

Folha 3: Sumarização de Dados - Soluções

1. Carregue o conjunto de dados Aids2 da package MASS. Responda às seguintes questões.

```
data(Aids2, package="MASS")
```

- (a) Qual a categoria de transmissão do vírus mais frequente nestes dados?
E a categoria menos frequente?
(Sugestão: use as funções `table()`, `names()`, `which.max()` e `which.min()`).

```
ts <- table(Aids2$T.categ)
names(ts[which.max(ts)])

## [1] "hs"

names(ts[which.min(ts)])

## [1] "mother"
```

- (b) Obtenha a tabela de contingência em função do sexo e categoria de transmissão do vírus.

```
table(Aids2$sex, Aids2$T.categ)

##
##      hs hsid  id  het haem blood mother other
## F    1   0   20   20   0    37      4      7
## M 2464  72   28   21   46   57      3     63
```

- (c) Obtenha um sumário das propriedades estatísticas básicas da idade em função da categoria de transmissão do vírus.

```
by(Aids2$age, Aids2$T.categ, summary)

## Aids2$T.categ: hs
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  17.00  31.00  37.00  37.52  43.00  82.00
## -----
## Aids2$T.categ: hsid
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  20.00  26.00  30.00  30.49  35.00  44.00
## -----
## Aids2$T.categ: id
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  18.00  27.00  30.00  31.06  34.00  51.00
## -----
## Aids2$T.categ: het
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

```
##      21.00    30.00    37.00    38.93    47.00    74.00
## -----
## Aids2$T.categ: haem
##      Min. 1st Qu.  Median      Mean 3rd Qu.     Max.
##      9.00   16.50   29.00   31.87   44.75   66.00
## -----
## Aids2$T.categ: blood
##      Min. 1st Qu.  Median      Mean 3rd Qu.     Max.
##      0.00   31.25   46.50   44.40   59.00   80.00
## -----
## Aids2$T.categ: mother
##      Min. 1st Qu.  Median      Mean 3rd Qu.     Max.
##      0.000    1.000    3.000    3.143    5.500    6.000
## -----
## Aids2$T.categ: other
##      Min. 1st Qu.  Median      Mean 3rd Qu.     Max.
##      23.00   31.00   39.00   41.83   49.75   78.00
```

- (d) Obtenha uma tabela com a categoria de transmissão do vírus ordenada por ordem decrescente da idade média.

```
library(dplyr)
Aids2 %>% group_by(T.categ) %>%
  summarise(idade.media=mean(age)) %>% arrange(desc(idade.media))

## Source: local data frame [8 x 2]
##
##   T.categ idade.media
## 1   blood    44.404255
## 2   other    41.828571
## 3    het     38.926829
## 4    hs      37.518053
## 5   haem     31.869565
## 6    id      31.062500
## 7   hsid     30.486111
## 8 mother     3.142857
```

- (e) Obtenha o número de indivíduos com idades entre 0-25, 25-50, 50-75 e 75-100. Verifique como eles se distribuem pelas diferentes categorias de transmissão do vírus.

```
table(cut(Aids2$age,c(0,25,50,75,100)))

##
##   (0,25] (25,50] (50,75] (75,100]
##      231    2335     268         5

table(cut(Aids2$age,c(0,25,50,75,100)),Aids2$T.categ)

##
##           hs hsid   id  het haem blood mother other
## (0,25]    155   16    8   5  20   16      6      5
## (25,50]   2113   56   39  30  17   32      0     48
## (50,75]    194    0    1   6   9   42      0     16
## (75,100]     3    0    0   0   0    1      0      1
```

- (f) Assuma que uma frequência inferior ou igual 1.5% de uma categoria numa variável discreta, representa um *outlier* nessa variável. Identifique as categorias de transmissão do vírus que podem ser consideradas *outliers*.

```

tcateg.out <- names(ts)[ts/nrow(Aids2) <= 0.015]
tcateg.out

## [1] "het"      "mother"

```

- (g) Liste o conjunto de exemplos cuja categoria de transmissão está nas condições anteriores. (Sugestão: use o operador %in%).

```

filter(Aids2, T.categ %in% tcateg.out)

##      state sex  diag death status T.categ age
## 1    NSW   M  10738 11504      A    het   26
## 2    NSW   M  11195 11504      A    het   39
## 3    NSW   M  11289 11504      A mother    1
## 4    NSW   F   9938 10780      D    het   35
## 5    NSW   M  10416 11504      A    het   31
## 6    NSW   M  10431 10676      D    het   74
## 7    NSW   F  10584 11238      D mother    5
## 8    NSW   F  10649 10705      D    het   63
## 9    NSW   M  10691 11504      A    het   42
## 10   NSW   M  10719 11472      D    het   49
## 11   NSW   M  10767 11504      A mother    3
## 12   NSW   F  10237 11087      D    het   50
## 13   NSW   M  11003 11504      A    het   34
## 14   NSW   M  10937 11133      D    het   35
## 15   NSW   M  11125 11441      D    het   58
## 16   NSW   F  11101 11504      A    het   33
## 17   NSW   F  10503 11197      D    het   27
## 18   NSW   F  11137 11504      A    het   65
## 19   NSW   M  11200 11504      A    het   42
## 20   NSW   M  11223 11504      A    het   29
## 21   NSW   M  11397 11504      A    het   37
## 22   QLD   F   9905 10801      D    het   25
## 23   QLD   M  10354 10583      D    het   37
## 24   QLD   F  11306 11504      A mother    6
## 25   QLD   M  11070 11504      A    het   25
## 26   QLD   F  10400 11485      D    het   47
## 27   QLD   M   9862 10777      D    het   33
## 28 Other   F   9447  9447      D mother    0
## 29 Other   M  11061 11504      A mother    6
## 30 Other   F  10966 11504      A    het   35
## 31 Other   M  11274 11504      A    het   40
## 32 Other   M  11376 11504      A    het   30
## 33 Other   F  10068 11005      D    het   21
## 34 VIC    F  10248 10298      D    het   50
## 35 VIC    F  10177 11085      D    het   24
## 36 VIC    F  10371 11504      A    het   25
## 37 VIC    F  10378 10640      D    het   49
## 38 VIC    M  10548 10823      D    het   38
## 39 VIC    M  10069 11504      A    het   37
## 40 VIC    F  10884 11504      A    het   30
## 41 VIC    F  11181 11504      A    het   54
## 42 VIC    M  11199 11504      A    het   52
## 43 VIC    F  11404 11504      A    het   26
## 44 VIC    F  11312 11327      D mother    1
## 45 Other   F   9294 11504      A    het   29
## 46 Other   F  10758 11504      A    het   40
## 47 Other   M  11475 11504      A    het   46
## 48 Other   F  11420 11504      A    het   34

```

- (h) Inspeccione se existem *outliers* extremos, isto é, valores anormalmente altos e/ou baixos em cada uma das variáveis contínuas.

```
boxplot.stats(Aids2$diag)

## $stats
## [1] 8760 10163 10665 11103 11503
##
## $n
## [1] 2843
##
## $conf
## [1] 10637.15 10692.85
##
## $out
## [1] 8302 8711 8726 8499 8582 8720

boxplot.stats(Aids2$death)

## $stats
## [1] 9424.0 10671.5 11235.0 11504.0 11504.0
##
## $n
## [1] 2843
##
## $conf
## [1] 11210.33 11259.67
##
## $out
## [1] 8469 8850 9254 8959 8879 9180 9175 9172 9109 9218 9254 9314 9350 9240
## [15] 9309 9262 9126 9268 9226 9207 9285 9152 9418 9082 9222 9379 9290 9392
## [29] 9250 9300 9259 9214 9315 9392 9318 9304 9380 9355 9388 9362 8945 9274
## [43] 9301 9039 8979 9215 9176 8815 9239 9224 9308 9395 9381 9163 8568 8630
## [57] 9045 8992 9084 9026 9075 9394 9140 9405 9310 9215 9343 9232

boxplot.stats(Aids2$age)

## $stats
## [1] 11 30 37 43 62
##
## $n
## [1] 2843
##
## $conf
## [1] 36.61478 37.38522
##
## $out
## [1] 1 70 69 64 82 71 66 69 78 67 64 64 3 69 70 63 65 65 5 68 67 66 65
## [24] 74 5 63 67 64 5 3 66 65 69 69 67 66 66 66 78 65 64 64 63 65 72 77
## [47] 73 80 69 73 63 0 0 1 0 66 3 73 70 66 6 63 63 0 6 9 66 66 69
## [70] 1 64 66
```

- (i) Faça uma função que, dado um qualquer data frame como argumento, imprime para cada atributo o seu nome e o número de exemplos com valores considerados *outliers* nesse atributo.

```

conta.out <- function(df) {

  for(i in 1:ncol(df)) {

    if(is.numeric(df[,i])) {
      out <- boxplot.stats(df[,i])$out
    } else {
      ts <- table(df[,i])
      out <- names(ts)[ts/nrow(df) <= 0.015]
    }
    print(paste("Numero de outliers do atributo",names(df)[i],": ",length(out)))
  }
}

conta.out(Aids2)

## [1] "Numero de outliers do atributo state : 0"
## [1] "Numero de outliers do atributo sex : 0"
## [1] "Numero de outliers do atributo diag : 6"
## [1] "Numero de outliers do atributo death : 68"
## [1] "Numero de outliers do atributo status : 0"
## [1] "Numero de outliers do atributo T.categ : 2"
## [1] "Numero de outliers do atributo age : 72"

```