Análise exploratória

Suellen Teixeira Zavadzki de Pauli

DAEST/UTFPR

O que é estatística?

Estatística é um conjunto de técnicas para, sistematicamente:

- Planejar a coleta de dados oriundos de estudos ou experimentos, realizados em qualquer área do conhecimento;
- Descrever, analisar e interpretar dados;
- Extrair informações para subsidiar decisões;
- Avaliar evidências empíricas sob hipóteses de interesse.

Exemplos de aplicações:

- Opinião da população brasileira sobre o novo governo.
- Avaliar a efetividade de uma nova droga para a cura do câncer.
- Entender os hábitos de compra dos clientes de uma loja virtual.
- Recomendação personalizada de produtos.
- Comparar a produtividade da soja sob diferentes formas de

Conceitos fundamentais

- **População:**Conjunto de todos os elementos sob investigação;
- **Amostra:**Subconjunto da população;
- **Variável:**Característica a ser observada com cada indivíduo da amostra.

Exemplos...

Divisões básicas da estatística

- **Estatística descritiva e exploratória:**
  - Consistência dos dados e interpretações iniciais.
  - Visualização dos dados e relações entre variáveis.
- **Probabilidade:**
  - Fornece ferramentas para lidar/quantificar incerteza.
- **Inferência estatística:**
  - Estimação de quantidades desconhecidas.
  - Formular e testar hipóteses.
  - Extrapolar para a população resultados obtidos na amostra.

Etapas da análise estatística

- **Definir a população de interesse;**
  - População factível
- **Estabelecer os objetivos da pesquisa;**
  - Definir critérios objetivos sobre quais dados coletar;
  - Postular a análise estatística a ser utilizada.
- **Definir o método para coletar as amostras**
  - Fonte de dados secundários (IBGE, IPEA, etc);
  - Banco de dados da empresa;
  - Pesquisas amostrais;
  - Experimentos em laboratórios, etc.
- **Análise dos dados**
  - Análise descritiva e exploratória (o que aconteceu na amostra?).
  - Análise inferencial (o que acontece na população?).

Exemplo

Uma pesquisa foi realizada com alunos. As variáveis são as seguintes:

- **Id**: identificação do aluno;
- **Turma**: A ou B;
- **Sexo**: feminino (F) ou masculino (M);
- **Idade**: em anos;
- **Alt**: altura em metros;
- **Peso**: em quilogramas; Filhos: nº de filhos na família;
- **Fuma**: hábito de fumar: sim (S) ou não (N);
- **Toler**: tolerância ao cigarro: (I) indiferente; (P) incomoda pouco; (M) incomoda muito;
- **Exerc.**: horas de atividade física, por semana;
- **Cine**: nº. de vezes que vai ao cinema por semana;
- **OpCine**: opinião a respeito das salas de cinema na cidade: (B) regular a boa; (M) muito boa;

Organização dos dados

- A partir de um conjunto de dados coletado, a questão é: Como extrair informações a respeito de uma ou mais características de interesse?
- Basicamente há duas opções:
  - Tabelas de frequências;
  - Gráficos.
- É importante levar em consideração a natureza dos dados.

Organização dos dados

- Uma típica tabela de dados brutos contém:
  - Variáveis (características, medições, etc) nas colunas.
  - Sujeito (indivíduo, objetos, etc) nas linhas.

	Id	Turma	Sexo	Idade	Alt	Peso	Filhos	Fuma	Toler	Exerc	Cine	OpCine	TV	OpTV
1	1	A	F	17	1.60	60.5	2	NAO	P	0	1	B	16	R
2	2	A	F	18	1.69	55.0	1	NAO	M	0	1	B	7	R
3	3	A	M	18	1.85	72.8	2	NAO	P	5	2	M	15	R
4	4	A	M	25	1.85	80.9	2	NAO	P	5	2	B	20	R
5	5	A	F	19	1.58	55.0	1	NAO	M	2	2	B	5	R
6	6	A	M	19	1.76	60.0	3	NAO	M	2	1	B	2	R

- Tipos de variáveis:
  - Qualitativa nominal: Turma, Sexo, Fuma;
  - Qualitativa ordinal: Toler, OpCine, OpTV;
  - Quantitativa discreta: Idade, Filhos, Exerc, Cine, TV.
  - Quantitativa contínua: Alt, Peso.

Tabela de frequências

- A tabela de dados brutos pode ser muito longa, portanto será difícil extrair alguma informação.
- As tabelas de frequência ajudam a resumir a informação da variável de interesse.
- Vamos usar 3 tipos de frequência:
  - Frequência absoluta: contagem de cada valor observado. Representado por n_i o número de indivíduos com a característica i.
  - Frequência relativa: número de indivíduos com a característica i dividido pelo total de indivíduos n, ou seja $f_i = \frac{n_i}{n}$.
  - Frequência acumulada: frequência (absoluta ou relativa) acumulada até um certo valor, obtida pela soma das frequências de todos os valores da variável, menores ou iguais ao valor considerado.

Tabela de frequência - qualitativa nominal

- Considerando a variável `Sexo`

	n_i	f_i
F	37	0.74
M	13	0.26
Total	50	1.00

- Neste caso não faz sentido usar frequência acumulada.

Tabela de frequência - qualitativa ordinal

- Considerando a variável `OpTV`

	n_i	f_i	f_{ac}
R	39	0.78	0.78
M	1	0.02	0.8
B	3	0.06	0.86
M	7	0.14	1.00
Total	50	1.00	

Tabela de frequência - quantitativa discreta

- Considerando a variável `Idade`

	n_i	f_i	f_{ac}
17	9	0.18	0.18
18	22	0.44	0.62
19	7	0.14	0.76
20	4	0.08	0.84
21	3	0.06	0.90
22	0	0.00	0.90
23	2	0.04	0.94
24	1	0.02	0.96
25	2	0.04	1.00
Total	50	1.00	

Tabela de frequência - quantitativa contínua

- No caso de quantitativas contínuas não faz sentido contar cada valor pois podem existir muitos (potencialmente infinito).
- A solução é criar classes ou faixas de valores, e contar o número de ocorrências dentro destas classes
- Para definir as classes:
  - Defina a amplitude da classe, de maneira que se obtenham de 5 a 8 classes (de mesma amplitude).
  - Identifique os valores máximo e mínimo da variável e construa as classes de maneira que inclua todos os valores
- As classes de valores podem seguir um dos formatos:

Classe	Notação	Denominação	Resultado
[a,b)	$a \vdash b$	Fechado em a, aberto em b	Inclui a, não inclui b
(a,b]	$a \dashv b$	Aberto em a, fechado em b	Não inclui a, inclui b

Tabela de frequência - quantitativa contínua

- Considerando a variável `Peso`.
- Foram construídas 6 classes de amplitude 10.
- As classes são do tipo [a, b) ou a $\vdash$ b.

	n_i	f_i	f_{ac}
$[40, 50)$	8	0.16	0.16
$[50, 60)$	22	0.44	0.60
$[60, 70)$	8	0.16	0.76
$[70, 80)$	6	0.12	0.88
$[80, 90)$	5	0.10	0.98
$[90, 100)$	1	0.02	1.00
Total	50	1.00	

Tabela de frequência - quantitativa discreta (muitos valores)

- Considerando a variável `TV`.
- Apesar de ser discreta, o número de valores únicos é muito grande e não seria útil contar as frequências de cada valor.
- Neste caso, utiliza-se o mesmo procedimento usado para quantitativas contínuas

	n_i	f_i	f_{ac}
$[0, 6)$	14	0.28	0.28
$[6, 12)$	17	0.34	0.62
$[12, 18)$	11	0.22	0.84
$[18, 24)$	4	0.08	0.92
$[24, 30)$	3	0.06	0.98
$[30, 36)$	1	0.02	1.00
Total	50	1.00	

Representação gráfica

- Podemos visualizar as tabelas através de gráficos.
- Existe um tipo de gráfico adequado para cada tipo de variável.
- Cuidado deve ser tomado com representações visuais pois um gráfico desproporcional pode gerar interpretações distorcidas.
- As principais representações gráficas são:
  - Diagrama circular (setores ou "pizza");
  - Gráfico de barras;
  - Histogramas;
  - Boxplots.

Diagrama circular

- Adequado para variáveis qualitativas nominal e ordinal.
- O uso deste tipo de gráfico deve ser evitado, pois pode ser de difícil interpretação.

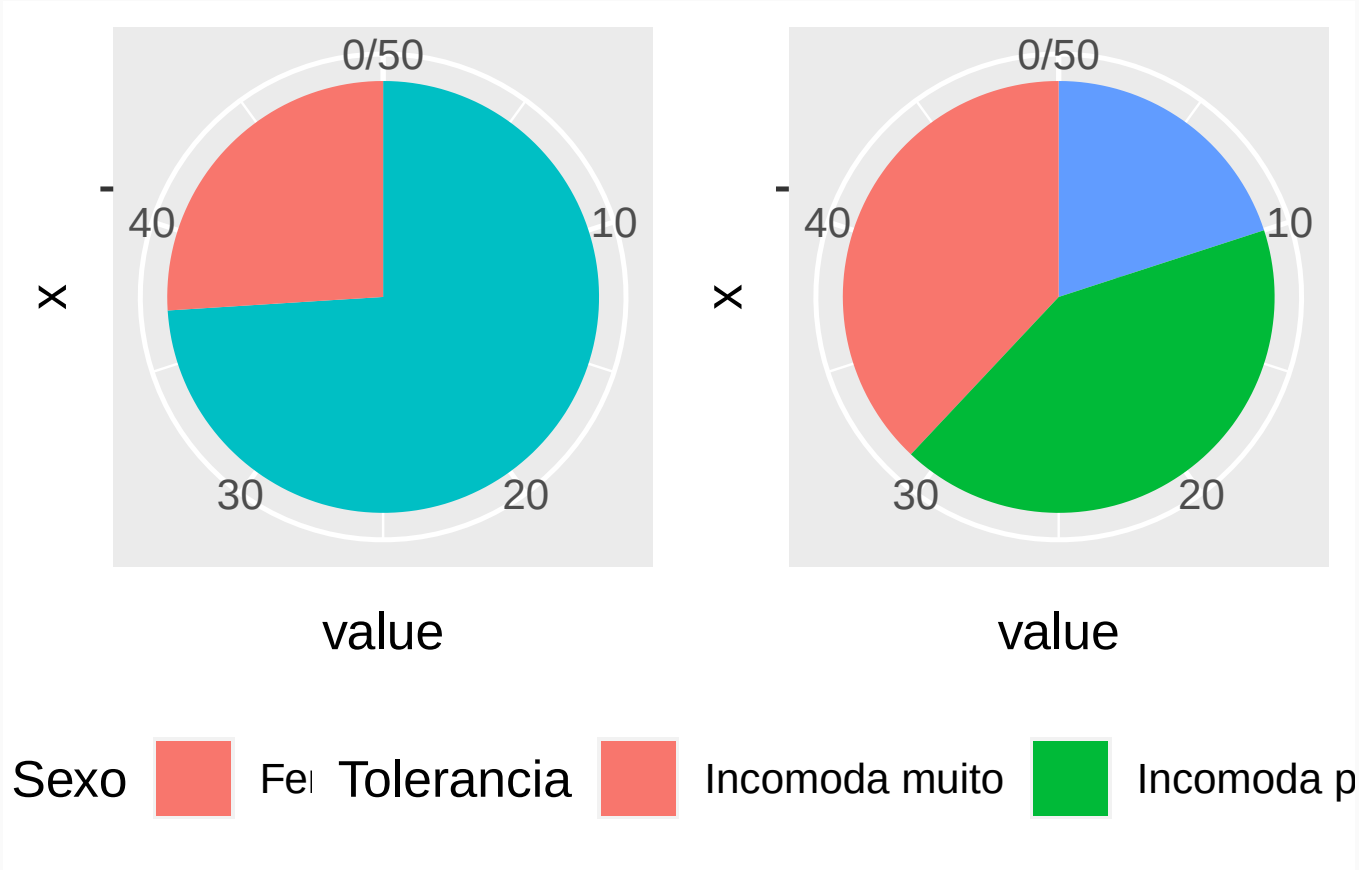
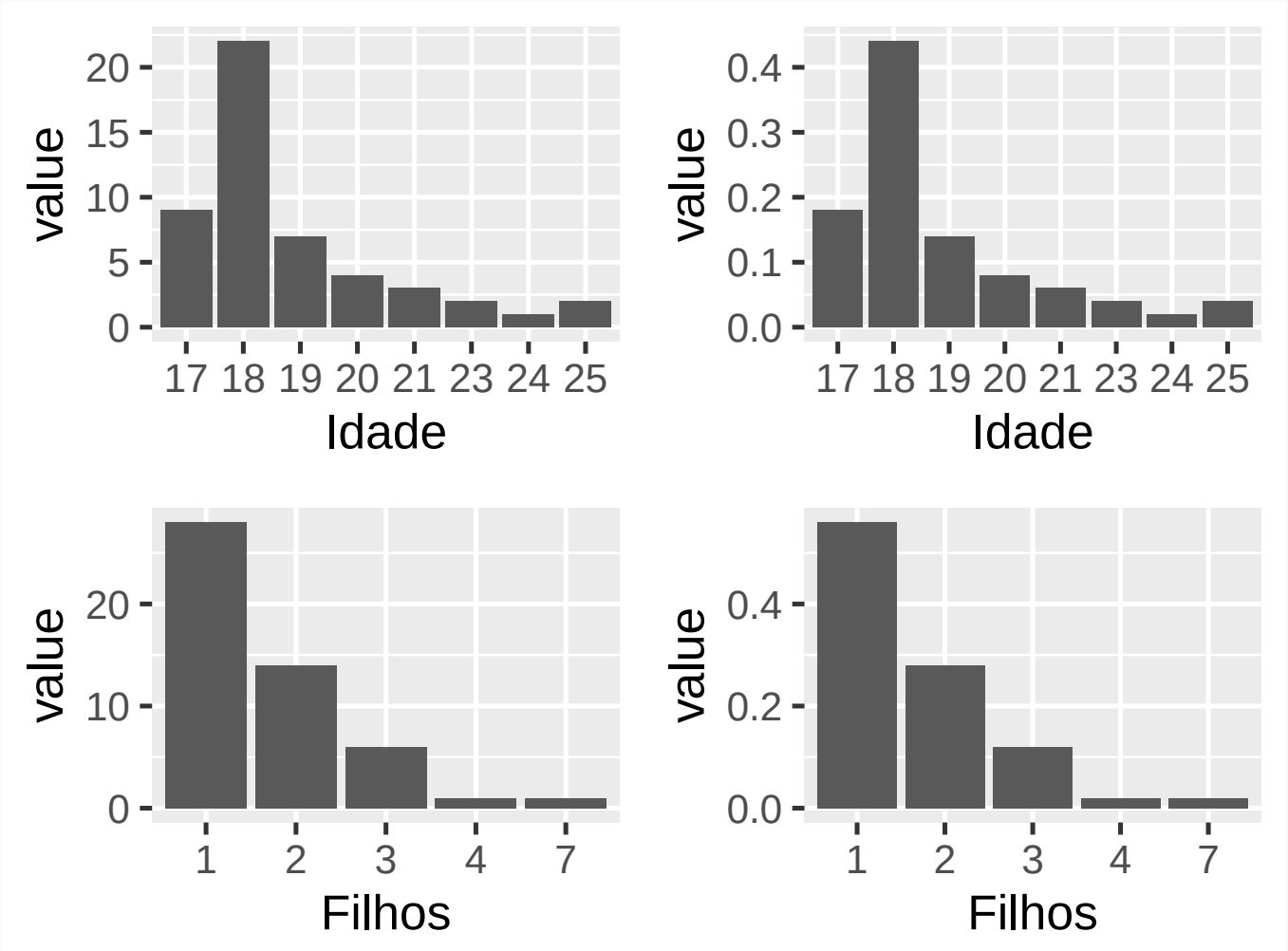


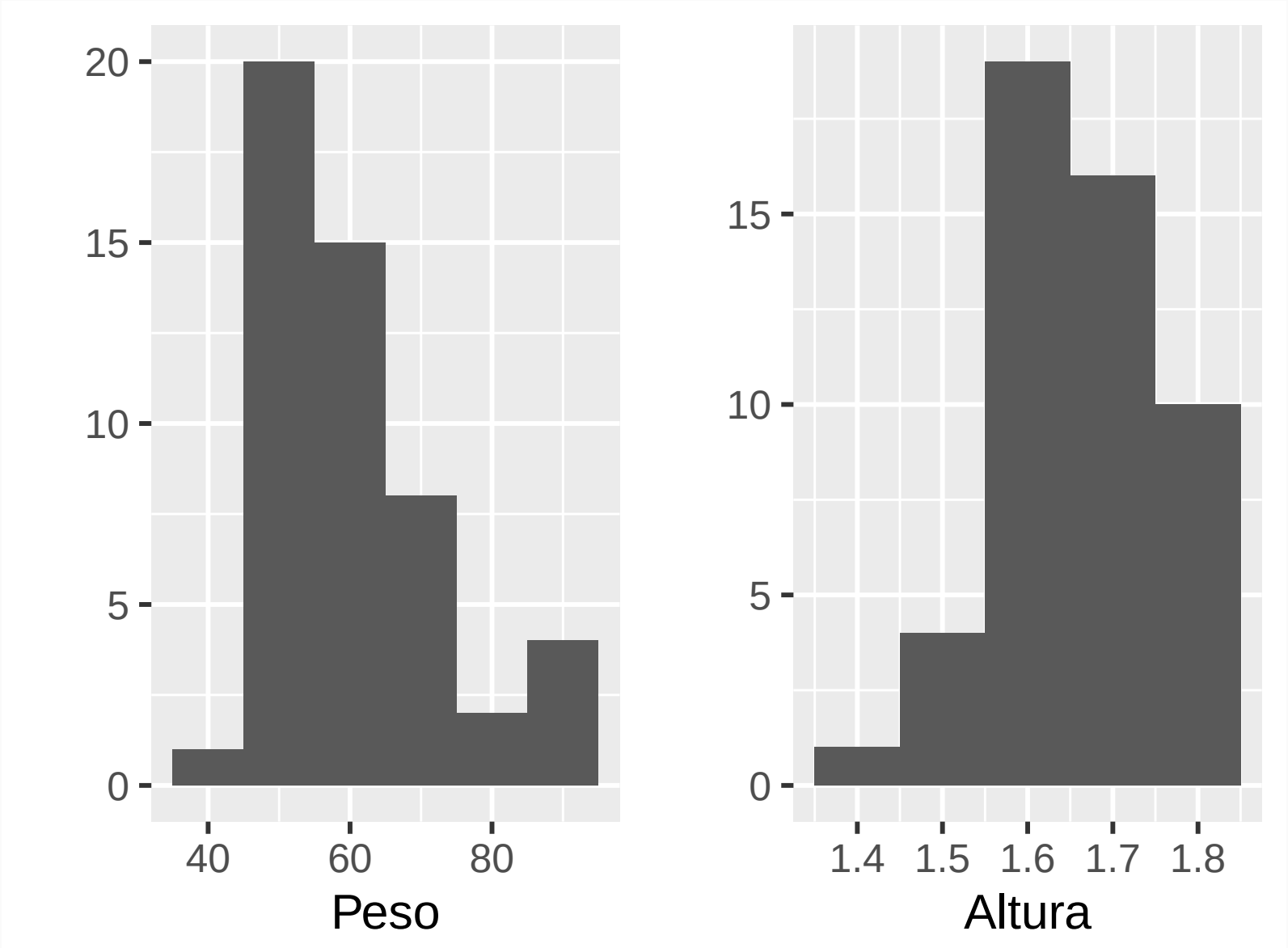
Gráfico de barras

- Adequado para variáveis qualitativas nominal/ordinal e quantitativa discreta (poucos valores distintos).
- Podem ser usadas as frequências absolutas ou relativas.



Histograma

- Adequado para quantitativa contínua.



Mediana e quartis

- Mediana: valor da variável que divide o conjunto de dados ordenado em dois subgrupos de mesmo tamanho.
- Quartis: valores da variável que divide o conjunto de dados ordenados em quatro subgrupos de mesmo tamanho.
- Posição dos quartis:
  - $Q_1 = 0.25 \cdot (\mathrm{N} + 1)$ e arredonde
  - Q_2 = média dos valores nas posições $(\mathrm{N}/2)$ e $(\mathrm{N}/2) + 1$ se N par e $Q_2 = (\mathrm{N} + 1)/2$ se N ímpar.
  - $Q_3 = 0.75 \cdot (\mathrm{N} + 1)$ e arredonde.

Boxplot

- Adequado para quantitativa contínua.
- Pode ser usado também para quantitative discreta com muitos valores.

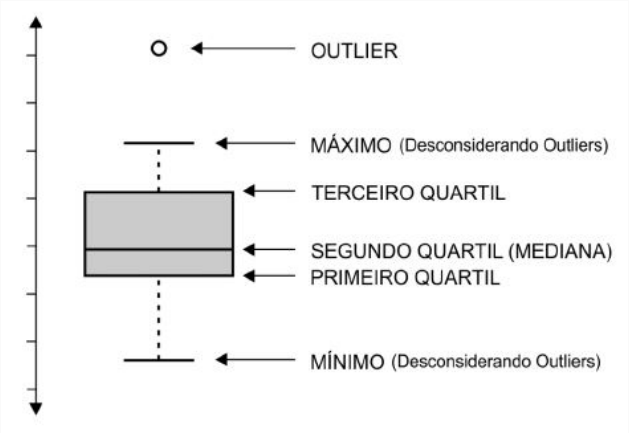


Figure 1: RMSE of all model configurations for PETR4, VALE and ITUB4.