Estatística Descritiva

Prof. Leticia Raposo

2019.2

Contents

1 Processamento dos dados

1.1 Lendo os dados

```
setwd("C:/Users/Leticia/Google Drive/UNIRIO/Disciplinas Ministradas/2019.2/Biologia - Biomedicina")
Titanic <- read.table("Titanic.txt", header = T)
```

1.2 Vendo as primeiras/últimas linhas do banco e suas dimensões

```
head(Titanic) #ver as primeiras linhas do banco
```

```
##   PassengerId Survived Pclass
## 1           1        0      3
## 2           2        1      1
## 3           3        1      3
## 4           4        1      1
## 5           5        0      3
## 6           6        0      3
##                                                  Name    Sex Age SibSp
## 1                             Braund, Mr. Owen Harris   male  22     1
## 2 Cumings, Mrs. John Bradley (Florence Briggs Thayer) female  38     1
## 3                              Heikkinen, Miss. Laina female  26     0
## 4        Futrelle, Mrs. Jacques Heath (Lily May Peel) female  35     1
## 5                            Allen, Mr. William Henry   male  35     0
## 6                                    Moran, Mr. James   male  NA     0
##   Parch           Ticket    Fare Cabin Embarked
## 1     0        A/5 21171  7.2500              S
```

```
## 2     0         PC 17599 71.2833   C85        C
## 3     0 STON/O2. 3101282  7.9250              S
## 4     0           113803 53.1000  C123        S
## 5     0           373450  8.0500              S
## 6     0           330877  8.4583              Q
```

```
tail(Titanic) #ver as últimas linhas do banco
```

```
##     PassengerId Survived Pclass                                      Name
## 886         886        0      3     Rice, Mrs. William (Margaret Norton)
## 887         887        0      2                     Montvila, Rev. Juozas
## 888         888        1      1              Graham, Miss. Margaret Edith
## 889         889        0      3 Johnston, Miss. Catherine Helen "Carrie"
## 890         890        1      1                     Behr, Mr. Karl Howell
## 891         891        0      3                       Dooley, Mr. Patrick
##        Sex Age SibSp Parch     Ticket   Fare Cabin Embarked
## 886 female  39     0     5     382652 29.125              Q
## 887   male  27     0     0     211536 13.000              S
## 888 female  19     0     0     112053 30.000   B42        S
## 889 female  NA     1     2 W./C. 6607 23.450              S
## 890   male  26     0     0     111369 30.000  C148        C
## 891   male  32     0     0     370376  7.750              Q
```

```
dim(Titanic) #ver as dimensões do banco
```

```
## [1] 891  12
```

1.3 Vendo um resumo dos dados

```
#ver a estrutura dos dados -
# bom para ver se as variáveis foram lidas corretamente
str(Titanic)
```

```
## 'data.frame':    891 obs. of  12 variables:
##  $ PassengerId: int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Survived   : int  0 1 1 1 0 0 0 0 1 1 ...
##  $ Pclass     : int  3 1 3 1 3 3 1 3 3 2 ...
##  $ Name       : Factor w/ 891 levels "Abbing, Mr. Anthony",..: 109 191 358 277 16 559 520 629 417 581
##  $ Sex        : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 2 2 1 1 ...
##  $ Age        : num  22 38 26 35 35 NA 54 2 27 14 ...
##  $ SibSp      : int  1 1 0 1 0 0 0 3 0 1 ...
##  $ Parch      : int  0 0 0 0 0 0 0 1 2 0 ...
##  $ Ticket     : Factor w/ 681 levels "110152","110413",..: 524 597 670 50 473 276 86 396 345 133 ...
##  $ Fare       : num  7.25 71.28 7.92 53.1 8.05 ...
##  $ Cabin      : Factor w/ 148 levels "","A10","A14",..: 1 83 1 57 1 1 131 1 1 1 ...
##  $ Embarked   : Factor w/ 4 levels "","C","Q","S": 4 2 4 4 4 3 4 4 4 2 ...
```

```
summary(Titanic) #mostra um resumo
```

```
##   PassengerId       Survived          Pclass
##  Min.   :  1.0   Min.   :0.0000   Min.   :1.000
##  1st Qu.:223.5   1st Qu.:0.0000   1st Qu.:2.000
##  Median :446.0   Median :0.0000   Median :3.000
##  Mean   :446.0   Mean   :0.3838   Mean   :2.309
##  3rd Qu.:668.5   3rd Qu.:1.0000   3rd Qu.:3.000
##  Max.   :891.0   Max.   :1.0000   Max.   :3.000
```

```
##
##                                     Name         Sex           Age
##  Abbing, Mr. Anthony                   :  1   female:314   Min.   : 0.42
##  Abbott, Mr. Rossmore Edward           :  1   male  :577   1st Qu.:20.12
##  Abbott, Mrs. Stanton (Rosa Hunt)      :  1                Median :28.00
##  Abelson, Mr. Samuel                   :  1                Mean   :29.70
##  Abelson, Mrs. Samuel (Hannah Wizosky):  1                3rd Qu.:38.00
##  Adahl, Mr. Mauritz Nils Martin        :  1                Max.   :80.00
##  (Other)                               :885                NA's   :177
##      SibSp           Parch             Ticket         Fare
##  Min.   :0.000   Min.   :0.0000   1601    :  7   Min.   :  0.00
##  1st Qu.:0.000   1st Qu.:0.0000   347082  :  7   1st Qu.:  7.91
##  Median :0.000   Median :0.0000   CA. 2343:  7   Median : 14.45
##  Mean   :0.523   Mean   :0.3816   3101295 :  6   Mean   : 32.20
##  3rd Qu.:1.000   3rd Qu.:0.0000   347088  :  6   3rd Qu.: 31.00
##  Max.   :8.000   Max.   :6.0000   CA 2144 :  6   Max.   :512.33
##                                   (Other) :852
##          Cabin     Embarked
##             :687    :  2
##  B96 B98    :  4   C:168
##  C23 C25 C27:  4   Q: 77
##  G6         :  4   S:644
##  C22 C26    :  3
##  D          :  3
##  (Other)    :186
```

```
# install.packages("summarytools")
library(summarytools)
# view(dfSummary(iris)) #para ver em uma janela a parte
dfSummary(tobacco)
```

```
## Data Frame Summary
## tobacco
## Dimensions: 1000 x 9
## Duplicates: 2
##
## ----------------------------------------------------------------------------------------------------
## No   Variable        Stats / Values             Freqs (% of Valid)    Graph
## ---- --------------- -------------------------- --------------------- ---------------------------
## 1    gender          1. F                       489 (50.0%)           IIIIIIIIII
##      [factor]        2. M                       489 (50.0%)           IIIIIIIIII
##
## 2    age             Mean (sd) : 49.6 (18.3)    63 distinct values    .     .     . . . :
##      [numeric]       min < med < max:                                 : : : : : . : : : :
##                      18 < 50 < 80                                     : : : : : : : : : :
##                      IQR (CV) : 32 (0.4)                              : : : : : : : : : :
##                                                                       : : : : : : : : : :
##
## 3    age.gr          1. 18-34                   258 (26.5%)           IIIII
##      [factor]        2. 35-50                   241 (24.7%)           IIII
##                      3. 51-70                   317 (32.5%)           IIIIII
##                      4. 71 +                    159 (16.3%)           III
##
## 4    BMI             Mean (sd) : 25.7 (4.5)     974 distinct values             :
##      [numeric]       min < med < max:                                         : : :
```

```
##                     8.8 < 25.6 < 39.4                                          : : :
##                     IQR (CV) : 5.7 (0.2)                                     : : : : :
##                                                                            . : : : : : .
##
## 5    smoker         1. Yes                   298 (29.8%)          IIIII
##      [factor]       2. No                    702 (70.2%)          IIIIIIIIIIIII
##
## 6    cigs.per.day   Mean (sd) : 6.8 (11.9)   37 distinct values   :
##      [numeric]      min < med < max:                              :
##                     0 < 0 < 40                                    :
##                     IQR (CV) : 11 (1.8)                           :
##                                                                   :   . . . . . .
##
## 7    diseased       1. Yes                   224 (22.4%)          IIII
##      [factor]       2. No                    776 (77.6%)          IIIIIIIIIIIIII
##
## 8    disease        1. Hypertension          36 (16.2%)           III
##      [character]    2. Cancer                34 (15.3%)           III
##                     3. Cholesterol           21 ( 9.5%)           I
##                     4. Heart                 20 ( 9.0%)           I
##                     5. Pulmonary             20 ( 9.0%)           I
##                     6. Musculoskeletal       19 ( 8.6%)           I
##                     7. Diabetes              14 ( 6.3%)           I
##                     8. Hearing               14 ( 6.3%)           I
##                     9. Digestive             12 ( 5.4%)           I
##                     10. Hypotension          11 ( 5.0%)
##                     [ 3 others ]             21 ( 9.5%)           I
##
## 9    samp.wgts      Mean (sd) : 1 (0.1)      0.86!: 267 (26.7%)   IIIII
##      [numeric]      min < med < max:         1.04!: 249 (24.9%)   IIII
##                     0.9 < 1 < 1.1            1.05!: 324 (32.4%)   IIIIII
##                     IQR (CV) : 0.2 (0.1)     1.06!: 160 (16.0%)   III
##                                              ! rounded
## ------------------------------------------------------------------------------------------------------------
```

1.4 Removendo variáveis

```r
#colocando os nomes dos passageiros como
# nome das linhas
rownames(Titanic) <- Titanic$Name
Titanic$Name <- NULL
Titanic$PassengerId <- NULL
Titanic$Ticket <- NULL
Titanic$Cabin <- NULL
```

1.5 Codificando corretamente as variáveis

```r
Titanic$Survived <- as.factor(Titanic$Survived) #de numérica para fator
Titanic$Pclass <- as.factor(Titanic$Pclass) #de numérica para fator
Titanic$Sex <- as.factor(Titanic$Sex) #de caracter para fator
Titanic$Embarked <- as.factor(Titanic$Embarked) #de caracter para fator
str(Titanic) #revendo se elas ficaram corretas
```

```
## 'data.frame':    891 obs. of  8 variables:
##  $ Survived: Factor w/ 2 levels "0","1": 1 2 2 2 1 1 1 1 2 2 ...
##  $ Pclass  : Factor w/ 3 levels "1","2","3": 3 1 3 1 3 3 1 3 3 2 ...
##  $ Sex     : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 2 2 1 1 ...
##  $ Age     : num  22 38 26 35 35 NA 54 2 27 14 ...
##  $ SibSp   : int  1 1 0 1 0 0 0 3 0 1 ...
##  $ Parch   : int  0 0 0 0 0 0 0 1 2 0 ...
##  $ Fare    : num  7.25 71.28 7.92 53.1 8.05 ...
##  $ Embarked: Factor w/ 4 levels "","C","Q","S": 4 2 4 4 4 3 4 4 4 2 ...
```

1.6 Removendo categorias

```
#removendo as linhas em que a variável Embarked possui ""
Titanic <- droplevels(Titanic[Titanic$Embarked != "",])
str(Titanic) #vendo se está tudo ok
```

```
## 'data.frame':    889 obs. of  8 variables:
##  $ Survived: Factor w/ 2 levels "0","1": 1 2 2 2 1 1 1 1 2 2 ...
##  $ Pclass  : Factor w/ 3 levels "1","2","3": 3 1 3 1 3 3 1 3 3 2 ...
##  $ Sex     : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 2 2 1 1 ...
##  $ Age     : num  22 38 26 35 35 NA 54 2 27 14 ...
##  $ SibSp   : int  1 1 0 1 0 0 0 3 0 1 ...
##  $ Parch   : int  0 0 0 0 0 0 0 1 2 0 ...
##  $ Fare    : num  7.25 71.28 7.92 53.1 8.05 ...
##  $ Embarked: Factor w/ 3 levels "C","Q","S": 3 1 3 3 3 2 3 3 3 1 ...
```

2 Estatística Descritiva Univariada

2.1 Variável Qualitativa

Vamos avaliar a variável **Sex** como exemplo.

2.1.1 Tabela de distribuição de frequências

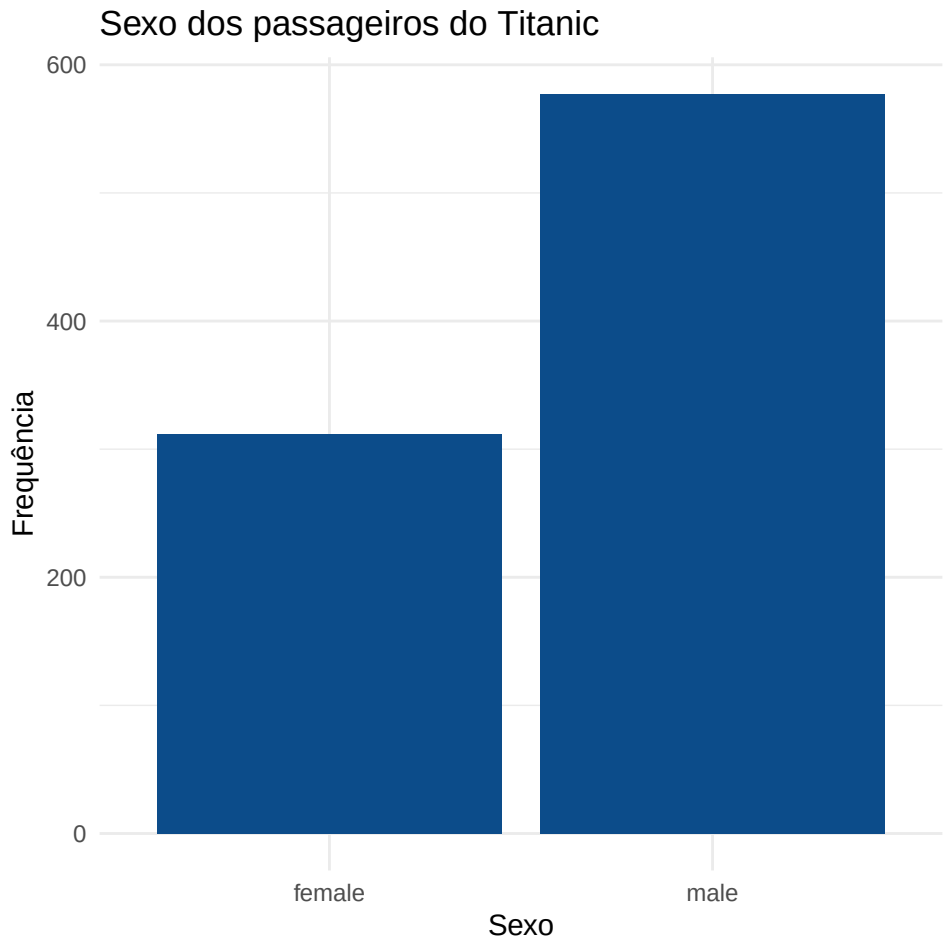
```
# freq(Titanic) #faria de todas as variáveis
freq(Titanic$Sex)
```

```
## Frequencies
## Titanic$Sex
## Type: Factor
##
##                Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
## ------------ ------ --------- -------------- --------- --------------
##       female    312     35.10          35.10     35.10          35.10
##         male    577     64.90         100.00     64.90         100.00
##         <NA>      0                               0.00         100.00
##        Total    889    100.00         100.00    100.00         100.00
```

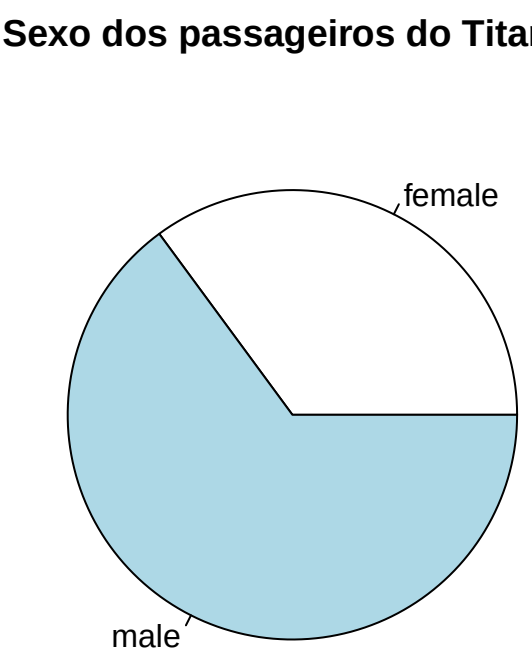
2.1.2 Gráficos

```
#Gráfico de barras

library(ggplot2)
```

```r
ggplot(Titanic) +
 aes(x = Sex) +
 geom_bar(fill = "#0c4c8a") +
 labs(x = "Sexo", y = "Frequência", title = "Sexo dos passageiros do Titanic") +
 theme_minimal()
```




```r
#Gráfico de setores
pie(table(Titanic$Sex), main = "Sexo dos passageiros do Titanic")
```



2.2 Variável Quantitativa

2.2.1 Medidas-resumo

Vamos avaliar a variável quantitativa **Age** como exemplo:

```
# descr(Titanic)
descr(Titanic$Age)
```

```
## Descriptive Statistics
## Titanic$Age
## N: 889
##
##                      Age
## ----------------- --------
##              Mean    29.64
##           Std.Dev    14.49
##               Min     0.42
##                Q1    20.00
##            Median    28.00
##                Q3    38.00
##               Max    80.00
##               MAD    12.97
##               IQR    18.00
```

```
##                CV     0.49
##          Skewness     0.39
##       SE.Skewness     0.09
##          Kurtosis     0.17
##           N.Valid   712.00
##         Pct.Valid    80.09
```

```r
# Mean: média
# Std.Dev: desvio-padrão
# Min: mínimo
# Q1: 1o quartil
# Median: mediana
# Q3: 3o quartil
# Max: máximo
# MAD: desvio médio absoluto
# IQR: intervalo interquartílico
# CV: coeficiente de variação (não está multiplicado por 100)
# Skewness: assimetria
# SE.Skewness: erro padrão da assimetria
# Kurtosis: curtose
# N.Valid: número de observações válidas
# Pct.Valid: porcentagem de observações válidas

# Moda
# install.packages("DescTools")
library(DescTools)
Mode(Titanic$Age, na.rm = T)
```

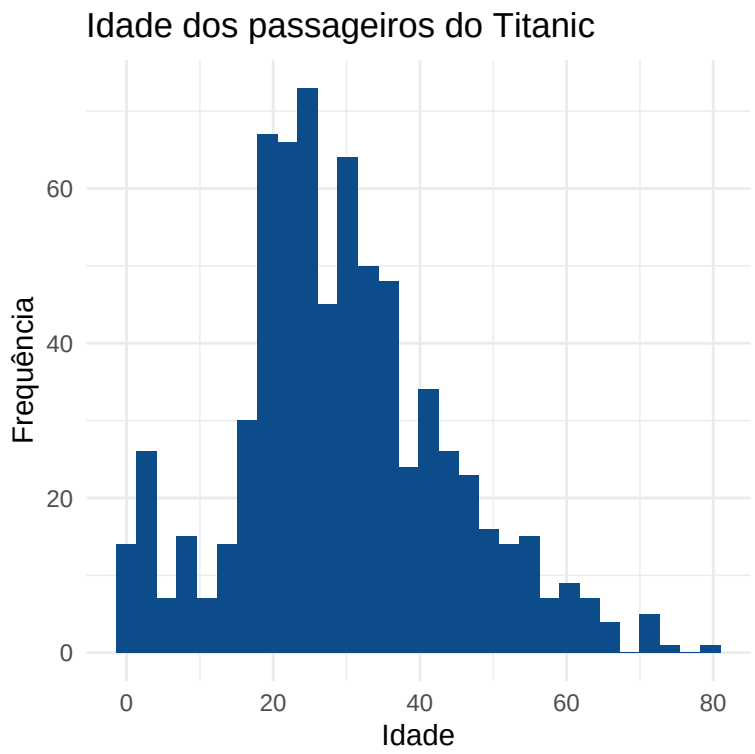
```
## [1] 24
```

```r
# Quantis
quantile(Titanic$Age, c(.25, .50, .75), na.rm = T)
```

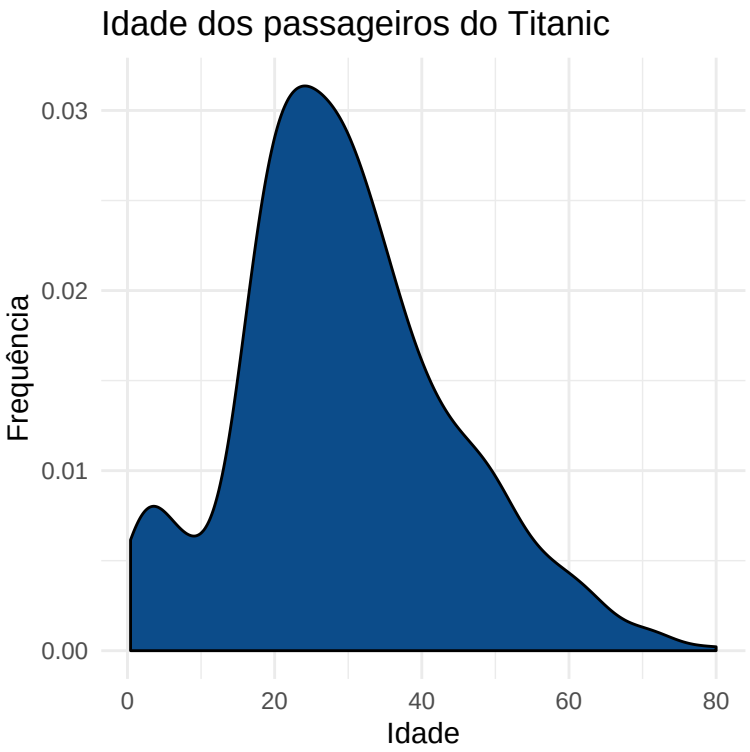
```
## 25% 50% 75%
##  20  28  38
```

2.2.2 Gráficos

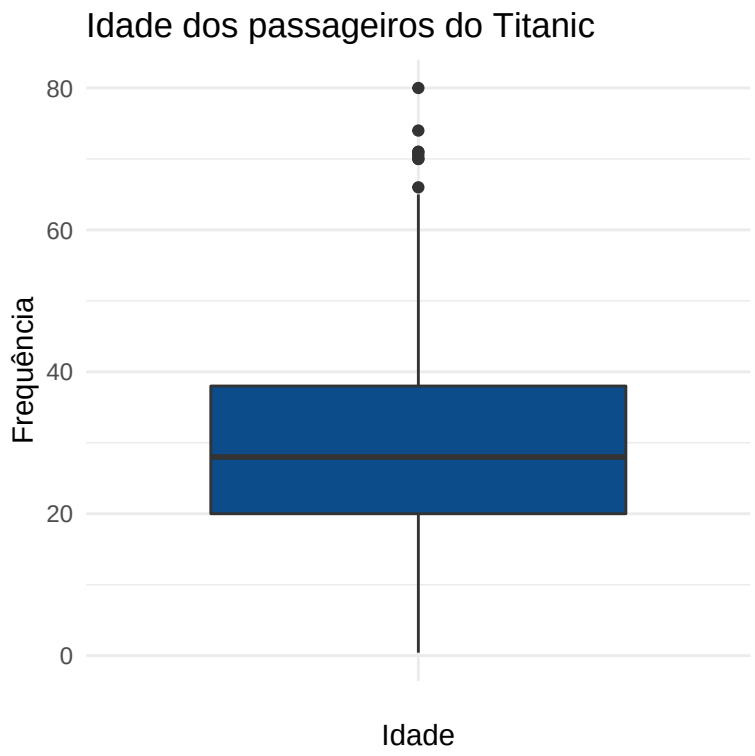
```r
library(ggplot2)

# Histograma
ggplot(Titanic) +
 aes(x = Age) +
 geom_histogram(bins = 30L, fill = "#0c4c8a") +
 labs(x = "Idade", y = "Frequência", title = "Idade dos passageiros do Titanic") +
 theme_minimal()
```



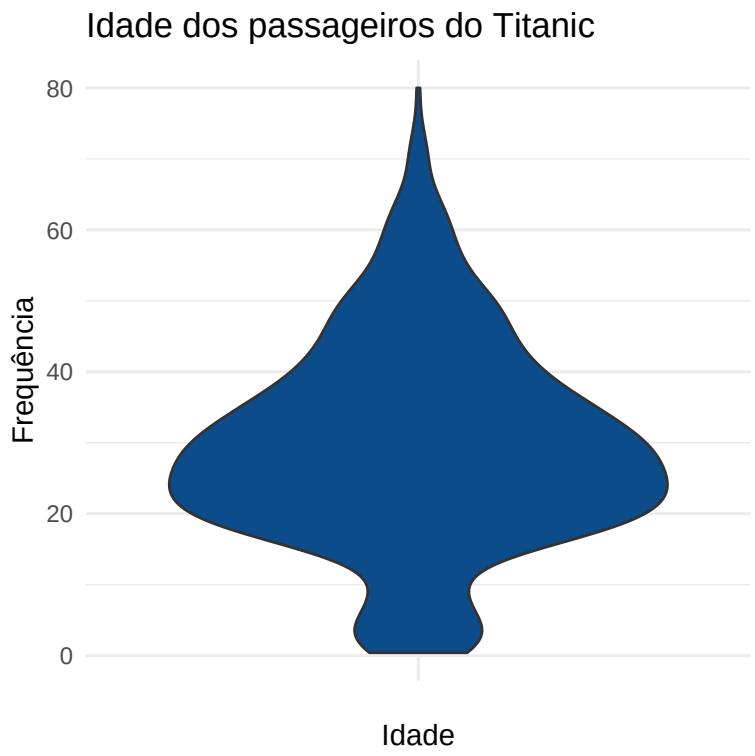

```
# Gráfico de densidades
ggplot(Titanic) +
 aes(x = Age) +
 geom_density(adjust = 1L, fill = "#0c4c8a") +
 labs(x = "Idade", y = "Frequência", title = "Idade dos passageiros do Titanic") +
 theme_minimal()
```




```
# Boxplot
ggplot(Titanic) +
 aes(x = "", y = Age) +
 geom_boxplot(fill = "#0c4c8a") +
 labs(x = "Idade", y = "Frequência", title = "Idade dos passageiros do Titanic") +
 theme_minimal()
```




```
# Gráfico de violino
ggplot(Titanic) +
 aes(x = "", y = Age) +
 geom_violin(adjust = 1L, scale = "area", fill = "#0c4c8a") +
 labs(x = "Idade", y = "Frequência", title = "Idade dos passageiros do Titanic") +
 theme_minimal()
```



3 Estatística Descritiva Bivariada

3.1 Variáveis Qualitativa x Qualitativa

Vamos avaliar as variáveis **Sex** e **Survived** como exemplo.

3.1.1 Tabela de distribuição de frequências

```
#total, com estatística qui-quadrado
ctable(Titanic$Sex, Titanic$Survived, prop = "t", chisq = T)
```

```
## Cross-Tabulation, Total Proportions
## Sex * Survived
## Data Frame: Titanic
##
##
## -------- ---------- ------------- ------------- --------------
##            Survived             0             1          Total
##      Sex
##   female               81 ( 9.1%)   231 (26.0%)   312 ( 35.1%)
##     male              468 (52.6%)   109 (12.3%)   577 ( 64.9%)
##    Total              549 (61.8%)   340 (38.2%)   889 (100.0%)
## -------- ---------- ------------- ------------- --------------
##
## ----------------------------
##  Chi.squared   df   p.value
## ------------- ---- ---------
##     258.4      1       0
## ----------------------------
```

```r
#perfil linha, com estatística qui-quadrado
ctable(Titanic$Sex, Titanic$Survived, prop = "r", chisq = T)
```

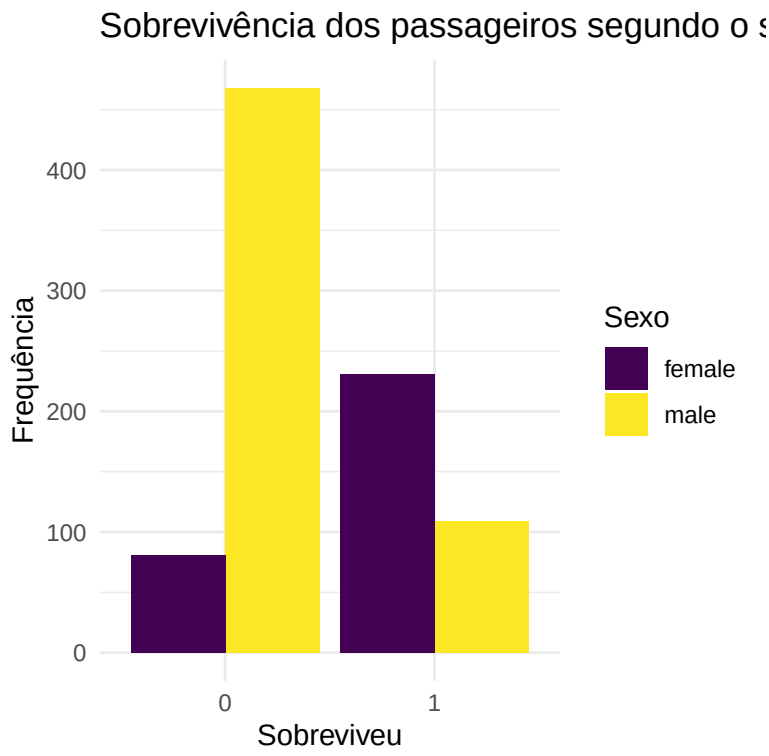
```
## Cross-Tabulation, Row Proportions
## Sex * Survived
## Data Frame: Titanic
##
##
## -------- ---------- ------------ ------------ --------------
##            Survived            0            1          Total
##      Sex
##   female               81 (26.0%)   231 (74.0%)   312 (100.0%)
##     male              468 (81.1%)   109 (18.9%)   577 (100.0%)
##    Total              549 (61.8%)   340 (38.2%)   889 (100.0%)
## -------- ---------- ------------ ------------ --------------
##
## ----------------------------
##  Chi.squared   df   p.value
## ------------- ---- ---------
##     258.4      1       0
## ----------------------------
```

```r
#perfil coluna, sem estatística qui-quadrado
ctable(Titanic$Sex, Titanic$Survived, prop = "c", chisq = F)
```

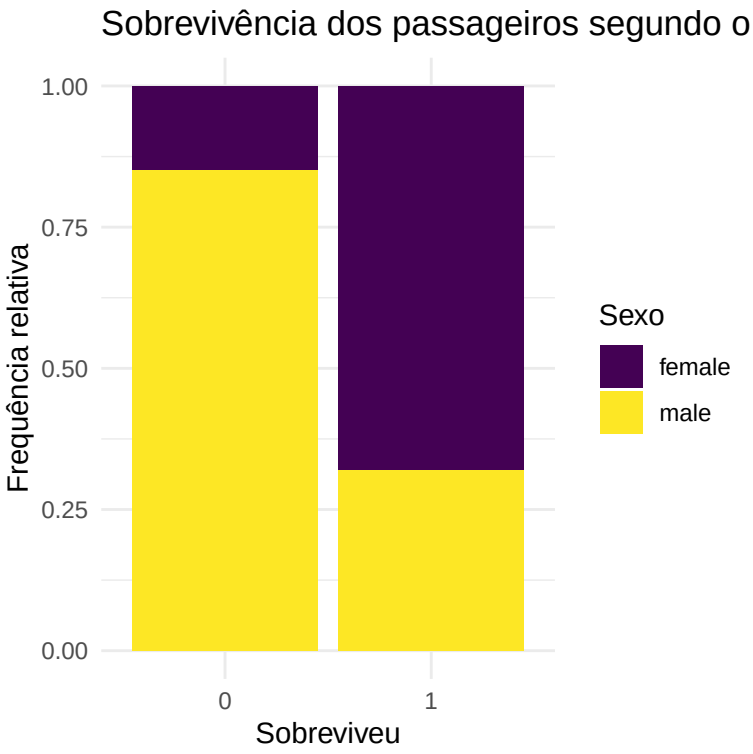
```
## Cross-Tabulation, Column Proportions
## Sex * Survived
## Data Frame: Titanic
##
## -------- ---------- -------------- -------------- --------------
##            Survived              0              1          Total
##      Sex
##   female               81 ( 14.8%)   231 ( 67.9%)   312 ( 35.1%)
##     male              468 ( 85.2%)   109 ( 32.1%)   577 ( 64.9%)
##    Total              549 (100.0%)   340 (100.0%)   889 (100.0%)
## -------- ---------- -------------- -------------- --------------
```

3.1.2 Gráficos

```r
#Gráfico de barras múltiplas

library(ggplot2)

ggplot(Titanic) +
 aes(x = Survived, fill = Sex) +
 geom_bar(position = "dodge") +
 scale_fill_viridis_d(option = "viridis") +
 labs(x = "Sobreviveu", y = "Frequência",
      title = "Sobrevivência dos passageiros segundo o sexo", fill = "Sexo") +
 theme_minimal()
```




```
#Gráfico de barras empilhadas
ggplot(Titanic) +
 aes(x = Survived, fill = Sex) +
 geom_bar(position = "fill") +
 scale_fill_viridis_d(option = "viridis") +
 labs(x = "Sobreviveu", y = "Frequência relativa",
      title = "Sobrevivência dos passageiros segundo o sexo", fill = "Sexo") +
 theme_minimal()
```



3.2 Variáveis Quantitativa x Quantitativa

Vamos avaliar as variáveis **Age** e **Fare** como exemplo.

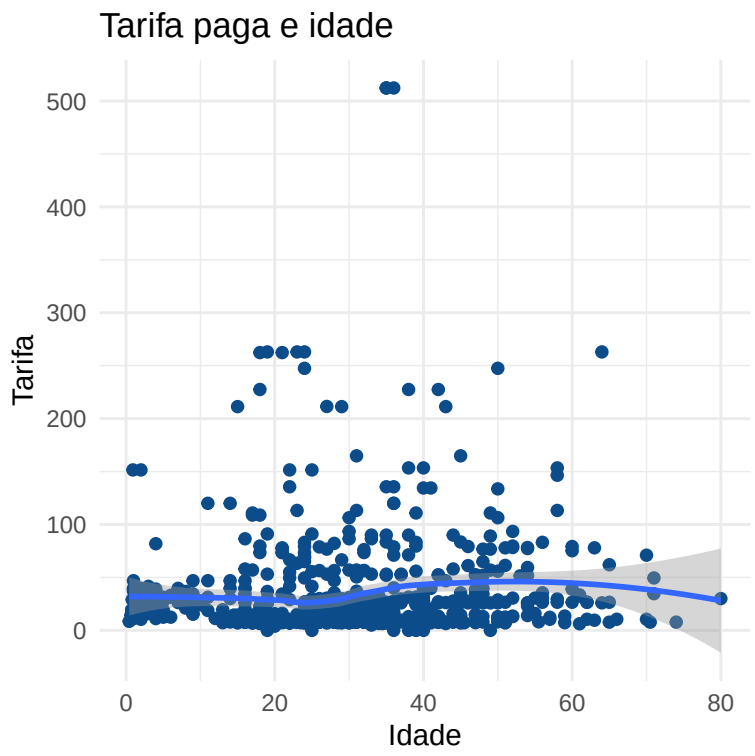
3.2.1 Correlação

```
cor(Titanic$Age, Titanic$Fare, use = "complete.obs")
```

```
## [1] 0.09314252
```

3.2.2 Gráficos

```
library(ggplot2)

ggplot(Titanic) +
 aes(x = Age, y = Fare) +
 geom_point(size = 1.7, colour = "#0c4c8a") +
 geom_smooth(span = 0.75) +
 labs(x = "Idade", y = "Tarifa", title = "Tarifa paga e idade") +
 theme_minimal()
```



3.3 Variáveis Quantitativa x Qualitativa

3.3.1 Medidas-resumo

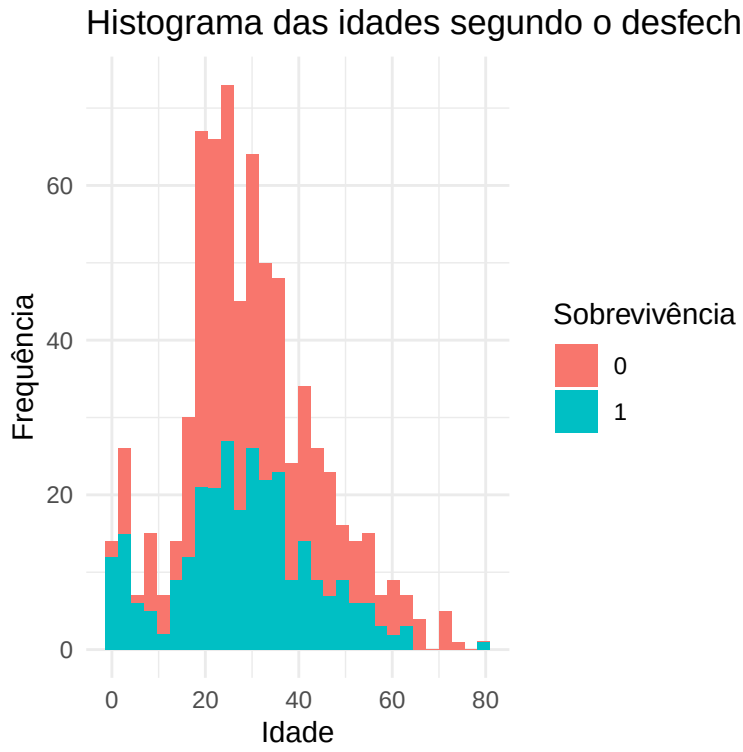
Vamos avaliar a variável quantitativa **Age** segundo o desfecho **Survived** como exemplo:

```
with(Titanic, stby(Age, Survived, descr))
```

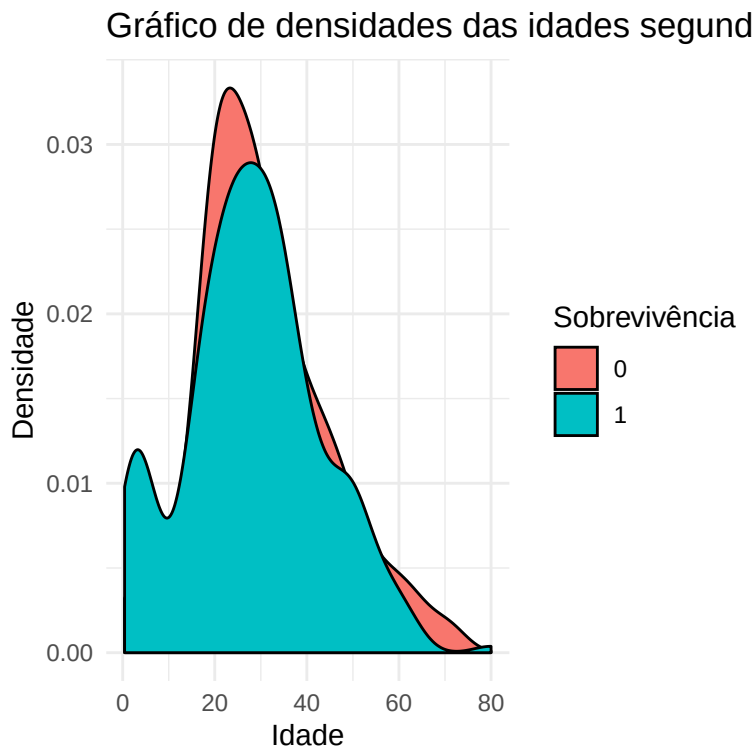
```
## Descriptive Statistics
## Age by Survived
## Data Frame: Titanic
## N: 549
##
##                          0        1
## ----------------- -------- --------
##              Mean    30.63    28.19
##           Std.Dev    14.17    14.86
##               Min     1.00     0.42
##                Q1    21.00    19.00
##            Median    28.00    28.00
##                Q3    39.00    36.00
##               Max    74.00    80.00
##               MAD    11.86    13.34
##               IQR    18.00    17.00
##                CV     0.46     0.53
##          Skewness     0.58     0.17
##       SE.Skewness     0.12     0.14
##          Kurtosis     0.25    -0.09
##           N.Valid   424.00   288.00
##         Pct.Valid    77.23    84.71
```

3.3.2 Gráficos

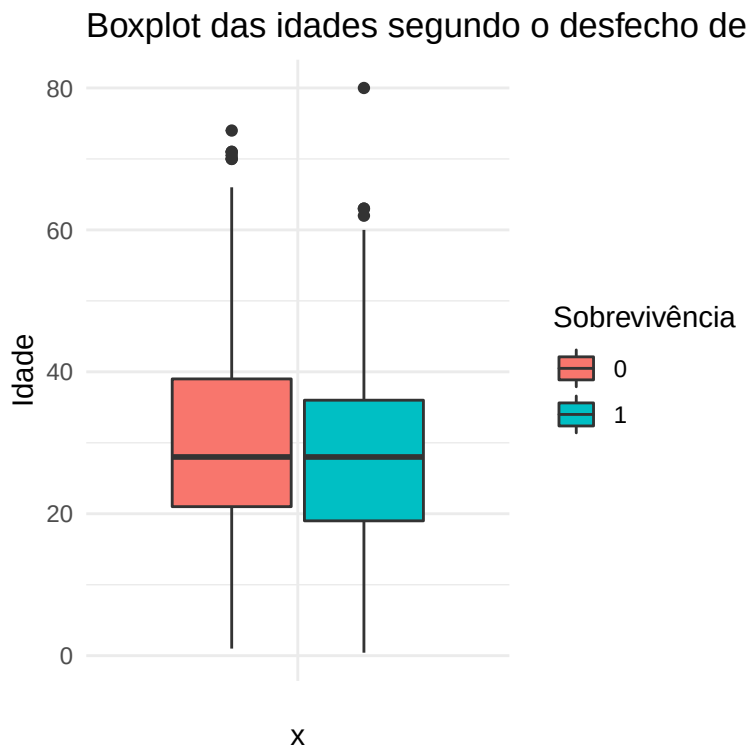
```
library(ggplot2)

# Histograma
ggplot(Titanic) +
 aes(x = Age, fill = Survived) +
 geom_histogram(bins = 30L) +
 scale_fill_hue() +
 labs(x = "Idade", y = "Frequência",
      title = "Histograma das idades segundo o desfecho de sobrevivência",
      fill = "Sobrevivência") +
 theme_minimal()
```




```
# Gráfico de densidades
ggplot(Titanic) +
 aes(x = Age, fill = Survived) +
 geom_density(adjust = 1L) +
 scale_fill_hue() +
 labs(x = "Idade", y = "Densidade",
      title = "Gráfico de densidades das idades segundo o desfecho de sobrevivência",
      fill = "Sobrevivência") +
 theme_minimal()
```




```
# Boxplot
ggplot(Titanic) +
 aes(x = "", y = Age, fill = Survived) +
 geom_boxplot() +
 scale_fill_hue() +
 labs(y = "Idade", title = "Boxplot das idades segundo o desfecho de sobrevivência",
      fill = "Sobrevivência") +
 theme_minimal()
```




```
#Gráfico de violino
ggplot(Titanic) +
 aes(x = "", y = Age, fill = Survived) +
 geom_violin(adjust = 1L, scale = "area") +
 scale_fill_hue() +
 labs(y = "Idade",
      title = "Gráfico de violino das idades segundo o desfecho de sobrevivência",
      fill = "Sobrevivência") +
 theme_minimal()
```

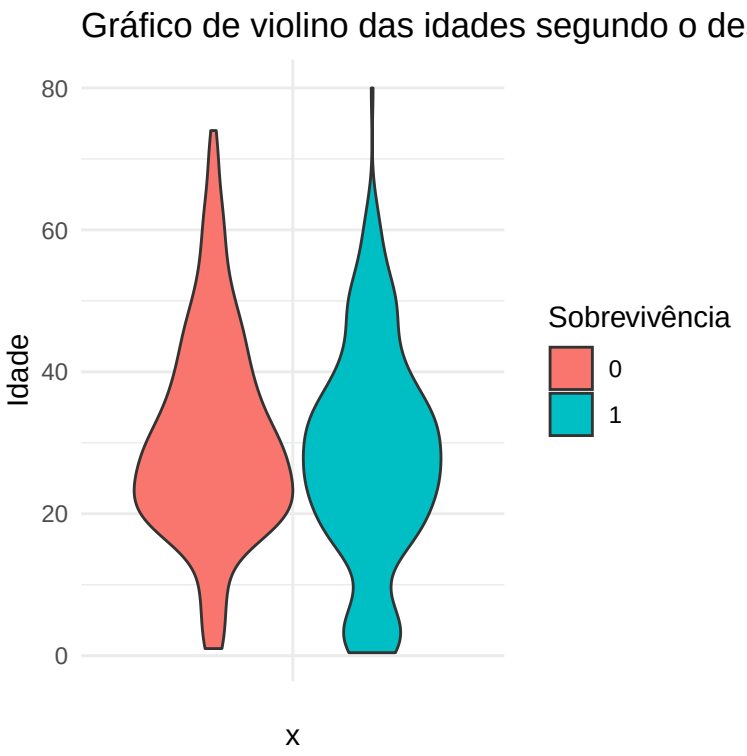


Gráfico de violino das idades segundo o des
80
60
40
20
0
Idade
x
Sobrevivência
0
1