Estatística Descritiva

Felipe Quintino

felipe.quintino@unb.br

Departamento de Estatística-EST
Universidade de Brasília-UnB

Minicurso: Introdução ao R para análise de dados de imigração

Semana Universitária de 2025

Sumário

1. Estatística descritiva
   - Distribuição de frequência
   - Gráficos apropriados para cada tipo de variável
   - Medidas de Posição (Média, Mediana, Moda)
   - Medidas de Dispersão (Variância, Desvio Padrão e Coeficiente de Variação)
   - Quantis e Boxplot

2. Referências
   - Referências

Distribuição de frequência

Definição 1

Uma **distribuição de frequência** *é uma tabela que mostra classes ou intervalos dos valores com a contagem do número de ocorrências em cada classe ou intervalo. A* **frequência** *f de uma classe é o número de ocorrências de dados na classe.*

Exemplo: Apresentação de Dados Numéricos em Tabelas

Como podemos resumir e apresentar as informações de idade dos alunos das turmas de Estatística Aplicada e Bioestatística?

Idades: 18, 18, 21, 18, 27, 45, 29, 18, 18, 22, 20, 22, 33, 25, 19, 23, 20, 23, 24

Tabela de Frequências

Table: Tabela de idades de uma amostra dos alunos das turmas de Estatística aplicada e Bioestatística. UnB, 2º/2022

Idade	Frequência
18	5
19	1
20	2
21	1
22	2
23	2
24	1
25	1
27	1
29	1
33	1
45	1

- Podemos resumir mais as informações de idades, agrupando as idades de forma conveniente para os interesses da análise.
- Além disso, podemos controlar a quantidade de categorias que aparecerão na tabela.
- Para isso, transformaremos a variável numérica em categórica.

Table: Tabela de grupos de idades de uma amostra dos alunos das turmas de Estatística aplicada e Bioestatística. UnB, 2º/2022

Idade (em anos)	Frequência
Até 18	5
De 19 a 24	9
De 25 a 29	3
30 ou mais	2

Atenção!

Apesar de obter uma representação tabular "mais amigável", perdemos a escala de mensauração de razão. Desse modo, não conseguimos fazer contas aritméticas diretamente com os dados apresentados na tabela.

Observação 2

- *Em uma distribuição de frequência, é melhor quando todas as classes têm a mesma amplitude.*
- *Normalmente, utiliza-se o valor mínimo dos dados para o limite inferior da primeira classe.*
- *Às vezes, pode ser mais conveniente escolher um valor que seja um pouco menor que o valor mínimo.*
- *A distribuição de frequência produzida irá variar levemente.*

- Cada classe tem um **limite inferior** de classe, que é o menor número que pode pertencer à classe, e um **limite superior** de classe, que é o maior número que pode pertencer à classe.
- A **amplitude de classe** é a distância entre os limites inferiores (ou superiores) de classes consecutivas.
- A diferença entre os valores máximo e mínimo dos dados é chamada de **amplitude**.

Depois de construir uma distribuição de frequência, você pode incluir diversas características adicionais que ajudarão a fornecer um melhor entendimento dos dados. Essas características podem ser incluídas como colunas adicionais em sua tabela.

Depois de construir uma distribuição de frequência, você pode incluir diversas características adicionais que ajudarão a fornecer um melhor entendimento dos dados. Essas características podem ser incluídas como colunas adicionais em sua tabela.

- O **ponto médio** de uma classe é a soma dos limites inferior e superior da classe dividida por dois;
  - O ponto médio também é chamado de **representante da classe**.
- A **frequência relativa** de uma classe é a fração, ou proporção, de dados que está nessa classe.
  - Para calcular a frequência relativa de uma classe, divida a frequência f pelo tamanho da amostra n.
- A **frequência acumulada** de uma classe é a soma das frequências dessa classe com todas as anteriores.
  - A frequência acumulada da última classe é igual ao tamanho n da amostra.

Table: Tabela de grupos de idades de uma amostra dos alunos das turmas de Estatística aplicada e Bioestatística. UnB, 2º/20222

Idade (em anos)	f	$fr(\%)$	fa
Até 18	5		
De 19 a 24	9		
De 25 a 29	3		
30 ou mais	2		
Total	19		

Table: Tabela de grupos de idades de uma amostra dos alunos das turmas de Estatística aplicada e Bioestatística. UnB, 2º/2022

Idade (em anos)	f	$fr(\%)$	fa
Até 18	5	26%	5
De 19 a 24	9	47%	14
De 25 a 29	3	16%	17
30 ou mais	2	11%	19
Total	19	100%	19

Legenda:

- f: frequência
- $fr(\%)$: frequência relativa
- fa: frequência acumulada

Atividade

O conjunto de dados a seguir lista os preços (em dólares) de 30 aparelhos GPS (global positioning system) portáteis. Construa uma distribuição de frequência com sete classes.

128	100	180	150	200	90	340	105	85	270
200	65	230	150	150	120	130	80	230	200
110	126	170	132	140	112	90	340	170	190

Gráficos apropriados para cada tipo de variável

As representações gráficas fornecem, em geral, uma visualização mais sugestiva do que as tabelas. Portanto, constituem-se numa forma alternativa de apresentação das distribuições de frequências.

Alguns exemplos de gráficos são:

- Gráfico de barras (ou de colunas)
- Gráficos de Pizza (setores)
- Histograma
- Boxplot
- Gráfico de linhas
- Mapas

Elementos presentes em gráficos

Alguns elementos importantes de serem observados nos gráficos são:

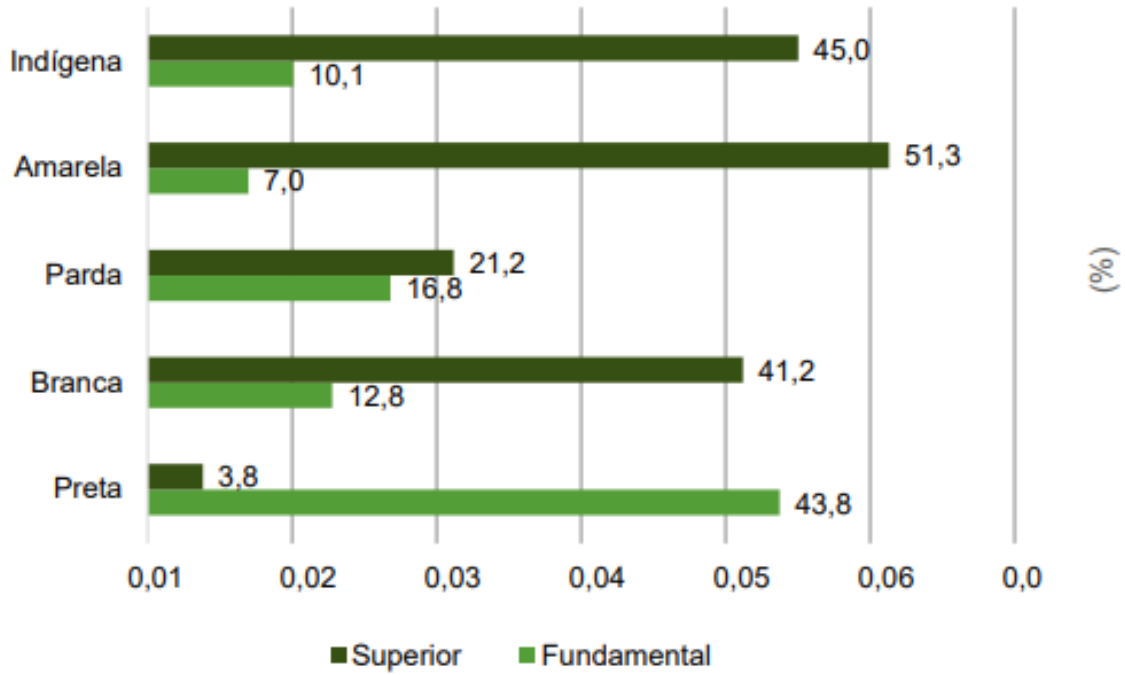
- Título do gráfico
  - O que?
  - Onde?
  - Quando?
- Legenda;
- Rótulos dos eixos e de dados;
- Linhas de grade (opcional);
- Títulos dos eixos de valores (vertical) e de categorias (horizontal);
- Fonte (quem elaborou o gráfico?).

Gráfico de barras (ou de colunas)

No **gráfico de barras (colunas)** cada categoria é representada por uma barra (coluna) de comprimento proporcional à sua frequência, conforme identificação do eixo horizontal (vertical).

Exemplo de Gráfico de barras

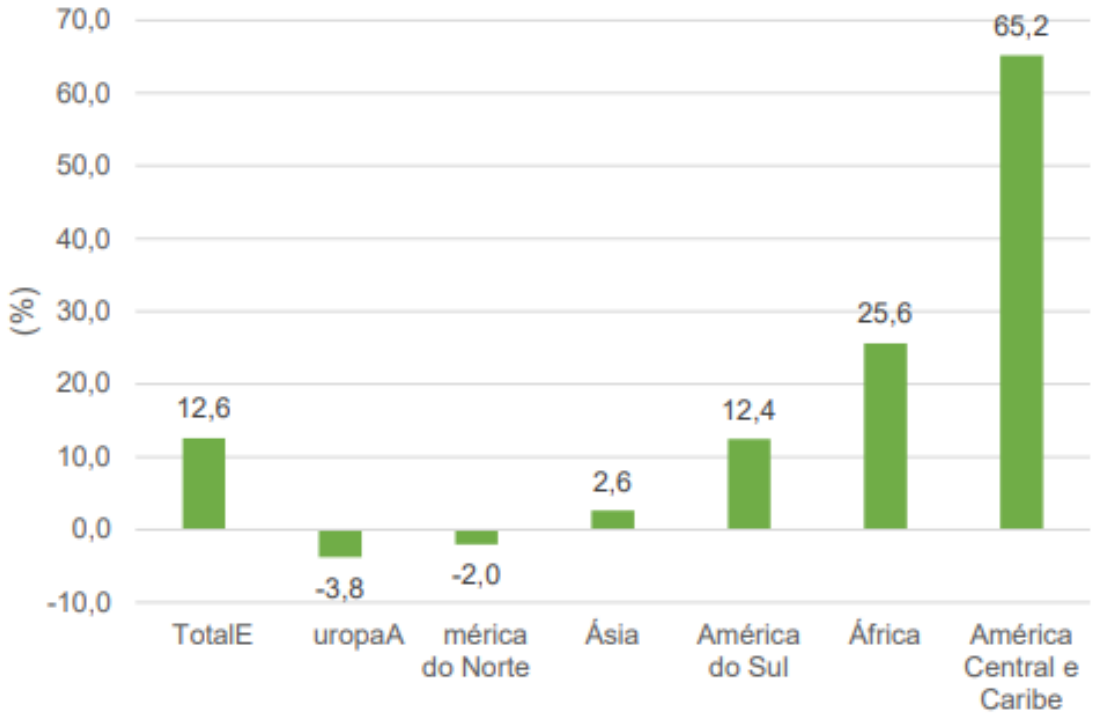
Gráfico 8. Distribuição percentual dos imigrantes no mercado formal de trabalho por nível de instrução, segundo cor ou raça – Brasil 2020



Fonte: Elaborado pelo OBMigra, a partir dos dados do Ministério da Economia, base harmonizada RAIS - CTPS estoque, 2020.

Exemplo de Gráfico de colunas

Gráfico 1. Taxas médias anuais de crescimento do número de imigrantes no mercado formal de trabalho brasileiro, total e continentes – 2011 a 2020



Fonte: Elaborado pelo OBMigra, a partir dos dados do Ministério da Economia, base harmonizada RAIS-CTPS estoque, 2011-2020.

Nota: Não são apresentados resultados para a categoria outros.

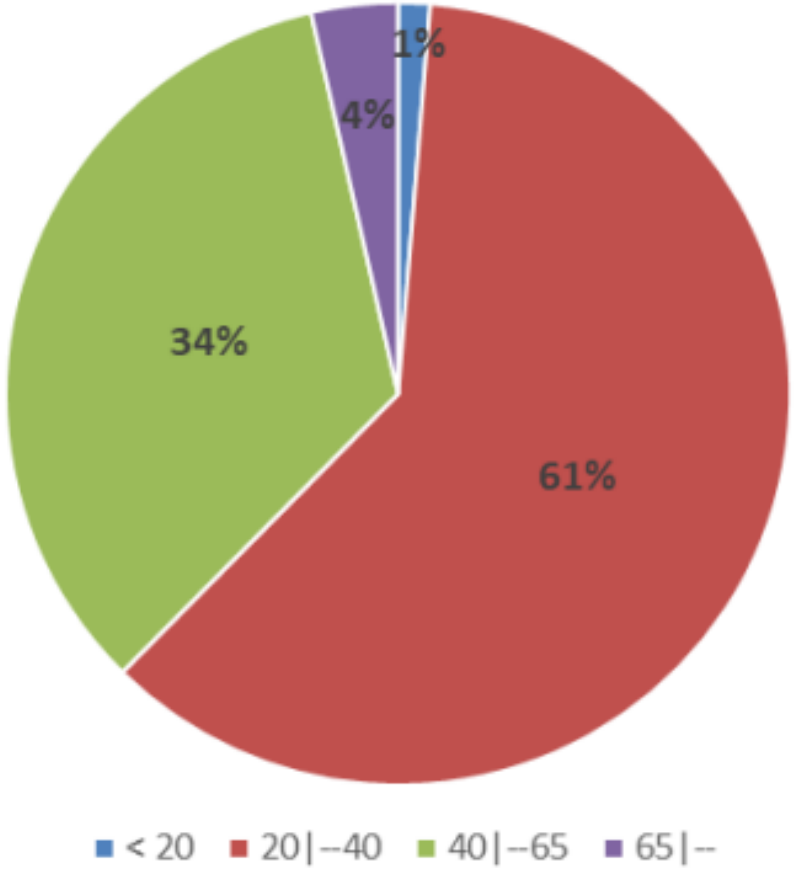
Gráfico de pizza (ou de setores)

Gráficos de pizza (ou de setores) fornecem uma maneira conveniente de apresentar graficamente dados qualitativos como percentagens de um todo.

- Um gráfico de pizza é um círculo dividido em setores que representam categorias.
- A área de cada setor é proporcional à frequência de cada categoria.

Exemplo de Gráfico de pizza (ou de setores)

Gráfico 6.2. Proporção de migrantes com vínculo formal de trabalho, por grupos de idade, segundo grupos de idades.



Fonte: Ministério do Trabalho, Relação Anual de Informações Sociais, 2017.

Exercício

Identifique a tabela associada a cada um dos gráficos anteriores.

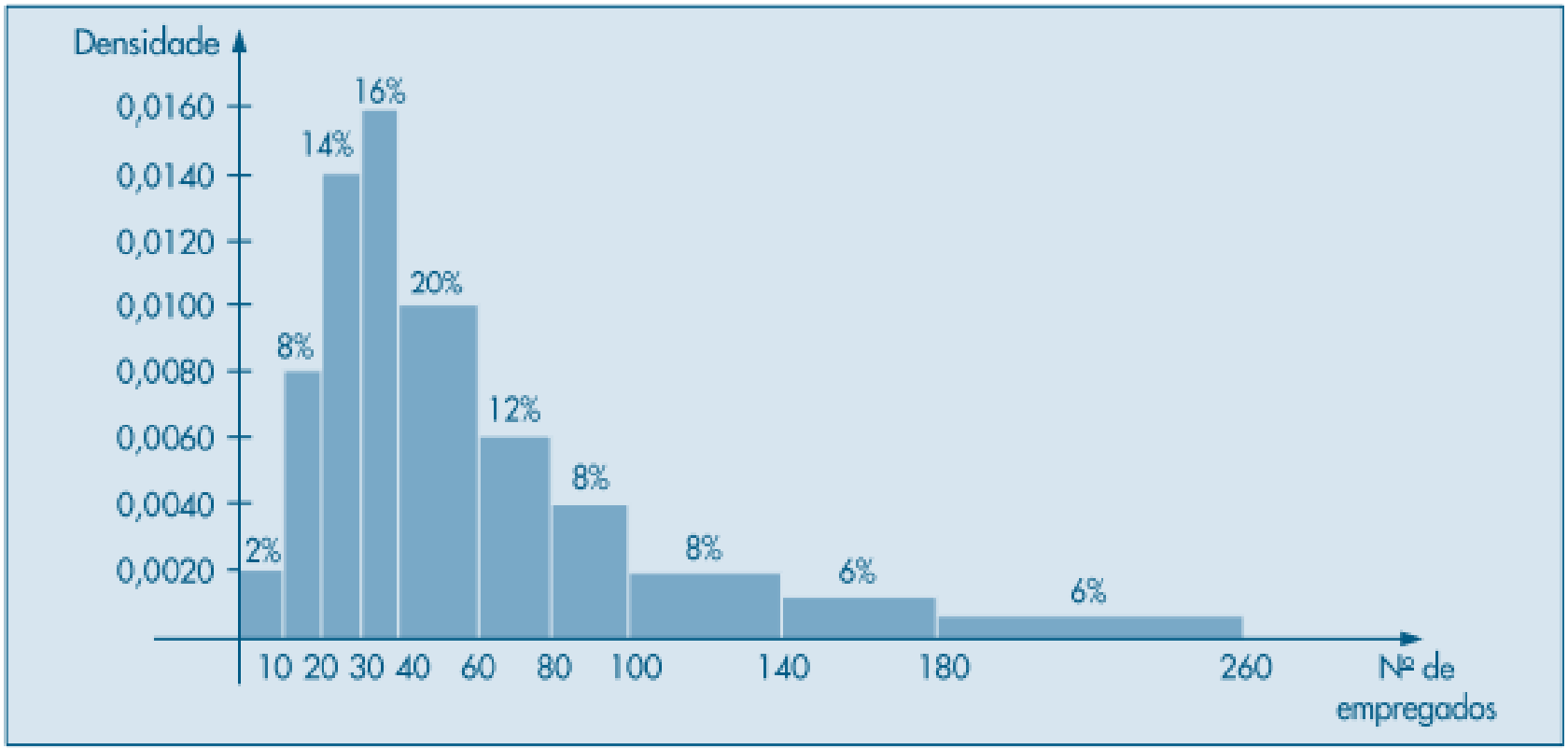
Histograma

Um **histograma de frequência** é um diagrama de barras (colunas) que representa a distribuição de frequência de um conjunto de dados.

Um histograma tem as seguintes propriedades:

1. A escala horizontal é quantitativa e indica os valores dos dados.
2. A escala vertical indica as frequências das classes.
3. Barras (colunas) consecutivas devem estar encostadas umas nas outras.

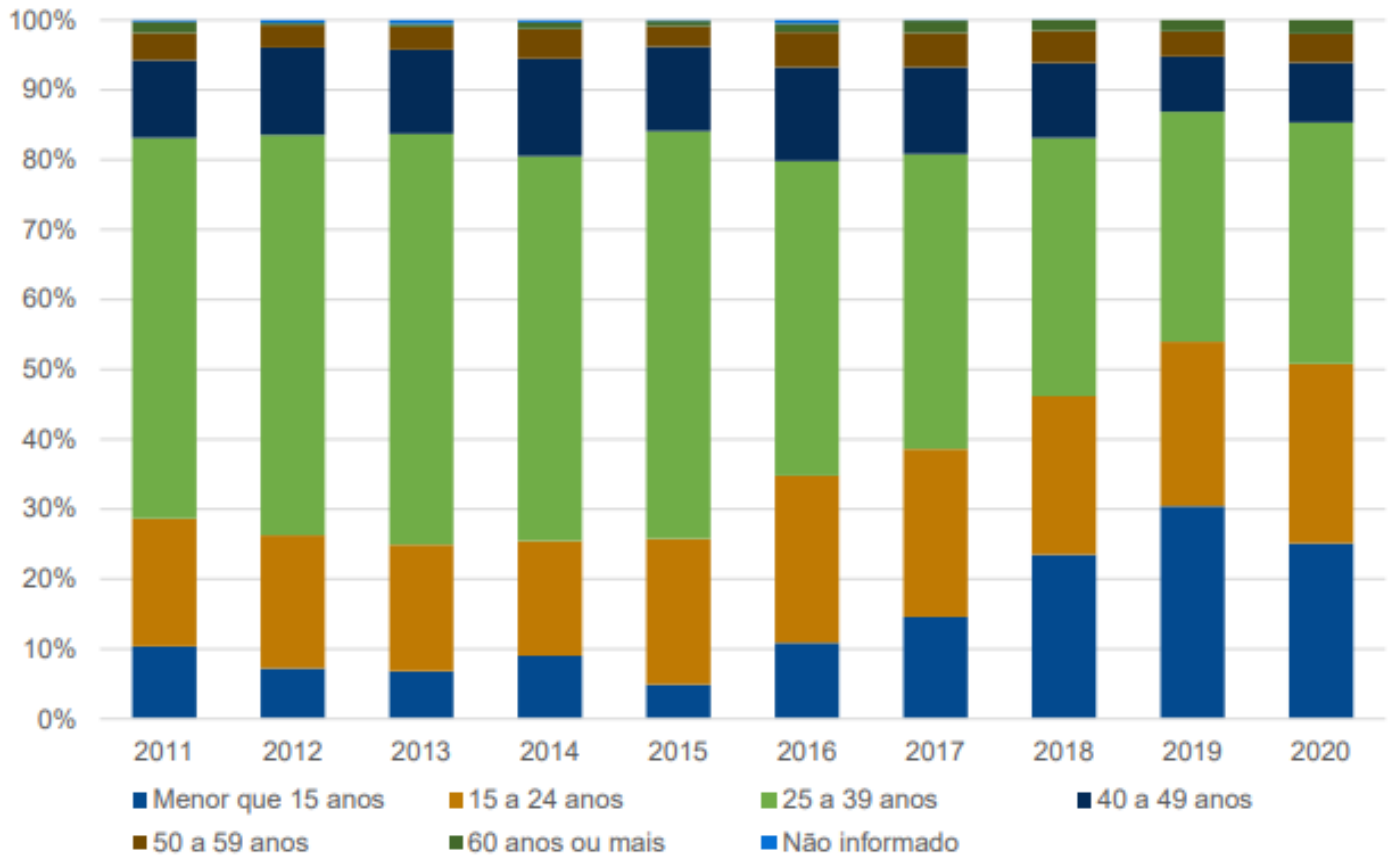
Exemplo de Histograma



Fonte: Figura 2.19 em Bussab e Moretin.

Exemplo de Gráfico de colunas

Gráfico 2.2. Distribuição relativa das solicitações de reconhecimento da condição de refugiado apresentadas por latino-americanos, por grupos de idade, Brasil, 2011-2020



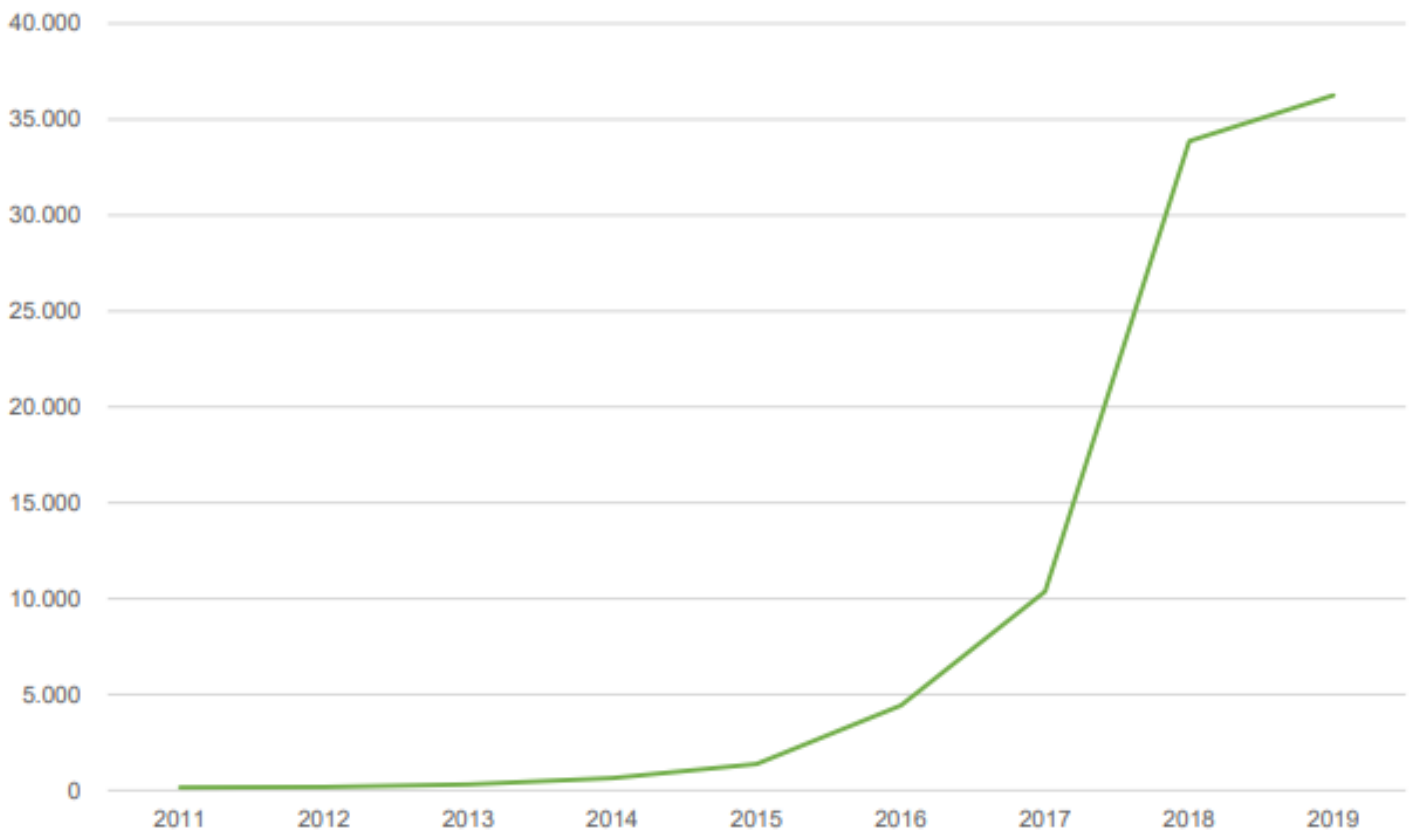
Fonte: Elaborado pelo OBMigra, a partir dos dados da Polícia Federal, Solicitações de refúgio (STI-MAR).

Gráfico de linhas (temporal)

Podemos utilizar o **gráfico de linhas** para representar fenômenos que evoluem ao longo do tempo.

Exemplo de Gráfico de linhas (temporal)

Gráfico 3.1.1. Número de carteiras de trabalho emitidas para solicitantes de refúgio e refugiados latino-americanos, Brasil, 2011-2019



Fonte: Elaborado pelo OBMigra, a partir dos dados do Ministério da Economia, CTPS, 2011-2019.

Mapas de calor

Informações geográficas, além das opções de barras, colunas e pizza, é possível fazer uma representação via **mapas de calor**, indicando a intensidade da frequência em cada região.

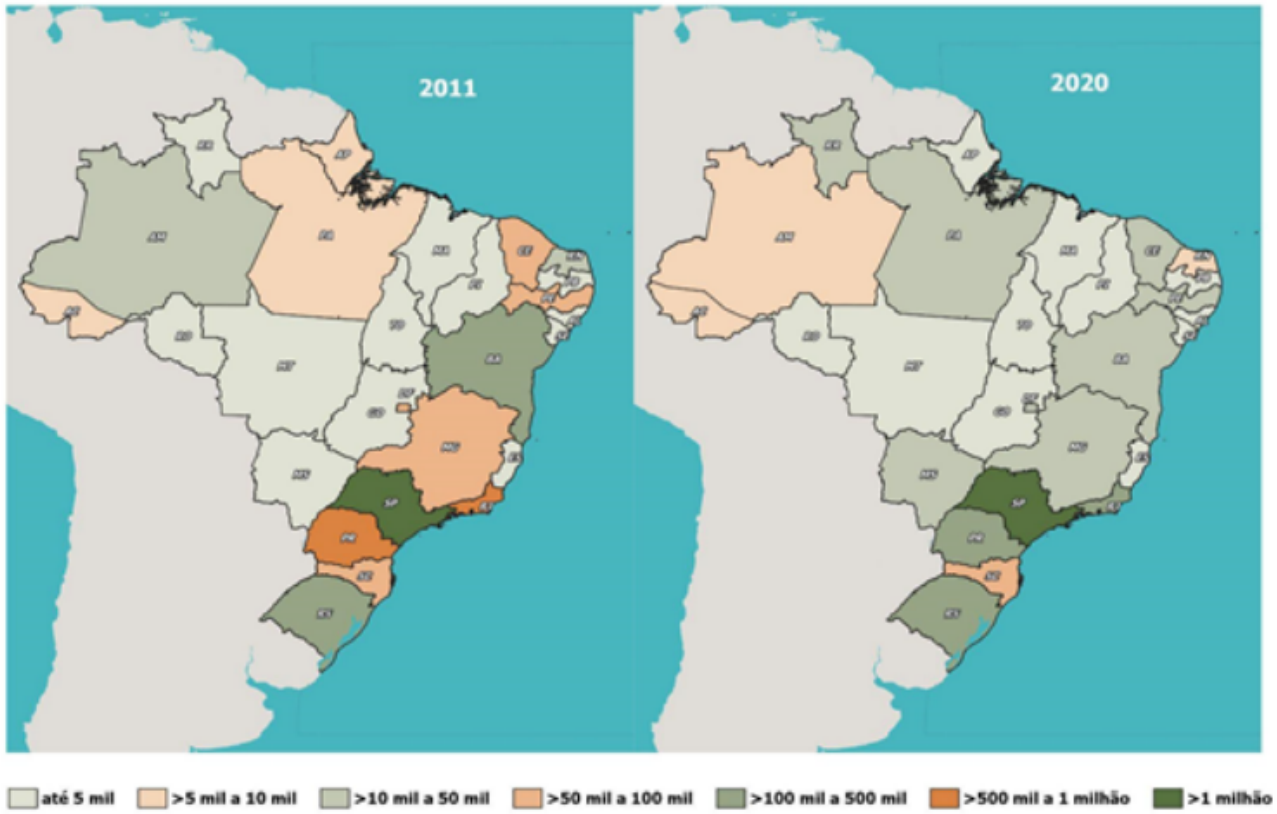
Exemplo de Mapa de Calor do Brasil

Mapa 3. Número de imigrantes solicitantes de residência, segundo Brasil, Grandes Regiões e Unidades da Federação, 2011 - 2020

Fonte: Elaborado pelo OBMigra, a partir dos dados da Polícia Federal – SisMigra, 2020.

Exemplo de Mapa de Calor do Brasil

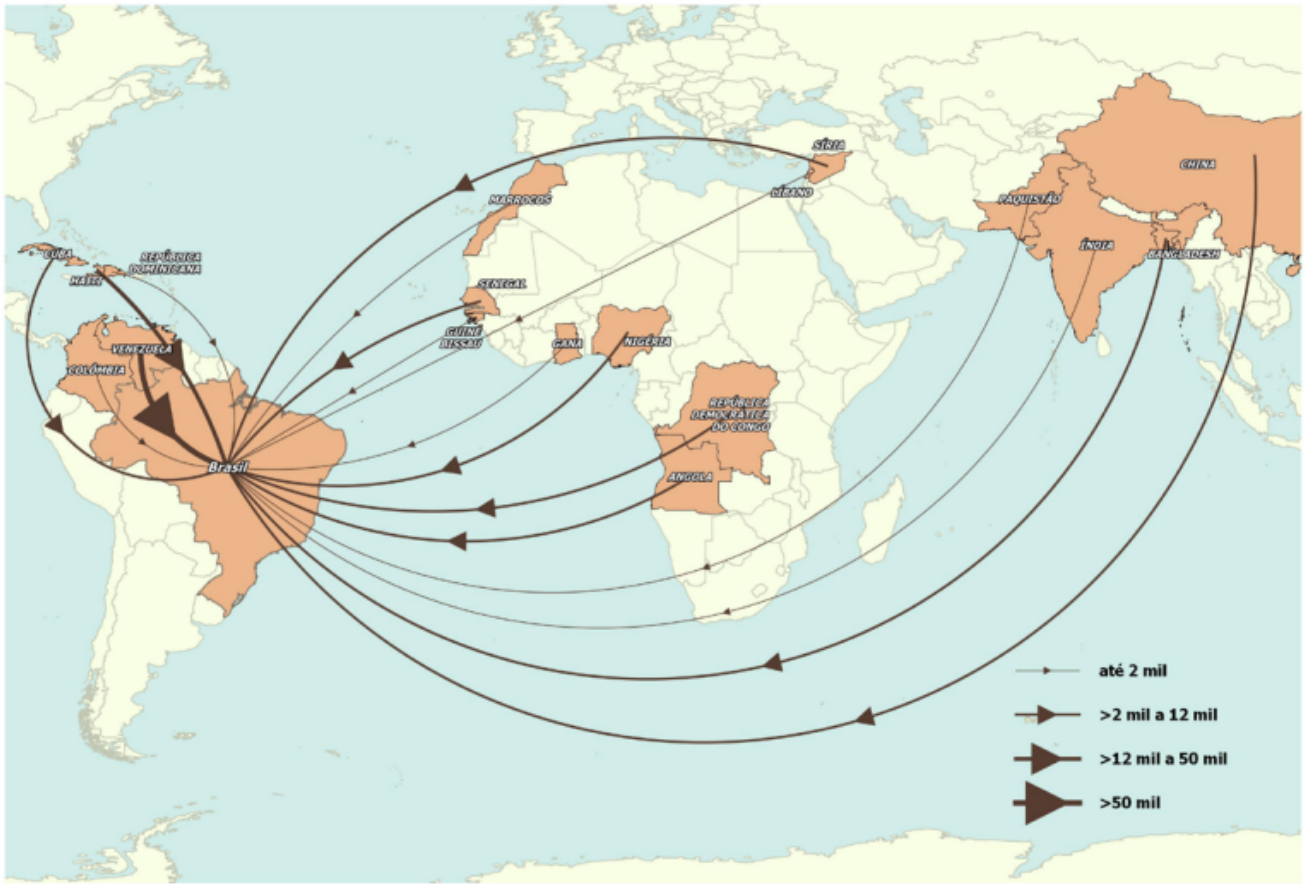
Mapa 1. Entradas de mulheres no território brasileiro nos pontos de fronteira, Unidades da Federação, por ano, Brasil, 2011 e 2020



Fonte: Elaborado pelo OBMigra, a partir dos dados da Polícia Federal, Sistema de Tráfego Internacional (STI), 2020.

Exemplo de Mapa de Fluxo

Mapa 1. Número de solicitantes do reconhecimento da condição de refugiado, segundo principais países de nascimento (*) - Brasil, 2011 - 2020



Fonte: Elaborado pelo OBMigra, a partir dos dados da Polícia Federal - STI-MAR, 2020.

Atividade

Faça uma análise de cada gráfico apresentado na aula.

Medidas de Posição

Vimos que o resumo de dados por meio de tabelas de frequências e gráficos fornece muito mais informações sobre o comportamento de uma variável do que a própria tabela original de dados. Muitas vezes, queremos resumir ainda mais estes dados, apresentando um ou alguns valores que sejam representativos da série toda.

Medidas de Posição

Vimos que o resumo de dados por meio de tabelas de frequências e gráficos fornece muito mais informações sobre o comportamento de uma variável do que a própria tabela original de dados. Muitas vezes, queremos resumir ainda mais estes dados, apresentando um ou alguns valores que sejam representativos da série toda.

- Média
- Mediana
- Moda

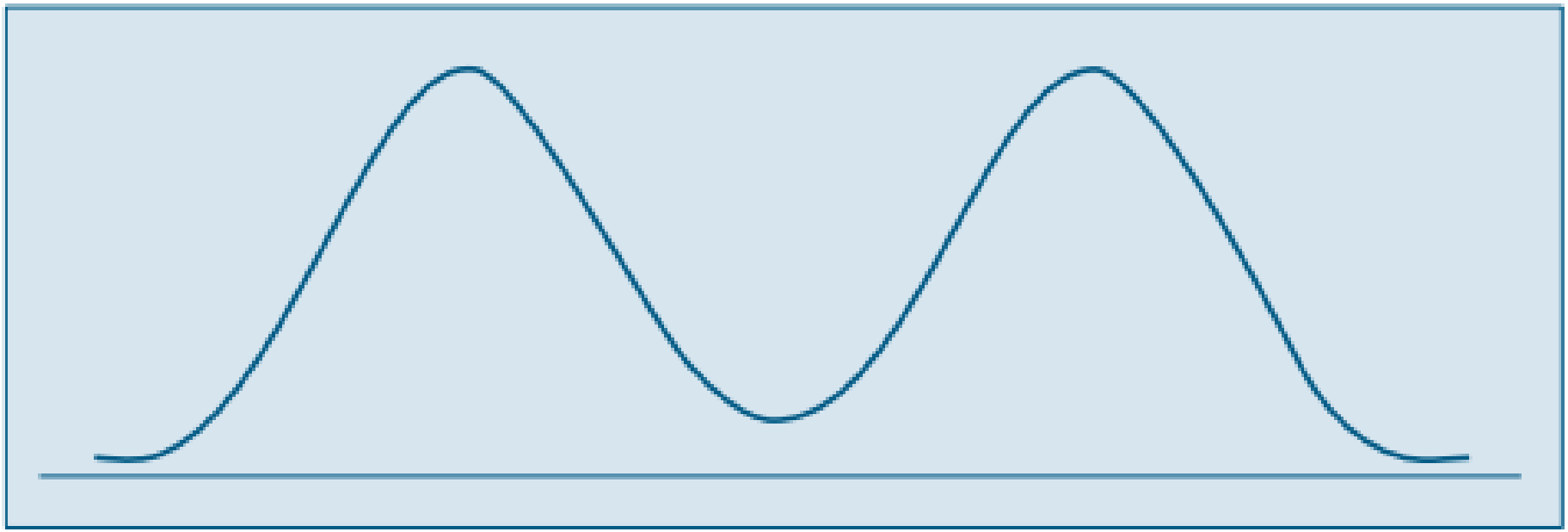
Moda

A **moda** é definida como a realização mais frequente do conjunto de valores observados.

Exemplo 3

1. *A moda do conjunto de dados* $\{1, 2, 2, 1, 3, 4, 1\}$ *é 1.*
2. *As modas do conjunto de dados* $\{1, 2, 2, 1, 3, 4, 1, 2\}$ *são 1 e 2.*

Um exemplo de distribuição de dados com duas modas



Fonte: ver Exercício 5, pág. 41, em Bussab e Morettin.

Mediana

A **mediana** é a realização que ocupa a posição central da série de observações, quando estão ordenadas em ordem crescente.

- Assim, se as cinco observações de uma variável forem $\{3, 4, 7, 8, 8\}$, a mediana é o valor 7, correspondendo à terceira observação.
- Quando o número de observações for par, usa-se como mediana a média aritmética das duas observações centrais. Acrescentando-se o valor 9 à série acima, a mediana será $(7 + 8)/2 = 7{,}5$.

Média aritmética

A **média aritmética** é a soma das observações dividida pelo número delas.

Por exemplo, a média aritmética de $\{3, 4, 7, 8, 8\}$ é

$$\frac{3+4+7+8+8}{5} = 6.$$

Notações

Considere n números $x_1, x_2, \cdots, x_n$. Definimos:

- o **somatório** é a soma dos n números e utilizamos a notação

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

- a **soma dos termos quadráticos** é o somatório dos valores $x_1^2, x_2^2, \cdots, x_n^2$, ou seja,

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

- o **produtório** é o produto dos números

$$\prod_{i=1}^{n} x_i = x_1 \cdot x_2 \cdot \cdots \cdot x_n.$$

A média $\bar{x}$ dos números $x_1, x_2, \cdots, x_n$ pode ser reescrita como

$$\begin{aligned} \bar{x} &= \frac{x_1 + x_2 + \cdots + x_n}{n} \\ &= \frac{1}{n} \sum_{i=1}^{n} x_i. \end{aligned}$$

Algumas propriedades do somatório

As seguintes expressões são válidas

1. $$\sum_{i=1}^{n} c = nc, \quad \text{para c constante;}$$

2. $$\sum_{i=1}^{n} cx_i = c\sum_{i=1}^{n} x_i, \quad \text{para c constante;}$$

3. $$\sum_{i=1}^{n}(x_i + y_i) = \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i;$$

4. $$\sum_{i=1}^{n}(x_i + y_i)^2 = \sum_{i=1}^{n} x_i^2 + 2\sum_{i=1}^{n} x_i y_i + \sum_{i=1}^{n} y_i^2;$$

Média ponderada

Nos casos em que for conveniente é possível utilizar pesos a cada valor do conjunto de dados que desejamos calcular a média. Chamaremos o resultado de **média ponderada**.

Por exemplo, a nota final (NF) do curso pode ser obtida por uma média ponderada das notas obtidas nas avaliações 1, 2 e 3 (A_1, A_2, A_3) utilizando, respectivamente, os pesos $30, 30$ e 40%.

$$NF = \frac{30 \cdot A_1 + 30 \cdot A_2 + 40 \cdot A_3}{100}.$$

Atividade

Para a amostra de idades dos alunos das turmas de Estatística Aplicada e Bioestatística do 2º/2022, determine as medidas de posição média, mediana e moda.

Idades: 18, 18, 21, 18, 27, 45, 29, 18, 18, 22, 20, 22, 33, 25, 19, 23, 20, 23, 24

Medidas de Dispersão

O resumo de um conjunto de dados por uma única medida representativa de posição central esconde toda a informação sobre a variabilidade do conjunto de observações.

Exemplo 4

Suponhamos que cinco grupos de alunos submeteram-se a um teste, obtendo-se as seguintes notas:

$$\begin{aligned} \textit{grupo A (variável X)}: &\quad 3,4,5,6,7 \\ \textit{grupo B (variável Y)}: &\quad 1,3,5,7,9 \\ \textit{grupo C (variável Z)}: &\quad 5,5,5,5,5 \\ \textit{grupo D (variável W)}: &\quad 3,3,5,7 \\ \textit{grupo E (variável V)}: &\quad 3,5,5,6,6 \end{aligned}$$

- Observe que no Exemplo 4, a identificação de cada uma destas séries por sua média (5, em todos os casos) nada informa sobre suas diferentes variabilidades.
- Notamos, então, a conveniência de serem criadas medidas que sumarizem a variabilidade de um conjunto de observações e que nos permita, por exemplo, comparar conjuntos diferentes de valores, como os dados acima, segundo algum critério estabelecido
- Um critério frequentemente usado para tal fim é aquele que mede a dispersão dos dados em torno de sua média, e duas medidas são as mais usadas: desvio médio e variância.

O princípio básico é analisar os desvios das observações em relação à média dessas observações.

Voltando ao Exemplo 4, note que a soma dos desvios $x_i - \bar{x}$ é zero!

$$\sum_{i=1}^{5}(x_i - \bar{x}) = -2 - 1 + 0 + 1 + 2 = 0.$$

O mesmo ocorre nos demais grupos!

Duas alternativas que evitam o problema exposto no exemplo anterior são considerar

(a) o total dos desvios absolutos

$$\sum_{i=1}^{5} |x_i - \bar{x}| = 2 + 1 + 0 + 1 + 2 = 6;$$

(b) considerar o quadrado dos resíduos

$$\sum_{i=1}^{5} (x_i - \bar{x})^2 = 4 + 1 + 0 + 1 + 4 = 10.$$

Desvio médio e variância

Definimos o **desvio médio** e a **variância**, respectivamente, por

$$dm(X) = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n}$$

e

$$var(X) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}.$$

Desvio padrão

- Sendo a variância uma medida de dimensão igual ao quadrado da dimensão dos dados (por exemplo, se os dados são expressos em cm, a variância será expressa em cm^2), pode causar problemas de interpretação.
- Costuma-se usar, então, o **desvio padrão**, que é definido como a raiz quadrada positiva da variância:

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

- Ambas as medidas de dispersão (dm e dp) indicam em média qual será o "erro" (desvio) cometido ao tentar substituir cada observação pela medida resumo do conjunto de dados (no caso, a média).

Observação 5

A variância também pode ser calculada pela seguinte fórmula:

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

Atividade

Quer se estudar o número de erros de impressão de um livro. Para isso escolheu-se uma amostra de 50 páginas, encontrando-se o número de erros por página da tabela abaixo.

(a) Qual o número médio de erros por página?

(b) E o número mediano?

(c) Qual é o desvio padrão?

(d) Faça uma representação gráfica para a distribuição.

(e) Se o livro tem 500 páginas, qual o número total de erros esperado no livro?

Erros	frequência
0	25
1	20
2	3
3	1
4	1

Quantis e Boxplot

- Quantis e Percentis
- Boxplot
- Assimetria

Tanto a média como o desvio padrão podem não ser medidas adequadas para representar um conjunto de dados, pois:

(a) são afetados, de forma exagerada, por valores extremos;

(b) apenas com estes dois valores não temos idéia da simetria ou assimetria da distribuição dos dados.

Definimos um **quantil de ordem** p ou p-**quantil**, indicada por $q(p)$, onde p é uma proporção qualquer, $0 < p < 1$, tal que $100p\%$ das observações sejam menores do que $q(p)$.

Definimos um **quantil de ordem** p ou p-**quantil**, indicada por $q(p)$, onde p é uma proporção qualquer, $0 < p < 1$, tal que $100p\%$ das observações sejam menores do que $q(p)$.

Indicamos, abaixo, alguns quantis e seus nomes particulares.

$$
\begin{array}{lllllll}
q(0,25) = & q_1: & 1º \text{ Quartil} & = & 25º \text{ Percentil} & & \\
q(0,50) = & q_2: & \text{Mediana} & = & 2º \text{ Quartil} & = & 50º \text{ Percentil} \\
q(0,75) = & q_3: & 3º \text{ Quartil} & = & 75º \text{ Percentil} & &
\end{array}
$$

Exemplo 6

Suponha que tenhamos os seguintes valores ordenados:

$$2\ \mathbf{3}\ \mathbf{5}\ 7\ \mathbf{8}\ 10\ \mathbf{11}\ \mathbf{12}\ 15.$$

Os valores da média, q_1, mediana(q_2) e q_3 são apresentados a seguir

Exemplo 6

Suponha que tenhamos os seguintes valores ordenados:

$$2\ \mathbf{3}\ \mathbf{5}\ 7\ \mathbf{8}\ 10\ \mathbf{11}\ \mathbf{12}\ 15.$$

Os valores da média, q_1, mediana(q_2) e q_3 são apresentados a seguir

$$\begin{aligned} \bar{x} &= 8,1 \\ q_1 &= \frac{3+5}{2} = 4 \\ q_2 &= 8 \\ q_3 &= \frac{11+12}{2} = 11,5 \end{aligned}$$

Exemplo 7

Acrescentemos, agora, o valor 67 à lista de nove valores do exemplo anterior.

$$2\ 3\ \mathbf{5}\ 7\ \mathbf{8}\ \mathbf{10}\ 11\ \mathbf{12}\ 15\ \ 67.$$

Reclaculando as medidas anteriores, obtemos

Exemplo 7

Acrescentemos, agora, o valor 67 à lista de nove valores do exemplo anterior.

$$2\ 3\ \mathbf{5}\ 7\ \mathbf{8}\ \mathbf{10}\ 11\ \mathbf{12}\ 15\ \ 67.$$

Reclaculando as medidas anteriores, obtemos

$$\begin{aligned} \bar{x} &= \mathbf{14} \\ q_1 &= 5 \\ q_2 &= \frac{8+10}{2} = 9 \\ q_3 &= 12 \end{aligned}$$

Observe que, ao acrescentar uma informação **discrepante** aos valores do Exemplo 6, o valor da média sofreu uma grande alteração se comparado à variação sofrida pela mediana.

Boxplot

O **Boxplot** é um diagrama que dá uma idéia da posição, dispersão, assimetria, caudas e dados discrepantes. A posição central é dada pela mediana e a dispersão pela distância entre quartis (dq)

$$dq = q_3 - q_1.$$

As posições relativas de q_1, q_2 e q_3 dão uma noção da assimetria da distribuição. Os comprimentos das caudas são dados pelas linhas que vão do retângulo aos valores remotos e pelos valores atípicos.

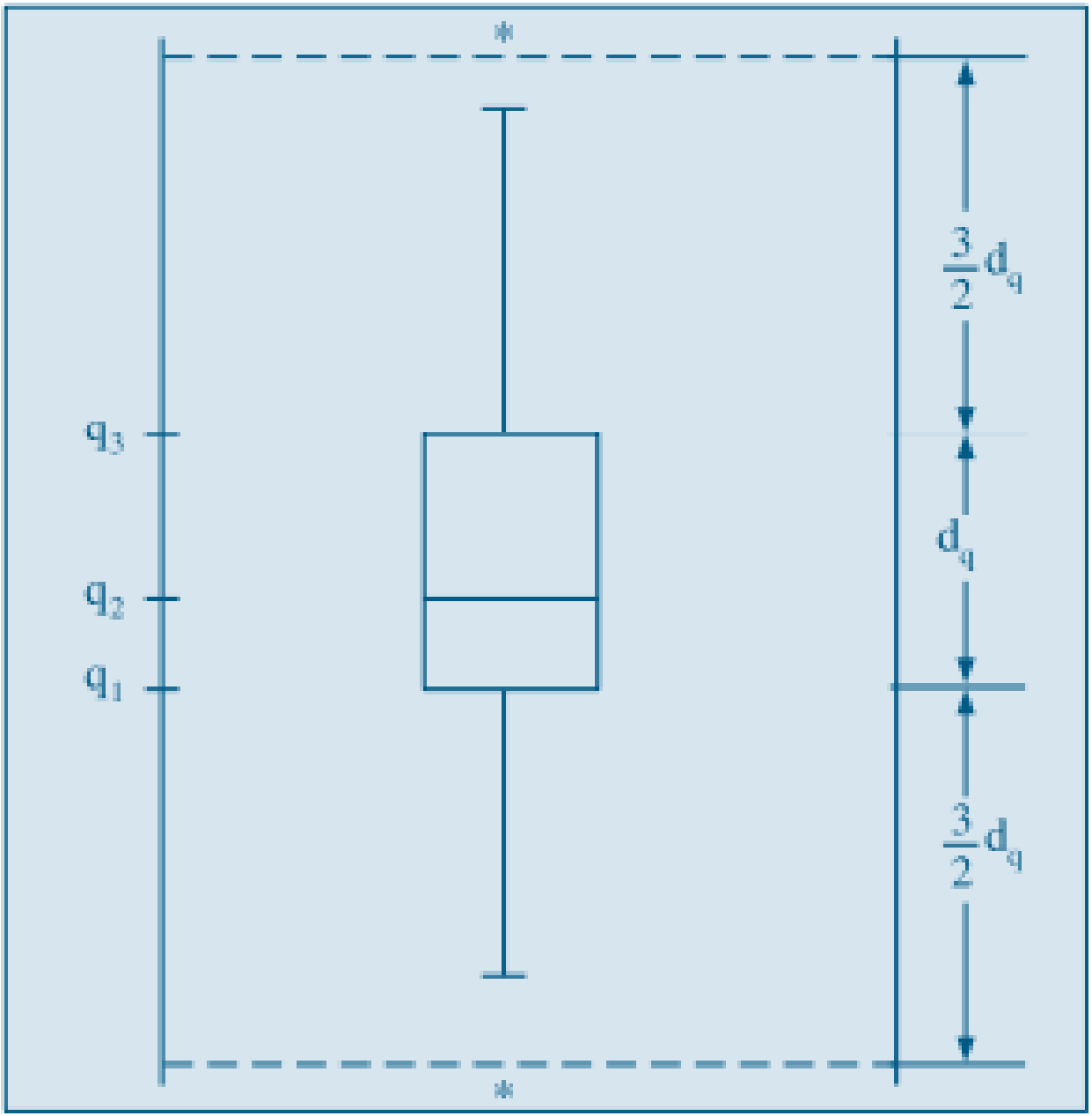
Os **limites superior e inferior** (LS e LI) nos auxiliarão na detecção de **outliers** (valores atípicos) e são definidos a seguir

$$\begin{aligned} LS &= q_3 + 1,5dq, \\ LI &= q_1 - 1,5dq. \end{aligned}$$

Os **limites superior e inferior** (LS e LI) nos auxiliarão na detecção de **outliers** (valores atípicos) e são definidos a seguir

$$\begin{aligned} LS &= q_3 + 1,5dq, \\ LI &= q_1 - 1,5dq. \end{aligned}$$

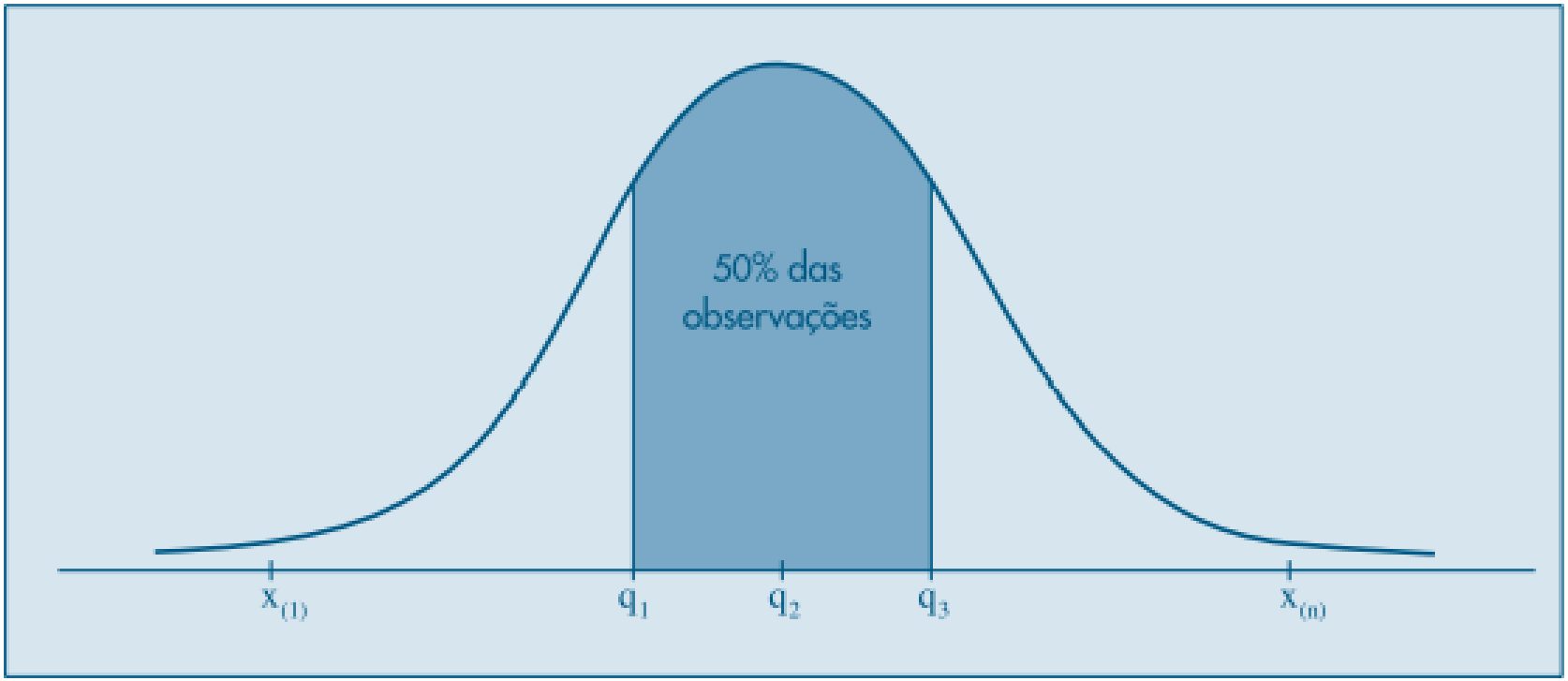
Os valores que estiverem abaixo do limite inferior ou acima do limite superior serão representados "fora da caixa" por um *.


q_3
q_2
q_1
$\frac{3}{2}d_q$
d_q
$\frac{3}{2}d_q$

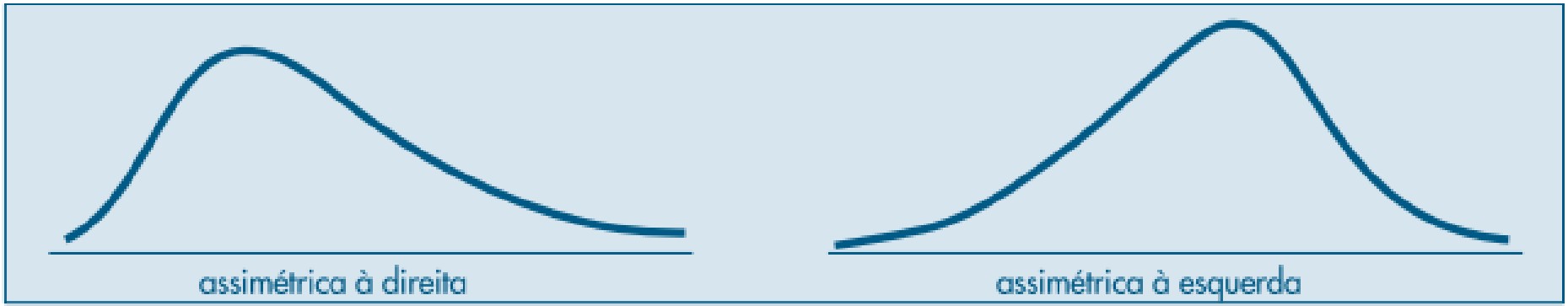
Assimetria

Além do Boxplot, o histograma e o polígono de frequências podem auxiliar visualmente nessa verificação.

Exemplo de Simetria



Exemplo de Assimetria



Atividade

Para a amostra de idades dos alunos das turmas de Estatística Aplicada e Bioestatística do 2º/2022, determine os quantis q_1, q_2 e q_3 e esboce o Boxplot desses dados.

Idades: 18, 18, 21, 18, 27, 45, 29, 18, 18, 22, 20, 22, 33, 25, 19, 23, 20, 23, 24

Referências Bibliográficas


P.A. Barbeta.
Estatística Aplicada às Ciências sociais.
Ed. da UFSC, Florianópolis. 2004.


P.A. Morettin, W.O. Bussab.
Estatística básica.
Saraiva Educação SA; 2017.


R. Lasson, B. Farber.
Estatística Aplicada,
4a edição, Ed. Pearson, São Paulo, 2010.


A.Q. Miah.
Applied statistics for social and management sciences.
Springer; 2016.


M.N. Magalhães, A.C.P. De Lima.
Noções de probabilidade e estatística.
Vol. 5. Editora da Universidade de São Paulo, 2002.