8$ para o turno da tarde (um erro relativo de 7,6%) em relação ao valor de s calculado para o turno da tarde;
- não existem notas atípicas nas duas distribuições;
- apesar do turno da manhã apresentar melhores notas, a menor nota foi observada neste turno.

8 Solução: Exercícios

13 a) Concordo, pois

$$\begin{aligned}\bar{y} &= \frac{1}{n} \sum_{i=1}^n y_i = \frac{1}{n} \sum_{i=1}^n [500 + 100 \cdot (x_i - \bar{x})] \\ &= \frac{1}{n} \left[n \cdot 500 + \frac{100}{s} \cdot \underbrace{\sum_{i=1}^n (x_i - \bar{x})}_{=0} \right] = 500\end{aligned}$$

b) Concordo, pois

$$\begin{aligned}s_y^n &= \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2 = \frac{1}{n-1} \sum_{i=1}^n \left[500 + \frac{100}{s} \cdot (x_i - \bar{x}) - 500 \right]^2 \\ &= \frac{1}{n-1} \sum_{i=1}^n \left[\frac{100}{s} \cdot (x_i - \bar{x}) \right]^2 = \frac{100^2}{s^2} \cdot \underbrace{\frac{1}{n-1} \cdot \sum_{i=1}^n (x_i - \bar{x})^2}_{\substack{=s^2 \\ \text{variância das} \\ \text{notas originais}}} = 100^2.\end{aligned}$$

Logo, o desvio padrão das notas transformadas é $s_y = \sqrt{100^2} = 100$

- Concordo, está de acordo com a regra empírica apresentada na seção [Para saber mais](#): o intervalo centrado na média mais ou menos um desvio padrão corresponde às notas entre 400 e 600.
- Concordo, está de acordo com a regra empírica apresentada na seção [Para saber mais](#): o intervalo centrado na média mais ou menos dois desvios padrões corresponde às notas entre 300 e 700.