

Introdução à análise exploratória de dados

Wagner H. Bonat
Elias T. Krainski
Fernando P. Mayer

Universidade Federal do Paraná
Departamento de Estatística
Laboratório de Estatística e Geoinformação

22/02/2018



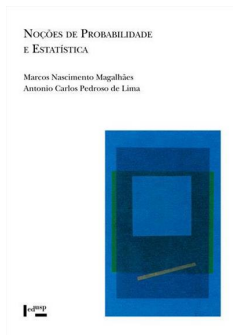
Sumário

- 1 Informações gerais
 - O que é estatística?
- 2 Análise exploratória de dados
 - Organização de Dados
 - Tabelas de frequência
 - Representação gráfica
- 3 Exercícios recomendados

Referência bibliográfica

Livro-texto:

- Marcos Nascimento Magalhães e Antonio Carlos Pedroso de Lima.
Noções de Probabilidade e Estatística. Editora: EDUSP.



Tópicos do curso

- 1 Análise exploratória de dados.
- 2 Probabilidades.
- 3 Variáveis aleatórias discretas.
- 4 Medidas resumo.
- 5 Variáveis bidimensionais.
- 6 Variáveis aleatórias contínuas.
- 7 Inferência estatística - Estimação.
- 8 Inferência estatística - Testes de hipóteses.
- 9 Tópicos especiais.

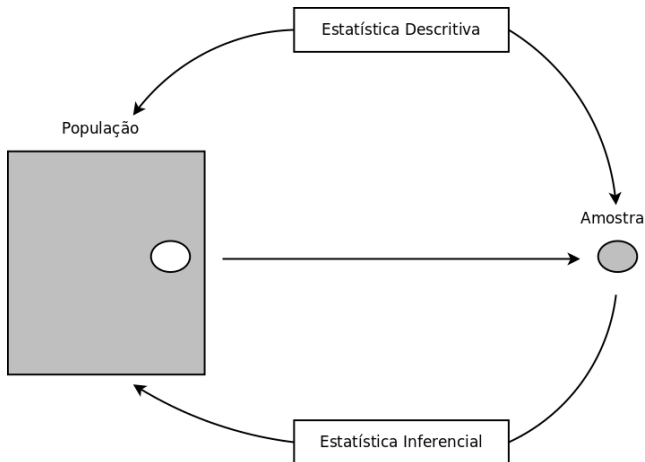
O que é estatística?

- 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 ou conclusões

Tópicos de estatística básica

- Conceitos essenciais em Estatística:
 - Estatística descritiva.
 - Probabilidade.
 - Inferência estatística.
- Conceitos fundamentais:
 - **População:** Conjunto de todos os elementos sob investigação.
 - **Amostra:** Subconjunto da população.
 - **Variável** de interesse: característica a ser observada em cada indivíduo da amostra

População e amostra



Etapas da análise estatística

- Definição do método de coleta de dados
 - estabelecer os objetivos (questões) de pesquisa
 - definir critérios objetivos de como e quais dados coletar
 - postular a análise estatística a ser utilizada

Etapas da análise estatística

- Definição do método de coleta de dados
 - estabelecer os objetivos (questões) de pesquisa
 - definir critérios objetivos de como e quais dados coletar
 - postular a análise estatística a ser utilizada
- Estatística Descritiva
 - depende do tipo de dado coletado
 - deve ser racionalizada
 - relacionada com os objetivos da pesquisa

Etapas da análise estatística

- Definição do método de coleta de dados
 - estabelecer os objetivos (questões) de pesquisa
 - definir critérios objetivos de como e quais dados coletar
 - postular a análise estatística a ser utilizada
- Estatística Descritiva
 - depende do tipo de dado coletado
 - deve ser racionalizada
 - relacionada com os objetivos da pesquisa
- Inferência estatística
 - depende do objetivo da pesquisa

Planejamento da coleta de dados

- definição do experimento
 - variáveis respostas
 - variáveis de controle
 - desenho do experimento e randomização

Planejamento da coleta de dados

- definição do experimento
 - variáveis respostas
 - variáveis de controle
 - desenho do experimento e randomização
- coleta de dados por amostragem
 - definição da população e característica de interesse
 - definição do plano amostral
 - Aleatória simples (com ou sem reposição) ou sistemática
 - Estratificada, por estratos da população (segundo uma característica)
 - Conglomerados, por grupos de indivíduos da população (subpopulações)
 - Amostragem complexa (combina anteriores)

Planejamento da coleta de dados

- definição do experimento
 - variáveis respostas
 - variáveis de controle
 - desenho do experimento e randomização
- coleta de dados por amostragem
 - definição da população e característica de interesse
 - definição do plano amostral
 - Aleatória simples (com ou sem reposição) ou sistemática
 - Estratificada, por estratos da população (segundo uma característica)
 - Conglomerados, por grupos de indivíduos da população (subpopulações)
 - Amostragem complexa (combina anteriores)
- coleta de dados observacionais. Exemplos:
 - população de plantas
 - presença de seres vivos num ambiente
 - fenômenos climáticos

Análise estatística

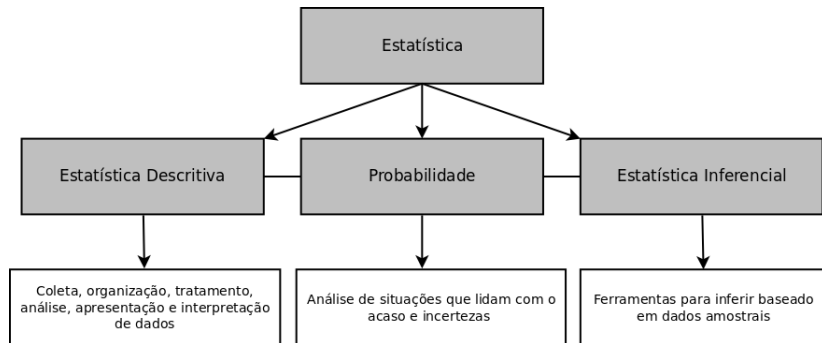
- Estatística Descritiva

- consistência e interpretações iniciais
- visualização dos dados e relações entre variáveis

- Inferência estatística

- estimação de quantidades desconhecidas
- formulação e teste de hipóteses
- extrapolar para a população, se os dados são de uma amostra.

Etapas da análise estatística



Sumário

- 1 Informações gerais
 - O que é estatística?
- 2 Análise exploratória de dados
 - Organização de Dados
 - Tabelas de frequência
 - Representação gráfica
- 3 Exercícios recomendados

Organização de 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 temos duas opções:
 - Tabelas de frequência
 - Gráficos
- O importante é levar em consideração a **natureza dos dados**.

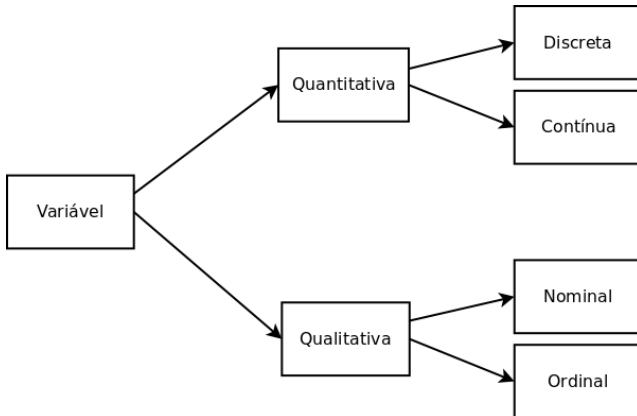
Organização de Dados

- 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	A	F	17	1.60	60.5	2	NAO	P	0	1	B 16	R	
2	A	F	18	1.69	55.0	1	NAO	M	0	1	B 7	R	
3	A	M	18	1.85	72.8	2	NAO	P	5	2	M 15	R	
4	A	M	25	1.85	80.9	2	NAO	P	5	2	B 20	R	
5	A	F	19	1.58	55.0	1	NAO	M	2	2	B 5	R	
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, Filh, Exer, Cine, TV.
 - Quantitativa contínua: Alt, Peso.

Tipos de variáveis



Tabelas de frequência

- 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 valores i , e n o número total
 - Frequência **relativa**: número de valores i dividido pelo total 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
Sum	50	1.00

- Não faz sentido usar frequência acumulada

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
Sum	50	1.00	

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	0.80
B	3	0.06	0.86
N	7	0.14	1.00
Sum	50	1.00	

Tabela de frequência - quantitativa contínua

- No caso de quantitativas contínuas não faz sentido contar cada valor pois podem existir muitos
- A solução é criar **classes** ou **faixas de valores**, e contar o número de ocorrências dentro destas classes.
- Para definir as classes:
 - 1 Defina a amplitude da classe, de maneira que se obtenham de 5 a 8 classes (de mesma amplitude)
 - 2 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
<i>Sum</i>	50	1.00	

Tabela de frequência - quantitativa discreta (muitos valores)

- Considerando a variável TV
- Apesar de ser discreta, a amplitude de valores é muito grande e não seria viável contar as frequências de cada valor
- Nesse caso, utiliza-se o mesmo procedimento para quantitativas contínuas
 - Foram construídas 6 classes de amplitude 6¹
 - As classes são do tipo $[a, b)$ ou $a \vdash b$

	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
<i>Sum</i>	50	1.00	

¹Obs.: no livro a tabela tem 5 classes, pois a última tem comprimento 12.

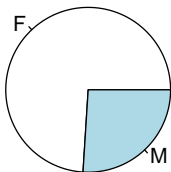
Representação gráfica

- As informações contidas nas tabelas podem ser visualizadas através de gráficos
- Assim como nas tabelas, 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
- Os principais são:
 - Diagrama circular (setores ou “pizza”)
 - Gráfico de barras
 - Histograma
 - Boxplot

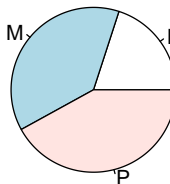
Diagrama circular

- Adequado para variáveis qualitativas nominal e ordinal.

Sexo



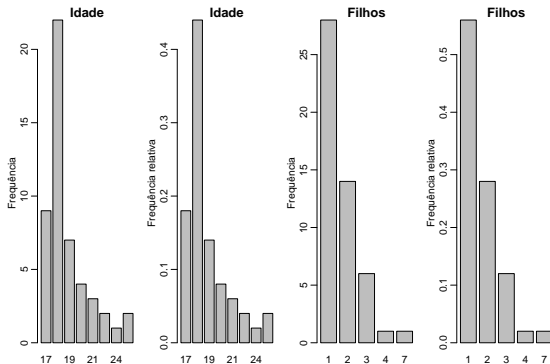
Toler



- O uso deste tipo de gráfico deve ser evitado, pois pode ser de difícil interpretação

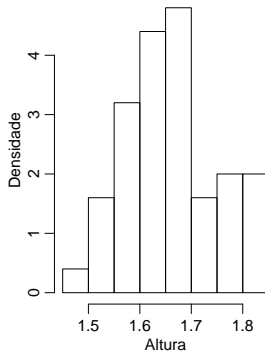
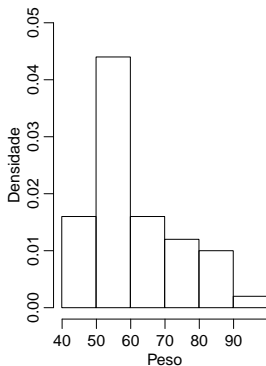
Gráfico de barras

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



Histograma

- Adequado para quantitativa contínua.



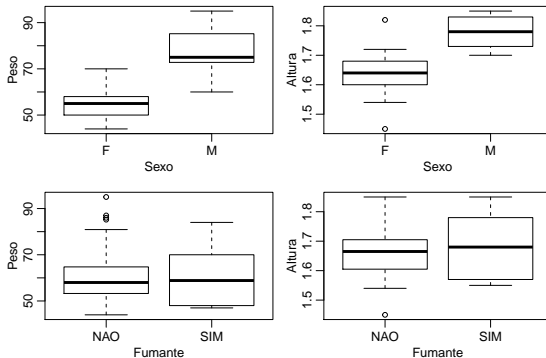
- Altura de cada retângulo é a densidade definida pelo quociente da área pela amplitude da faixa, $h = \frac{f_i}{AMP}$.

Mediana e quartis

- **Mediana:** valor da variável que divide o conjunto de dados ordenados 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:**
 - $Q_1 = 0.25 \cdot (N + 1)$ e arredonde.
 - $Q_2 =$ média dos valores nas posições $(N/2)$ e $(N/2) + 1$ se N par e $Q_2 = (N + 1)/2$ se N ímpar.
 - $Q_3 = 0.75 \cdot (N + 1)$ e arredonde.
- **Exemplo:** Considere o conjunto de dados: 8.43(1), 8.65(2), 9.96(3), 10.91(4), 10.46(5) e 10.83(6).
 - $Q_1 = 0.25 \cdot 7 = 1.75 \approx 2$, ou seja 8.65.
 - $Q_2 =$ média dos valores nas posições 3 e 4, ou seja, $(9.96 + 10.91)/2 = 10.43$.
 - $Q_3 = 0.75 \cdot 7 = 5.25 \approx 5$, ou seja, 10.46.

Boxplots

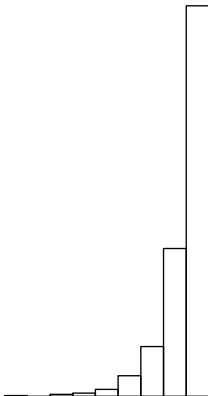
- Adequado para quantitativa contínua.



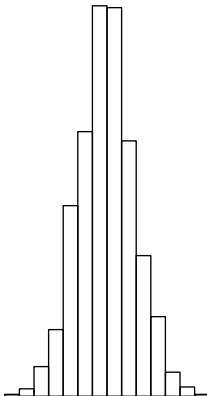
- Excelente para explorar relações entre variáveis quantitativas e qualitativas.

Tipos de simetria

Assimétrico à esquerda



Simétrico



Assimétrico à direita

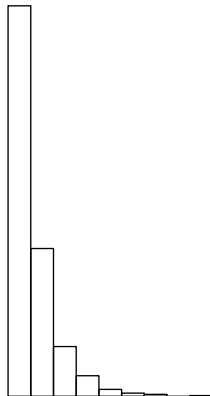


Diagrama de dispersão

- Adequado para verificar relação entre variáveis quantitativas.

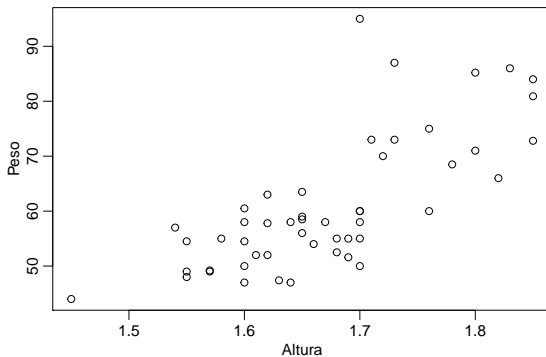
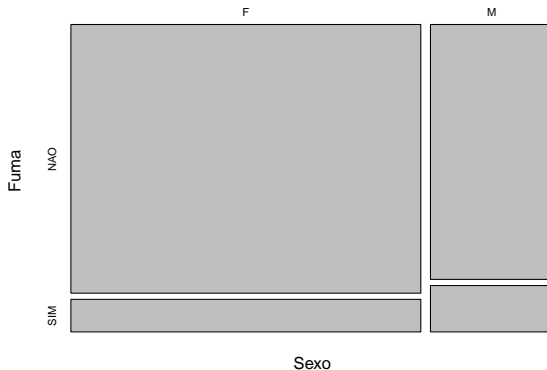


Gráfico de mosaico

- Adequado para verificar relação entre variáveis qualitativas (nominais ou ordinais).



Sumário

- 1 Informações gerais
 - O que é estatística?
- 2 Análise exploratória de dados
 - Organização de Dados
 - Tabelas de frequência
 - Representação gráfica
- 3 Exercícios recomendados

Exercícios recomendados

- Seção 1.1: Ex. 1, 2 e 3.
- Seção 1.2: Ex. 1 e 4.
- Seção 1.4: Ex. 1, 3, 5 (troque diagrama circular pro gráfico de barras), 8, 9, 12, 18 e 20.