# Análise exploratória I

Motivação e análise descritiva univariada para variáveis quantitativas.

Prof. Me. Lineu Alberto Cavazani de Freitas

**CE003 – Estatística II**

Departamento de Estatística
Laboratório de Estatística e Geoinformação

# Análise exploratória

- Parte primordial de qualquer análise estatística é chamada **análise descritiva** ou **exploratória**.
- Consiste basicamente de **tabelas**, **resumos numéricos** e **análises gráficas** das variáveis disponíveis em um conjunto de dados.
- Trata-se de uma etapa de extrema importância e deve preceder qualquer análise mais sofisticada.
- As técnicas de análise exploratória visam **resumir** e **apresentar** as informações de um conjunto de dados brutos.

# Análise exploratória

- Tentar compreender um conjunto de dados sem algum método que permita resumir as informações é inviável.
- A análise exploratória é a primeira forma de tentarmos enteder o que acontece nos nossos dados.
- Uma das tarefas é a etapa de consistência dos dados, isto é, verificar se os dados coletados são condizentes com a realidade.

Figura 1. Extraído de pixabay.com.

# Análise exploratória

- O conjunto de técnicas aplicáveis está diretamente associado ao **tipo das variáveis de interesse** (quantitativas x qualitativas) e suas ramificações.
- Podemos conduzir análises focadas nas variáveis uma a uma (**análises univariadas**).
- Bem como conduzir análises focadas em avaliar a relação entre as variáveis (**análises multivariadas**).

Figura 2. Extraído de pixabay.com.

# Análise exploratória

Podemos fazer uso diversas técnicas, tais como

- Tabelas de frequência absolutas.
- Tabelas de frequência relativas.
- Tabelas de frequência acumuladas.
- Tabelas para múltiplas variáveis.
- Gráficos (para análise uni e multivariada).
- Medidas de posição central.
- Medidas de posição relativa.
- Medidas de forma.
- Medidas de dispersão.
- Medidas de associação.

# Análise descritiva univariada para variáveis quantitativas

# Análise descritiva univariada para variáveis quantitativas

- Uma variável quantitativa é uma **característica** que pode ser **representada numericamente**.
- Podem ser classificadas em **discretas** (finitos valores em um dado intervalo) ou **contínuas** (infinitos valores em um dado intervalo).
- Quando estamos lidando com **variáveis quantitativas discretas com poucos possíveis valores**, as técnicas apresentadas para variáveis qualitativas se aplicam.

# Tabelas de frequência

Tabela 1. Tabela de frequências para o número de irmãos.

| Irmãos | Frequência | Percentual | Freq. Acumulada | Percentual Acumulado |
|---|---|---|---|---|
| 0 | 4 | 15.4 % | 4 | 15.4 % |
| 1 | 8 | 30.8 % | 12 | 46.2 % |
| 2 | 8 | 30.8 % | 20 | 77 % |
| 3 | 3 | 11.5 % | 23 | 88.5 % |
| 4 | 2 | 7.7 % | 25 | 96.2 % |
| 5 | 1 | 3.8 % | 26 | 100 % |
| Total | 26 | 100 % | 26 | 100 % |

# Gráfico de barras verticais

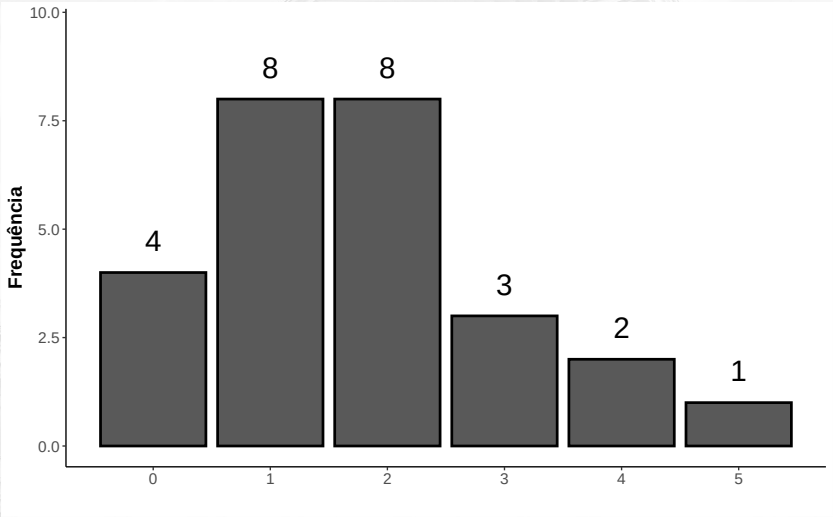


Figura 3. Gráfico de barras verticais para o número de irmãos.

# Análise descritiva univariada para variáveis quantitativas

- Para variáveis quantitativas contínuas ou discretas com muitos possíveis valores, precisamos de técnicas específicas.
- Uma estratégia comum é o **agrupamento em faixas de valores**, e avaliação das frequências nestas faixas.
- Podem ser usadas tabelas de frequências absolutas, relativas e acumuladas para as faixas de valores.
- Utilizando a **razão entre frequência relativa e a amplitude das faixas** de valores, geramos a **densidade**.

# Análise descritiva univariada para variáveis quantitativas

**Faixas de valores**

- Cuidados devem ser tomados quanto às notações e tipos de faixas (aberto e fechado à esquerda ou direita).
- Em geral definimos intervalos **abertos à esquerda** e **fechados à direita**.
- Considerando dois valores $a$ e $b$, em que $a < b$, os intervalos consideram que $a$ **não** está incluído na faixa, $b$ está.

- Notações usuais:
  - $a < y \leq b$
  - $a \vdash b$
  - $(a,b]$
- $5 < y \leq 10$ ou $5 \vdash 10$ ou $[5,10)$
  - Valores maiores que 5 até valores menores ou iguais a 10. 5 não está no intervalo.

# Análise descritiva univariada para variáveis quantitativas

- Como agrupar em classes?
- Qual o tamanho ideal das faixas de valores?
- Classes definidas com a mesma amplitude é o procedimento mais usual.
- Existem procedimentos que podem ser usados para obter a amplitude, como **Sturges**.
- Em geral, 5 a 15 faixas são suficientes.

# Tabelas de frequência para uma variável quantitativa

Tabela 2. Tabela de frequências usando faixas de altura.

| Faixas | Frequência | Freq. Relativa | Freq. Acumulada | Freq. Rel. Acumulada |
|---|---|---|---|---|
| [160,165] | 6 | 0.23 | 6 | 0.23 |
| (165,170] | 2 | 0.08 | 8 | 0.31 |
| (170,175] | 6 | 0.23 | 14 | 0.54 |
| (175,180] | 6 | 0.23 | 20 | 0.77 |
| (180,185] | 4 | 0.15 | 24 | 0.92 |
| (185,190] | 1 | 0.04 | 25 | 0.96 |
| (190,195] | 1 | 0.04 | 26 | 1.00 |

# Tabelas de frequência para uma variável quantitativa

Tabela 3. Tabela de frequências usando faixas de altura.

| Faixas | Frequência | Percentual | Freq. Acumulada | Percentual Acumulado |
|---|---|---|---|---|
| [160,165] | 6 | 23 % | 6 | 23 % |
| (165,170] | 2 | 8 % | 8 | 31 % |
| (170,175] | 6 | 23 % | 14 | 54 % |
| (175,180] | 6 | 23 % | 20 | 77 % |
| (180,185] | 4 | 15 % | 24 | 92 % |
| (185,190] | 1 | 4 % | 25 | 96 % |
| (190,195] | 1 | 4 % | 26 | 100 % |

# Tabelas de frequência para uma variável quantitativa

Tabela 4. Tabela de frequências usando faixas de altura.

| Faixas | Frequência | Percentual | Freq. Acum. | Perc. Acum. | Amplitude | Densidade |
|---|---|---|---|---|---|---|
| [160,165] | 6 | 23 % | 6 | 23 % | 5 | 0.046 |
| (165,170] | 2 | 8 % | 8 | 31 % | 5 | 0.016 |
| (170,175] | 6 | 23 % | 14 | 54 % | 5 | 0.046 |
| (175,180] | 6 | 23 % | 20 | 77 % | 5 | 0.046 |
| (180,185] | 4 | 15 % | 24 | 92 % | 5 | 0.030 |
| (185,190] | 1 | 4 % | 25 | 96 % | 5 | 0.008 |
| (190,195] | 1 | 4 % | 26 | 100 % | 5 | 0.008 |

# Gráficos para representação de frequências de uma variável quantitativa

- Assim como no caso de variáveis qualitativas ou quantitativas discretas com poucos possíveis valores, a representação por meio de gráficos pode ser bastante benéfica para análise de variáveis quantitativas.

Algumas possibilidades são

- Histograma.
- Gráfico de densidade empírica.
- Box-plot

# Histograma

- Consiste em **retângulos contíguos** de base dada pelas faixas de valores definindas para uma variável.
- Algumas possibilidades são:
  - A área representar a frequência da rescpectiva faixa.
  - A altura representar a frequência absoluta na faixa.
  - A altura representar o quociente da área pela amplitude da faixa: a densidade.

# Histograma

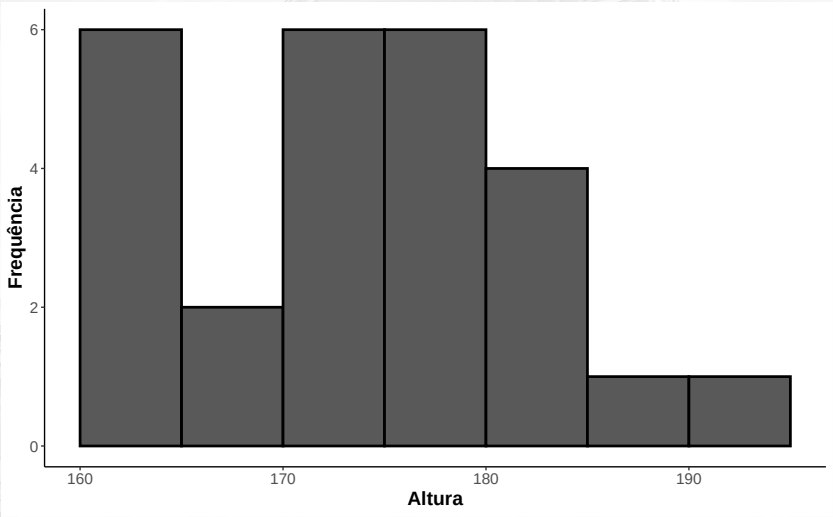


Figura 4. Histograma das alturas dos alunos.

# Efeito do número de classes

- O número de classes pode afetar diretamente as tabelas e gráficos.
- Com poucas classes, os dados ficam excessivamente resumidos e as classes ficam muito heterogêneas.
- Com muitas classes, os dados ficam segmentados em excesso e as representações são comprometidas.

# Efeito do número de classes

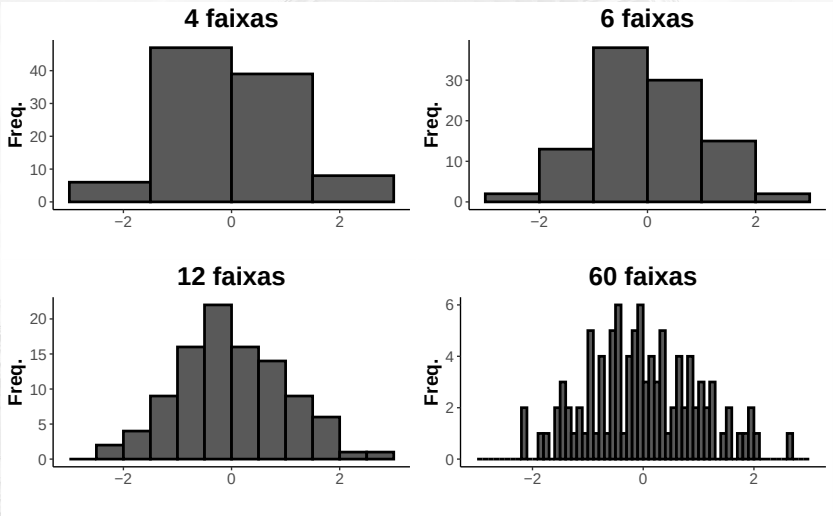


Figura 5. Efeito do número de classes em histogramas.

# Gráfico de densidade empírica

## Intuição

- Imagine uma sequência de histogramas de densidade em que o número de observações aumenta, juntamente com o número de faixas.
- No limite, teremos uma curva.
- Esta curva é chamada de gráfico de densidade empírica.
- É um gráfico "computacionalmente intensivo", depende da definição de uma função kernel e do tamanho da banda.
- A área sob a curva é igual a 1.

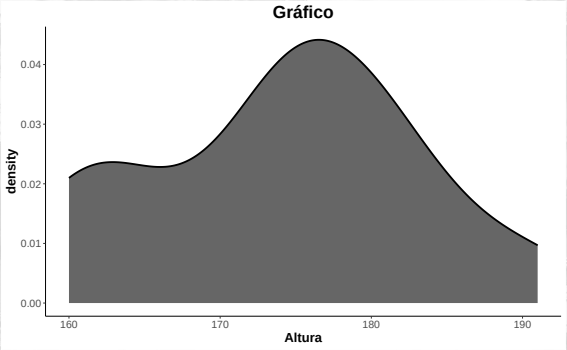


Figura 6. Gráfico de densidade para as alturas dos alunos.

# Box-plot

- Outra importante visualização é o box-plot.
- É possível analisar a distribuição dos dados, aspectos quanto a posição, variabilidade, assimetria e também a presença de valores atípicos.
- Retomaremos o box-plot após estudar quartis, em medidas descritivas.

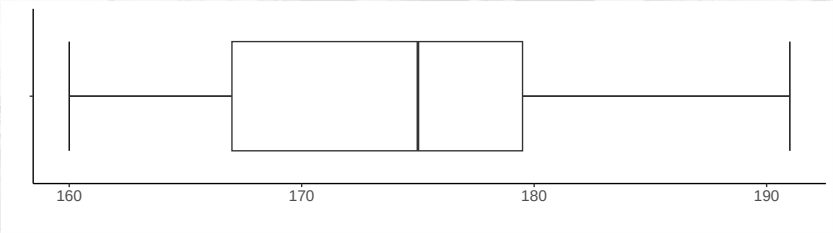


Figura 7. Box-plot das alturas dos alunos.

# Histograma, densidade e box-plot

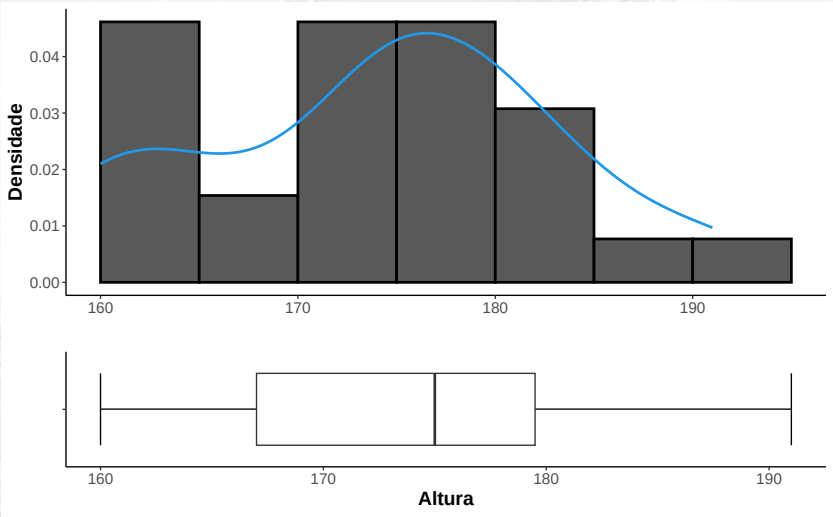


Figura 8. Combinação de representações.

# Assimetria

- Um conjunto pode ser aproximadamente **simétrico**, **assimétrico** à esquerda ou à direita.
- Tais características são facilmente diagnosticadas por meio de análise gráfica usando um histograma, gráfico de densidade ou box-plot.
- Futuramente veremos como diagnosticar assimetria por meio de medidas descritivas.

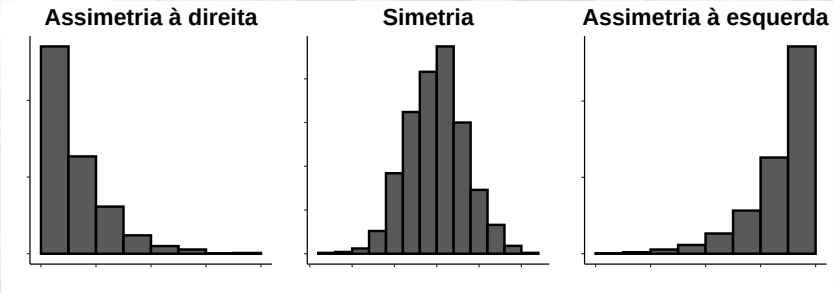


Figura 9. Gráfico de setores para a variável...

**O que foi visto:**

- Introdução à análise exploratória.
- Análise exploratória univariada para variáveis qualitativas.
- Análise exploratória univariada para variáveis quantitativas.

**Próximos assuntos:**

- Resumos numéricos.
- Medidas de posição central.
- Medidas de posição relativa.
- Medidas de dispersão.