Métodos Quantitativos II - Lista 2

Professor Manoel Galdino e Monitor Davi Veronese

September 4, 2023

Esta lista destina-se a revisar questões de estatística básica. Nos exercícios, vamos utilizar dados da Pesquisa Nacional por Amostra de Domicílios Contínua (PNADc).

Os alunos devem entregar um arquivo PDF contendo as respostas e o script para replicação.

```
# Material de apoio para esta lista:
# https://jonnyphillips.github.io/Analise_de_Dados_2022/
```

1

O pacote PNADcIBGE permite importar bases de dados diretamente para o environment do R. Primeiro, instale e ative o pacote. Depois, importe os dados do último trimestre de 2017 (variáveis selecionadas) por meio do código abaixo.

```
# Ver: https://cran.r-project.org/web/packages/PNADcIBGE/PNADcIBGE.pdf
# Instale o pacote
# install.packages("PNADcIBGE")

# Carregue o pacote
library(PNADcIBGE)

# Importe os dados desejados
data <- get_pnadc(year=2017,
                  quarter=4,
                  selected=FALSE,
                  vars=c("Ano", "Trimestre", "UF", "V2007", "VD4020", "VD4035"),
                  design=FALSE,
```

```
                    savedir=tempdir())
# Por razões didáticas, selecionamos "design=FALSE" para ignorar o plano amostral.
# Não faça isso em sua pesquisa.

# Selecione apenas as variáveis úteis para esta lista:
library(tidyverse)
library(tidylog)

data <- data %>%
  select(Ano, Trimestre, UF, V2007, VD4020, VD4035)

# Renomeie as variáveis:
data <- data %>%
  rename(Sexo = V2007,
         Renda = VD4020,
         Horas_trabalhadas = VD4035)
```

2

Calcule:

i) a renda média;

ii) a variância da renda;

iii) a renda média dos homens e das mulheres;

iv) a renda média em cada estado brasileiro;

v) a covariância entre a renda e o número de horas trabalhadas.

```
library(knitr)

# Renda média e variância
data %>% summarize(renda_media = mean(Renda, na.rm = TRUE),
                   variancia_renda = var(Renda, na.rm = TRUE)) %>% kable()
```

renda_media	variancia_renda
1931.283	9543677

```
# Renda média por sexo
data %>% group_by(Sexo) %>%
  summarize(renda_media = mean(Renda, na.rm = TRUE)) %>% kable()
```

Sexo	renda_media
Homem	2077.956
Mulher	1720.783

```
# Renda por estado
data %>% group_by(UF) %>%
  summarize(renda_media = mean(Renda, na.rm = TRUE)) %>% kable()
```

UF	renda_media
Rondônia	1811.234
Acre	1573.968
Amazonas	1593.292
Roraima	2034.655
Pará	1393.895
Amapá	2108.920
Tocantins	1779.919
Maranhão	1075.140
Piauí	1291.965
Ceará	1263.629
Rio Grande do Norte	1355.121
Paraíba	1598.286
Pernambuco	1507.282
Alagoas	1219.109
Sergipe	1355.237
Bahia	1250.915
Minas Gerais	1864.039
Espírito Santo	2046.777
Rio de Janeiro	2254.184
São Paulo	2594.433
Paraná	2273.570
Santa Catarina	2353.374
Rio Grande do Sul	2360.197
Mato Grosso do Sul	2127.319
Mato Grosso	2173.382
Goiás	2081.978
Distrito Federal	3842.553

```r
# Covariância entre renda e número de horas trabalhadas
cov(data$Renda, data$Horas_trabalhadas, use = "complete.obs")

## [1] 5776.864
```

3

Exemplifique a veracidade da equação, considerando X = Renda, Y = Horas trabalhadas, a = 2 e b = 3.

$$E[aX + bY] = a \times E[X] + b \times E[Y] \tag{1}$$

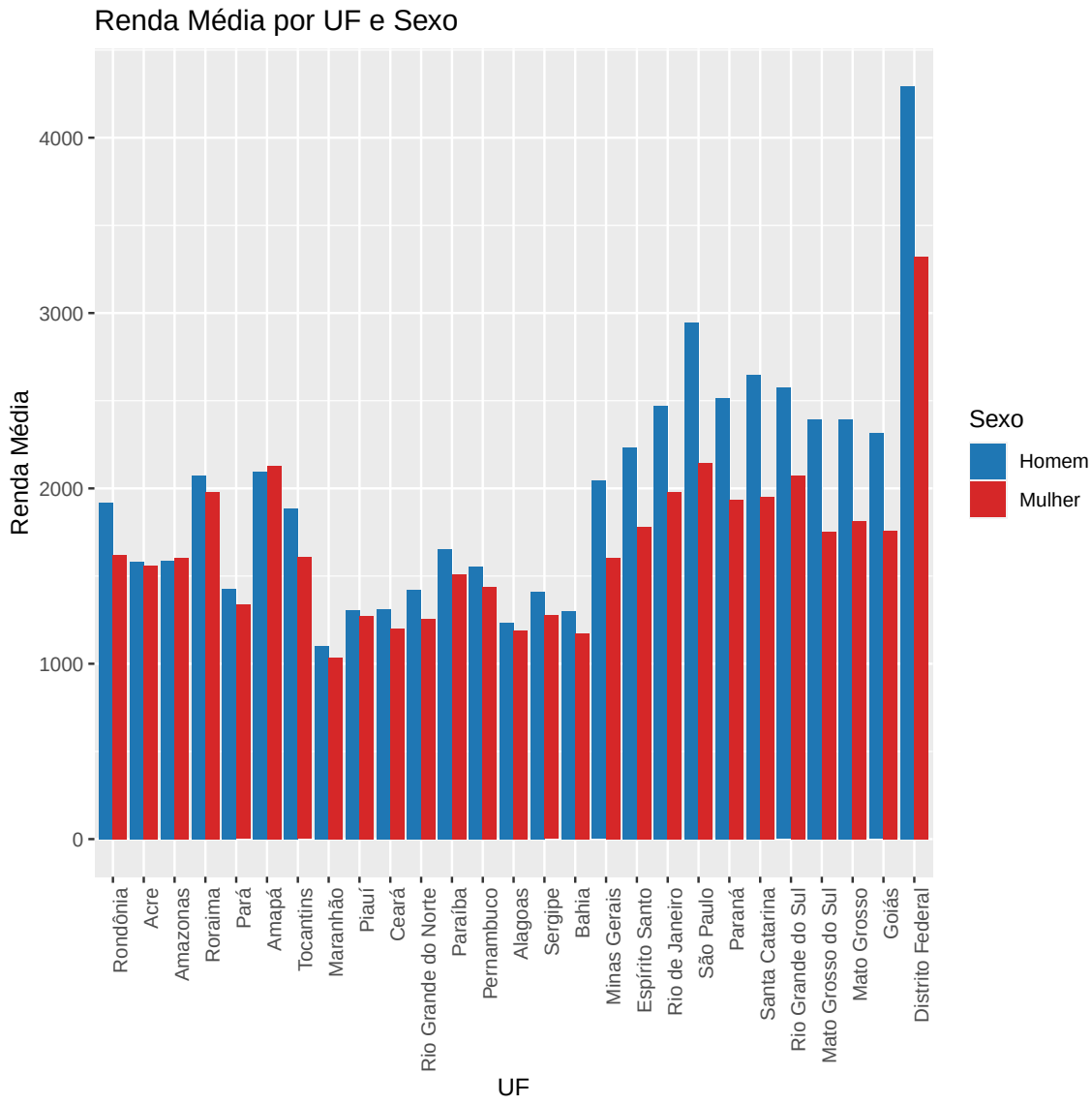
```r
data %>%
 mutate(renda_3 = 2*Renda,
        horas_2 = 3*Horas_trabalhadas,
        Renda_horas = renda_3 + horas_2, na.rm = TRUE) %>%
 summarize(Renda_horas_mean = mean(Renda_horas, na.rm = TRUE),
           mean_renda = mean(Renda, na.rm = TRUE),
           mean_horas = mean(Horas_trabalhadas, na.rm = TRUE)) %>%
 mutate(Renda_horas_mean_2 = 2*mean_renda + 3*mean_horas)

## # A tibble: 1 x 4
##   Renda_horas_mean mean_renda mean_horas Renda_horas_mean_2
##              <dbl>      <dbl>      <dbl>              <dbl>
## 1            3975.      1931.       37.2              3974.
```

4

Apresente um gráfico que permita visualização adequada da média da renda por estado brasileiro e sexo.

```r
library(ggplot2)

data %>%
  group_by(UF, Sexo) %>%
  summarize(renda_media = mean(Renda, na.rm = TRUE)) %>%
```

```r
ggplot() +
geom_col(aes(x = UF, y = renda_media, fill = Sexo), position = "dodge") +
labs(x = "UF", y = "Renda Média", title = "Renda Média por UF e Sexo") +
theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
scale_fill_manual(values = c("#1f77b4", "#d62728"))
```



5

Agora trabalharemos explicitamente com a esperança condicional. Note que essa lógica estava implícita nas questões anteriores. Assuma duas variáveis aleatórias, X e Y, tais que X = renda e Y = horas trabalhadas.

Calcule:

i)

$$E[X|10 \leq Y \leq 20] \tag{2}$$

ii)

$$E[X|Y \geq 20] \tag{3}$$

```
# i)
data %>%
 filter(Horas_trabalhadas >= 10 & Horas_trabalhadas <= 20) %>%
 summarize(renda_media = mean(Renda, na.rm = TRUE)) # 940

## # A tibble: 1 x 1
##   renda_media
##         <dbl>
## 1        940.

# ii)

data %>%
 filter(Horas_trabalhadas >= 20) %>%
 summarize(renda_media = mean(Renda, na.rm = TRUE)) # 2015

## # A tibble: 1 x 1
##   renda_media
##         <dbl>
## 1       2015.
```

6

Para os itens seguintes (i a iv), remova todas as observações cuja renda seja superior a 10.000 reais.

i) apresente um gráfico de densidade da variável renda. Interprete;

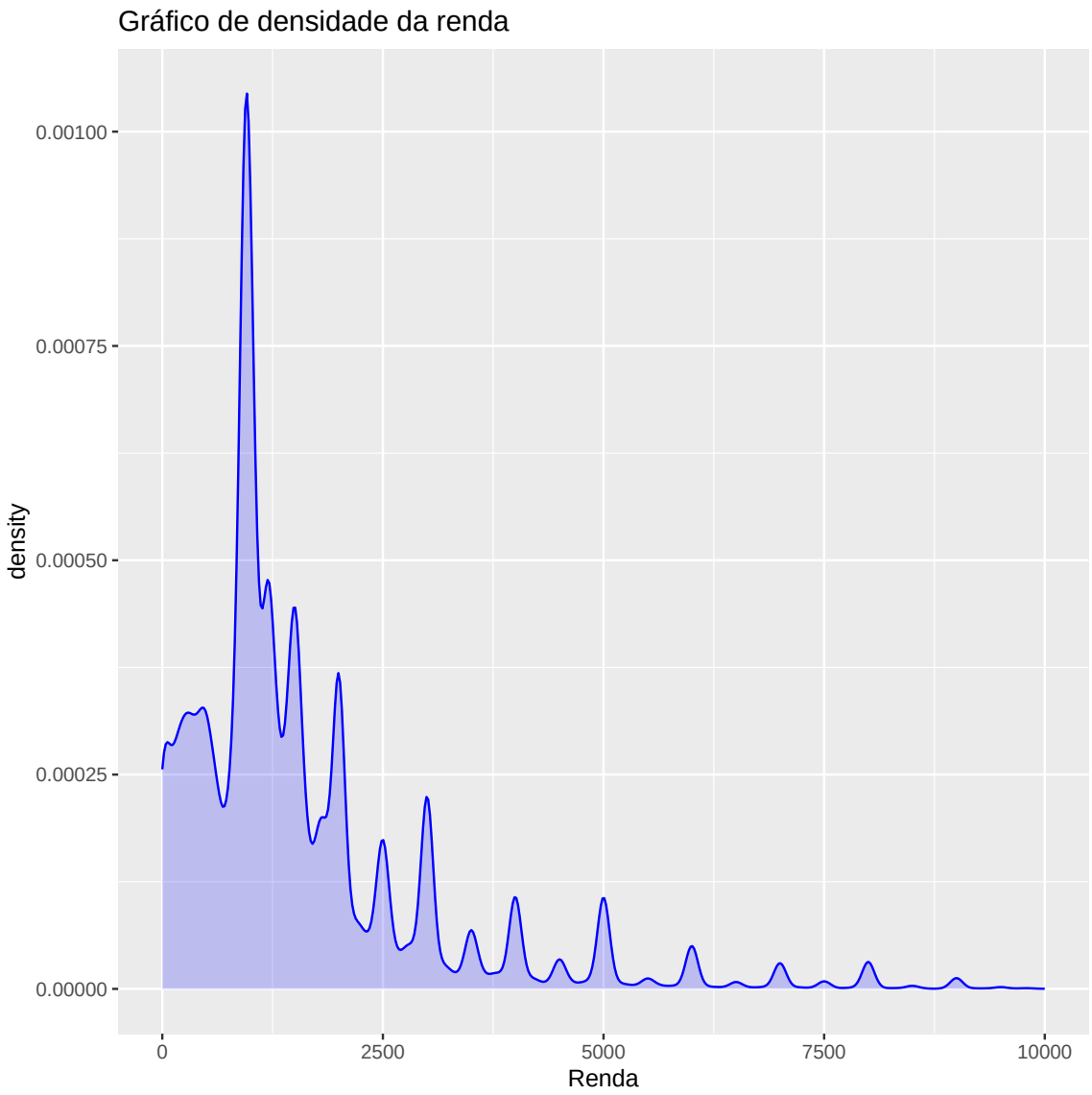
ii) qual é a probabilidade de que, ao retirarmos aleatoriamente uma observação (um indivíduo) dessa base de dados, sua renda seja estritamete maior do que 1000 e estritamente menor do que 2000 reais? Apenas para propósitos didáticos, ignore o erro amostral e trate a sua base de dados como uma população (não faça isso em sua pesquisa);

iii) apresente um gráfico de densidade da renda dado que as horas trabalhadas (Y) sejam menores ou iguais a 20;

iv) calcule:

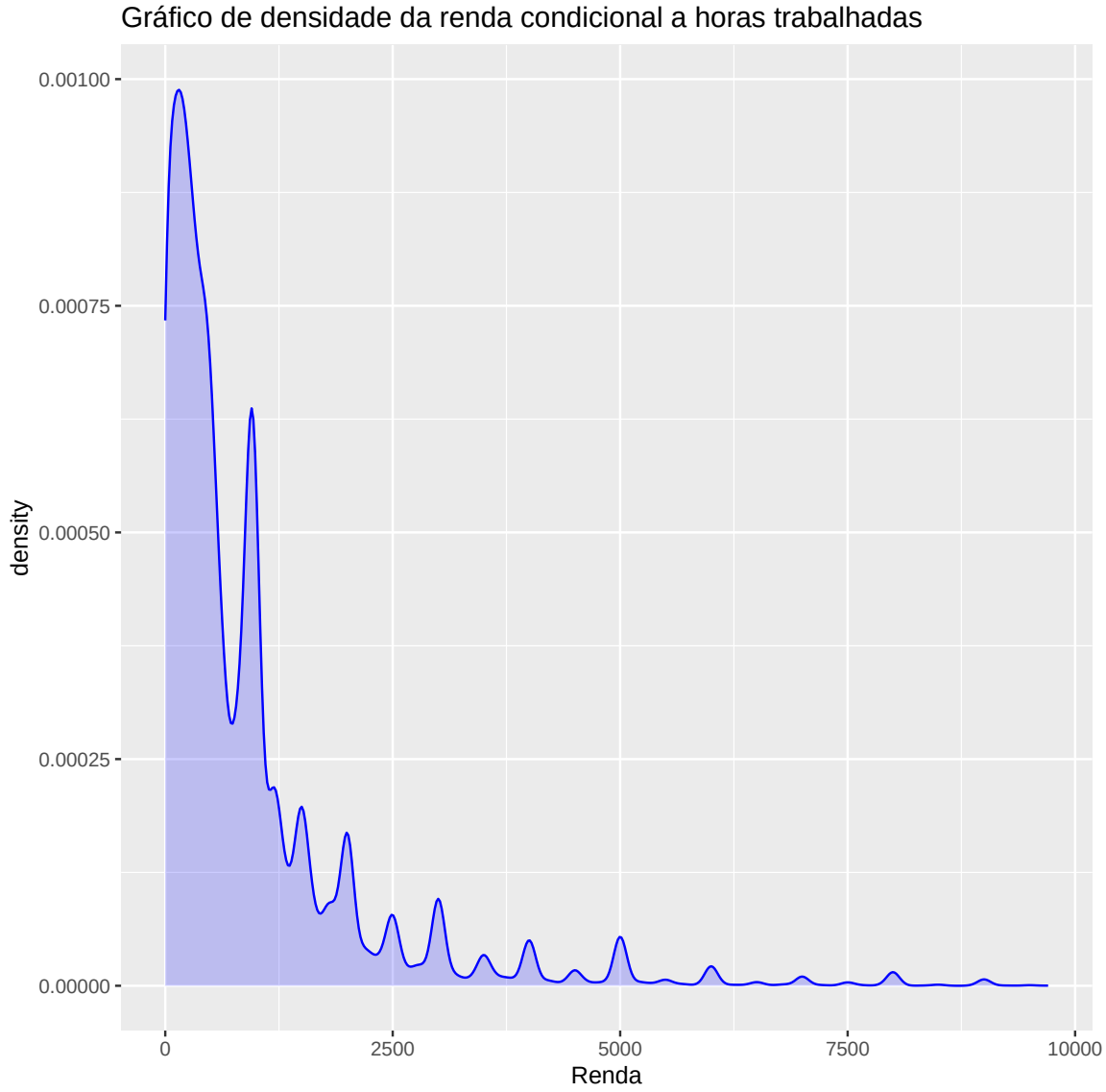
$$P(1000 < X < 2000 | Y \leq 20) \tag{4}$$

```r
library(ggplot2)

# i)

data %>%
  filter(Renda < 10000) %>%
  ggplot() +
  geom_density(aes(x=Renda), colour="blue", fill="blue", alpha=0.2) +
  ggtitle("Gráfico de densidade da renda")
```




```r
# ii)

m <- data %>% filter(Renda < 10000) %>% tally()

n <- data %>% filter(Renda < 10000) %>%
  filter(Renda > 1000 & Renda < 2000) %>%
  tally()

n[[1]]/m[[1]] # 27,34%
```

```
## [1] 0.2734149

# iii)

data %>%
 filter(Renda < 10000) %>%
 filter(Horas_trabalhadas <= 20) %>%
  ggplot() +
  geom_density(aes(x=Renda), colour="blue", fill="blue", alpha=0.2) +
  ggtitle("Gráfico de densidade da renda condicional a horas trabalhadas")
```



```r
# iv)

w <- data %>% filter(Renda < 10000) %>% tally()
z <- data %>% filter(Renda < 10000) %>%
  filter(Renda > 1000 & Renda < 2000 & Horas_trabalhadas <= 20) %>%
  tally()

z[[1]]/w[[1]] # 1,93%

## [1] 0.01928995
```

7

Mostre que:

i) Se

$$\frac{\sum_{i=1}^{n} X_i}{n} = \overline{X} \tag{5}$$

Então

$$E[7 \times X_i - 7 \times \overline{X}] = 0 \tag{6}$$

ii)

$$E[(X_i - E[X_i])^2] = E[X_i^2] - (E[X_i])^2 \tag{7}$$

8

Apresente seus resultados em um arquivo PDF. Garanta que seu arquivo esteja limpo, contendo as respostas, os gráficos e as tabelas, mas não eventuais mensagens e erros. O arquivo PDF pode ser gerado diretamente a partir do R por meio do RMarkdown ou do RSweave. Para os alunos de graduação, isso é recomendado, mas não obrigatório. Adicionalmente, forneça o script para replicação.