Métodos Quantitativos

Módulo 2: Estatística descritiva - parte 1

Pedro H. G. F. de Souza
Sergei Soares

23/09/2019

Introdução

Estatística descritiva

Hoje: primeira aula prática de fato. Foco em manipulação, visualização e análise descritiva… e também **programação**.

Motivação: “recebi uma base de dados… e agora?”

Para dar os primeiros passos, vamos deixar toda a parte de inferênciad e lado, esquecer a incerteza, e colocar o foco em explorar e entender os dados – etapa **sempre** necessária.

Leituras recomendadas

Agresti A.; Finlay B. *Statistical methods for the social sciences* (4^a edição). Nova Jersey: Prentice Hall, 2009. (p. 31-71)

Bussab W.; Morettin P. *Estatística Básica*. São Paulo: Editora Saraiva, 2010. (p. 9-67)

Field A.; Miles J.; Field Z. *Discovering Statistics using R*. Londres; Thousand Oaks, CA: Sage, 2012. (p. 19-27, p. 62-136)

Lyman R.; Longnecker M. *An Introduction to Statistical Methods and Data Analysis* (6^a edição). Belmont, CA: Brooks/Cole, Cengage Learning, 2010. (p. 88-140)

Triola, M. *Elementary Statistics* (11^a edição). Boston, Nova York: Addison-Wesley, 2012. (p. 82-134)

Venables W.; Smith D.M.; et al. *An Introduction to R*, 2019.

… todos esses textos estão no(s) site(s).

Softwares

Usar ou não o Excel?

Prós:

- Familiar e fácil de usar, com interface *point-and-click*
- Muitas funções nativas para manipulação de dados, análises estatísticas simples e visualização de gráficos

Contras:

- Software comercial
- Não aceita comandos por script, o que torna muito fácil cometer erros bobos
- Não dá conta de procedimentos mais complexos ou bases de dados muito grandes

Softwares estatísticos

SPSS

- **Prós**: simples, tradicional e fácil de aprender
- **Contras**: comercial, lento, funcionalidades limitadas

Stata

- **Prós**: excelente para estatística e muito bom para manipulação de dados e gráficos; rápido e eficiente; aprendizado gradual; muitas funções criadas pela comunidade
- **Contra**: comercial, ruim para dados “não tradicionais”, evolução lenta

R

- **Prós**: gratuito; eficiente; excelente para tudo (em especial dados “não tradicionais”); evolução fulminante por ser *open-source*; infinitas funções criadas pela comunidade
- **Contras**: mais difícil de aprender (código verborrágico e documentação chinfrim)

Guia para o R

Instalação:

- R → http://www.r-project.org/
- R Studio → http://www.rstudio.com/

Livros

- Field et al, "Discovering Statistics Using R" → http://gen.lib.rus.ec ("dizem")
- Grolemund & Wickham, "R for data science" → http://r4ds.had.co.nz/

Tutoriais

- Venables et al → http://cran.r-project.org/doc/manuals/r-release/R-intro.pdf
- r-Tutor.com → http://www.r-tutor.com/r-introduction
- computerworld.com → http://www.computerworld.com/article/2497143/business-intelligence-beginner-s-guide-to-r-introduction.html

… tem muita, muita coisa no Google, inclusive em português.

Primeiro contato

```
#--- Pasta de trabalho ---#

# Limpar o espaco de trabalho
  rm(list=ls())

# Cria objeto com o caminho da pasta de trabalho (por ex., se for )
# "c:\metodos\", altere para file.path("c:","metodos")
  path <- file.path("d:","onedrive","work","incompletos",
                    "aulas","metodos_quant","modulo2")

# Definir  pasta de trabalho
  setwd(path)

# Mostrar a pasta de trabalho atual
  getwd()
```

```
## [1] "d:/onedrive/work/incompletos/aulas/metodos_quant/modulo2"
```

Pacotes

Os pacotes feitos por usuário são a alma do R. Eles precisam ser instalados uma única vez, mas devem ser carregados a cada execução.

```
# Instalação de alguns pacotes muito úteis:
  #install.packages(c("dplyr","ggplot2","png","readr",readxl",
  #                   "scales","summarytools","writexl"))

# Carrega os pacotes
  library(dplyr)
  library(ggplot2)
  library(png)
  library(readr)
  library(readxl)
  library(scales)
  library(summarytools)
  library(writexl)
```

Abrindo uma base de dados

```
# Importa XLSX da pesquisa da avaliacao que esta em './dados/modulo2'
# (a funcao read_xlsx é do pacote 'readxl')
  pesqaval <- read_xlsx("./dados/pesquisa_avaliacao.xlsx")

# Lista primeiras cinco linhas da pesquisa
  head(pesqaval, n = 5)

# Mostra estrutura de variaveis
  str(pesqaval)

# Mostra tipo de dados
  class(pesqaval)
```

Para dados em csv, usar a função “read_csv” do pacote *readr*.

Para consultar a documentação de qualquer função carregada, digitar “?funcao” no console.

head(pesqaval, n = 5)

```
## # A tibble: 5 x 21
##      id q1_horas_trabal~ q2_horas_estudo q3_cursos_anter~ q4_mestrado_qua~
##   <dbl>            <dbl>           <dbl> <chr>            <chr>
## 1     1               60               8 Sim, apenas uma  Talvez, não dec~
## 2     2               40               3 Não, nenhuma     Talvez, não dec~
## 3     3               32               2 Sim, mais de uma Talvez, não dec~
## 4     4               40               7 Sim, apenas uma  Sim, foco princ~
## 5     5               40               2 Sim, apenas uma  Talvez, não dec~
## # ... with 16 more variables: q5_conhecimento_atual <dbl>,
## #   q6_vontade_aprender <dbl>, q7_avaliacao <dbl>, q8_aprendizado <dbl>,
## #   q9a_objetivos_claros <chr>, q9b_organizacao_adequada <chr>,
## #   q9c_professores_preparados <chr>, q9d_explicacao_clara <chr>,
## #   q9e_correspondeu_expectativas <chr>,
## #   q10a_dificuldade_nivelamento <chr>, q10b_ritmo_aulas <chr>,
## #   q10c_carga_horaria <chr>, q10d_num_exercicios <chr>,
## #   q11_assunto_mais_util <chr>, q12_assunto_menos_util <chr>,
## #   q13_sugestoes <chr>
```

str(pesqaval)

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    25 obs. of  21 variables:
##  $ id                              : num  1 2 3 4 5 6 7 8 9 10 ...
##  $ q1_horas_trabalho               : num  60 40 32 40 40 35 35 40 40 36 ...
##  $ q2_horas_estudo                 : num  8 3 2 7 2 8 2 8 2 3 ...
##  $ q3_cursos_anteriores            : chr  "Sim, apenas uma" "Não, nenhuma" "Sim, mais de uma
##  $ q4_mestrado_quant               : chr  "Talvez, não decidi" "Talvez, não decidi" "Talvez,
##  $ q5_conhecimento_atual           : num  2 3 4 2 3 4 4 3 3 2 ...
##  $ q6_vontade_aprender             : num  7 4 6 7 6 7 6 7 7 7 ...
##  $ q7_avaliacao                    : num  7 5 6 6 4 7 7 4 5 6 ...
##  $ q8_aprendizado                  : num  6 5 6 7 4 7 6 3 5 5 ...
##  $ q9a_objetivos_claros            : chr  "Concordo fortemente" "Concordo" "Concordo forteme
##  $ q9b_organizacao_adequada        : chr  "Concordo fortemente" "Discordo" "Concordo" "Conco
##  $ q9c_professores_preparados      : chr  "Concordo fortemente" "Concordo" "Concordo" "Conco
##  $ q9d_explicacao_clara            : chr  "Concordo fortemente" "Discordo" "Concordo" "Conco
##  $ q9e_correspondeu_expectativas: chr  "Concordo fortemente" "Discordo" "Concordo" "Conco
##  $ q10a_dificuldade_nivelamento : chr  "Adequado" "Adequado" "Adequado" "Adequado" ...
##  $ q10b_ritmo_aulas                : chr  "Adequado" NA "Adequado" "Lento demais" ...
##  $ q10c_carga_horaria              : chr  "Adequado" "Insuficiente" "Insuficiente" "Adequado
##  $ q10d_num_exercicios             : chr  "Adequado" "Adequado" "Insuficiente" "Adequado" ..
##  $ q11_assunto_mais_util           : chr  "Somatório" NA "Lei de Zipf" "Desenhos experimenta
```

Pesquisa de avaliação

"View(pesqaval)" abre a janela com os dados:

Filter

	id	q1_horas_trabalho	q2_horas_estudo	q3_cursos_anteriores	q4_mestrado_quant	q5_conhecimento_atual	q6_vontade_aprender	q7_ava
1	1	60	8.0	Sim, apenas uma	Talvez, não decidi	2	7	7
2	2	40	3.0	Não, nenhuma	Talvez, não decidi	3	4	5
3	3	32	2.0	Sim, mais de uma	Talvez, não decidi	4	6	6
4	4	40	7.0	Sim, apenas uma	Sim, como foco principal	2	7	6
5	5	40	2.0	Sim, apenas uma	Talvez, não decidi	3	6	4
6	6	35	8.0	Não, nenhuma	Sim, de modo secundário	4	7	7
7	7	35	2.0	Sim, mais de uma	Sim, como foco principal	4	6	7
8	8	40	8.0	Sim, apenas uma	Sim, de modo secundário	3	7	4
9	9	40	2.0	Sim, mais de uma	Talvez, não decidi	3	7	5
10	10	36	3.0	Sim, apenas uma	Não, de jeito nenhum	2	7	6
11	11	40	3.0	Sim, mais de uma	Talvez, não decidi	4	5	4
12	12	40	4.0	Sim, apenas uma	Talvez, não decidi	3	7	6
13	13	40	4.0	Sim, mais de uma	Talvez, não decidi	3	5	5
14	14	40	4.0	Sim, mais de uma	Talvez, não decidi	4	7	6
15	15	40	3.0	Sim, apenas uma	Sim, como foco principal	3	6	6
16	16	30	10.0	Não, nenhuma	Talvez, não decidi	3	7	6
17	17	40	2.0	Não, nenhuma	Talvez, não decidi	1	4	3

Showing 1 to 20 of 25 entries, 21 total columns

Distribuição de frequências

Variáveis qualitativas

```
# Q4. Pretende fazer análises quantitativas no mestrado?
  freq(pesqaval$q4_mestrado_quant, headings=FALSE, round.digits=1)
```

```
##
##                          Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
## ------------------------- ------ --------- -------------- --------- --------------
##       Não, de jeito nenhum      1       4.5            4.5       4.0            4.0
##        Sim, foco principal      7      31.8           36.4      28.0           32.0
##       Sim, foco secundário      3      13.6           50.0      12.0           44.0
##         Talvez, não decidi     11      50.0          100.0      44.0           88.0
##                       <NA>      3                                12.0          100.0
##                      Total     25     100.0          100.0     100.0          100.0
```

Variáveis quantitativas discretas

```
# Q5. Como avalia seu conhecimento atual?
  freq(pesqaval$q5_conhecimento_atual,headings=FALSE,
                                      round.digits=1)
```

```
##
##              Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
## ----------- ------ --------- -------------- --------- --------------
##           1      1       4.5            4.5       4.0            4.0
##           2      3      13.6           18.2      12.0           16.0
##           3      9      40.9           59.1      36.0           52.0
##           4      7      31.8           90.9      28.0           80.0
##           5      1       4.5           95.5       4.0           84.0
##           6      1       4.5          100.0       4.0           88.0
##        <NA>      3                              12.0          100.0
##       Total     25     100.0          100.0     100.0          100.0
```

Variáveis no R: character vs. *factor*

```
# Q6. Qual sua vontade de aprender mais? (primeira tentativa)
  freq(pesqaval$q6_vontade_aprender, headings=FALSE,
                                     round.digits=1)
```

```
##
##              Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
## ----------- ------ --------- -------------- --------- --------------
##           4      2       9.1            9.1       8.0            8.0
##           5      3      13.6           22.7      12.0           20.0
##           6      5      22.7           45.5      20.0           40.0
##           7     12      54.5          100.0      48.0           88.0
##        <NA>      3                                12.0          100.0
##       Total     25     100.0          100.0     100.0          100.0
```

Variáveis no R: *character* vs. factor

```
# Q6. Qual sua vontade de aprender mais? (segunda tentativa)
  pesqaval <- pesqaval %>%
                mutate(q6_factor = factor(q6_vontade_aprender,
                                          levels = 1:7))
  freq(pesqaval$q6_factor, headings=FALSE, round.digits=1)
```

```
##
##              Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
## ----------- ------ --------- -------------- --------- --------------
##           1      0       0.0            0.0       0.0            0.0
##           2      0       0.0            0.0       0.0            0.0
##           3      0       0.0            0.0       0.0            0.0
##           4      2       9.1            9.1       8.0            8.0
##           5      3      13.6           22.7      12.0           20.0
##           6      5      22.7           45.5      20.0           40.0
##           7     12      54.5          100.0      48.0           88.0
##        <NA>      3                                12.0          100.0
##       Total     25     100.0          100.0     100.0          100.0
```

Variáveis contínuas discretizadas (chr)

```
# Recodificando variavel de horas de estudo e pedindo frequencias
pesqaval <- pesqaval %>%
              mutate(q2_recode =
                     ifelse(q2_horas_estudo<3,"Pouco",
                     ifelse(q2_horas_estudo<7,"Médio","Muito")))

freq(pesqaval$q2_recode, headings=FALSE, round.digits=1)
```

```
##
##              Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
## ----------- ------ --------- -------------- --------- --------------
##       Médio      9      42.9           42.9      36.0           36.0
##       Muito      6      28.6           71.4      24.0           60.0
##       Pouco      6      28.6          100.0      24.0           84.0
##        <NA>      4                                16.0          100.0
##       Total     25     100.0          100.0     100.0          100.0
```

Vars contínuas discretizadas (factor)

```
# Recodificando, convertendo em factor e pedindo frequencias
pesqaval <- pesqaval %>%
              mutate(q2_recode =
                     ifelse(q2_horas_estudo<3,"Pouco",
                     ifelse(q2_horas_estudo<7,"Médio","Muito"))) %>%
              mutate(q2_recode = factor(q2_recode, levels = c("Pouco",
                                        "Médio","Muito")))
freq(pesqaval$q2_recode, headings=FALSE, round.digits=1)
```

```
##
##              Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
## ----------- ------ --------- -------------- --------- --------------
##       Pouco      6      28.6           28.6      24.0           24.0
##       Médio      9      42.9           71.4      36.0           60.0
##       Muito      6      28.6          100.0      24.0           84.0
##        <NA>      4                                16.0          100.0
##       Total     25     100.0          100.0     100.0          100.0
```

Exercício

- Recodificar a variável “q1_horas_trabalho” criando uma nova variável “q1_recode” com duas três categorias:
  - “Não trabalha” = zero horas
  - “Parcial” = entre 1 e 35 horas
  - “Integral” = 36 horas ou mais
- Converter a variável “q1_recode” em *factor* com o nome “q1_recfac”
- Obter tabela de frequências de “q1_recode” e “q1_recfac”

Exercício

```
# Recodificando

q1 <- pesqaval$q1_horas_trabalho

pesqaval <-
  pesqaval %>%
    mutate(q1_recode = ifelse(q1==0,"Não trabalha",
                        ifelse(q1<=35,"Parcial","Integral"))) %>%
    mutate(q1_recfac = factor(q1_recode, levels =
                              c("Não trabalha","Parcial","Integral")))
```

Exercício

```
freq(pesqaval$q1_recode, headings=FALSE, round.digits=1, report.nas = FALSE)
```

```
##
##                 Freq       %   % Cum.
## -------------- ------ ------- --------
##       Integral     14    63.6     63.6
##        Parcial      8    36.4    100.0
##          Total     22   100.0    100.0
```

```
freq(pesqaval$q1_recfac, headings=FALSE, round.digits=1, report.nas = FALSE)
```

```
##
##                   Freq       %   % Cum.
## ----------------- ------ ------- --------
##       Não trabalha      0     0.0      0.0
##            Parcial      8    36.4     36.4
##           Integral     14    63.6    100.0
##              Total     22   100.0    100.0
```

Visualização: gráficos univariados

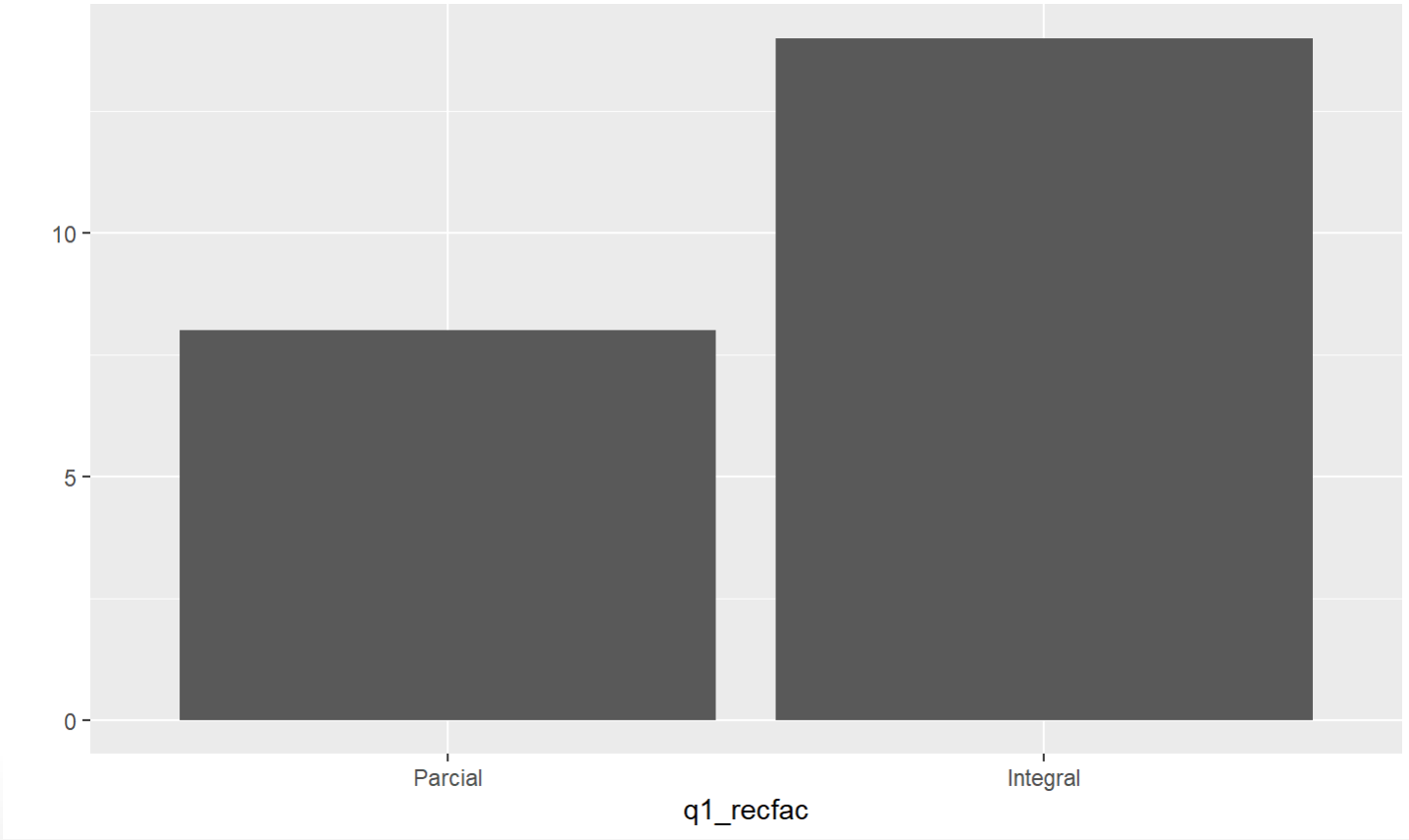
Excel ou R?

Tabelas de frequências são inconvenientes se tivermos muitas categorias → gráficos são mais fáceis de ler. Para produzi-los, temos duas opções: usar o próprio R ou o Excel, seja na base do copia-e-cola, seja exportando uma planilha do R (recomendado).

```
# Funcao freq() para frequência
  figura1 <- freq(pesqaval$q1_recfac, headings=FALSE,round.digits=1)
# Converte em data.frame
  figura1 <- data.frame(figura1, categorias = row.names(figura1))
# Renomeia colunas para facilitar
  figura1 <- figura1 %>%
               rename(freq = Freq, pct_valido = X..Valid,
                      pct_valido_acum = X..Valid.Cum.,
                      pct_total = X..Total,
                      pct_total_acum = X..Total.Cum.)
# Converte em arquivo Excel com o pacote 'writexl'
  write_xlsx(figura1, path = "./figuras/figura1_para_excel.xlsx")
```

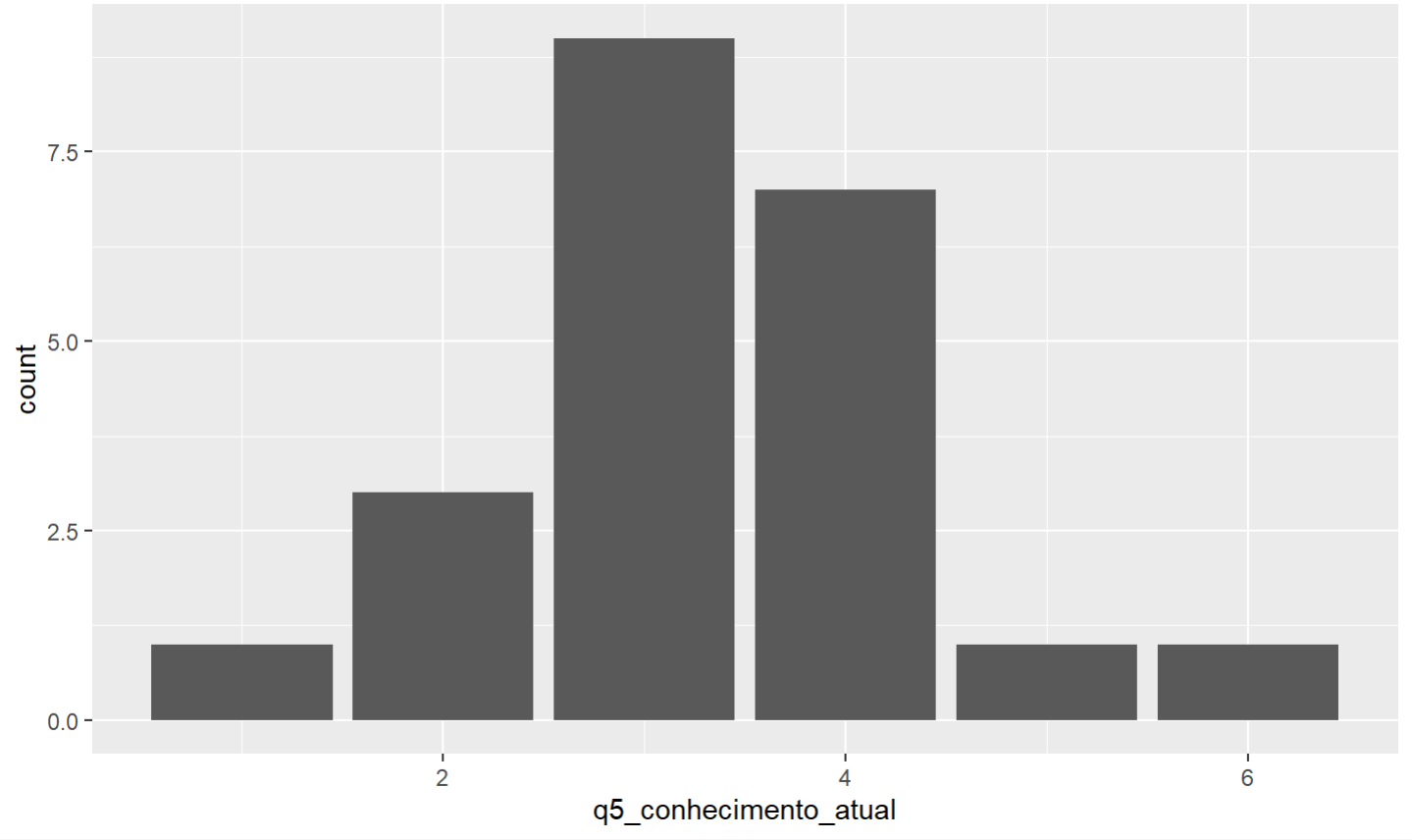
Gráficos de barras no R: qplot

```
qplot(data = subset(pesqaval, !is.na(q1_recfac)), x = q1_recfac,
      geom="bar")
```



Gráficos de barras no R: ggplot simples

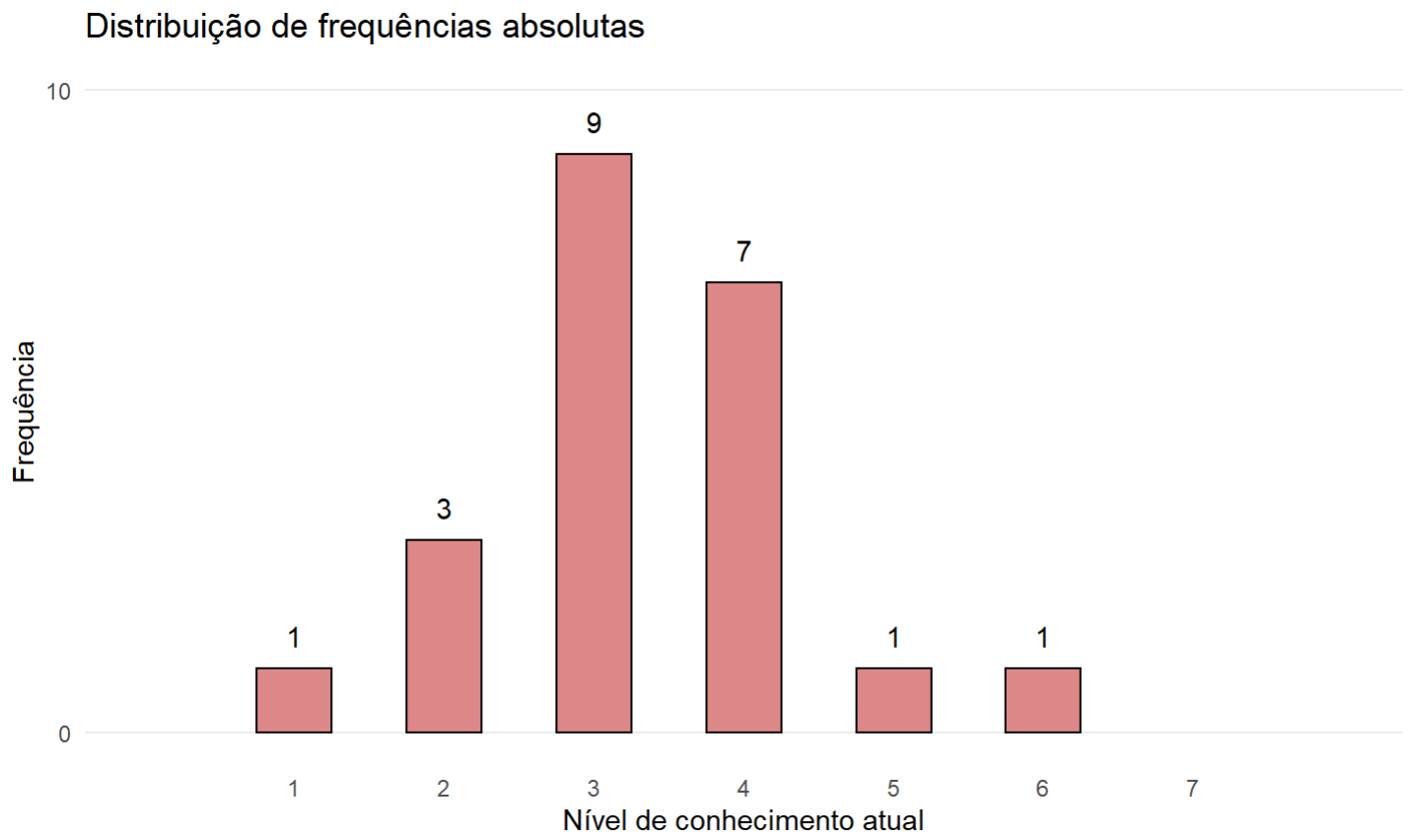
```
pesqaval %>%
  ggplot( aes(x = q5_conhecimento_atual) ) + geom_bar()
```



Gráficos de barras no R: “publicável”

```
pesqaval %>%
  ggplot( aes(x = q5_conhecimento_atual) ) +
    geom_bar( width = 0.5, color = 'black', fill = "#DD8888") +
    geom_text(stat='count', aes(label=..count..), vjust=-1) +
    scale_y_continuous(name="Frequência",limits = c(0,10),
                       breaks = c(0,10)) +
    scale_x_continuous(name="Nível de conhecimento atual",
                         limits=c(0,8),
                       breaks = c(1,2,3,4,5,6,7)) +
    ggtitle("Distribuição de frequências absolutas") +
    theme_minimal() +
    theme(panel.grid.major.x = element_blank(),
          panel.grid.minor.x = element_blank(),
          panel.grid.minor.y = element_blank())
```

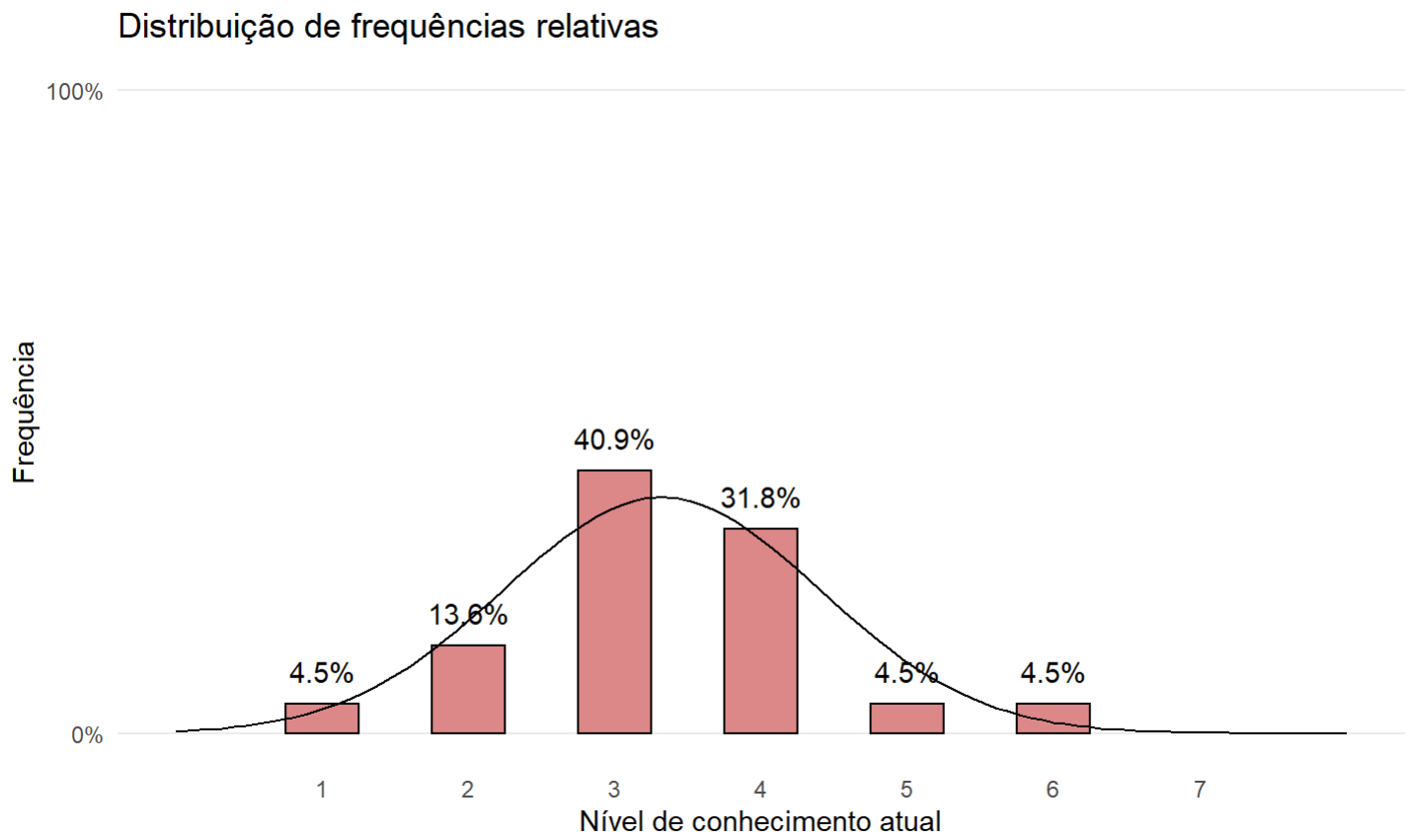
Gráficos de barras no R: “publicável”



Frequências relativas (%)

```
q5 <- pesqaval$q5_conhecimento_atual
ggplot(data = pesqaval, aes(x = q5) ) +
  geom_bar(aes(y = (..count..)/sum(..count..)),
           width = 0.5, color = 'black', fill = "#DD8888") +
  geom_text(aes(y = (..count..)/sum(..count..),
                label = percent((..count..)/sum(..count..))),
            stat='count', vjust=-1) +
  scale_y_continuous(name="Frequência",label=percent_format(),
                     limits = c(0,1), breaks = c(0,1)) +
  scale_x_continuous(name="Nível de conhecimento atual",
                     limits=c(0,8),breaks = c(1,2,3,4,5,6,7)) +
  ggtitle("Distribuição de frequências relativas") +
  theme_minimal() +
  theme(panel.grid.major.x = element_blank(),
        panel.grid.minor.x = element_blank(),
        panel.grid.minor.y = element_blank()) +
  stat_function(fun = dnorm, args = list(mean = mean(q5, na.rm=TRUE),
                                         sd = sd(q5, na.rm=TRUE)))
```

Frequências relativas (%)



Exercício

- Recodifique a variável “q7_avaliacao” e crie uma variável “q7_recode” com as categorias:
  - “Ruim”: entre 1 e 3
  - “Médio”: entre 4 e 5
  - “Bom”: entre 6 e 7
- Converta a variável “q7_recode” em *factor* com o nome “q7_recfac”
- Obtenha a tabela de frequências dessa variável
- No R ou no Excel, faça um gráfico de barras de “q7_recfac” (sem os NA’s)

Exercício

```
# Recodificando e convertendo
  pesqaval <- pesqaval %>%
                mutate(q7_recode = ifelse(q7_avaliacao<=3,"Ruim",
                                          ifelse(q7_avaliacao<=5,"Médio",
                                                 "Bom"))) %>%
                mutate(q7_recfac = factor(q7_recode,
                                          levels = c("Ruim",
                                                     "Médio",
                                                     "Bom")))
```

Exercício

```
# Frequências
  freq_q7 <- freq(pesqaval$q7_recfac, headings = FALSE, round.digits=1)
  freq_q7
```

```
##
##              Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
## ----------- ------ --------- -------------- --------- --------------
##        Ruim      1       4.5            4.5       4.0            4.0
##       Médio      8      36.4           40.9      32.0           36.0
##         Bom     13      59.1          100.0      52.0           88.0
##        <NA>      3                                12.0          100.0
##       Total     25     100.0          100.0     100.0          100.0
```

Exercício

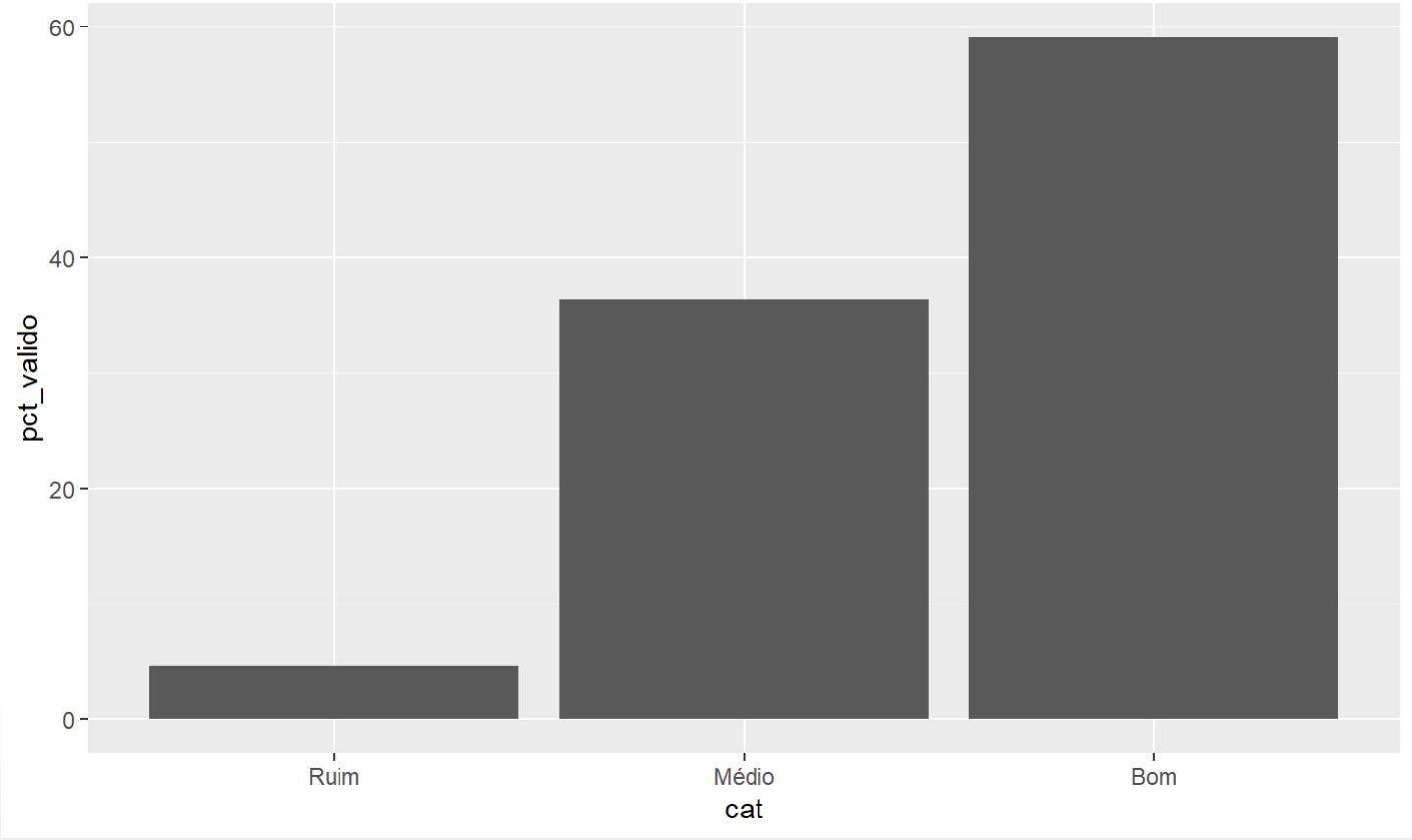
```
# Converte em data.frame
  freq_q7.df <- data.frame(freq_q7, categorias = row.names(freq_q7))

# Renomeia colunas para facilitar
  freq_q7.df <- freq_q7.df %>%
                rename(freq = Freq, pct_valido = X..Valid,
                       pct_valido_acum = X..Valid.Cum.,
                       pct_total = X..Total,
                       pct_total_acum = X..Total.Cum.)

# Converte em arquivo Excel com o pacote 'writexl'
  write_xlsx(freq_q7.df, path = "./figuras/figura2_para_excel.xlsx")
```

Exercício

```
freq_q7.df %>% filter(categorias != "<NA>" & categorias != "Total") %>%
    mutate(cat = factor(categorias, levels=c("Ruim","Médio","Bom"))) %>%
    ggplot(aes(x=cat, y=pct_valido)) + geom_bar(stat='identity')
```



Gráficos de barras vs histogramas

Para variáveis quantitativas contínuas, gráficos de barras são pouco informativos.

Nesses casos, histogramas são uma opção melhor – mas cuidado com a escolha do tamanho dos "bins".

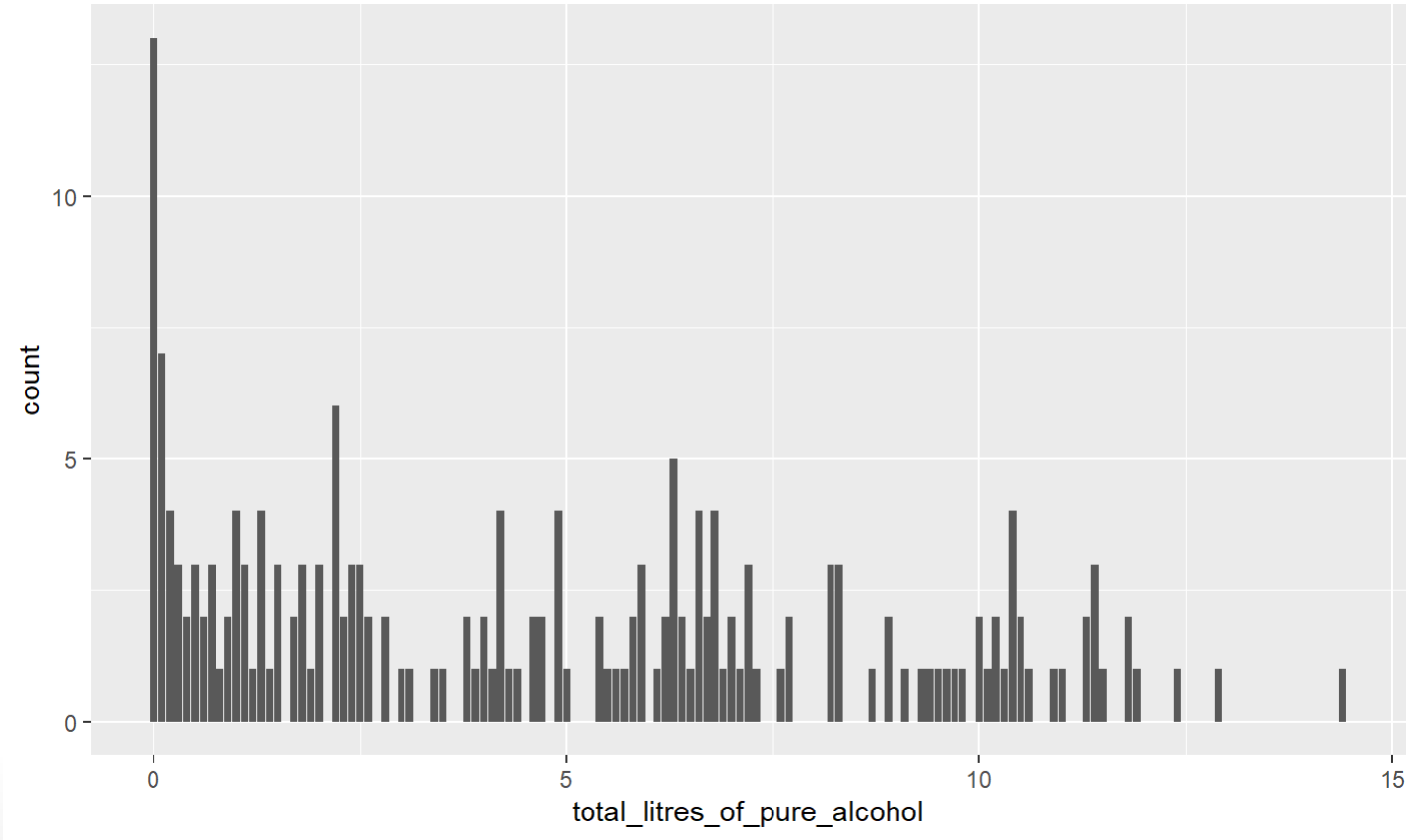
Para mostrar isso, vamos brincar com os dados compilados pelo site 538 sobre consumo anual de álcool per capita no mundo

- Reportagem: https://fivethirtyeight.com/features/dear-mona-followup-where-do-people-drink-the-most-beer-wine-and-spirits/
- Dados: https://github.com/fivethirtyeight/data/tree/master/alcohol-consumption

```
# Importando usando o 'readr'
  alcool <- read_csv("./dados/538_drinks2010.csv")
```

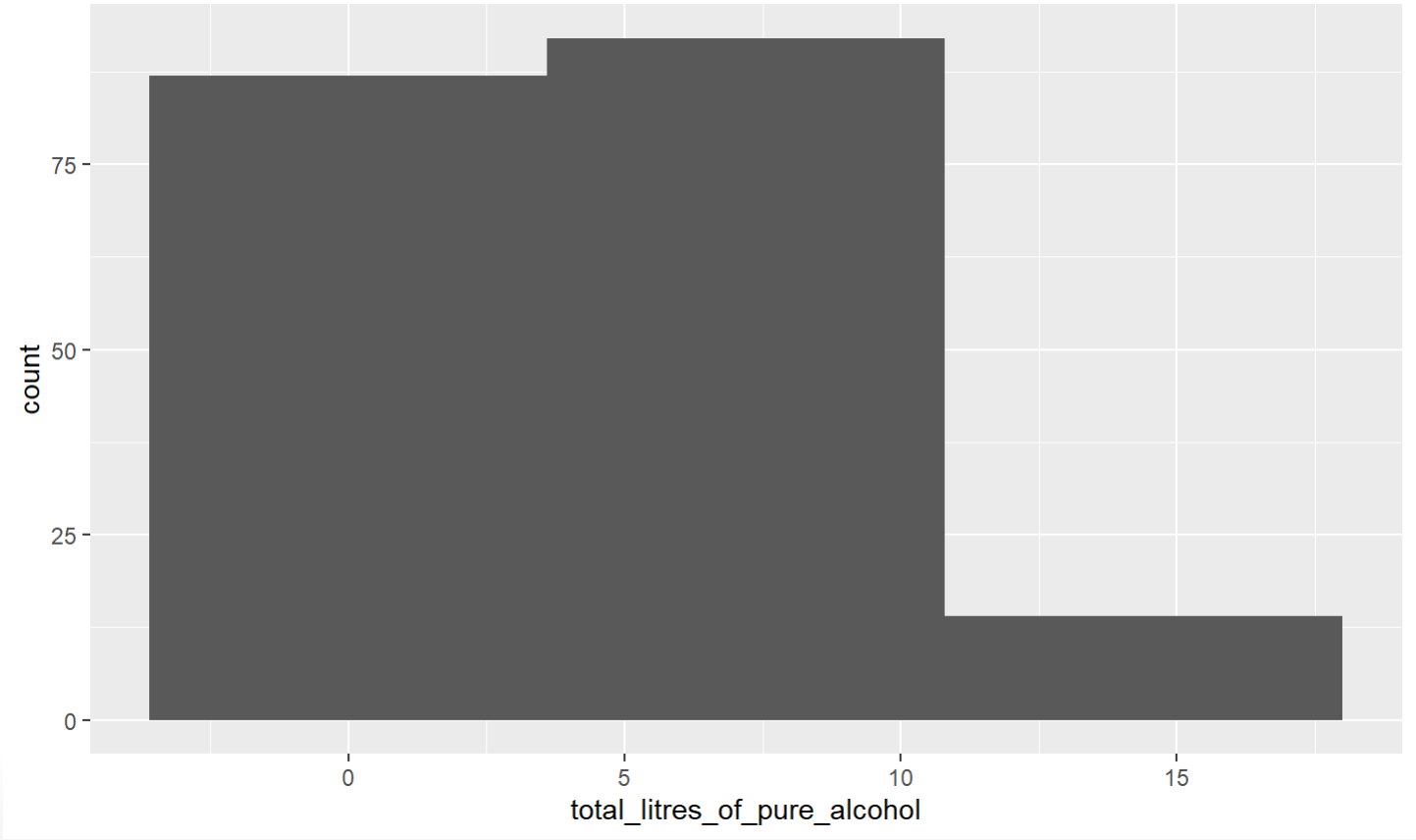
Gráfico de barras

```
ggplot(data = alcool, aes(x =  total_litres_of_pure_alcohol)) +
  geom_bar()
```



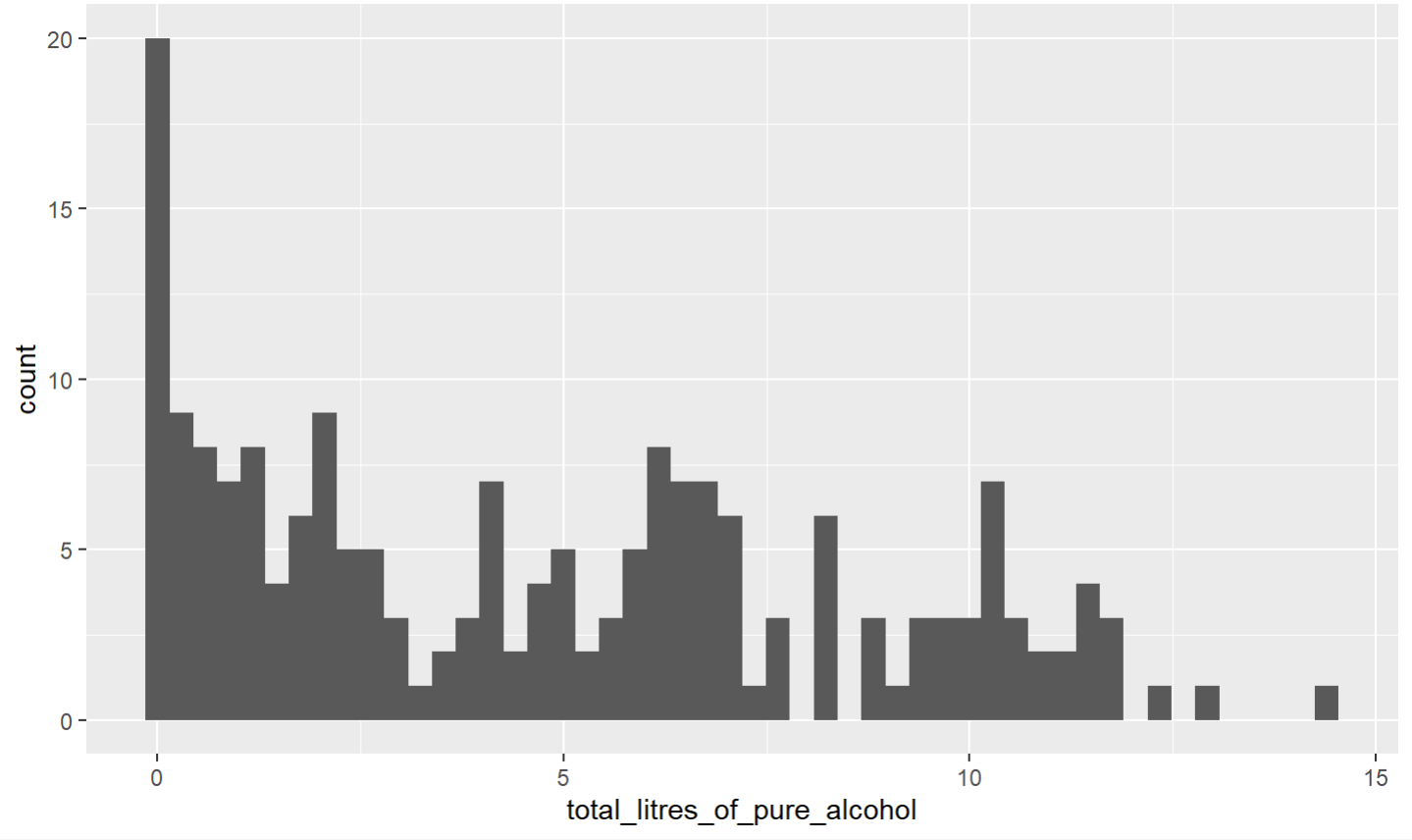
Histograma com 3 bins

```
ggplot(data = alcool, aes(x =  total_litres_of_pure_alcohol)) +
  geom_histogram(bins = 3)
```



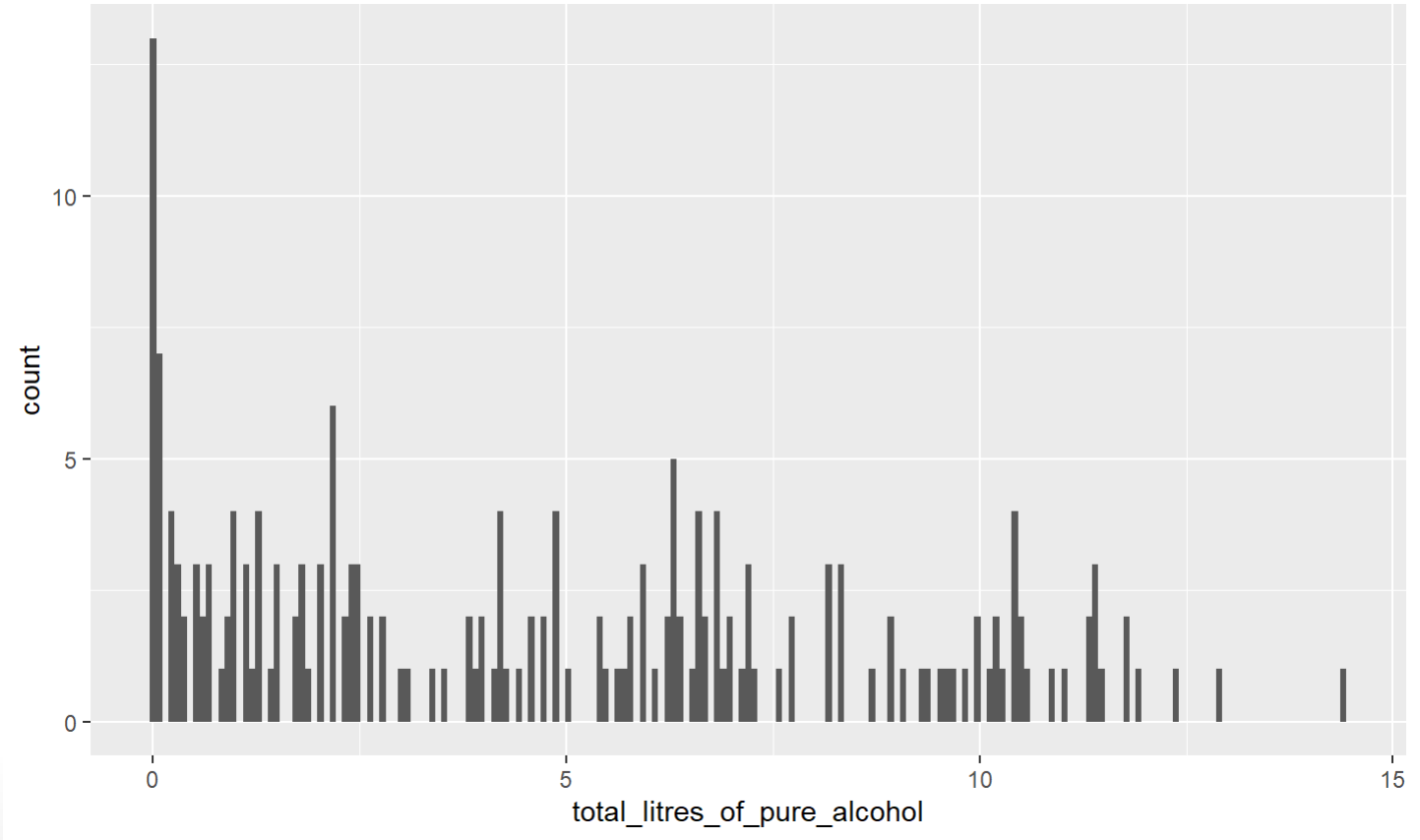
Histograma com 50 bins

```
ggplot(data = alcool, aes(x =  total_litres_of_pure_alcohol)) +
  geom_histogram(bins = 50)
```



Histograma com 193 bins

```
ggplot(data = alcool, aes(x =  total_litres_of_pure_alcohol)) +
  geom_histogram(bins = 193)
```



Estimativa de densidade kernel

Além de histogramas, podemos visualizar dados quantitativos suavizados pela técnica de *estimativa de densidade kernel*.

Basicamente, trata-se de suavizar as frequências observadas a partir de dois parâmetros: *bandwith* e *função kernel*.

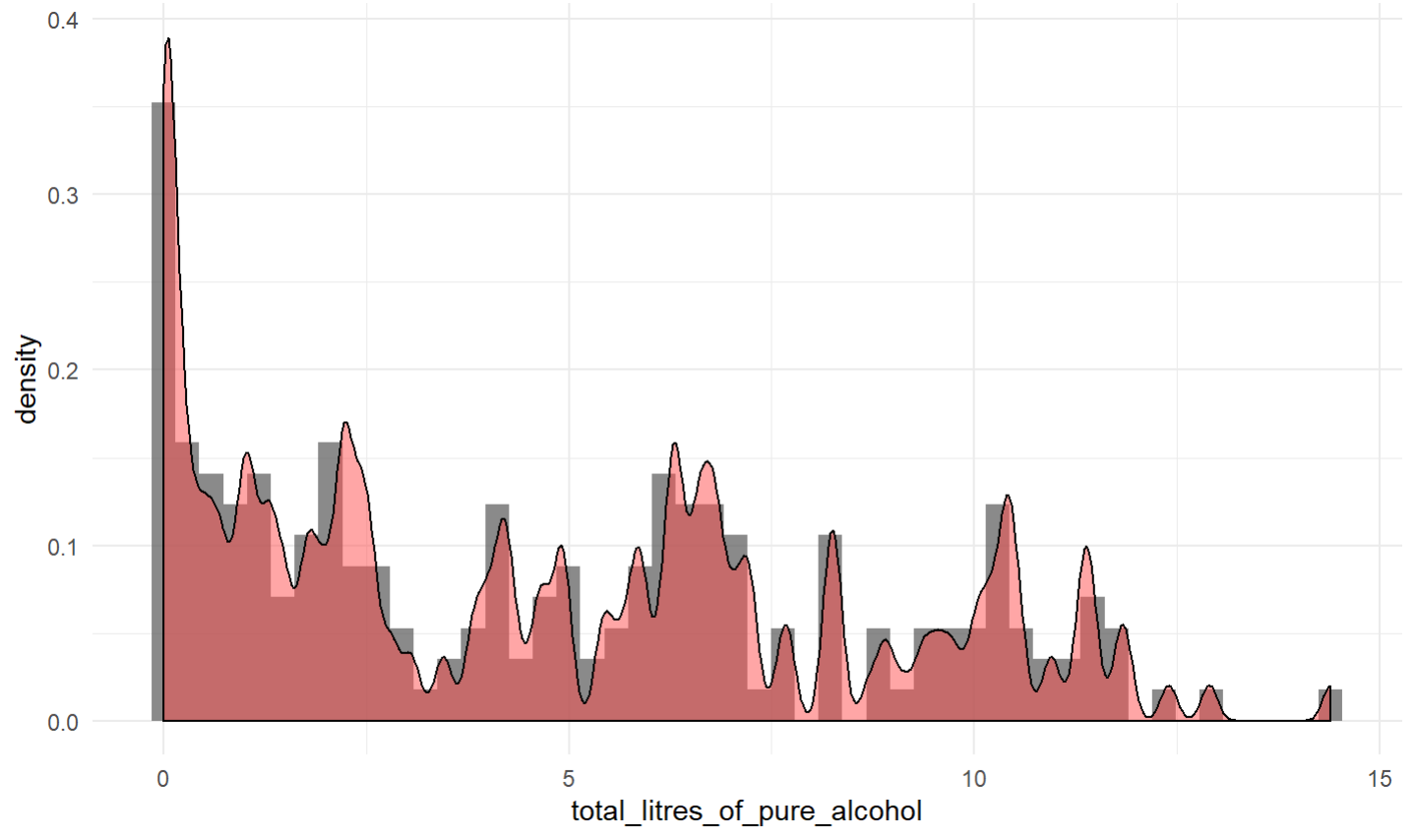
Recomendo muito o tutorial interativo de https://mathisonian.github.io/kde/.

Exemplo:

```
ggplot(data = alcool, aes(x =  total_litres_of_pure_alcohol)) +
 geom_histogram(aes(y = ..density..), alpha = 0.7,
                bins = 50, fill = "#333333") +
  geom_density(fill = "#ff4d4d", alpha = 0.5, kernel="XXXX", bw = X) +
  theme_minimal()
```

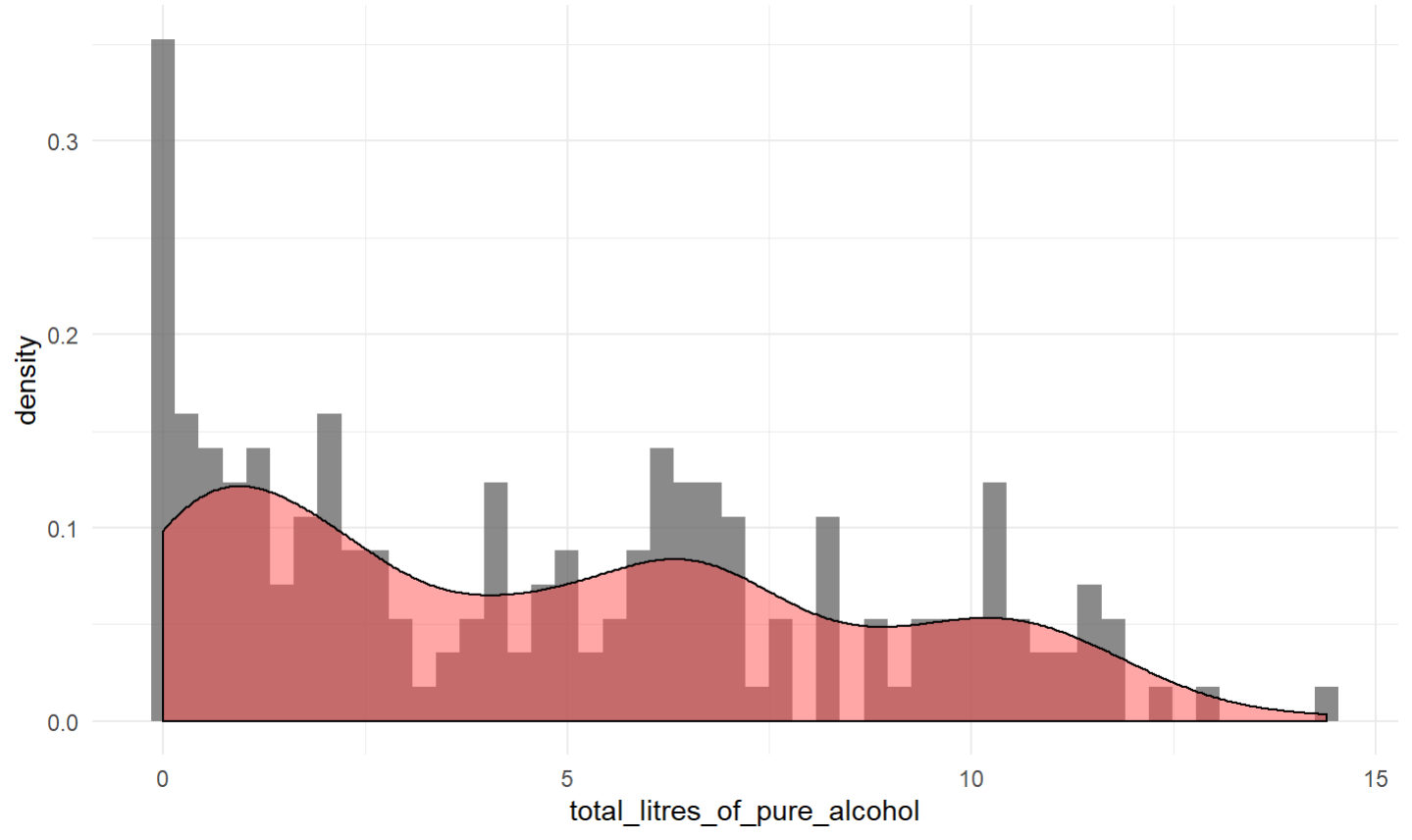
KD gaussiano, bw = 0.1

```
ggplot(data = alcool, aes(x =  total_litres_of_pure_alcohol)) +
 geom_histogram(aes(y = ..density..), alpha = 0.7, bins = 50) +
  geom_density(fill="#ff4d4d",alpha=0.5,kernel="gaussian",bw=0.1) + theme_minimal()
```



KD gaussiano, bw = 1

```
ggplot(data = alcool, aes(x =  total_litres_of_pure_alcohol)) +
 geom_histogram(aes(y = ..density..), alpha = 0.7, bins = 50) +
  geom_density(fill="#ff4d4d",alpha=0.5,kernel="gaussian",bw=1) + theme_minimal()
```



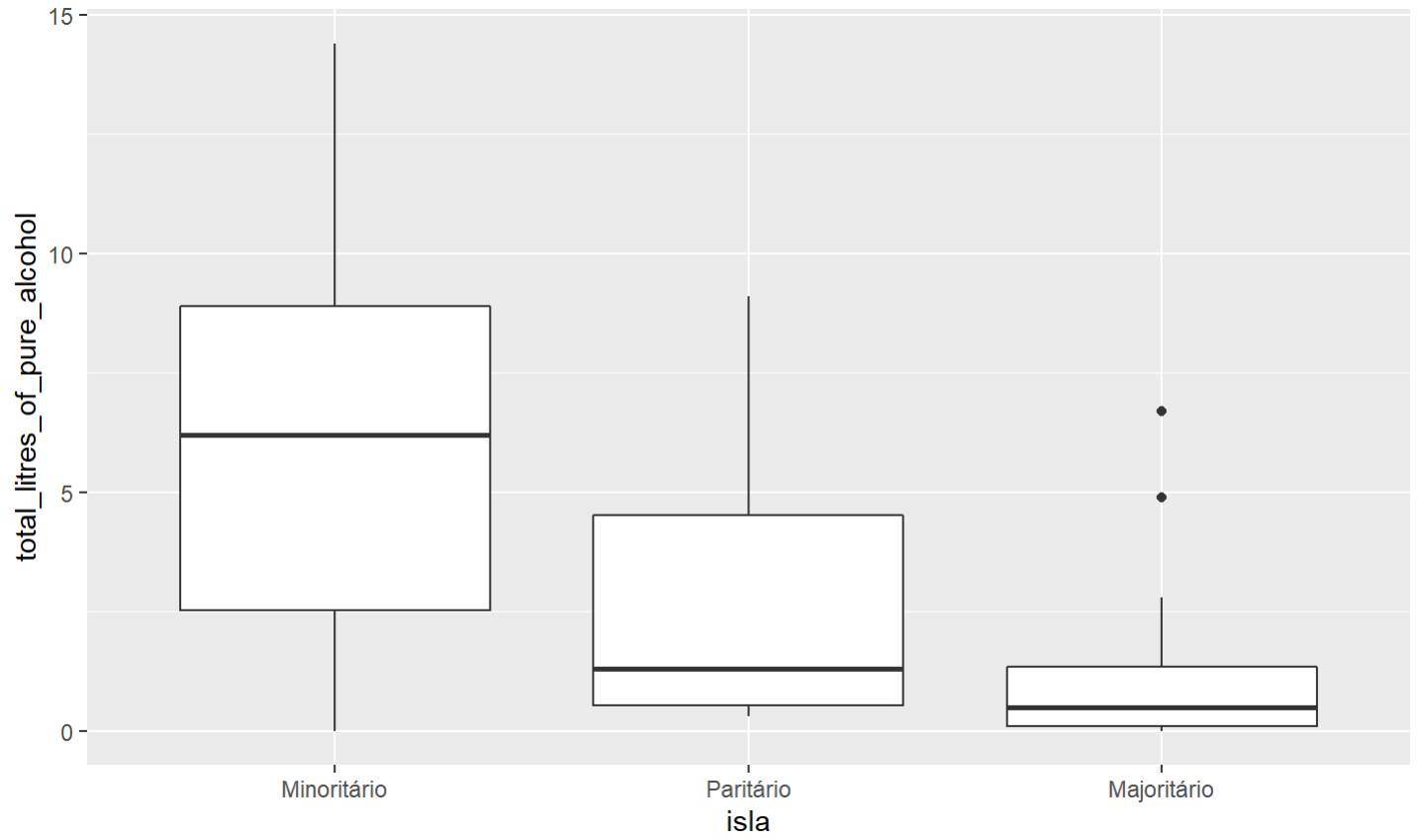
Acrescentando variável

```
isla <- read_xlsx("./dados/pew_islamismo.xlsx")
alcool <- left_join(alcool, isla, key = country)
alcool$isla <- ifelse(alcool$pct_islamico<=33,"Minoritário",
                ifelse(alcool$pct_islamico<=66,"Paritário","Majoritário"))
alcool$isla <- factor(alcool$isla,
                      levels = c("Minoritário","Paritário","Majoritário"))
freq(alcool$isla, headings = FALSE)
```

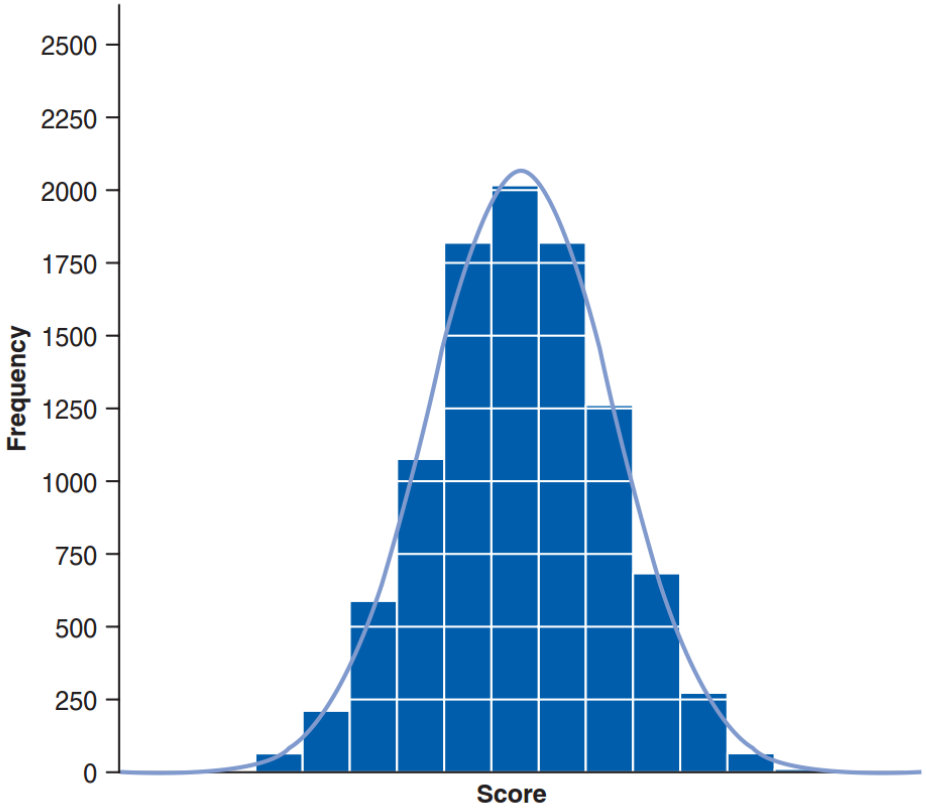
```
##
##                    Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
## ----------------- ------ --------- -------------- --------- --------------
##       Minoritário    130     72.63          72.63     67.36          67.36
##         Paritário     10      5.59          78.21      5.18          72.54
##       Majoritário     39     21.79         100.00     20.21          92.75
##              <NA>     14                                7.25         100.00
##             Total    193    100.00         100.00    100.00         100.00
```

Box plots com a variável nova

```
ggplot(data = subset(alcool, !is.na(isla)),
       aes(x=isla,y=total_litres_of_pure_alcohol)) + geom_boxplot()
```

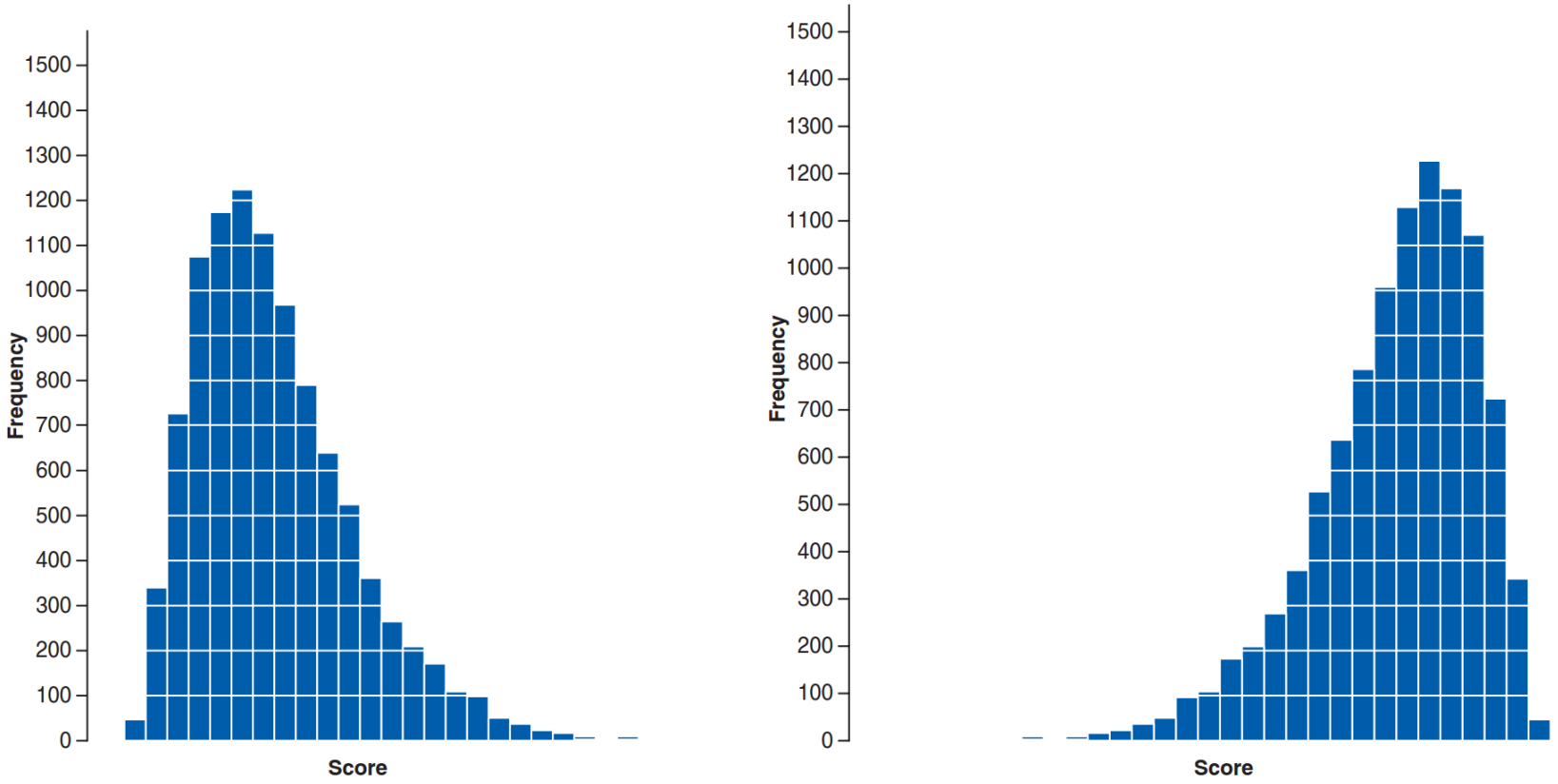


Simetria



Fonte: Field et al (2012), “Discovering statistics with R”, p. 20.

Assimetria (*skewness*)



Fonte: Field et al (2012), “Discovering statistics with R”, p. 20.

Medidas escalares

Medidas de tendência central

Média aritmética $\rightarrow \frac{1}{n}\sum_{i=1}^{n} x_i$

- Sensível a *outliers*, boa para dados lineares, aditivos, "simétricos", independência
- "Se todos os números fossem iguais, qual seria o número que gera a mesma soma total?"

Média geométrica $\rightarrow \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n}$

- Boa para dados multiplicativos, não independentes, variáveis exponenciais, lognormais, assimétricas, taxas de juros, crescimento etc (não pode ter zeros, números negativos etc)
- "Se todos os números fossem iguais, qual seria o que gera o mesmo produto total?"

Mediana $\rightarrow x_{\frac{n+1}{2}}$

- $P(X \geq m) = P(X \leq m) = 0.50$
- Insensível a *outliers*, igual à MA em distribuições simétricas

Moda $\rightarrow$ valor mais comum

- Útil para variáveis discretas, inclusive nominais/categóricas.

Dica: https://towardsdatascience.com/on-average-youre-using-the-wrong-average-geometric-harmonic-means-in-data-analysis-2a703e21ea0

Medidas de dispersão

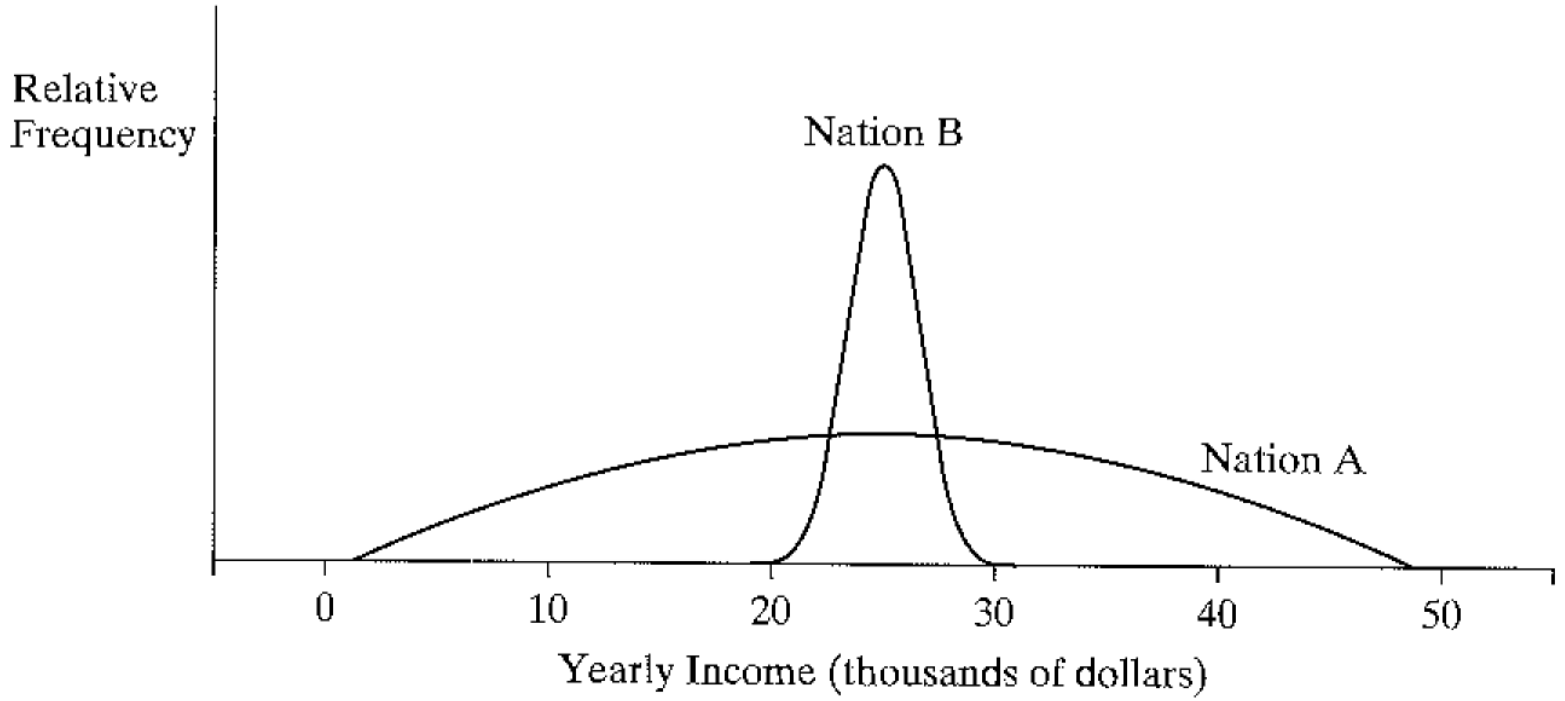


FIGURE 3.12: Distributions with the Same Mean but Different Variability

Medidas de dispersão

Amplitude (*range*) $\rightarrow max(X) - min(X)$

Amp. interquartil (*IQR*) $\rightarrow CDF^{-1}(.75) - CDF^{-1}(.25)$

Desvio médio absoluto $\rightarrow \frac{1}{n} \cdot \sum_{i=1}^{n} |x_i - \bar{x}|$

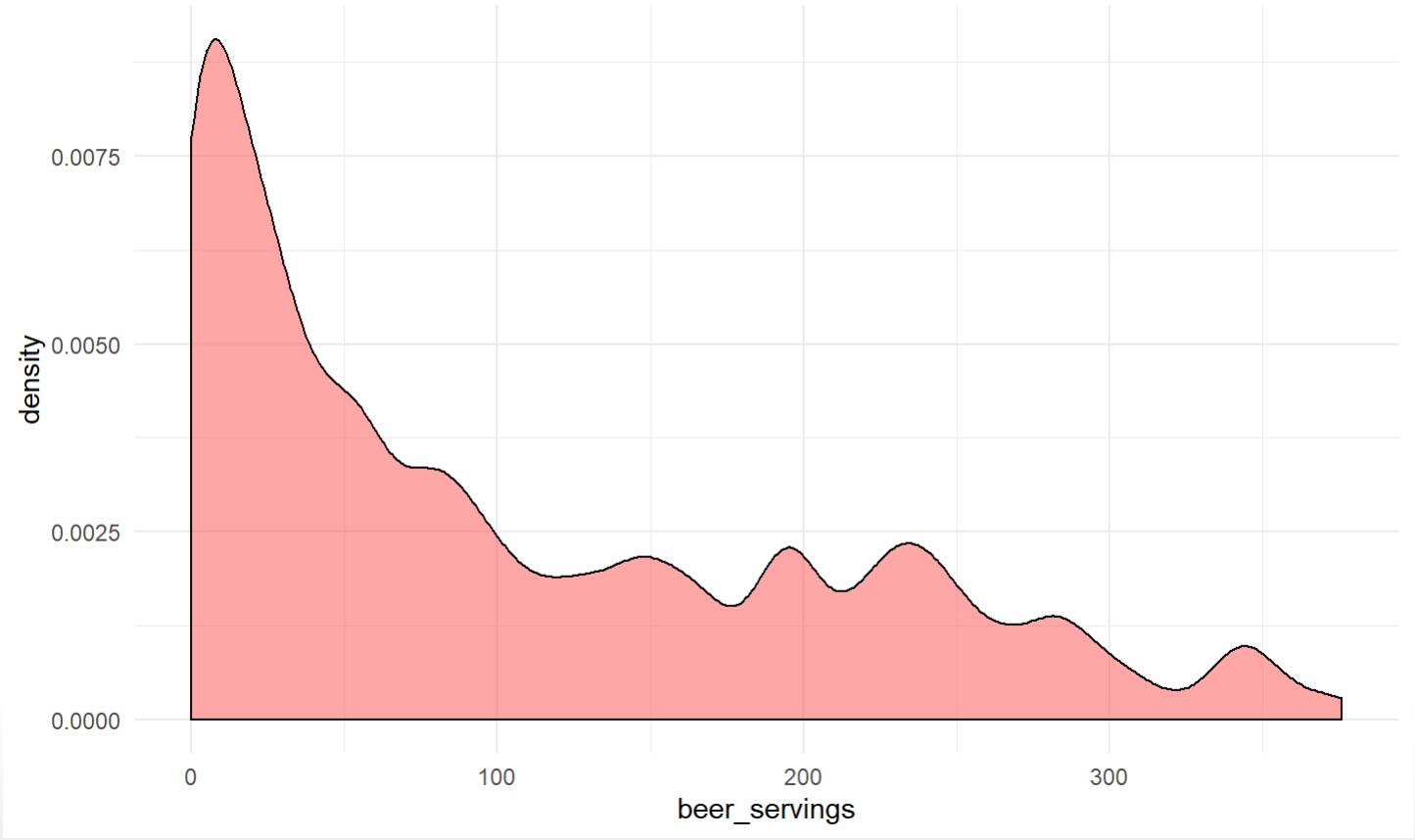
Variância $\rightarrow s^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2$

Desvio padrão $\rightarrow s = \sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2}$

Coeficiente de variação $\rightarrow CV = \frac{s}{\bar{x}}$

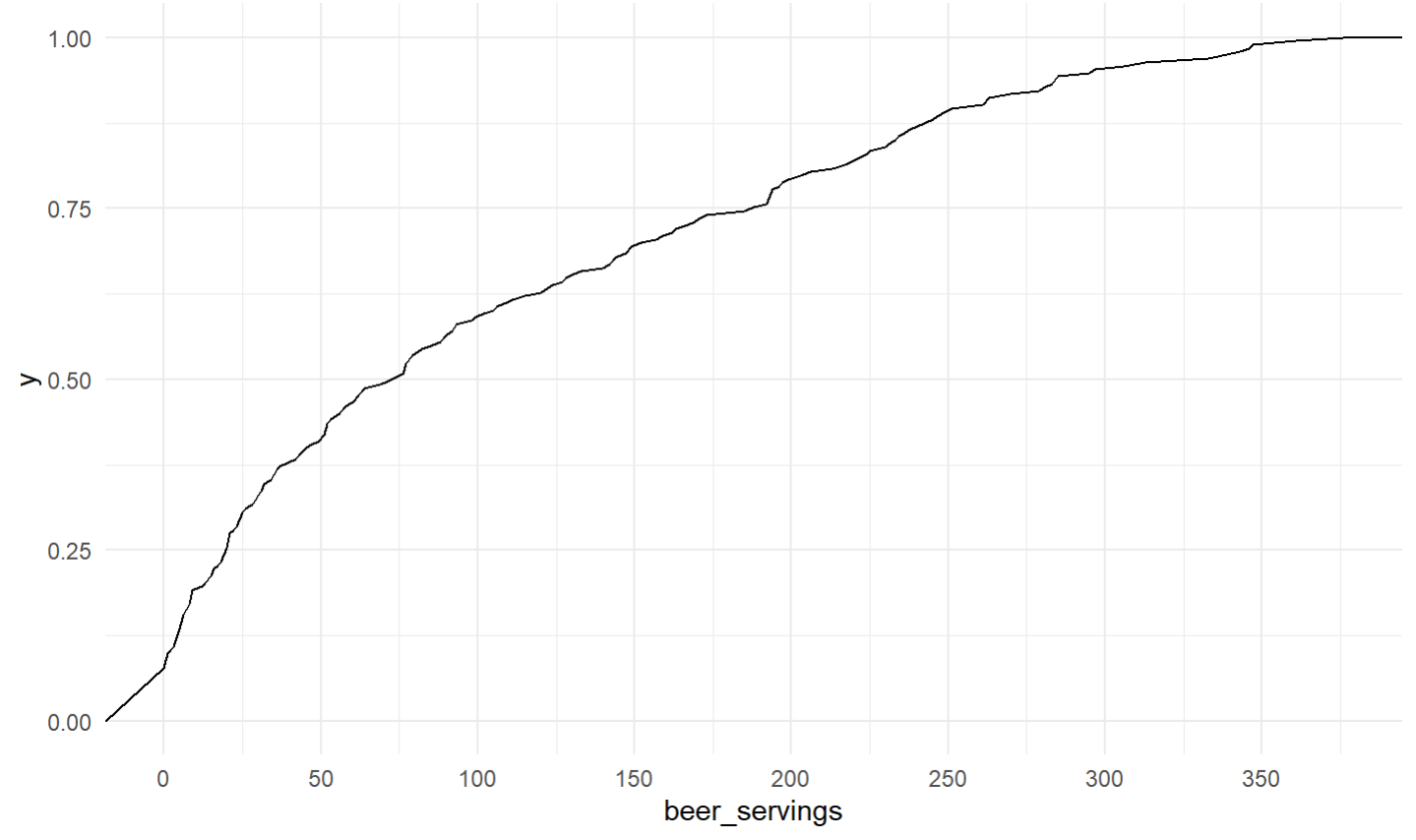
Consumo de cerveja per capita (ePDF)

```
ggplot(data = alcool, aes(x =  beer_servings)) +
  geom_density(fill = "#ff4d4d", alpha = 0.5,
               kernel="gaussian", bw = 10) + theme_minimal()
```



Consumo de cerveja per capita (eCDF)

```
ggplot(data = alcool, aes(x =  beer_servings)) +
  geom_line(aes(y = ..y..), stat='ecdf') + theme_minimal() +
  scale_x_continuous( breaks = c(0,50,100,150,200,250,300,350,400))
```



Consumo de cerveja per capita

```
alcool %>% select(ends_with("servings")) %>% descr(headings = FALSE)
```

```
##
##                   beer_servings   spirit_servings   wine_servings
## ---------------- -------------- ----------------- ---------------
##              Mean         106.16             80.99           49.45
##           Std.Dev         101.14             88.28           79.70
##               Min           0.00              0.00            0.00
##                Q1          20.00              4.00            1.00
##            Median          76.00             56.00            8.00
##                Q3         188.00            128.00           59.00
##               Max         376.00            438.00          370.00
##               MAD          99.33             80.06           11.86
##               IQR         168.00            124.00           58.00
##                CV           0.95              1.09            1.61
##          Skewness           0.80              1.27            1.88
##       SE.Skewness           0.17              0.17            0.17
##          Kurtosis          -0.51              1.36            2.78
##           N.Valid         193.00            193.00          193.00
##         Pct.Valid         100.00            100.00          100.00
```

Exercício

- Calcule medidas de tendência central e de dispersão para as variáveis de consumo de cerveja, destilados e vinho separadamente para países em que o islã é minoritário, paritário e majoritário.

Exercício

```
alcool %>% filter(isla == "Minoritário") %>%
  select(ends_with("servings")) %>% descr(headings = FALSE)

alcool %>% filter(isla == "Paritário") %>%
  select(ends_with("servings")) %>% descr(headings = FALSE)

alcool %>% filter(isla == "Majoritário") %>%
  select(ends_with("servings")) %>% descr(headings = FALSE)
```

Exercício

```
alcool %>% filter(!is.na(isla)) %>% group_by(isla) %>%
  select(ends_with("servings")) %>% descr(headings = FALSE)
```

Próxima aula

Tópicos

Transformações de variáveis;

Estatísticas bivariadas e multivariadas: covariância, correlação, associação entre variáveis categóricas.

Gráficos bivariados.

Indicadores sociais, números índices, deflatores. Mensuração da desigualdade e da pobreza.

Leituras sugeridas

Agresti A.; Finlay B. *Statistical methods for the social sciences* (4^a edição). Nova Jersey: Prentice Hall, 2009. (p. 55-58)

Bussab W.; Morettin P. *Estatística Básica*. São Paulo: Editora Saraiva, 2010. (p. 68-101)

Lyman R.; Longnecker M. *An Introduction to Statistical Methods and Data Analysis* (6^a edição). Belmont, CA: Brooks/Cole, Cengage Learning, 2010. (p. 102-140)

Medeiros M. *Uma introdução às representaçòes gráficas da desigualdade de renda*. Brasília: Ipea, 2006. (Texto para Discussão n. 1202)

Soares, S. Análise de bem-estar e decomposição por fatores da queda na desigualdade entre 1995 e 2004. *Econômica*, v. 8, n. 1, p. 83–115, 2006. Soares S. *Metodologias par estabelecer a linha de pobreza: objetivas, subjetivas, relativas, multidimensionais*. Brasília: Ipea, 2009. (Texto para Discussão n. 1381)