Jeidsan A. da C. Pereira

Estatística e Ciência de Dados

Notas e solução dos exercícios

Conteúdo

Prefácio

Esta página contém notas e solução para os exercícios propostos no livro **Estatística e Ciência de Dados**, de autoria de Pedro Alberto Morettin e Júlio da Motta Singer, publicado pela LTC em 2022 [Morettin and Singer, 2022].

É importante destacar que trata-se de um produto não oficial, as anotações e soluções de exercícios aqui apresentadas são de cunho pessoal e não possuem qualquer revisão ou análise por parte dos autores da obra ou da editora. Dessa forma e por se tratar de um produto construído durante o processo de aprendizagem, o conteúdo pode conter erros, tanto no texto em si, como na lógica utilizada para solução dos exercícios.

Dúvidas ou sugestões de melhoria podem ser encaminhadas para o e-mail *jeidsan.pereira@gmail.com*[1].

Pendências

- Exercício 2.2;
-

[1] mailto:jeidsan.pereira@gmail.com

1

Estatística, Ciência de Dados e Megadados

1.1 Introdução

Atualmente, os termos *Data Science* (**Ciência de Dados**) e *Big Data* (**Megadadas**) são utilizados em profusão como se envolvessem conceitos novos, distintos daqueles com que os estatísticos lidam há cerca de dois séculos [Morettin and Singer, 2022, p. 1].

1.2 Aprendizado com estatística

O **aprendizado supervisionado** está relacionado com metodologia desenvolvida essencialmente para **previsão** e **classificação**. No âmbito da previsão, o objetivo é utilizar **variáveis preditivas** (sexo, classe social, renda, por exemplo) observadas em várias **unidades** (clientes de um banco, por exemplo) para "advinhar" valores de uma **variável resposta numérica** (saldo médio, por exemplo) de novas unidades. O problema de classificação consiste em qual categoria de uma **variável resposta qualitativa** (bons e maus pagadores, por exemplo) as novas unidades são classificadas [Morettin and Singer, 2022, p. 3].

> No **aprendizado não supervisionado**, dispomos apenas um conjunto de dados, sem distinção entre preditoras e respostas, e o objetivo é descrever **associações** e **padrões** entre essas variáveis e **agrupá-las** com o objetivo de identificar características comuns e conjuntos de unidades de investigação ou desenvolver métodos para combiná-las e assim **reduzir sua dimensionalidade** [Morettin and Singer, 2022, p. 3].

> Além de aprendizado supervisionado e não supervisionado, podemos acrescentar um terceiro tipo, denominado **aprendizado com reforço** (*reinforcement learning*), segundo o qual um algoritmo "aprende" a realizar determinadas tarefas por meio de repetições com o fim de maximizar um prêmio sujeito a um valor máximo [Morettin and Singer, 2022, p. 3].

> Embora tanto o aprendizado supervisionado quanto o aprendizado com reforço utilizem um mapeamento entre entradas (*inputs*) e saídas (*outputs*), no primeiro caso a retroalimentação (*feedback*) fornecida ao algoritmo é um conjunto de ações corretas necessárias para a realização de uma tarefa; no aprendizado com reforço, por outro lado, a retroalimentação é baseada num sistema com prêmios e punições como indicativos de ações corretas ou incorretas [Morettin and Singer, 2022, p. 3].

1.3 Aprendizado automático

Jordan [2019 *apud* Morettin and Singer, 2022, p. 4] distingue três tipos de inteligência artificial: i) inteligência artificial imitativa humana; ii) aumento de inteligência; e iii) infraestrutura inteligente.

> De modo informal, a inteligência artificial está relacionada com um esforço para automatizar tarefas intelectuais usualmente realizadas por seres humanos (Chollet, 2018) e consequentemente, intimamente ligada ao desenvolvimento da computação (ou programação de computadores) [Morettin and Singer, 2022, p. 4].

> Convém ressaltar que o objetivo do aprendizado automático não é o mesmo daquele consi- derado na análise de regressão usual, em que se pretende entender como cada variável preditora Ro está associada com a variável resposta. O objetivo do aprendizado automático é selecionar o mo- delo que produz melhores previsões, mesmo que as variáveis selecionadas com essa finalidade não sejam aquelas consideradas numa análise padrão [Morettin and Singer, 2022, p. 5].

1.4 Uma cronologia do desenvolvimento da estatística

Sem notas para esta seção.

1.5 Notação e tipos de dados

Sem notas para esta seção.

1.6 Paradigmas para o aprendizado com estatística

Sem notas para esta seção.

1.7 Este livro

> Independentemente do volume de dados disponíveis para análise, Ciência de Dados é uma atividade multidisciplinar que envolve: i) um problema a ser resolvido com questões claramente especificadas; ii) um conjunto de dados (seja ele volumoso ou não); iii) os meios para sua obtenção; iv) sua organização; v) a especificação do problema original em termos das variáveis desse conjunto de dados; vi) a descrição e resumo dos dados à luz do problema a ser resolvido; vii) a escolha das técnicas estatísticas apropriadas para a resolução desse problema; viii) os algoritmos computacionais necessários para a implementação dessas técnicas; ix) a apresentação dos resultados [Morettin and Singer, 2022, p. 11].

1.8 Conjuntos de dados

Sem notas para esta seção.

1.9 Notas do capítulo

Sem notas para esta seção.

Parte I

Análise Exploratória de Dados

2

Preparação dos dados

2.1 Considerações preliminares

> O ramo da Estatística conhecido como **Análise Exploratória de Dados** se ocupa da organização e resumo dos dados de uma amostra ou, eventualmente, de toda a população e o ramo conhecido como **Inferência Estatística** se refere ao processo de se tirar conclusães sobre uma população com base em uma amostra dela [Morettin and Singer, 2022, p. 21].

2.2 Planilhas de dados

Sem notas para esta seção.

2.3 Contrução de tabelas

Sem notas para esta seção.

2.4 Construção de gráficos

Sem notas para esta seção.

2.5 Notas de capítulo

Sem notas para esta seção.

2.6 Exercícios

Exercício 2.1

O objetivo de um estudo da Faculdade de Medicina da USP foi avaliar a associação entre a quantidade de morfina administrada a pacientes com dores intensas provenientes de lesões medulares ou radiculares e a dosagem dessa substância em seus cabelos. Três medidas foram realizadas em cada paciente, a primeira logo após o início do tratamento e as demais após 30 e 60 dias. Detalhes podem ser obtidos no documento disponível no arquivo `morfina.doc`.

A planilha `morfina.xls`, disponível no arquivo morfina foi entregue ao estatístico para análise e contém resumos de características demográficas além dos dados do estudo.

a) Com base nessa planilha, apresente um dicionário com a especificação das variáveis segundo as indicações da Seção 2.2 e construa a planilha correspondente.
b) Com as informações disponíveis, construa tabelas para as variáveis sexo, raça, grau de instrução e tipo de lesão segundo as sugestões da Seção 2.3.

Solução. Utilizando o arquivo `morfina.doc` chegamos à seguinte solução para o item `a`:

Tabela 2.1: Tabela 2.1: Dicionários para as variáveis do estudo `morfina.doc` [1]

Rótulo	Variável	Unidade de medida
id	Identificação do paciente	
data	Data de avaliação do paciente	
idade	Idade do paciente	anos
sexo	Sexo do paciente	1 - masculino 2 - feminino
raça	Raça e/ou etnia do paciente	1 - pardo 2 - negro 3 - branco 4 - indígena 5 - amarelo
religião	Religião do paciente	1 - catolico 2 - sem religião 3 - evangélico 4 - espírita 5 - judeu 6 - outra
peso	Peso do paciente	quilogramas (kg)
altura	Altura do paciente	metros (m)
instrução	Grau de instrução do paciente	1 - Analfabeto 2 - Alfabetizado 3 - Ens. Fundamental 4 - Ens. Médio 5 - Ens. Superior
enfermidade	Enfermidade primária do paciente	1 - Lesão medular 2 - Lesão radicular 3 - Trauma 4 - FAF 5 - Pós cirúrgico
tipo	Tipo de cabelo do paciente	1 - Natural 2 - Artificial
cor	Cor do cabelo do paciente	1 - Marrom 2 - Vermelho 3 - Preto 4 - Louro 5 - Cinza 6 - Branco 7 - Outro
forma	Forma do cabelo do paciente	1 - Caucasiana 2 - Asiática 3 - Negróide 4 - Outra
medicamentos	Medicamentos em uso pelo paciente	
composicao	Composição da solução do reservatório	
ampola	Número da ampola de morfina	

Rótulo	Variável	Unidade de medida
sf	Quantidade de soro fisiológico para diluição	mililitros (ml)
concentracao	Concentração da solução	percentual (%)
disp_prop	Número de disparos proposto por dia	
disp_real	Número de disparos realizados por dia	
vol_desprez	Quantidade de solução desprezada no reservatório	mililitros (ml)
con_desp	Concentração da solução desprezada no reservatório	percentual (%)
obstrucao	Ocorrência de obstrução no cateter	1 - Sim 2 - Não
infeccao	Ocorrência de infecção do sistema	1 - Sim 2 - Não
nausea	Ocorrência de nausea como efeito colateral	1 - Sim 2 - Não
sonolencia	Ocorrência de sonolência como efeito colateral	1 - Sim 2 - Não
constipação	Ocorrência de constipação como efeito colateral	1 - Sim 2 - Não
tontura	Ocorrência de tontura como efeito colateral	1 - Sim 2 - Não
prurido	Ocorrência de prurido como efeito colateral	1 - Sim 2 - Não
retencao	Ocorrência de retenção urinária como efeito colateral	1 - Sim 2 - Não
outros	Ocorrência de outros efeitos colaterais	1 - Sim 2 - Não
d0	Dose inicial de morfina	miligrama (mg)
d30	Dose de morfina após 30 dias	miligrama (mg)
d60	Dose de morfina após 60 dias	miligrama (mg)
t0	Quantidade inicial de morfina no cabelo	miligrama (mg)
t30	Quantidade de morfina no cabelo após 30 dias	miligrama (mg)
t60	Quantidade de morfina no cabelo após 60 dias	miligrama (mg)

Para o item `b`, como teremos quatro tabelas distintas e os dados não estão cruzados, optamos por incluir o rótulo da primeira coluna no título do gráfico, a fim de não termos títulos repetidos. Temos o seguinte:

Tabela 2.2: Tabela 2.2: Distribuição dos pacientes conforme o sexo

Sexo	Número de pacientes	Percentual (%)
Masculino	19	54
Feminino	16	46
Total	**35**	**100**

[1]Trata-se de uma versão inicial para o dicionário de dados. O mesmo será revisado posteriormente para se adequar à especificação do documento `morfina.doc`.

Tabela 2.3: Tabela 2.3: Distribuição dos pacientes conforme raça/etnia

Raça	Número de pacientes	Percentual (%)
Pardo	20	57
Branco	13	37
Negro	2	6
Total	**35**	**100**

Tabela 2.4: Tabela 2.4: Distribuição dos pacientes conforme o grau de instrução

Grau de Instrução	Número de pacientes	Percentual (%)
Ensino Fundamental	25	71
Ensino Médio	9	26
Ensino Superior	1	3
Total	**35**	**100**

Tabela 2.5: Tabela 2.5: Distribuição dos pacientes conforme o tipo de lesão

Tipo de lesão	Número de pacientes	Percentual (%)
Medular	10	29
Radicular	25	71
Total	**35**	**100**

Exercício 2.2

A Figura 2.6 foi extraída de um estudo sobre atitudes de profissionais de saúde com relação a cuidados com infecção hospitalar. Critique-a e reformule-a para facilitar sua leitura, lembrando que a comparação de maior interesse é entre as diferentes categorias profissionais.

Solução.

Tabela 2.6: Tabela 2.6: ABC

A	B	C
A	B	C
A	B	C

Exercício 2.3

Utilize as sugestões para construção de planilhas apresentadas na Seção 2.2 com a finalidade de preparar os dados do arquivo empresa para análise estatística.

Solução. Vamos iniciar construindo um dicionário para os dados

Tabela 2.7: Tabela 2.7: Dicionário de dados para a planilha `empresa.xls`

Rótulo	Descrição	Unidade de medida
id	Identificador ddo funcionário	
estado	Estado civil do funcionário	1 - Solteiro 2 - Casado
instrucao	Grau de instrução do funcionário	1 - Ensino Fundamental 2 - Ensino Médio 3 - Ensino Superior
filhos	Número de filhos do funcionário	
salario	Salário do funcionário	salário mínimo
anos	Idade do funcionário	anos
meses	Fração da idade do funcionário	meses
regiao	Região de procedência do funcionário	1 - Interior2 - Capital 3 - Outra

Com o dicionário de dados em mãos, podemos atualizar a planilha:

Tabela 2.8: Tabela 2.8: Dados de funcionários de uma empresa (parte)

id	estado	instrucao	filhos	salario	anos	meses	regiao
1	1	1		4.00	26	3	1
2	2	1	1	4.56	32	10	2
3	2	1	2	5.25	36	5	2
4	1	2		5.73	20	10	3
5	1	1		6.26	40	7	3
6	2	1	0	6.66	28	0	1
7	1	1		6.86	41	0	1
8	1	1		7.39	43	4	2
9	2	2	1	7.59	34	10	2
10	1	2		7.44	23	6	3

Exercício 2.4

Num estudo planejado para avaliar o consumo médio de combustível de veículos em diferentes velocidades foram utilizados 4 automóveis da marca A e 3 automóveis da marca B selecionados ao acaso das respectivas linhas de produção. O consumo (em

L/km) de cada um dos 7 automóveis foi observado em 3 velocidades diferentes (40 km/h, 80 km/h e 110km/h) Delineie uma planilha apropriada para a coleta e análise estatística dos dados, rotulando-a adequadamente.

Solução. Vamos começar construindo um dicionário de dados para a planilha:

Tabela 2.9: Tabela 2.9: Dicionário para as variáveis do estudo sobre consumo de combustível

Rótulo	Descrição	Unidade
id	Identificador do veículo	
marca	Marca fabricante do veículo	
modelo	Modelo do veículo	
consumo40	Consumo de combustível a 40 km/h	L/km
consumo80	Consumo de combustível a 80 km/h	L/km
consumo110	Consumo de combustível a 110 km/h	L/km

Agora vamos montar uma planilha (fictícia) seguindo a padronização definida acima:

Tabela 2.10: Tabela 2.10: Planilha de dados para estudo sobre consumo de combustível

id	marca	modelo	consumo40	consumo80	consumo110
1	A	XPTO 1			
2	A	XPTO 2			
3	A	XPTO 3			
4	A	XPTO 4			
5	B	XYZ 1			
6	B	XYZ 2			
7	B	XYZ 3			

Exercício 2.5

Utilizando os dados do arquivo `enforco.xls`, prepare uma planilha Excel num formato conveniente para análise pelo `R`. Inclua apenas as variáveis `Idade`, `Altura`, `Peso`, `Frequência cardiaca` e `VO2` no repouso além do quociente `VE/VCO2`, as correspondentes porcentagens relativamente ao máximo, o quociente `VO2/FC` no pico do exercício e `data do óbito`. Importe a planilha Excel que você criou utilizando comandos `R` e obtenha as características do arquivo importado (número de casos, número de observações omissas etc.)

Solução. Conforme especificação enunciada, criamos o arquivo `esforco.csv` que poderá ser carregado da seguinte maneira:

```
(esforco <- read_csv(paste0(data_dir, "esforco.csv")))
```

```
## Rows: 127 Columns: 8
## -- Column specification --------------------------------------------------------
---
## Delimiter: ","
## chr (1): obito
## dbl (7): id, idade, altura, peso, fc_repouso, vo2_repouco, ve_vo2_pico
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
## # A tibble: 127 x 8
##       id idade altura  peso fc_repouso vo2_repouco ve_vo2_pico obito
##    <dbl> <dbl>  <dbl> <dbl>      <dbl>       <dbl>       <dbl> <chr>
##  1     1    38    149    54         89         5.9        65.6 26/07/1991
##  2     2    49    167    80         69         3.4        37.3 30/07/1995
##  3     3    65    153    56         82         3          59.7 21/08/1993
##  4     4    52    175    78         89         3.8        52.4 14/11/1992
##  5     5    52    157    59         82         3.2        48.8 30/07/1994
##  6     6    58    150    62         75         3.8        54.1 Não
##  7     7    24    155    42         89         3.5       102.  17/10/1991
##  8     8    39    149    55         91         3.9        67.8 31/08/1992
##  9     9    48    160    77        101         2.5        59.5 Não
## 10    10    50    171    81        120         3          47.8 Não
## # i 117 more rows
```

Temos 127 casos e 8 variáveis.

Exercício 2.6

A Figura 2.7 contém uma planilha encaminhada pelos investigadores responsáveis por um estudo sobre AIDS para análise estatistica. Organize-a de forma a permitir sua análise por meio de um pacote computacional como o R.

Solução. Os dados foram reorganizados no arquivo `aids.csv`.

```
(aids <- read_csv(paste0(data_dir, "aids.csv")))
```

```
## Rows: 19 Columns: 7
```

```
## -- Column specification --------------------------------------------------------
---
## Delimiter: ","
## chr (3): id, dst, mac
## dbl (4): grupo, diagnostico, peso, tempo_peso
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.


## # A tibble: 19 x 7
##    id        grupo diagnostico dst                mac     peso tempo_peso
##    <chr>     <dbl>       <dbl> <chr>              <chr>  <dbl>      <dbl>
##  1 2847111D      1         0   <NA>               pilula    11         37
##  2 3034048F      1         0.5 <NA>               pilula    NA         NA
##  3 3244701J      1         1   <NA>               condon    NA         NA
##  4 2943791B      1         0   <NA>               não        8         39
##  5 3000327F      1         4   condiloma/sífilis  não        9         39
##  6 3232893D      1         1   <NA>               diu        3         39
##  7 3028772E      1         3   <NA>               não        3         38
##  8 3240047G      1         0   <NA>               pilula     9         38
##  9 3017222G      1        NA   HPV                condon    NA         NA
## 10 3015834J      1         2   <NA>               condon    14         40
## 11 3173611E      2         0.4 abcesso ovariano   condon    15         40
## 12 3296159D      2         0   <NA>               condon    NA         NA
## 13 3147820D1     2         2   <NA>               <NA>       4         37
## 14 3274750K      2         3   <NA>               condon     8         38
## 15 3274447H      2         0   sífilis com 3 meses condon    NA         NA
## 16 2960066D      2         5   <NA>               <NA>      13         36
## 17 3235727J      2         7   <NA>               condon    -2         38
## 18 3264897E      2         0   condiloma/sífilis  condon     0         NA
## 19 3044120J      2         5   HPV                <NA>       3         39
```

Uma possível melhoria seria a transformação das variáveis `grupo` e `mac` em fatores.

Exercício 2.7

A planilha apresentada na Figura 2.8 contém dados de um estudo em que o limiar auditivo foi avaliado nas orelhas direita (OD) e esquerda (OE) de 13 pacientes em 3 ocasiões (Limiar, Teste 1 e Teste 2). Reformate-a segundo as recomendações da Seção 2.2, indicando claramente

a) a definição das variáveis.
b) os rótulos para as colunas da planilha.

Solução. Precisamos inicialmente definir um dicionário para as variáveis e, na sequência, refatorar a planilha.

Tabela 2.11: Tabela 2.11: Dicionário de dados para o estudo sobre limiar auditivo [2]

Rótulo	Descrição da variável	Unidade de medida
id	Identificador do paciente	
od0	Limiar auditivo da orelha direita no início do estudo	%
oe0	Limiar auditivo da orelha esquerda no início do estudo	%
od1	Limiar auditivo da orelha direita no primeiro teste	%
oe1	Limiar auditivo da orelha esquerda no primeiro teste	%
od2	Limiar auditivo da orelha direita no segundo teste	%
oe2	Limiar auditivo da orelha esquerda no segundo teste	%

Tabela 2.12: Tabela 2.12: Limiar auditivo de pacientes observados em 3 ocasiões

id	od0	oe0	od1	oe1	od2	oe2
1	50.00	50.00	50.00	50.00	80.00	80.00
2	41.00	40.00	45.00	50.00	68.00	80.00
3	41.25	41.25	45.00	45.00	64.00	72.00
4	45.00	43.75	60.00	50.00	76.00	88.00
5	51.25	47.50	50.00	50.00	80.00	80.00
6	45.00	52.50	50.00	50.00	84.00	96.00
7	52.50	50.00	55.00	45.00	40.00	28.00
8	42.15	48.75	50.00	50.00	80.00	76.00
9	50.00	48.75	50.00	50.00	72.00	80.00
10	47.50	46.25	55.00	60.00	84.00	84.00
11	55.00	56.25	40.00	35.00	80.00	84.00
12	46.25	46.25	45.00	45.00	72.00	84.00
13	50.00	47.50	40.00	50.00	76.00	76.00

Exercício 2.8

A planilha disponível no arquivo `cidades.xls` contém informações demográficas de 3554 municípios brasileiros.

[2] Como consideramos o limite de detecção como sendo 0.05, foram utilizadas duas casas decimais para representar os limites auditivos observados.

a) Importe-a para permitir a análise por meio do software R, indicando os problemas encontrados nesse processo além de sua solução.
b) Use o comando `summary` para obter um resumo das variáveis do arquivo.
c) Classifique cada variável como numérica ou alfanumérica e indique o número de observações omissas de cada uma delas.

Solução. Ao tentar realizar a leitura utilizando a função `read_csv` que já conhecemos, obteríamos um erro devido a planilha conter formatações.

```
cidades <- read_csv(paste0(data_dir, "cidades.xls"))
```

```
## Error in vroom_(file, delim = delim %||% col_types$delim, col_names = col_names, : cadeia de caracteres com nul inc
```

Podemos limpar a formatação da planilha e tentar a importação novamente ou utilizar a função `read_xls` do pacore **readxl**.

```
cidades <- readxl::read_xls(paste0(data_dir, "cidades.xls"), na = '-')
```

Note que, como os valores faltantes na planilha estão indicado com um hífem, utilizamos o argumeno `na = '-'` para convertelos automaticamente para `NA`. Note também que as últimas duas linhas do data frame `cidades` contém os totalizadores e não observações. Vamos removê-las!

```
cidades <- head(cidades, -2)
```

Note que o rótulo das variáveis está em maiúsculo, vamos colocá-los em minústuco com a ajuda da função `str_to_lower()` do pacore **stringr**:

```
colnames(cidades) <- str_to_lower(colnames(cidades))
```

Podemos agora ver que temos 17 variáveis e 3554 unidades de análise. Um resumo das variáveis do conjunto de dados é mostrado pelo comando `summary`:

```
summary(cidades)
```

```
##     munic               uf               código         poptot
##  Length:3554        Length:3554        Min.   :1001   Min.   :     795
```

```
##  Class :character   Class :character   1st Qu.:1889   1st Qu.:    7995
##  Mode  :character   Mode  :character   Median :3720   Median :   15632
##                                        Mean   :3440   Mean   :   43650
##                                        3rd Qu.:4609   3rd Qu.:   30655
##                                        Max.   :5497   Max.   :10406166
##
##     cres_pop           popurb            pibtot             cres_pib
##  Min.   :-13.330   Min.   :    423   Min.   :     0.90   Min.   : 0.0000
##  1st Qu.:  0.020   1st Qu.:   4388   1st Qu.:    13.48   1st Qu.: 0.6936
##  Median :  1.145   Median :   9232   Median :    26.79   Median : 1.0372
##  Mean   :  1.283   Mean   :  36908   Mean   :   177.93   Mean   : 1.1607
##  3rd Qu.:  2.310   3rd Qu.:  20732   3rd Qu.:    66.80   3rd Qu.: 1.4493
##  Max.   : 23.630   Max.   :9785640   Max.   :105906.65   Max.   :24.6598
##                                      NA's   :14          NA's   :14
##      grau1             grau2            superior          11oumais
##  Min.   :    469   Min.   :     47   Min.   :      0   Min.   :     37
##  1st Qu.:   4738   1st Qu.:    495   1st Qu.:     75   1st Qu.:    407
##  Median :   8491   Median :    950   Median :    178   Median :    786
##  Mean   :  22833   Mean   :   5060   Mean   :   2064   Mean   :   5407
##  3rd Qu.:  16057   3rd Qu.:   2272   3rd Qu.:    522   3rd Qu.:   1963
##  Max.   :5322497   Max.   :1606381   Max.   :1076916   Max.   :2142313
##  NA's   :13        NA's   :13        NA's   :13        NA's   :13
##     empregad          microemp           peqemp             medemp
##  Min.   :     10   Min.   :     3.0   Min.   :    0.00   Min.   :   0.000
##  1st Qu.:    414   1st Qu.:    94.0   1st Qu.:    1.00   1st Qu.:   1.000
##  Median :    926   Median :   207.0   Median :    3.00   Median :   1.000
##  Mean   :   7778   Mean   :   916.9   Mean   :   36.69   Mean   :   6.929
##  3rd Qu.:   2743   3rd Qu.:   503.0   3rd Qu.:   13.00   3rd Qu.:   2.000
##  Max.   :3986021   Max.   :377600.0   Max.   :18494.00   Max.   :3198.000
##  NA's   :14        NA's   :14         NA's   :14         NA's   :14
##      graenp
##  Min.   :  0.000
##  1st Qu.:  0.000
##  Median :  0.000
##  Mean   :  1.341
##  3rd Qu.:  1.000
##  Max.   :568.000
##  NA's   :14
```

Esse comando também nos permite perceber os tipos de cada uma das variáveis e se as mesmas contém valores faltantes. Essas informações estão resumidas na Tabela 2.13.

Tabela 2.13: Tabela 2.13: Resumo das observações da tabela `cidades.xls`

Variável	Tipo	Número de observações faltantes
munic	Alfanumérica	0
uf	Alfanumérica	0
codigo	Numérica	0
poptot	Numérica	0
cres_pop	Numérica	0
popurb	Numérica	0
pibtot	Numérica	14
cres_pib	Numérica	14
grau1	Numérica	13
grau2	Numérica	13
superior	Numérica	13
11oumais	Numérica	13
empregad	Numérica	14
microemp	Numérica	14
peqemp	Numérica	14
medemp	Numérica	14
graemp	Numérica	14

Exercício 2.9

Preencha a ficha de inscrição do Centro de Estatistica Aplicada www.ime.usp.br/-cea[3] com as informações de um estudo em que você está envolvido.

Solução. Não se aplica.

[3] http://www.ime.usp.br/-cea

3

Análise de dados de uma variável

3.1 Introdução

A ideia de uma análise descritiva de dados é tentar responder as seguintes questões:

i) Qual a frequência com que cada valor (ou intervalo de valores) aparece no conjunto de dados ou seja, qual a distribuição de frequências dos dados?
ii) Quais são alguns valores típicos do conjunto de dados, como mínimo e máximo?
iii) Qual seria um valor para representar a posição (ou localização) central do conjunto de dados?
iv) Qual seria uma medida da variabilidade ou dispersão dos dados?
v) Existem valores atípicos ou discrepantes (outliers) no conjunto de dados?
vi) A distribuição de frequências dos dados pode ser considerada simétrica?

[Morettin and Singer, 2022, p. 37]

3.2 Distribuição de frequências

Sem notas para esta seção.

3.3 Medidas resumo

Dado um número $0 < \alpha < 1$, a **média aparada** de ordem α, $\overline{x}(\alpha)$, é definida como a média do conjunto de dados obtido após a eliminação das 100% primeiras observações ordenadas e das 100% últimas observações ordenadas do conjunto original [...]. Para $\alpha = 0,25$ obtemos a chamada **meia média**. [Morettin and Singer, 2022, p. 47].

Dado um número natural $k \geq 2$ e um conjunto $X = \{x_1, \cdots, x_n\}$ com $n \in \mathbb{N}$, o k-ésimo **momento centrado** de X é dado por [Morettin and Singer, 2022, p. 50]:

$$m_k = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^k$$

Dentre as medidas de assimetria, as mais comuns são:

a) o coeficiente de assimetria de Fisher-Pearson:
$$g_1 = \frac{m_3}{m_2^{(3/2)}}$$

b) o coeficiente de assimetria de Fisher-Pearson ajustado:
$$\frac{\sqrt{n(n-1)}}{n-2} g_1$$

[Morettin and Singer, 2022, p. 50]

As principais propriedades desses coeficientes são

i) seu sinal reflete a direção da assimetria (sinal negativo corresponde a assimetria à direita e sinal positivo corresponde a assimetria à esquerda);

ii) comparam a assimetria dos dados com aquela da distribuição normal, que é simétrica,
iii) valores mais afastados do zero indicam maiores magnitudes de assimetria e consequente- mente, maior afastamento da distribuição normal;
iv) a estatística indicada em (3.15) tem um ajuste para o tamanho amostral;
v) esse ajuste tem pequeno impacto em grandes amostras. [Morettin and Singer, 2022, p. 51]

Outro coeficiente de assimetria mais intuitivo é o chamado **coeficiente de assimetria de Pearson 2**, estimado por

$$Sk_2 = \frac{3 \cdot [\bar{x} - med(x_1, \cdots, x_n)]}{S}$$

[Morettin and Singer, 2022, p. 51].

Seja X uma variável aleatória qualquer, com média μ e variância σ^2. A **curtose** de X é definida por

$$K(X) = E\left[\frac{(X-\mu)^4}{\sigma^4}\right]$$

[Morettin and Singer, 2022, p. 53]

3.4 *Boxplots*

Sem notas para esta seção.

3.5 Modelos probabilísticos

Sem notas para esta seção.

3.6 Dados amostrais

Sem notas para esta seção.

3.7 Gráficos QQ

> Uma das questões fundamentais na especificação de um modelo para inferência estatística é a escolha de um modelo probabilístico para representar a distribuição (desconhecida) da variável de interesse na população. Uma possível estratégia para isso é examinar o histograma dos dados amostrais e compará-lo com histogramas teóricos associados a modelos probabilísticos candidatos. Alternativamente, os gráficos QQ (QQ plots) também podem ser utilizados com essa finalidade [Morettin and Singer, 2022, p. 59].

3.8 Desvio padrão e erro padrão

Sem notas para esta seção.

3.9 Intervalo de confiança e tamanho da amostra

[...] **margem de erro**, que, essencialmente, é uma medida de nossa incerteza na extrapolação dos resultados obtidos para a população de onde assumimos que foi obtida [Morettin and Singer, 2022, p. 65].

A margem de erro depende do processo amostral, do desvio padrão amostral S, do tamanho da amostra n e é dado por $me = \frac{kS}{\sqrt{n}}$ em que k é uma constante que depende do modelo probabilístico adotado e da confiança com que pretendemos fazer a inferência [Morettin and Singer, 2022, p. 65].

Especificamente no caso da estimação da média (populacional) μ de uma variável X, a pergunta seria *Qual é o tamanho da amostra necessário para que a estimativa $\overline{X}$ da média μ tenha uma precisão ε?* A resposta pode ser obtida da expressão (3.21), fazendo $\varepsilon = \frac{1{,}96 \cdot S}{\sqrt{n}}$ [Morettin and Singer, 2022, p. 66].

3.10 Transformação de variáveis

Se quisermos utilizar os procedimentos talhados para análise de dados com distribuição normal em situações nas quais a distribuição

dos dados amostrais é sabidamente assimétrica, pode-se considerar uma transformação das observações com a finalidade de se obter uma distribuição "mais simétrica" e portanto, mais próxima da distribuição normal. Uma transformação bastante usada com esse propósito é

$$x^{(p)} = \begin{cases} x^p, & se\ p > 0 \\ log(x), & se\ p = 0 \\ -x^p, & se\ p < 0 \end{cases}$$

Essa transformação com $0 < p < 1$ apropriada para distribuições assimétricas à direita, pois valores grandes decrescem de x decrescem mais relativamente a valores pequenos. Para distribuições assimétricas à esquerda, basta tomar $p > 1$. Normalmente, consideramos valores de p na sequência

$$\cdots, -3, -2, -1, -1/2, -1/3, -1/4, 0, 1/4, 1/3, 1/2, 1, 2, 3, \cdots$$

e para cada um deles construímos gráficos apropriados (histogramas, boxplots) com os dados originais transformados, com a finalidade de escolher o valor mais adequado para p. Hinkley (1977) sugere que para cada valor de p na sequência acima se calcule a média, a mediana e um estimador de escala (esvio padrão ou algum estimador robusto) e então se escolha o valor que minimiza

$$d_p = \frac{mdia - mediana}{medida\ de\ escala}$$

[Morettin and Singer, 2022, p. 67].

A transformação (3.23) [acima] é um caso particular das **transformações de Box-Cox** que são da forma

$$g(x) = \begin{cases} \frac{(x^p - 1)}{p}, & se\ p \neq 0 \\ log(x), & se\ p = 0 \end{cases}$$

[Morettin and Singer, 2022, p. 69].

3.11 Notas de capítulo

Sem notas para esta seção.

3.12 Exercícios

4

Análise de dados de duas variáveis

5

Análise de dados de várias variáveis

6

Análise de Regressão

7

Análise de Sobrevivência

Parte II

Aprendizado Supervisionado

8

Regularização e Modelos Aditivos Generalizados

9

Classificação por meio de técnicas clássicas

10

Algoritmos de Suporte Vetorial

11

Árvores e Florestas

12

Redes neurais

Parte III

Aprendizado não Supervisionado

13

Análise de Agrupamentos

14

Redução de dimensionalidade

A

Otimização numérica

B

Noções de simulação

C

Algoritmos para dados aumentados

Bibliografia

Pedro Alberto Morettin and Julio da Motta Singer. *Estatística e Ciência de Dados*. LTC, Rio de Janeiro, 2022.