Atividade 4. Estatísticas descritivas no R

Pedro H. G. Ferreira de Souza

2022-11-10

Instruções

Para responder aos exercícios abaixo, simplesmente substituam as reticências (...) pelas funções e expressões adequadas e respondam às perguntas inseridas.

Pontuação máxima em cada parte

Parte	Pontos
1	10
2	30
3	30
4	30

Parte 1. Preparação

Limpando o workspace

```
# Para iniciar, limparemos o workspace, apagando os objetos carregados
rm(list = ls())
```

Identificação do aluno

Exercício:

```
# Crie um vetor com a forma abaixo para identificar seu nome e email
aluno <- c(...)
names(aluno) <- c('nome', 'email')
print(aluno)
```

Resposta:

```
# Crie um vetor com a forma abaixo para identificar seu nome e email
aluno <- c("Pedro Souza", "pedro.ferreira@ipea.gov.br")
names(aluno) <- c('nome', 'email')
print(aluno)
```

```
##                         nome                        email
##               "Pedro Souza" "pedro.ferreira@ipea.gov.br"
```

Pacotes

Exercício:

```
# Instale (se necessario) e carregue os pacotes: 'tidyverse' e 'summarytools'
library(tidyverse)
library(...)
```

Resposta:

```
# Crie um vetor com a forma abaixo para identificar seu nome e email
library(tidyverse)
library(summarytools)
```

Data Frames

```
# Utilizaremos data frames que vem com pacotes. E' so executar o codigo abaixo.
crimes.df <- USArrests
tempestades.df <- storms
titanic.df <- data.frame(Titanic)
```

Parte 2. Passageiros do Titanic

O data frame `titanic.df` contém informações sobre o destino dos passageiros do Titanic, conforme classe, sexo, idade e sobrevivência.

Dicionário:

- `class`: classe no navio
  - *1st, 2nd, 3rd, Crew* (ou seja, 1a, 2a, 3a, Tripulação)
- `sex`: sexo
  - *Male, Female* (ou seja, Homem, Mulher)
- `age`: faixa etária
  - *Child, Adult* (ou seja, Criança, Adulto)
- `Survived`: sobreviveu?
  - *Yes, No* * (ou seja, Sim, Não)
- `Freq`: número de passageiros na categoria

Tabelas de frequência com e sem ponderação

Execute os dois comandos abaixo:

```
freq(titanic.df$Class, headings = FALSE)
```

```
##
##               Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
## ----------- ------ --------- -------------- --------- --------------
##         1st      8     25.00          25.00     25.00          25.00
##         2nd      8     25.00          50.00     25.00          50.00
##         3rd      8     25.00          75.00     25.00          75.00
##        Crew      8     25.00         100.00     25.00         100.00
##        <NA>      0                                0.00         100.00
##       Total     32    100.00         100.00    100.00         100.00
```

```
freq(titanic.df$Class, weights = titanic.df$Freq, headings = FALSE)
```

```
##
```

```
##                 Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
## ----------- --------- --------- -------------- --------- --------------
##         1st    325.00     14.77          14.77     14.77          14.77
##         2nd    285.00     12.95          27.71     12.95          27.71
##         3rd    706.00     32.08          59.79     32.08          59.79
##        Crew    885.00     40.21         100.00     40.21         100.00
##        <NA>      0.00                               0.00         100.00
##       Total   2201.00    100.00         100.00    100.00         100.00
```

Pergunta: por que é necessário ponderar pela frequência como no segundo comando?

```
Porque os dados estão agregados por categorias e as frequências estao em uma coluna.
```

Pergunta: Qual a categoria com maior percentual de pessoas?

```
Tripulação
```

Gráfico de barras

*Faça um gráfico de barras com a distribuição **relativa** de sobreviventes por classe.*

Dica: veja os slides 28 e 29 da aula 05, mas não se esqueça de incluir os pesos.

Exercício:

```
Class_pct.df <- freq(..., weights = ...) %>% tb(na.rm = TRUE)
qplot(data = ...f, x = ..., y = ..., geom = 'col')
```

Resposta:

```
Class_pct.df <- titanic.df %>% filter(Survived == 'Yes')
Class_pct.df <- freq(Class_pct.df$Class,
                     weights = Class_pct.df$Freq) %>% tb(na.rm = TRUE)
qplot(data = Class_pct.df, x = Class, y = pct, geom = 'col')
```

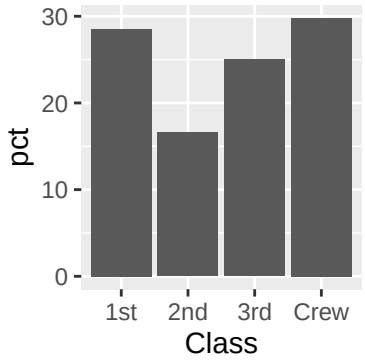


Tabela cruzada

Faça duas tabelas cruzadas de `Class` com `Survived`: uma somando 100% nas colunas e outras somando 100% nas linhas. Não se esqueça de ponderar pelos pesos.

Dica; veja os slides 54 e 55 da aula 5.

Exercício:

```
with(titanic.df, ctable(x = ..., y = ..., weights = ..., prop = ...))
with(..., ctable(...))
```

Resposta:

```
with(titanic.df, ctable(x = Class, y = Survived, weights = Freq, prop = 'r'))
```

```
## Cross-Tabulation, Row Proportions
## Class * Survived
## Data Frame: titanic.df
##
## ------- ---------- ---------------- --------------- -----------------
##           Survived               No             Yes             Total
##   Class
##     1st               122.0 (37.5%)   203.0 (62.5%)    325.0 (100.0%)
##     2nd               167.0 (58.6%)   118.0 (41.4%)    285.0 (100.0%)
##     3rd               528.0 (74.8%)   178.0 (25.2%)    706.0 (100.0%)
##    Crew               673.0 (76.0%)   212.0 (24.0%)    885.0 (100.0%)
##   Total              1490.0 (67.7%)   711.0 (32.3%)   2201.0 (100.0%)
## ------- ---------- ---------------- --------------- -----------------
```

```
with(titanic.df, ctable(x = Class, y = Survived, weights = Freq, prop = 'c'))
```

```
## Cross-Tabulation, Column Proportions
## Class * Survived
## Data Frame: titanic.df
##
## ------- ---------- ---------------- ---------------- -----------------
##           Survived               No              Yes             Total
##   Class
##     1st               122.0 (  8.2%)   203.0 ( 28.6%)    325.0 ( 14.8%)
##     2nd               167.0 ( 11.2%)   118.0 ( 16.6%)    285.0 ( 12.9%)
##     3rd               528.0 ( 35.4%)   178.0 ( 25.0%)    706.0 ( 32.1%)
##    Crew               673.0 ( 45.2%)   212.0 ( 29.8%)    885.0 ( 40.2%)
##   Total              1490.0 (100.0%)   711.0 (100.0%)   2201.0 (100.0%)
## ------- ---------- ---------------- ---------------- -----------------
```

Pergunta: qual categoria de `Class` teve o maior percentual de sobreviventes? Qual foi esse percentual?

1a Classe, com 62.5%

Pergunta: entre os que morreram, qual categoria de `Class` foi mais numerosa? Com qual percentual?

Tripulação, com 45.2%

Parte 3. Crimes nos estados americanos

O data frame `crimes.df` contém estatísticas de prisões para os 50 estados americanos em 1973.

Dicionário:

- `Murder`: taxa de prisões por homicídio (por 100,000 habitantes)
- `Assault`: taxa de prisões por agressão (por 100,000 habitantes)
- `Rape`: taxa de prisões por estupro (por 100,000 habitantes)
- `UrbanPop`: percentual da população que vive em áreas urbanas

Coluna com nomes dos estados

Para começar, vamos transformar os nomes das linhas em uma coluna chamada `states`.

```
crimes.df <- rownames_to_column(crimes.df, var = 'states')
glimpse(crimes.df)
```

```
## Rows: 50
## Columns: 5
## $ states   <chr> "Alabama", "Alaska", "Arizona", "Arkansas", "California", "Co~
## $ Murder   <dbl> 13.2, 10.0, 8.1, 8.8, 9.0, 7.9, 3.3, 5.9, 15.4, 17.4, 5.3, 2.~
## $ Assault  <int> 236, 263, 294, 190, 276, 204, 110, 238, 335, 211, 46, 120, 24~
## $ UrbanPop <int> 58, 48, 80, 50, 91, 78, 77, 72, 80, 60, 83, 54, 83, 65, 57, 6~
## $ Rape     <dbl> 21.2, 44.5, 31.0, 19.5, 40.6, 38.7, 11.1, 15.8, 31.9, 25.8, 2~
```

Média, mediana e desvio padrão (parte 1)

Calcule a média, mediana e desvio padrão das variáveis quantitativas.

Exercício:

```
crimes.df %>% ...
```

Resposta:

```
crimes.df %>% descr(stats = c('mean', 'med', 'sd'), headings = FALSE)
```

```
##
##                 Assault   Murder    Rape   UrbanPop
## ------------- --------- -------- ------- ----------
##          Mean    170.76     7.79   21.23      65.54
##        Median    159.00     7.25   20.10      66.00
##       Std.Dev     83.34     4.36    9.37      14.47
```

Nova variável recodificando `UrbanPop`

Crie uma variável LOGICAL com valor TRUE se `UrbanPop` estiver acima da medina.

Exercício:

```
crimes.df <- crimes.df %>% mutate(...)
```

Resposta:

```
crimes.df <- crimes.df %>% mutate(muito_urbanizado = UrbanPop > median(UrbanPop))
freq(crimes.df$muito_urbanizado, headings = FALSE)
```

```
##
##               Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
## ----------- ------ --------- -------------- --------- --------------
##       FALSE     26     52.00          52.00     52.00          52.00
##        TRUE     24     48.00         100.00     48.00         100.00
##        <NA>      0                                0.00         100.00
##       Total     50    100.00         100.00    100.00         100.00
```

Média, mediana e desvio padrão (parte 2)

Produza tabela com o número de estados e a média de `UrbanPop` pelas cateogrias da variável binária criada acima.

Exercício:

```
crimes.df %>% group_by(...) %>% descr(...)
```

Resposta:

```
crimes.df %>%
  group_by(muito_urbanizado) %>%
    descr(UrbanPop, stats = c('n.valid', 'mean'), headings = FALSE)
```

```
##
##                 FALSE    TRUE
## ------------- ------- -------
##       N.Valid   26.00   24.00
##          Mean   54.19   77.83
```

Calcule a média, mediana e desvio padrão de `Murder`, `Assault` e `Rape` pelas categorias dessa nova variável.

Exercício:

```
crimes.df %>% group_by(...) %>% select(...) %>% descr(...)
```

Resposta:

```
crimes.df %>%
  group_by(muito_urbanizado) %>%
  select(Murder, Assault, Rape) %>%
  descr(stats = c('mean','med','sd'), headings = FALSE)
```

```
## Group: muito_urbanizado = FALSE
##
##                 Assault   Murder    Rape
## ------------- --------- -------- -------
##          Mean    145.31     7.57   17.28
##        Median    114.00     6.40   16.35
##       Std.Dev     84.08     5.11    7.68
##
## Group: muito_urbanizado = TRUE
##
##                 Assault   Murder    Rape
## ------------- --------- -------- -------
##          Mean    198.33     8.03   25.51
##        Median    189.50     7.65   25.80
##       Std.Dev     74.75     3.45    9.28
```

Pergunta: estados mais urbanizados eram MAIS ou MENOS violentos que estados não urbanizados?

```
Mais violentos para as três variaveis.
```

Correlação

Qual a correlação entre as taxas de homicídio, agressão e estupro?

Exercício:

```
crimes.df %>% select(...) %>% cor()
```

Resposta:

```
crimes.df %>% select(Murder, Assault, Rape) %>% cor()
```

```
##            Murder   Assault      Rape
## Murder  1.0000000 0.8018733 0.5635788
## Assault 0.8018733 1.0000000 0.6652412
## Rape    0.5635788 0.6652412 1.0000000
```

Pergunta: as três correlações sao positivas ou negativas? O que isso significa?

Todas são positivas e relativamente fortes, acima de 0.5. Isso significa que estados com maior taxa de homicídios também têm maiores taxas de agressão e estupro.

Histograma

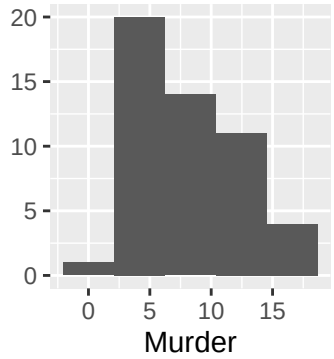
Faça histogramas da taxa de homicídio com 5, 10 e 15 bins.

Exercício:

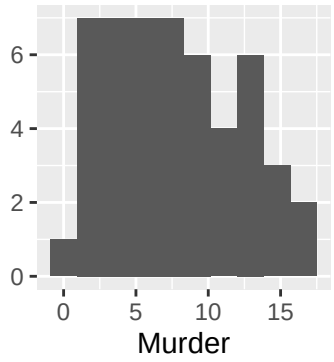
```
qplot(data = ..., x = ..., geom = '...', bins = 5)
qplot(..., ..., ..., bins = 10)
qplot(..., ..., ..., ...)
```

Resposta:

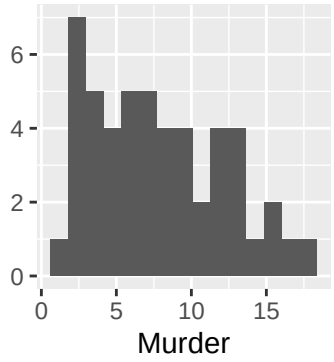
```
qplot(data = crimes.df, x = Murder, geom = 'histogram', bins = 5)
```




```
qplot(data = crimes.df, x = Murder, geom = 'histogram', bins = 10)
```




```
qplot(data = crimes.df, x = Murder, geom = 'histogram', bins = 15)
```



Pergunta: qual dos três gráficos acima é o seu preferido? Por quê?

```
Prefiro o de 15 bins, pois mostra um padrão sem simplificar excessivamente.
```

Densidade

Faça um gráfico de densidade kernel para homicídios.

Exercício:

```
qplot(data = ..., x = ..., geom = 'density', bw = 1)
```

Resposta:

```
# Faca um grafico de densidade kernel para homicidios
qplot(data = crimes.df, x= Murder, geom = 'density', bw = 1)
```

Scatter plot

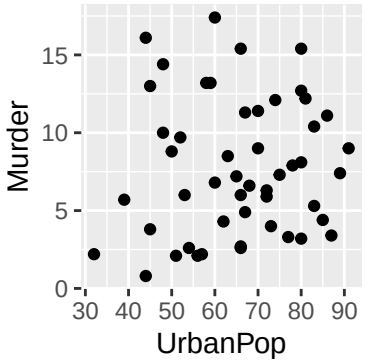
Faça um scatter plot da taxa de urbanização no eixo X e da taxa de homicídios no eixo Y, depois calcule a correlação entre ambas.

Exercício:

```
qplot(data = crimes.df, x = ..., y = ..., geom = 'point')
crimes.df %>% select(...) %>% ...
```

Resposta:

```
qplot(data = crimes.df, x = UrbanPop, y = Murder, geom = 'point')
```




```
crimes.df %>% select(UrbanPop, Murder) %>% cor()
```

```
##            UrbanPop     Murder
## UrbanPop 1.00000000 0.06957262
## Murder   0.06957262 1.00000000
```

Pergunta: qual o grau de correlação linear entre as variáveis? Em termos descritivos, o que isso implica?

```
Correlação de 0.07, bem fraco. Não há nenhuma relação linear bivariada aparente entre ambas.
```

Parte 4. Tempestades

O data frame `tempestades.df` contém uma amostra de tempestades atlânticas entre 1975 e 2000.

Dicionário:

- `name`: nome da tempestade
- `year`, `month`, `day`, `hour`: data de mensuração
- `lat`, `long`: latitude e longitude
- `status`: status da tempestade
  - *Tropical Depression, Tropical Storm ou Hurricane* (ou seja, depressão trópica, tempestade tropical, furacão)
- `category`: intensidade da tempestade
  - Entre -1 (mais fraco) e 4 (furacão mais forte)
- `wind`: velocidade máxima do vento em nós
- `pressure`: pressão no centro da tempestade, em milibars
- `tropicalstorm_force_diameter`: diâmetro em milhas náuticas da área afetada por ventos de 34 nós ou mais
- `hurricane_force_diameter`: diâmetro em milhas náuticas da área afetada por ventos de 64 nós ou mais

Filtrando e condensando os dados

A partir de `tempestades.df`, crie um novo data frame `furacoes.df` seguindo essas etapas:

a. Filtre o data frame de tempestades para manter **somente** os furacões (ver variável `status`)
b. Calcule o valor médio de `wind` e `pressure`e o valor máximo de `category` para cada furacão

Observe que cada furacão será uma linha no novo data frame, que terá as colunas `wind`, `pressure` e `category`.

Exercício:

```
furacoes.df <-
  tempestades.df %>%
```

```
  filter(status == ...) %>%
  group_by(name) %>%
  summarise(wind = mean(wind),
            ... = ...),
            category = max(as.numeric(category)))
```

Resposta:

```
furacoes.df <-
  tempestades.df %>%
    filter(status == 'hurricane') %>%
      group_by(name) %>%
        summarise(wind = mean(wind),
                  pressure = mean(pressure),
                  category = max(as.numeric(category)))
head(furacoes.df)
```

```
## # A tibble: 6 x 4
##   name      wind pressure category
##   <chr>    <dbl>    <dbl>    <dbl>
## 1 AL121991  65       980.        3
## 2 Alberto   78.3     978.        5
## 3 Alex      80.6     970.        5
## 4 Alicia    84.4     974.        5
## 5 Allison   65       988.        3
## 6 Andrew   118.      947.        7
```

Factor

Converta a variável `category` em FACTOR com `as.factor()`.

```
furacoes.df$category <- as.factor(furacoes.df$category)
```

Estatísticas descritivas (parte 1)

Calcule `n.valid`, média, desvio padrão, mediana e mínimo e máximo para `wind` e `pressure`.

Exercício:

```
furacoes.df %>% select(...) %>% ...(stats = c(...))
```

Resposta:

```
furacoes.df %>%
  select(wind, pressure) %>%
  descr(stats = c('n.valid','mean','sd','med', 'min', 'max'), headings = FALSE)
```

```
##
##                 pressure     wind
## ------------- ---------- --------
##       N.Valid     137.00   137.00
##          Mean     972.78    82.35
##       Std.Dev      12.99    12.84
##        Median     975.61    79.42
##           Min     931.63    65.00
##           Max     996.75   118.26
```

Estatísticas descritivas por categorias de furacão

Obtenha as mesmas estatísticas por categoria de furacão.

Exercício:

```
furacoes.df %>%
  select(...) %>%
    ...(category) %>%
      ...(stats = ...)
```

Resposta:

```
furacoes.df %>%
  select(category,wind, pressure) %>%
    group_by(category) %>%
      descr(stats = c('n.valid','mean','sd','med', 'min', 'max'), headings = FALSE)
```

```
## Group: category = 3
##
##                 pressure    wind
## ------------- ---------- -------
##       N.Valid      37.00   37.00
##          Mean     985.11   68.68
##       Std.Dev       5.53    2.77
##        Median     985.41   68.68
##           Min     974.70   65.00
##           Max     996.75   74.38
##
## Group: category = 4
##
##                 pressure    wind
## ------------- ---------- -------
##       N.Valid      15.00   15.00
##          Mean     980.46   76.51
##       Std.Dev       4.97    2.58
##        Median     980.21   76.43
##           Min     970.47   72.63
##           Max     988.40   82.50
##
## Group: category = 5
##
##                 pressure    wind
## ------------- ---------- -------
##       N.Valid      26.00   26.00
##          Mean     975.41   80.08
##       Std.Dev       5.00    3.95
##        Median     975.87   79.39
##           Min     960.09   71.79
##           Max     987.00   91.00
##
## Group: category = 6
##
##                 pressure     wind
## ------------- ---------- --------
##       N.Valid      40.00    40.00
```

```
##           Mean     965.62    89.26
##        Std.Dev       9.52     8.26
##         Median     968.09    88.93
##            Min     946.42    76.43
##            Max     980.49   110.60
##
## Group: category = 7
##
##                  pressure     wind
## ------------- ---------- --------
##        N.Valid      19.00    19.00
##           Mean     954.15   102.17
##        Std.Dev      10.55    10.72
##         Median     953.82   105.11
##            Min     931.63    85.36
##            Max     971.43   118.26
```

Pergunta: qual a razão entre a mediana de `wind` da categoria 7 e da categoria 3?

```
print(105.11 / 68.68)
```

```
## [1] 1.530431
```

Pergunta: o que isso significa?

```
O furacão "típico" da categoria 7 tem ventos mais de 50% mais fortes do que o da categoria 3
```

Percentis 10, 25, 75 e 90

Calcule os percentis acima para a variável `wind`.

Exercício:

```
# Calcule os percentis acima para a variavel wind
furacoes.df %>% select(...) %>% ...(p10 = quantile(..., prob = .10),
                                    ...,
                                    ...,
                                    ...)
```

Resposta:

```
furacoes.df %>% select(wind, pressure) %>% summarise(p10 = quantile(wind, prob = .10),
                                                     p25 = quantile(wind, prob = .25),
                                                     p75 = quantile(wind, prob = .75),
                                                     p90 = quantile(wind, prob = .90))
```

```
## # A tibble: 1 x 4
##     p10   p25   p75   p90
##   <dbl> <dbl> <dbl> <dbl>
## 1  67.5  72.7  90.3  102.
```

Box plot (parte 1)

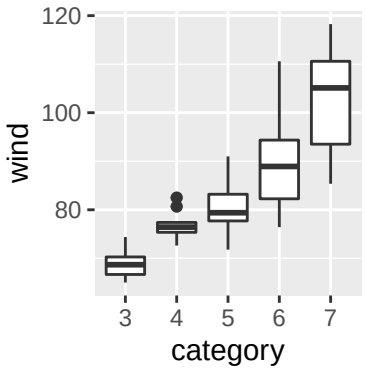
Faça um box plot da variável `wind` pelas categorias da variável `category`.

Exercício:

```
qplot(data = ..., x = ..., y = ..., geom = 'boxplot')
```

Resposta:

```
qplot(data = furacoes.df, x = category, y = wind, geom = 'boxplot')
```



Box plot (parte 2)

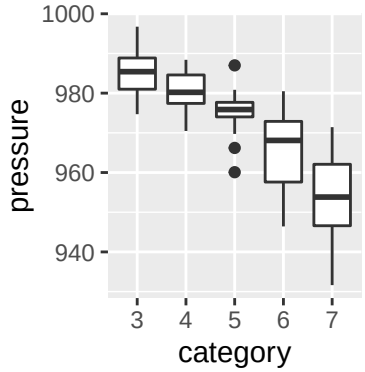
Faça um box plot da variável `pressure` pelas categorias da variável `category`.

Exercício:

```
...
```

Resposta:

```
qplot(data = furacoes.df, x = category, y = pressure, geom = 'boxplot')
```



Pergunta: a comparação entre os dois box plots sugere que tipo de correlação entre `wind` e `pressure`?

```
Correlação negativa forte.
```

Correlação (parte 2)

Calcule então a correlação entre `wind` e `pressure`.

Exercício:

```
...
```

Resposta:

```
cor(furacoes.df$wind, furacoes.df$pressure)
```

```
## [1] -0.9299381
```

Scatter plot (parte 2)

Faça um scatter plot entre `wind` (eixo X) e `pressure` (eixo Y) para furacões da categoria 4 ou superior.

Exercício:

```
furacoes4mais.df <- furacoes.df %>% filter(as.numeric(...) >= 4)
qplot(data = ..., x = ..., y = ..., geom = '...')
```

Resposta:

```
furacoes4mais.df <- furacoes.df %>% filter(as.numeric(category) >= 4)
qplot(data = furacoes4mais.df, x = wind, y = pressure, geom = 'point')
```