Análise exploratória

Gráficos e tabelas para variáveis quantitativas

Prof. Me. Lineu Alberto Cavazani de Freitas

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

- A análise exploratória de dados é uma área relativamente nova.
- Nasceu do clássico livro **Exploratory Data Analysis** de **John Tukey** em 1977.
- Algo curioso é que Tukey tinha uma relação próxima com a Ciência da Computação e definiu os termos **bit** e **software**.

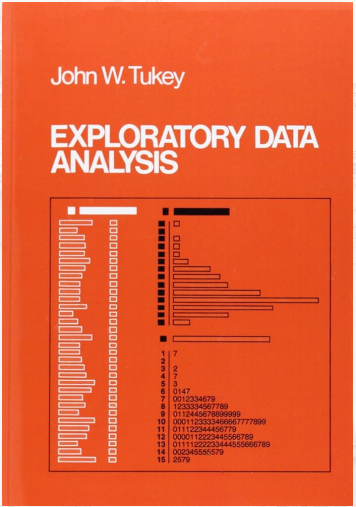

Figura 1. Capa do livro Exploratory Data Analysis de John Tukey.

Análise exploratória

- Como quase tudo em análise de dados, o **avanço computacional** permitiu com que a análise exploratória evoluísse substancialmente.
- Por exemplo: historicamente o processo de criação de um gráfico era reservado a pessoas qualificadas pois a produção de uma visualização era difícil.
- Hoje qualquer pessoa pode inserir dados em um aplicativo e gerar um gráfico.
- Este tipo de facilidade é importante para disseminação e democratização dos métodos, porém abre margem para certas práticas inadequadas.

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 entender 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 2. 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**).
- Também podemos conduzir análises focadas em avaliar a relação entre as variáveis (**análises multivariadas**).

Figura 3. 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.
- Medidas de posição central.
- Medidas de posição relativa.
- Medidas de forma.
- Medidas de dispersão.
- Medidas de associação.

Análise exploratória

- Para ilustrar as técnicas de análise exploratória de dados, usaremos o conjunto de dados “milsa”.
- Este conjunto de dados aparece no livro “Estatística Básica” de W. O. Bussab e P. A. Morettin.
- Conjunto de dados hipotético de atributos de 36 funcionários da companhia “Milsa”.

O conjunto possui as seguintes variáveis:

- **Funcionário**: identificadora de funcionário.
- **Estado civil**: casado ou solteiro.
- **Instrução**: 1º grau, 2º grau, superior.
- **Filhos**: número de filhos.
- **Salário**: salário do funcionário.
- **Anos**: idade em anos completos.
- **Meses**: meses além dos anos completos.
- **Região**: capital, interior, outro.

Análise exploratória

Tabela 1. Primeiras linhas do conjunto de dados Milsa.

Funcionário	Estado civil	Instrução	Filhos	Salário	Anos	Meses	Região
1	solteiro	1o Grau	NA	4.00	26	3	interior
2	casado	1o Grau	1	4.56	32	10	capital
3	casado	1o Grau	2	5.25	36	5	capital
4	solteiro	2o Grau	NA	5.73	20	10	outro
5	solteiro	1o Grau	NA	6.26	40	7	outro
6	casado	1o Grau	0	6.66	28	0	interior
7	solteiro	1o Grau	NA	6.86	41	0	interior
8	solteiro	1o Grau	NA	7.39	43	4	capital
9	casado	2o Grau	1	7.59	34	10	capital
10	solteiro	2o Grau	NA	7.44	23	6	outro

Análise descritiva univariada para variáveis quantitativas

Análise descritiva univariada para variáveis quantitativas

- Uma variável quantitativa é uma **característica** que pode ser **mensurada** e 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 2. Tabela de frequências para o número de filhos (desconsiderando dados ausentes).

Filhos	Frequência	Percentual	Freq. Acumulada	Percentual Acumulado
0	4	20 %	4	20 %
1	5	25 %	9	45 %
2	7	35 %	16	80 %
3	3	15 %	19	95 %
4	0	0 %	19	95 %
5	1	5 %	20	100 %
Total	20	100 %	20	100 %

Gráfico de barras verticais

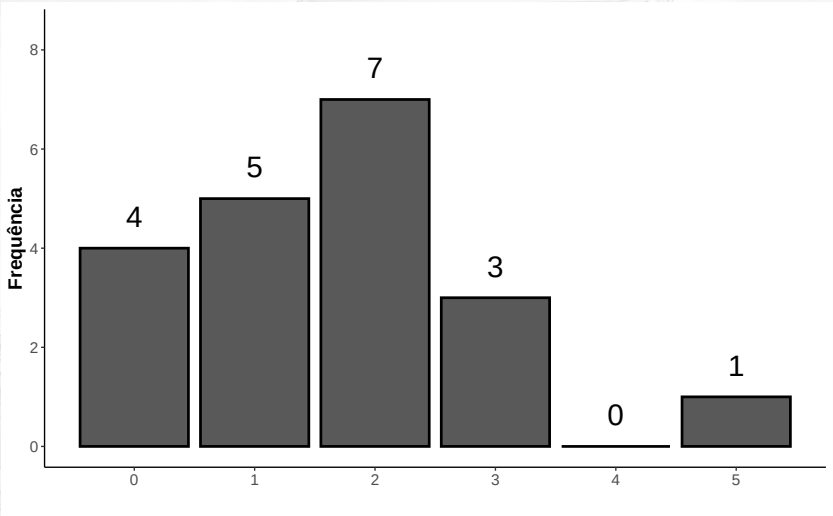

Figura 4. Gráfico de barras verticais para o número de filhos.

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).
- Diferentes pessoas e softwares podem usar intervalos distintos.
- Em geral usaremos intervalos **fechados à esquerda** e **abertos 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:
  - $a \leq y < b$.
  - $a \vdash b$.
  - $[a,b)$.
  - $[a,b[$.
- Exemplo:
  - $5 \leq y < 10$.
  - $5 \vdash 10$.
  - $[5,10)$.
  - $[5,10[$.
  - Valores maiores ou iguais a 5 até valores menores que 10 (10 não está no intervalo).

Análise descritiva univariada para variáveis quantitativas

Perguntas que surgem são:

- Como agrupar em classes?
- Qual o tamanho ideal das faixas de valores?

- Classes definidas com a **mesma amplitude** é o procedimento mais usual, apesar de ser possível definir classes com tamanhos diferentes.
- 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 3. Tabela de frequências usando faixas de salários.

Faixas	Frequência	Freq. Relativa	Freq. Acumulada	Freq. Rel. Acumulada
[4,6)	4	0.11	4	0.11
[6,8)	6	0.17	10	0.28
[8,10)	8	0.22	18	0.5
[10,12)	4	0.11	22	0.61
[12,14)	5	0.14	27	0.75
[14,16)	3	0.08	30	0.83
[16,18)	3	0.08	33	0.91
[18,20)	2	0.06	35	0.97
[20,22)	0	0.00	35	0.97
[22,24]	1	0.03	36	1
Total	36	1.00		

Tabelas de frequência para uma variável quantitativa

Tabela 4. Tabela de frequências usando faixas de salários.

Faixas	Frequência	Percentual	Freq. Acumulada	Percentual Acumulado
[4,6)	4	11 %	4	11 %
[6,8)	6	17 %	10	28 %
[8,10)	8	22 %	18	50 %
[10,12)	4	11 %	22	61 %
[12,14)	5	14 %	27	75 %
[14,16)	3	8 %	30	83 %
[16,18)	3	8 %	33	91 %
[18,20)	2	6 %	35	97 %
[20,22)	0	0 %	35	97 %
[22,24]	1	3 %	36	100 %
Total	36	100%		

Tabelas de frequência para uma variável quantitativa

Tabela 5. Tabela de frequências usando faixas de salários.

Faixas	Frequência	Percentual	Freq. Acum.	Perc. Acum.	Amplitude	Densidade
[4,6)	4	11 %	4	11 %	2	0.055
[6,8)	6	17 %	10	28 %	2	0.085
[8,10)	8	22 %	18	50 %	2	0.11
[10,12)	4	11 %	22	61 %	2	0.055
[12,14)	5	14 %	27	75 %	2	0.07
[14,16)	3	8 %	30	83 %	2	0.04
[16,18)	3	8 %	33	91 %	2	0.04
[18,20)	2	6 %	35	97 %	2	0.03
[20,22)	0	0 %	35	97 %	2	0
[22,24]	1	3 %	36	100 %	2	0.015
Total	36	100%				

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 respectiva faixa.
  - A **altura** representar a **frequência** absoluta na faixa.
  - A **altura** representar o quociente da área pela amplitude da faixa: a **densidade**.

Histograma

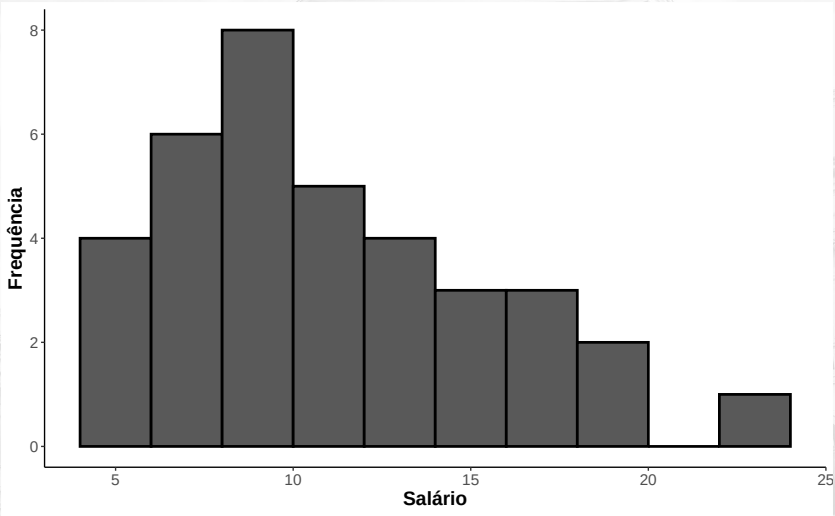

Figura 5. Histograma dos salários.

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

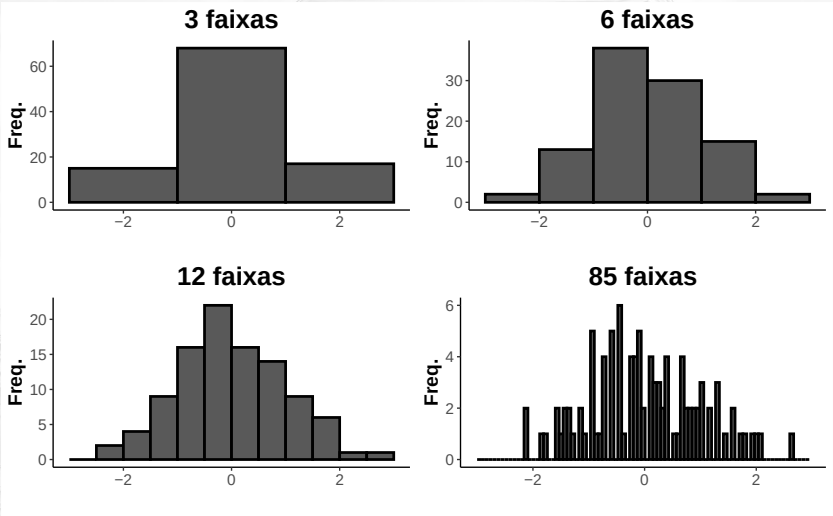

Figura 6. Efeito do número de classes em histogramas.

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.
- Outra forma de ver o gráfico de densidade empírica é como um **histograma suavizado**.

Gráfico de densidade empírica

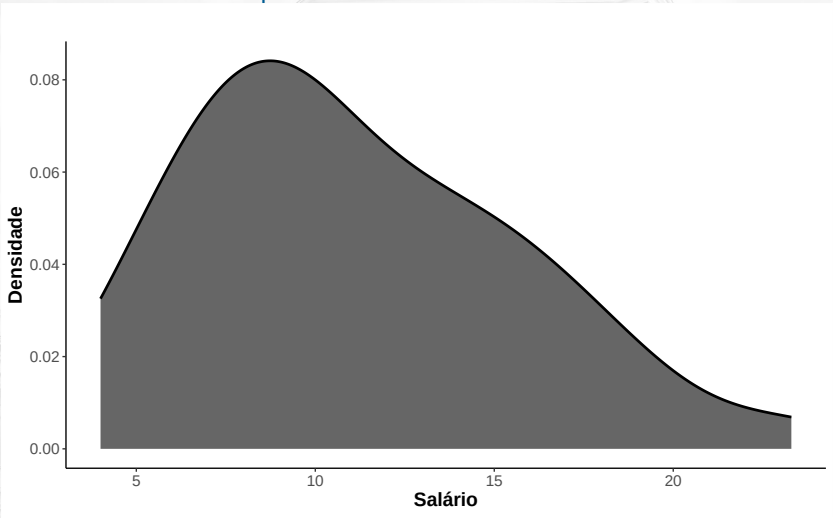

Figura 7. Gráfico de densidade dos salários.

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.

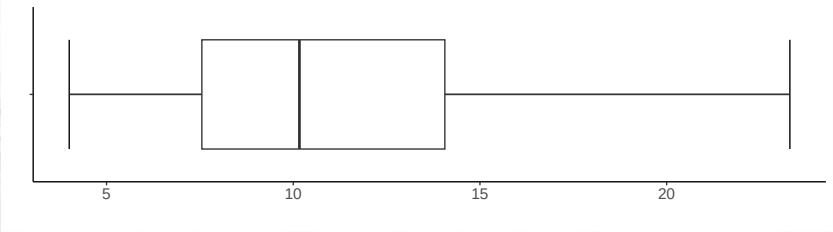

Figura 8. Box-plot dos salários.

Histograma, densidade e box-plot

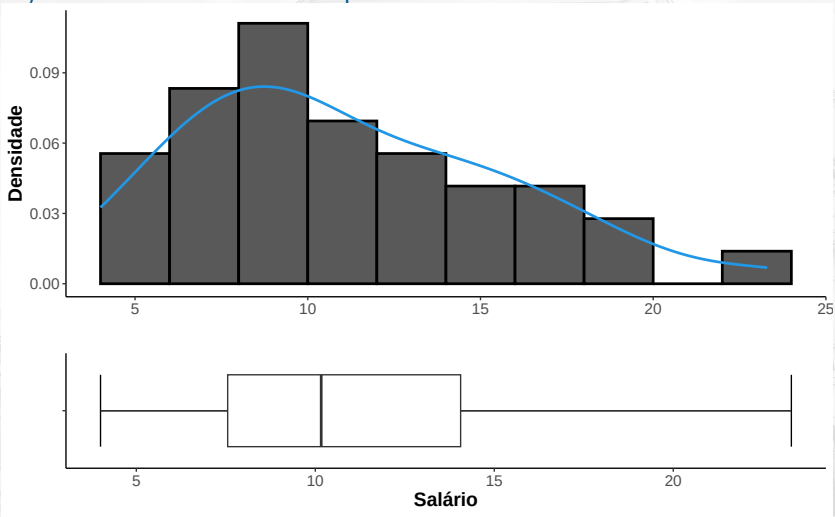

Figura 9. Combinação de representações.

Assimetria

- Um conjunto de valores 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**.

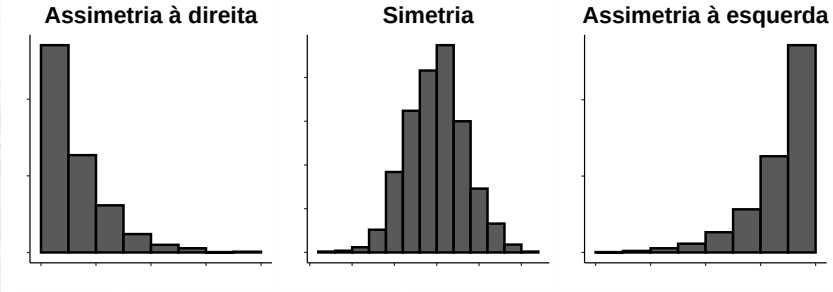

Figura 10. Ilustração assimetria.

Gráfico de frequências acumuladas

- Outra possibilidade para visualização de variáveis quantitativas é o **gráfico de frequências acumuladas**.
- A frequência acumulada indica quantos elementos estão abaixo de um certo valor.
- É um interessante recurso para obtenção de **separatrizes**.

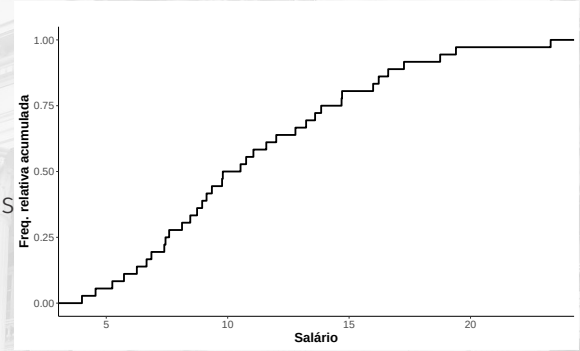

Figura 11. Gráfico de frequências acumuladas para o salário.

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.