Análise exploratória I

Motivação e análise descritiva univariada para variáveis qualitativas e 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 qualitativas

Análise descritiva univariada para variáveis qualitativas

- Uma variável qualitativa representa um atributo que pode ser expresso por meio de **rótulos** ou **categorias**.
- Podem ser classificadas em **nominais** (sem ordenação natural entre os níveis) ou **ordinais** (com ordenação natural entre os níveis).
- As categorias também são chamadas de **classes** ou **níveis**.
- Na análise descritiva de uma variável qualitativa estamos interessados em avaliar as **frequências** das classes.

Tipos de frequência

- **Frequência absoluta** (f_a): número de observações no conjunto de dados que pertence a uma determinada classe.
- **Frequência relativa** (f_r): frequência de classe dividida pelo número total de observações no conjunto de dados.
  - Pode ser apresentada em forma de percentual, quando multiplicada por 100.
- **Frequência acumulada** (F_a ou F_r): frequência absoluta ou relativa acumulada conforme disposição das classes.
  - Não faz muito sentido para variáveis qualitativas nominais.

Tipos de frequência

Exemplos

- Frequência absoluta e relativa dos alunos por gênero.
  - XX do sexo masculino.
  - XX do sexo feminino.
  - $XX/n = 0,XX$ do sexo masculino ($XX\%$).
  - $XX/n = 0,XX$ do sexo feminino ($XX\%$).

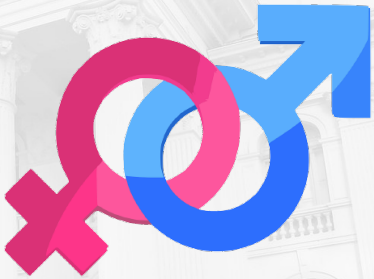

Figura 3. Extraído de pixabay.com.

Tabelas de frequência para uma variável qualitativa

- Utlizando apenas os dados brutos é difícil responder questões de interesse.
- Para reduzir os dados originais de forma que fique mais claro o entendimento dos mesmos são utilizadas as **tabelas de frequência**.
- No caso de variáveis qualitativas consiste em listar os possíveis níveis da variável e fazer a contagem de quantas vezes cada nível aparece nos dados brutos.

Figura 4. Extraído de pixabay.com.

Tabelas de frequência para uma variável qualitativa

- Cada **linha** da tabela diz respeito a um **nível** da variável categórica.
- As **colunas** podem apresentar diferentes tipos de **frequência** (absoluta, relativa, acumulada).
- Alguns cuidados para a apresentação dos resultados dizem respeito ao tipo de variável categórica em questão: nominal ou ordinal.
- Os níveis de variáveis **nominais não apresentam uma ordenação natural**, portanto, na apresentação dos resultados pode ser interessante ordenar os níveis por frequência ou por ordem alfabética.
- Esta estratégia não é recomendada para variáveis **ordinais**, pois estas **apresentam uma ordenação natural** e esta ordenação deve ser preferencialmente mantida na exposição dos resultados.

Tabelas de frequência para uma variável qualitativa nominal

Tabela 1. Tabela de frequências para...

Níveis	Frequência	Freq. Relativa
A	55	0.110
B	55	0.110
C	52	0.104
D	46	0.092
E	40	0.080
F	50	0.100
G	54	0.108
H	46	0.092
I	49	0.098
J	53	0.106
Total	500	1.000

Tabelas de frequência para uma variável qualitativa nominal

Tabela 2. Tabela de frequências para...

Níveis	Frequência	Freq. Relativa
A	55	0.110
B	55	0.110
G	54	0.108
J	53	0.106
C	52	0.104
F	50	0.100
I	49	0.098
D	46	0.092
H	46	0.092
E	40	0.080
Total	500	1.000

Tabelas de frequência para uma variável qualitativa nominal

Tabela 3. Tabela de frequências para...

Níveis	Frequência	Percentual
A	55	11 %
B	55	11 %
G	54	10.8 %
J	53	10.6 %
C	52	10.4 %
F	50	10 %
I	49	9.8 %
D	46	9.2 %
H	46	9.2 %
E	40	8 %
Total	500	100 %

Tabelas de frequência para uma variável qualitativa ordinal

Tabela 4. Tabela de frequências para...

Níveis	Frequência	Freq. Relativa	Freq. Acumulada	Freq. Rel. Acumulada
muito baixo	109	0.218	109	0.218
baixo	93	0.186	202	0.404
neutro	91	0.182	293	0.586
alto	106	0.212	399	0.798
muito alto	101	0.202	500	1.000
Total	500	1.000	500	1.000

Tabelas de frequência para uma variável qualitativa ordinal

Tabela 5. Tabela de frequências para...

Níveis	Frequência	Percentual	Freq. Acumulada	Percentual Acumulado
muito baixo	109	21.8 %	109	21.8 %
baixo	93	18.6 %	202	40.4 %
neutro	91	18.2 %	293	58.6 %
alto	106	21.2 %	399	79.8 %
muito alto	101	20.2 %	500	100 %
Total	500	100 %	500	100 %

Gráficos para representação de frequências de uma variável qualitativa

- A representação por meio de tabelas é útil mas nem sempre eficiente.
- Em diversos casos pode ser mais conveniente utilizar um **gráfico**.
- “Uma imagem vale mais que mil palavras”.
- Os cuidados com a ordenação dos níveis de acordo com o tipo da variável se mantém.

Algumas possibilidades são:

- Gráfico de barras verticais.
- Gráfico de barras horizontais.
- Gráfico de barras empilhadas.
- Gráfico de setores.

Gráfico de barras

- **Gráfico de barras verticais ou horizontais.**
  - Utiliza os possíveis **níveis** das variáveis **em um eixo**.
  - As **frequências ou porcentagens** ficam **no outro eixo**.
  - O tamanho da barra correspondente à frequência ou percentual.
- **Gráfico de barras empilhadas**.
  - Usa-se **uma única barra**.
  - A barra é dividida de acordo com a **contribuição relativa** de cada nível da variável.
  - Representa-se a frequência relativa ou o percentual.

Gráfico de barras verticais

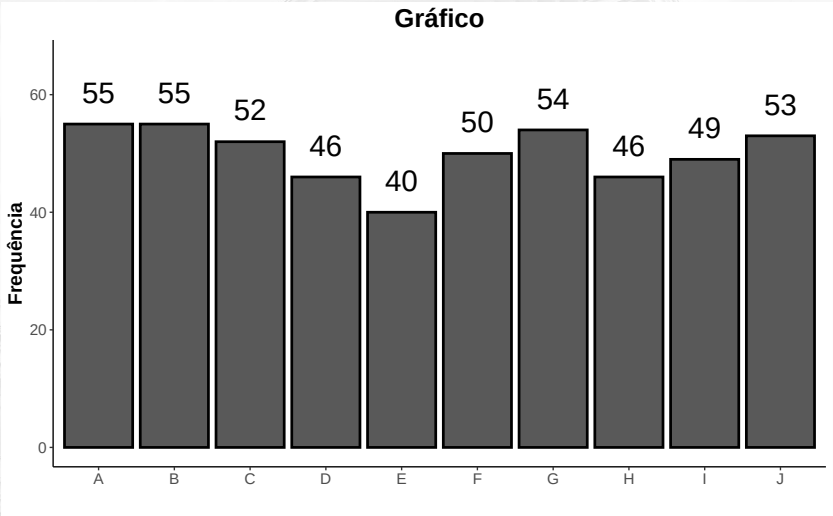

Figura 5. Gráfico de barras verticais para a variável...

Gráfico de barras verticais

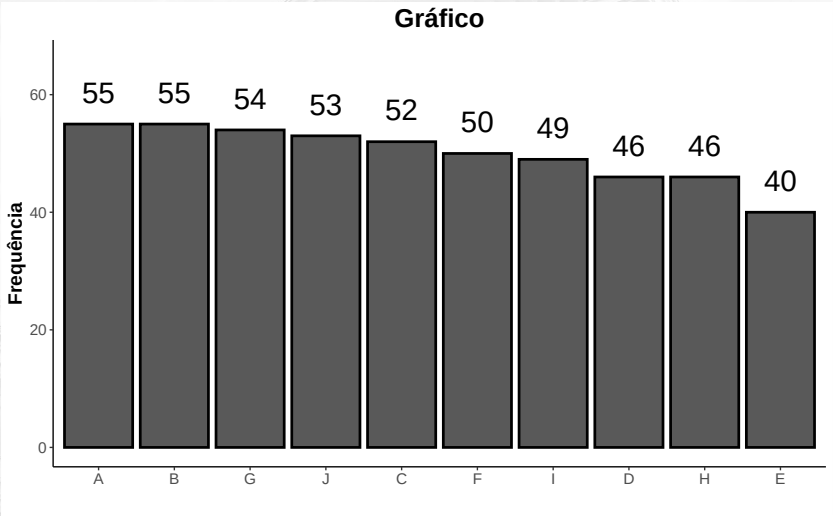

Figura 6. Gráfico de barras verticais para a variável...

Gráfico de barras horizontais

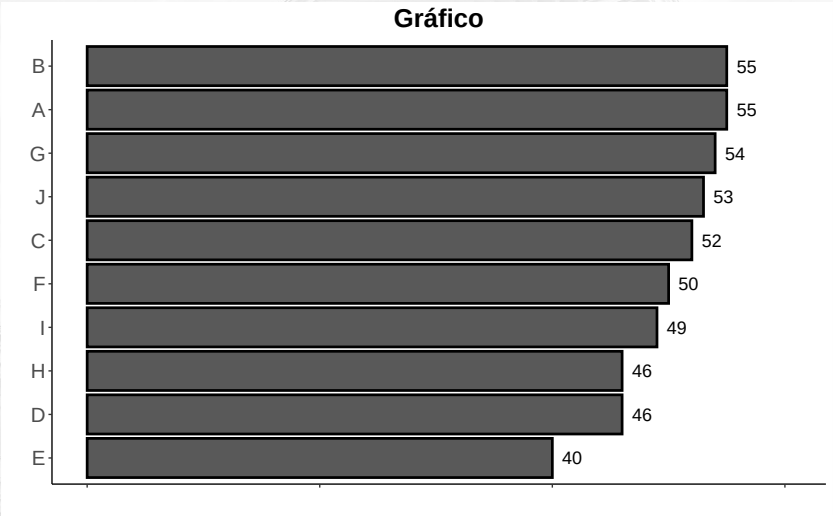

Figura 7. Gráfico de barras horizontais para a variável...

Gráfico de barras empilhadas

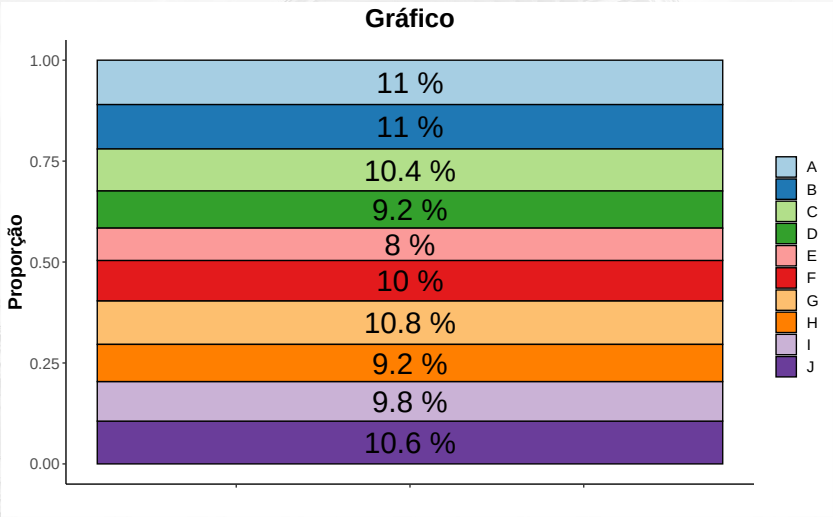

Figura 8. Gráfico de barras empilhadas para a variável...

Gráfico de setores

- Consiste em **repartir um círculo** em setores de tamanhos proporcionais às **frequências relativas** ou às **porcentagens** de cada valor.
- Pode ser usados para representar variáveis com **poucos níveis**.
- Apesar de muito usado e preferido em diversas áreas, deve ser evitado.
- O cérebro humano tem dificuldade em relacionar frequências com áreas relativas.
- Para variáveis com muitos níveis, o gráfico tende a ficar visualmente poluído e pouco informativo.
- Outro problema é que níveis com frequências iguais a 0 deixam de aparecer no gráfico, diferente de um gráfico de barras.

Gráfico de setores

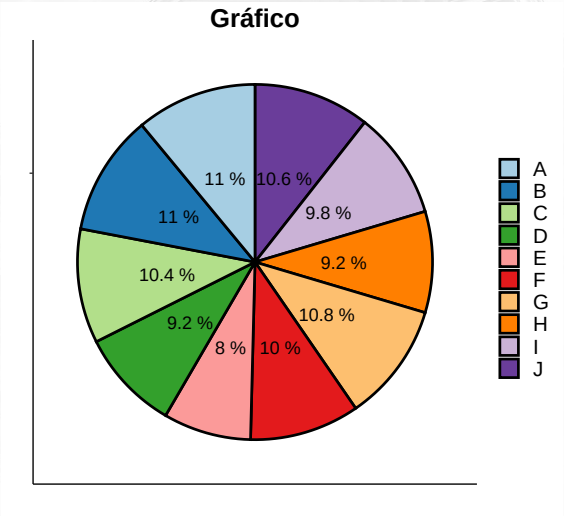

Figura 9. Gráfico de setores para a variável...

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 6. Tabela de frequências para...

Valores	Frequência	Percentual	Freq. Acumulada	Percentual Acumulado
0	47	9.4 %	47	9.4 %
1	71	14.2 %	118	23.6 %
2	65	13 %	183	36.6 %
3	51	10.2 %	234	46.8 %
4	54	10.8 %	288	57.6 %
5	55	11 %	343	68.6 %
6	52	10.4 %	395	79 %
7	52	10.4 %	447	89.4 %
8	53	10.6 %	500	100 %
Total	500	100 %	500	100 %

Gráfico de barras verticais

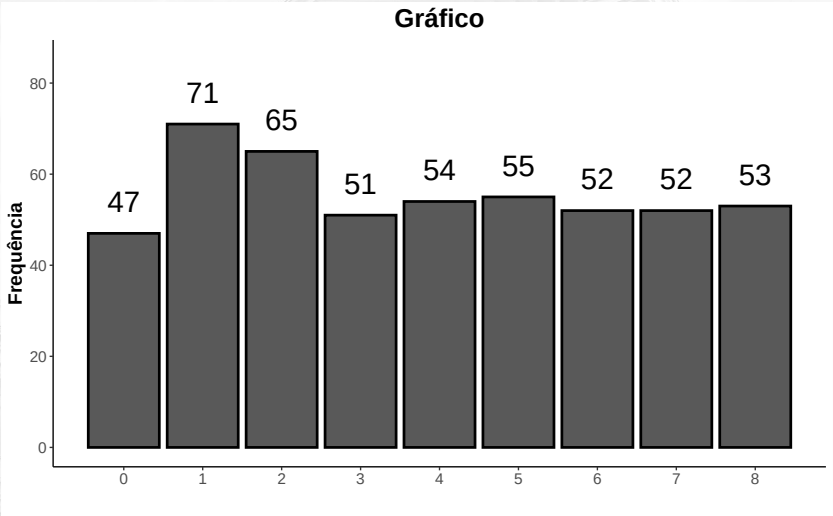

Figura 10. Gráfico de barras verticais para a variável...

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 7. Tabela de frequências para...

Faixas	Frequência	Freq. Relativa	Freq. Acumulada	Freq. Rel. Acumulada
[2,4]	1	0.002	1	0.002
(4,6]	46	0.092	47	0.094
(6,8]	102	0.204	149	0.298
(8,10]	111	0.222	260	0.520
(10,12]	111	0.222	371	0.742
(12,14]	82	0.164	453	0.906
(14,16]	27	0.054	480	0.960
(16,18]	15	0.030	495	0.990
(18,20]	3	0.006	498	0.996
(20,22]	2	0.004	500	1.000

Tabelas de frequência para uma variável quantitativa

Tabela 8. Tabela de frequências para...

Faixas	Frequência	Percentual	Freq. Acumulada	Percentual Acumulado
[2,4]	1	0.2 %	1	0.2 %
(4,6]	46	9.2 %	47	9.4 %
(6,8]	102	20.4 %	149	29.8 %
(8,10]	111	22.2 %	260	52 %
(10,12]	111	22.2 %	371	74.2 %
(12,14]	82	16.4 %	453	90.6 %
(14,16]	27	5.4 %	480	96 %
(16,18]	15	3 %	495	99 %
(18,20]	3	0.6 %	498	99.6 %
(20,22]	2	0.4 %	500	100 %

Tabelas de frequência para uma variável quantitativa

Tabela 9. Tabela de frequências para...

Faixas	Frequência	Percentual	Freq. Acum.	Perc. Acum.	Amplitude	Densidade
[2,4]	1	0.2 %	1	0.2 %	2	0.001
(4,6]	46	9.2 %	47	9.4 %	2	0.046
(6,8]	102	20.4 %	149	29.8 %	2	0.102
(8,10]	111	22.2 %	260	52 %	2	0.111
(10,12]	111	22.2 %	371	74.2 %	2	0.111
(12,14]	82	16.4 %	453	90.6 %	2	0.082
(14,16]	27	5.4 %	480	96 %	2	0.027
(16,18]	15	3 %	495	99 %	2	0.015
(18,20]	3	0.6 %	498	99.6 %	2	0.003
(20,22]	2	0.4 %	500	100 %	2	0.002

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

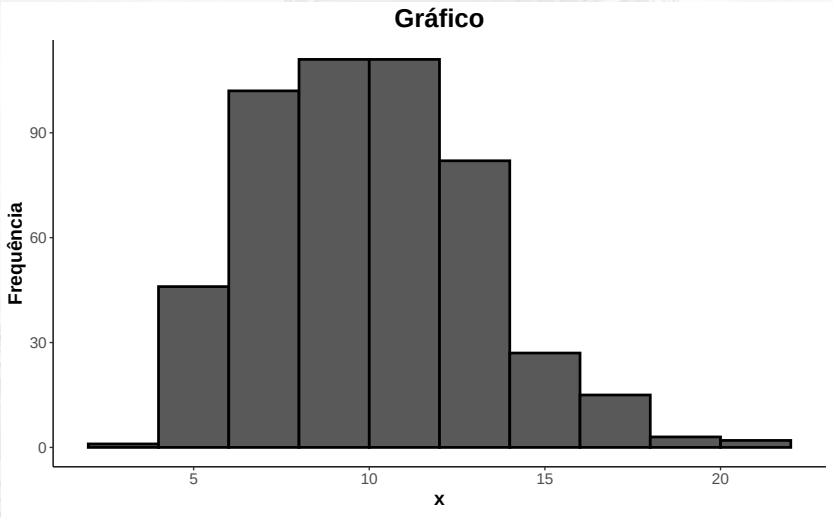

Figura 11. Gráfico de setores para a variável...

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

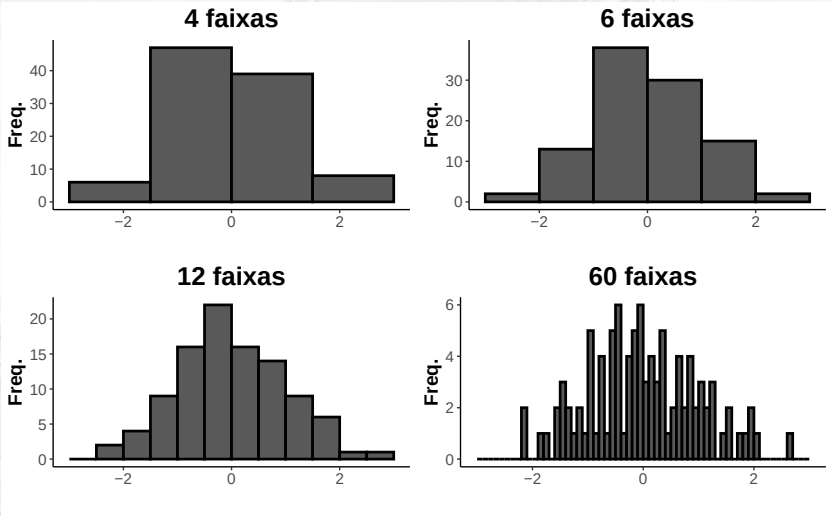

Figura 12. Gráfico de setores para a variável...

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.

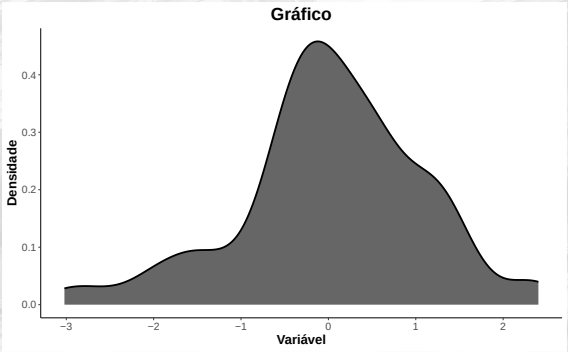

Figura 13. Gráfico de setores para a variável...

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.

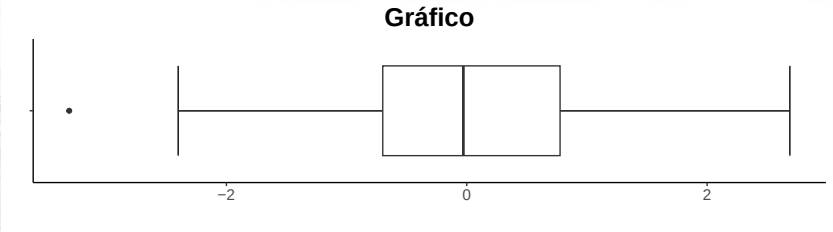

Figura 14. Gráfico de setores para a variável...

Histograma, densidade e box-plot

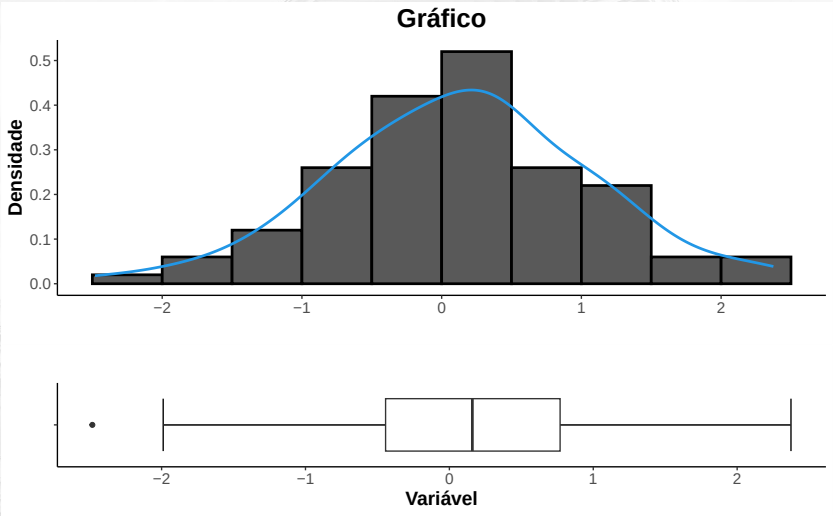

Figura 15. Gráfico de setores para a variável...

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.

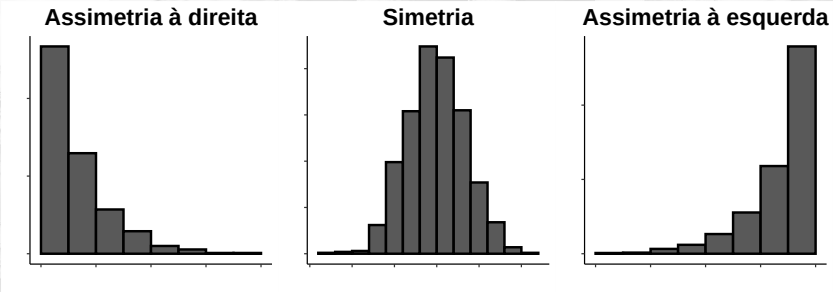

Figura 16. Gráfico de setores para a variável...

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.