

MAE0217 - Estatística Descritiva - Lista 1

Natalia Koza¹
Rafael Gonçalves Pereira da Silva²
Ricardo Geraldês Tolesano³
Rubens Kushimizo Rodrigues Xavier⁴
Rubens Gomes Neto⁵
Rubens Santos Andrade Filho⁶
Thamires dos Santos Matos⁷

Abril de 2021

Sumário

Exercício 1	2
a)	2
b)	2
Exercício 2	2
Exercício 4	3
Exercício 5	4
Exercício 7	4
Exercício 8	6
a)	7
b)	8
c)	9

¹Número USP: 10698432

²Número USP: 9009600

³Número USP: 10734557

⁴Número USP: 8626718

⁵Número USP: 9318484

⁶Número USP: 10370336

⁷Número USP: 9402940

Exercício 1

a)

Para a construção da planilha, primeiro notamos alguns problemas: os dados faltantes estavam codificados como "ZERO", um dado com a letra "O" no lugar de 0, campos numéricos com sufixo "MG", um dado com "," como separador decimal. Após corrigir todos esses problemas, deixamos os dados no formato longo.

Tabela 1: Primeiras 10 linhas dos dados corrigidos e no formato longo.

paciente	tempo	cabelo	morfina
1-JS	0	9,75	50,4
1-JS	30	10,95	33,6
1-JS	60	14,26	53,2
2-OHP	0	3,32	46,0
2-OHP	30	4,52	39,2
2-OHP	60	5,72	35,0
3-VVR	0	5,08	88,0
3-VVR	30	6,20	15,0
3-VVR	60	8,30	81,0
4-LCCS	0	4,28	88,0

E, a seguir, construímos o dicionário desses dados:

Tabela 2: Dicionário dos dados.

Rótulos	Variável	Unidade de medida
paciente	Identificador único do paciente	Código alfanumérico
tempo	Tempo desde o início	0: 0 dias
		30: 30 dias
		60: 60 dias
cabelo	quantidade de morfina nos cabelos	ng
morfina	dose de morfina administrada	mg

b)

Exercício 2

A tabela é derivada de um estudo cujo o objetivo era comparar resultados da Avaliação de Diagnóstico do estudo com os resultados da macro-avaliação Prova São Paulo e verificar a possível influência de variáveis socioeconômicas no aprendizado de Matemática dos alunos da quarta série do Ensino Fundamental

da rede de ensino do município de São Paulo. Pelo fato dos alunos estarem agrupados em escolas construiu-se um modelo de regressão que considerou o agrupamento dos alunos por subprefeitura que considera a relação linear da nota da Avaliação Diagnóstico com a nota da Prova São Paulo e a idade do aluno, e as variações entre subprefeituras.

A tabela visa comparar as notas médias da Avaliação Diagnóstico com os da Prova São Paulo, tal que tais valores fossem padronizados no intervalo $[0,1]$. Para facilitar a leitura, é interessante que seja simplificada e ordenada pelas notas da Avaliação Diagnóstico, além de ter os valores truncados para 2 casas decimais, precisão suficiente já que os resultados são semelhantes e que facilitam a visualização.

Tabela 3: Tabela comparativa das notas médias da avaliação das subprefeituras no modelo padronizado para a escala $[0,1]$.

Subprefeitura	Nota Avaliação	Nota Modelo Ajustado	Nota na Prova São Paulo
V.Prudente/ Sapopemba	0,65	0,62	0,41
São Miguel	0,60	0,58	0,40
Socorro	0,60	0,52	0,36
Aricanduva	0,59	0,59	0,40
Freguesia/Brasilândia	0,59	0,54	0,37
Ipiranga	0,59	0,54	0,37
Itaim Paulista	0,57	0,56	0,37
M'Boi Mirim	0,57	0,55	0,37
Campo Limpo	0,56	0,53	0,36
Casa Verde/ Cachoeirinha	0,54	0,55	0,38
Cidade Tiradentes	0,54	0,54	0,37
Jabaquara	0,53	0,53	0,36
São Mateus	0,52	0,51	0,35
Butantã	0,48	0,47	0,33
Itaquera	0,40	0,56	0,38

Exercício 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 110 km/h). Delineamos uma planilha apropriada para a coleta e análise estatística dos dados e rotulamos-a adequadamente. A planilha encontra-se no formato longo, isto é, medidas de uma mesma variável encontram-se em uma única coluna.

Tabela 4: Delineamento da Planilha para Coleta

automovel	marca	velocidade	consumo
1	A	40	
1	A	80	
1	A	110	
2	A	40	
2	A	80	
2	A	110	
3	A	40	
3	A	80	
3	A	110	
4	A	40	
4	A	80	
4	A	110	
5	B	40	
5	B	80	
5	B	110	
6	B	40	
6	B	80	
6	B	110	
7	B	40	
7	B	80	
7	B	110	

Exercício 5

Exercício 7

Reformatamos a planilha com 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) segundo as recomendações da Seção 2.2 e indicamos claramente a definição das variáveis e os rótulos para as colunas da planilha.

Tabela 5: Dicionário dos dados.

Rótulos	Variável	Unidade de medida
id	Identificador único do paciente	Número inteiro
ocasio	Ocasão da avaliação	0: Limiar
		1: Teste 1
		2: Teste 2
od	Avaliação da Orelha Direita	Percentual
oe	Avaliação da Orelha Esquerda	Percentual

Tabela 6: 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).

id	ocasio	oe (%)	od (%)
1	0	55,00	50,00
1	1	50,00	50,00
1	2	80,00	80,00
2	0	40,00	41,00
2	1	50,00	45,00
2	2	80,00	68,00
3	0	41,25	41,25
3	1	45,00	45,00
3	2	72,00	64,00
4	0	43,75	45,00
4	1	50,00	60,00
4	2	88,00	76,00
5	0	47,50	51,25
5	1	50,00	50,00
5	2	88,00	80,00
6	0	52,50	45,00
6	1	50,00	50,00
6	2	96,00	84,00
7	0	50,00	52,50
7	1	45,00	55,00
7	2	28,00	40,00
8	0	48,75	42,15
8	1	50,00	40,00
8	2	76,00	80,00
9	0	48,75	50,00
9	1	50,00	50,00
9	2	80,00	72,00
10	0	46,25	47,50
10	1	50,00	50,00
10	2	84,00	84,00
11	0	56,25	55,00
11	1	60,00	55,00
11	2	84,00	80,00
12	0	46,25	46,25
12	1	35,00	40,00
12	2	84,00	72,00
13	0	47,50	50,00
13	1	45,00	45,00
13	2	76,00	76,00

Exercício 8

a)

Importamos o arquivo `idades.xls` e usamos o comando `str` para olhar a estrutura dos dados:

```
dados <- readxl::read_xls('data/cidades.xls')
str(dados, strict.width="wrap", width=80)

## tibble[,17] [3,556 x 17] (S3: tbl_df/tbl/data.frame)
## $ MUNIC : chr [1:3556] "SAO PAULO-SP" "RIO DE JANEIRO-RJ" "SALVADOR-BA" "BELO
##   HORIZONTE-MG" ...
## $ UF : chr [1:3556] "SP" "RJ" "BA" "MG" ...
## $ CÓDIGO : num [1:3556] 1001 1002 1003 1004 1005 ...
## $ POPTOT : num [1:3556] 10406166 5850544 2440886 2229697 2138234 ...
## $ CRES_POP: num [1:3556] 1.41 1.32 2.5 1.61 2.13 2.91 1.82 1.38 4.94 1.35 ...
## $ POPURB : num [1:3556] 9785640 5850544 2439881 2229697 2138234 ...
## $ PIBTOT : chr [1:3556] "105906.65014758587" "47171.514842028657"
##   "12028.532892949122" "18572.982040672556" ...
## $ CRES_PIB: chr [1:3556] "1.7512161470303413" "1.321351601458381"
##   "3.0807993763343071" "2.8526224256715946" ...
## $ GRAU1 : chr [1:3556] "5322497" "2731075" "1139181" "1107558" ...
## $ GRAU2 : chr [1:3556] "1606381" "1110059" "465685" "373858" ...
## $ SUPERIOR: chr [1:3556] "1076916" "731746" "148887" "224303" ...
## $ 110UMAI: chr [1:3556] "2142313" "1562602" "485962" "488029" ...
## $ EMPREGAD: chr [1:3556] "3986021" "2041470" "563139" "990843" ...
## $ MICROEMP: chr [1:3556] "377600" "133165" "38922" "75665" ...
## $ PEQEMP : chr [1:3556] "18494" "9521" "2494" "4108" ...
## $ MEDEMP : chr [1:3556] "3198" "1804" "437" "690" ...
## $ GRAENP : chr [1:3556] "568" "380" "93" "142" ...
```

Observamos que muitas colunas numéricas foram lidas como do tipo caractere, mesmo a função tendo identificado corretamente que o arquivo usa virgula como separador decimal.

Apenas as colunas `MUNIC` e `UF` são de caracteres, as outras deveriam ser do tipo numéricas. Após analisar os dados, descobrimos que os dados faltantes estão codificados com um traço "-". Dessa forma, vamos ler novamente o arquivo e dizer ao R para tratar os traços como dados faltantes.

```
dados <- readxl::read_xls('data/cidades.xls', na='-')
str(dados, strict.width="wrap", width=80)

## tibble[,17] [3,556 x 17] (S3: tbl_df/tbl/data.frame)
## $ MUNIC : chr [1:3556] "SAO PAULO-SP" "RIO DE JANEIRO-RJ" "SALVADOR-BA" "BELO
##   HORIZONTE-MG" ...
## $ UF : chr [1:3556] "SP" "RJ" "BA" "MG" ...
## $ CÓDIGO : num [1:3556] 1001 1002 1003 1004 1005 ...
## $ POPTOT : num [1:3556] 10406166 5850544 2440886 2229697 2138234 ...
## $ CRES_POP: num [1:3556] 1.41 1.32 2.5 1.61 2.13 2.91 1.82 1.38 4.94 1.35 ...
## $ POPURB : num [1:3556] 9785640 5850544 2439881 2229697 2138234 ...
## $ PIBTOT : num [1:3556] 105907 47172 12029 18573 6478 ...
## $ CRES_PIB: num [1:3556] 1.75 1.32 3.08 2.85 1.49 ...
## $ GRAU1 : num [1:3556] 5322497 2731075 1139181 1107558 1004021 ...
## $ GRAU2 : num [1:3556] 1606381 1110059 465685 373858 317977 ...
```

```
## $ SUPERIOR: num [1:3556] 1076916 731746 148887 224303 112762 ...
## $ 11OUMAIS: num [1:3556] 2142313 1562602 485962 488029 340635 ...
## $ EMPREGAD: num [1:3556] 3986021 2041470 563139 990843 447896 ...
## $ MICROEMP: num [1:3556] 377600 133165 38922 75665 45504 ...
## $ PEQEMP : num [1:3556] 18494 9521 2494 4108 2152 ...
## $ MEDEMP : num [1:3556] 3198 1804 437 690 418 ...
## $ GRAENP : num [1:3556] 568 380 93 142 78 133 104 87 56 85 ...
```

Com isso os tipos de dados estão agora corretos.

b)

A seguir, apresentamos um resumo das variáveis dos dados como o comando `summary`:

```
summary(dados)
```

```
##      MUNIC              UF          CÓDIGO
## Length:3556      Length:3556      Min.   :1001
## Class :character  Class :character  1st Qu.:1889
## Mode  :character  Mode  :character  Median :3720
##                                         Mean  :3440
##                                         3rd Qu.:4609
##                                         Max.   :5497
##                                         NA's   :2
##      POPTOT          CRES_POP      POPURE
## Min.   :    795      Min.   : -13.330      Min.   :    423
## 1st Qu.:   7995      1st Qu.:  0.020      1st Qu.:   4389
## Median :  15650      Median :  1.150      Median :   9243
## Mean   :  134929      Mean   :  1.284      Mean   :  112497
## 3rd Qu.:  30724      3rd Qu.:  2.310      3rd Qu.:  20746
## Max.   :169544443      Max.   : 23.630      Max.   :137697439
##                                         NA's   :1
##      PIBTOT          CRES_PIB      GRAU1
## Min.   :    0.9      Min.   : 0.0000      Min.   :    469
## 1st Qu.:   13.5      1st Qu.: 0.6937      1st Qu.:   4744
## Median :   26.8      Median : 1.0373      Median :   8494
## Mean   :   536.9      Mean   : 1.1608      Mean   :   69987
## 3rd Qu.:   66.9      3rd Qu.: 1.4490      3rd Qu.:  16102
## Max.   :641969.4      Max.   :24.6598      Max.   :86262616
## NA's   :14          NA's   :14          NA's   :13
##      GRAU2          SUPERIOR      11OUMAIS
## Min.   :    47      Min.   :    0      Min.   :    37
## 1st Qu.:   495      1st Qu.:   75      1st Qu.:   407
## Median :   950      Median :   178      Median :   786
## Mean   :  15293      Mean   :   6202      Mean   :  16302
## 3rd Qu.:  2282      3rd Qu.:   522      3rd Qu.:  1969
## Max.   :18351427      Max.   :7358947      Max.   :19465358
## NA's   :13          NA's   :13          NA's   :13
##      EMPREGAD      MICROEMP      PEQEMP
## Min.   :    10      Min.   :    3      Min.   :    0.0
## 1st Qu.:   414      1st Qu.:   94      1st Qu.:    1.0
```



```

## Median :    927   Median :    207   Median :     3.0
## Mean   :   23433   Mean    :   2769   Mean    :   110.5
## 3rd Qu.:    2754   3rd Qu.:    504   3rd Qu.:    13.0
## Max.   :27933651   Max.    :3315552   Max.    :131536.0
## NA's   :14        NA's    :14        NA's    :14
##      MEDEMP      GRAENP
## Min.    :    0   Min.    :  0.000
## 1st Qu.:    1   1st Qu.:  0.000
## Median :    1   Median :  0.000
## Mean    :   21   Mean    :  4.031
## 3rd Qu.:    2   3rd Qu.:  1.000
## Max.    :25310   Max.    :4787.000
## NA's    :14     NA's    :14

```

c)

Observando o resumo do item anterior, notamos que as variáveis **MUNIC** e **UF** são alfanuméricas enquanto que as demais são numéricas. A seguir indicamos o número de observações omissas em cada variável:

Tabela 7: Classificação e Observações Omissas

Variável	Tipo	Obs. Omissas
MUNIC	Alfanumérica	0
UF	Alfanumérica	2
CÓDIGO	Numérica	2
POPTOT	Numérica	0
CRES_POP	Numérica	1
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
GRAENP	Numérica	14