

PROGRAMAÇÃO

3. Sumarização de Dados

Rita P. Ribeiro

Mestrado em Informática Médica

Curso de Especialização em Informática da Saúde

2016/2017



FACULDADE DE MEDICINA
UNIVERSIDADE DO PORTO



FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Programa

1. Conceitos Básicos do R
2. Manipulação de Dados
3. Sumarização de Dados
4. Visualização de Dados
5. Geração de Relatórios

Motivação

- ▶ Com conjuntos de dados minimamente grandes é, por vezes, difícil ter uma ideia do que "se está a passar nos dados"
- ▶ O objetivo da sumarização de dados é descrever as propriedades chave de um conjunto de dados
- ▶ Concretamente, descrever **propriedades importantes da distribuição dos valores** das variáveis do conjunto de dados

Caracterização dos Dados

▶ Conjunto de Dados

- ▶ **Univariados**: descritos por uma única variável
- ▶ **Multivariados**: descritos por mais do que uma variável

▶ Variáveis

- ▶ **Discretas** / Qualitativas: expressam valores por categorias (ordenadas ou não ordenadas)
- ▶ **Contínuas** / Quantitativas: expressam valores numa escala numérica

Tipos de Sumários

- ▶ Qual o valor mais comum de uma variável?
- ▶ Qual a variabilidade dos valores de uma variável?
- ▶ Existem valores "estranhos"/inesperados no conjunto de dados?
 - ▶ Outliers
 - ▶ Valores desconhecidos

Caracterização de Dados Univariados

Variáveis Discretas

- ▶ As variáveis discretas podem ser descritas através:
 - ▶ moda
 - ▶ tabelas de frequência

Caracterização de Dados Univariados

Variáveis Discretas - Moda e Tabelas de Frequência

- A **moda** é o valor dos dados que ocorre com maior frequência.

```
data(Melanoma, package="MASS")  
library(DMwR)  
centralValue(Melanoma$status)  
## [1] 2
```

- As **tabelas de frequência** apresentam a forma como os dados se distribuem pelas diferentes categorias.

```
#1 died from melanoma, 2 alive, 3 dead from other causes.  
table(Melanoma$status)  
  
##  
##    1    2    3  
##  57 134  14
```

Caracterização de Dados Univariados

Variáveis Contínuas

- ▶ As variáveis contínuas podem ser descritas por:
 - ▶ estatísticas de centralidade
 - ▶ estatísticas de dispersão
 - ▶ tabelas de frequência

Caracterização de Dados Univariados

Variáveis Contínuas - Estatísticas de Centralidade

► A média

$$\mu_x = \frac{1}{n} \sum_{i=1}^n x_i$$

► A mediana

- Valor acima (abaixo) do qual estão 50% dos valores no conjunto de dados
- Obtida ordenando os valores observados e escolhendo o valor na posição central

► A moda

- É o valor mais frequente no conjunto de dados
 - De notar que esta estatística pode ser aplicada a variáveis discretas

Caracterização de Dados Univariados

Variáveis Contínuas - Estatísticas de Centralidade (2)

► Ilustrações em R

```
mean(Melanoma$time)
```

```
## [1] 2152.8
```

```
median(Melanoma$age)
```

```
## [1] 54
```

```
library(DMwR)
```

```
centralValue(Melanoma$year)
```

```
## [1] 1970
```

```
library(dplyr)
```

```
m <- tbl_df(Melanoma)
```

```
m %>% summarise(avg.time = mean(time),  
                 med.age = median(age),  
                 cen.year = centralValue(year))
```

```
## Source: local data frame [1 x 3]
```

```
##
```

```
##   avg.time med.age cen.year
```

```
## 1   2152.8     54   1970
```

Caracterização de Dados Univariados

Variáveis Contínuas - Estatísticas de Dispersão

- ▶ A **variância**

$$\sigma_x^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \mu_x)^2$$

- ▶ O **desvio padrão**

$$\sigma_x = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \mu_x)^2}$$

- ▶ O **intervalo inter-quartil**: calculado como a diferença entre o 3º e 1º quartis
 - ▶ O 1º quartil é o nº abaixo do qual estão 25% dos valores
 - ▶ O 3º quartil é o valor acima do qual estão 25% dos valores
- ▶ O **range**: a diferença entre os valores máximo e mínimo

Caracterização de Dados Univariados

Variáveis Contínuas - Estatísticas de Dispersão (2)

► Ilustrações em R

```
var(Melanoma$thickness)

## [1] 8.758242

sd(Melanoma$time)

## [1] 1122.061

IQR(Melanoma$age)

## [1] 23

quantile(Melanoma$age)

##    0%   25%   50%   75%  100%
##    4    42    54    65    95
```

```
quantile(Melanoma$age, probs=c(0.1,0.9))

##   10%   90%
##  29.4  72.0

fivenum(Melanoma$age)

## [1]   4  42  54  65  95

range(Melanoma$age)

## [1]   4  95
```

Caracterização de Dados Univariados

Variáveis Contínuas - Estatísticas de Dispersão (3)

► Ilustrações em R com dplyr

```
library(dplyr)
m <- tbl_df(Melanoma)
m %>% summarise(var.thick = var(thickness),
                 sd.time = sd(time),
                 iqr.age = IQR(age))

## Source: local data frame [1 x 3]
##
##   var.thick  sd.time iqr.age
## 1   8.758242 1122.061     23
```

Caracterização de Dados Univariados

Variáveis Contínuas - Outliers

- Informalmente, um **outlier** é um valor que se desvia tanto dos outros ao ponto de gerar suspeitas de que foi gerado por um mecanismo diferente

- Uma definição usada para outliers são valores fora do intervalo,

$$[Q_1 - 1.5 \times IQR, Q_3 + 1.5 \times IQR]$$

em que Q_1 (Q_3) é o 1º (3º) quartil e IQR é o intervalo inter-quartil

```
boxplot.stats(Melanoma$time)

## $stats
## [1]    10 1525 2005 3042 4926
##
## $n
## [1] 205
##
## $conf
## [1] 1837.596 2172.404
##
## $out
## [1] 5565
```

Caracterização de Dados Univariados

Variáveis Contínuas - Tabelas de Frequência

- ▶ As tabelas de frequência apresentam como os valores da variável contínua se distribuem por um conjunto de diferentes "classes" de valores.

```
table(Melanoma$age)
```

```
##
##  4 12 14 15 16 19 20 21 22 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39
##  1  1  1  1  1  2  2  1  1  1  2  2  1  1  3  1  3  2  1  3  4  3  2  1  1
## 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64
##  4  3  6  4  4  3  6  4  2  6  3  1  6  2  8  6  8  3  8  1  8  4  3  3  4
## 65 66 67 68 69 70 71 72 73 74 75 76 77 78 80 83 84 86 89 95
##  5  3  4  7  5  2  3  5  1  2  2  2  4  1  1  2  1  1  1  1
```

```
table(cut(Melanoma$age,seq(0,100,by=20)))
```

```
##
##  (0,20]  (20,40]  (40,60]  (60,80]  (80,100]
##         9        37        92        61         6
```

Caracterização de Dados Multivariados

Sumários

- Sumário global das propriedades estatísticas básicas de um conjunto de dados.

```
summary(Melanoma)
```

```
##           time           status sex           age           year
## Min.      : 10    died       : 57 M:126    Min.      : 4.00    Min.      :1962
## 1st Qu.:1525    alive       :134 F: 79    1st Qu.:42.00    1st Qu.:1968
## Median :2005    deadOther: 14                Median :54.00    Median :1970
## Mean     :2153                                Mean     :52.46    Mean     :1970
## 3rd Qu.:3042                                3rd Qu.:65.00    3rd Qu.:1972
## Max.     :5565                                Max.     :95.00    Max.     :1977
##
## thickness           ulcer
## Min.      : 0.10    absence :115
## 1st Qu.: 0.97    presence: 90
## Median   : 1.94
## Mean     : 2.92
## 3rd Qu.: 3.56
## Max.     :17.42
```

Caracterização de Dados Multivariados

Sumários (2)

```
library(Hmisc); describe(Melanoma)
```

```
## Melanoma[, 1:2]
##
##  2  Variables      205  Observations
## -----
## time
##      n missing  unique    Mean    .05    .10    .25    .50    .75
##    205      0    194    2153   307.0   687.4  1525.0  2005.0  3042.0
##    .90    .95
##  3695.0  4115.8
##
## lowest :   10   30   35   99  185, highest: 4492 4668 4688 4926 5565
## -----
## status
##      n missing  unique
##    205      0      3
##
## died (57, 28%), alive (134, 65%), deadOther (14, 7%)
## -----
```

...

Caracterização de Dados Multivariados

Sumários Condicionais

```
by(Melanoma$time,Melanoma$sex,summary)
```

```
## Melanoma$sex: M
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       99   1636   2059   2283   3131   5565
```

```
## -----
```

```
## Melanoma$sex: F
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       10   1052   1860   1946   2784   4492
```

```
Melanoma %>% group_by(sex,status) %>%
```

```
  summarise(thick.avg = mean(thickness),thick.sd = sd(thickness))
```

```
## Source: local data frame [6 x 4]
```

```
## Groups: sex
```

```
##
```

```
##   sex      status thick.avg thick.sd
## 1  M        died  3.985357  3.710697
## 2  M        alive  2.016374  2.223752
## 3  M deadOther  2.601429  2.844664
## 4  F        died  4.625517  3.472504
## 5  F        alive  2.727907  2.487260
## 6  F deadOther  4.834286  4.192254
```

Caracterização de Dados Multivariados

Tabelas de Contingência

- ▶ A forma mais comum de descrever dados expressos por um conjunto de uma variáveis discretas é através de tabelas de contingência ou informação cruzada.
- ▶ As **tabelas de contingência** apresentam a frequência dos valores de uma variável em função das categorias de outra variável.

```
fable(Melanoma$sex,
      Melanoma$status)
```

```
##      died alive deadOther
##
## M      28     91         7
## F      29     43         7
```

```
fable(Melanoma$sex,Melanoma$sulcer,
      Melanoma$status)
```

```
##              died alive deadOther
##
## M absence         8     68         3
## presence        20     23         4
## F absence         8     24         4
## presence        21     19         3
```