Café com estatística e R

Treinamento 1 - Tipos de variáveis, escalas e uma introdução ao R

Marcelo Teixeira Paiva

2025-09-29

Abstract

Relatório do primeiro treinamento onde foi apresentado uma introdução ao R e os conceitos de tipos de variáveis e escalas.

Índice

Lista de Figuras

Lista de Tabelas

Chapter 1

Fundamentos de R para análise de dados

1.1 Conhecimentos básicos de R

1.1.1 Instalação do R e RStudio

R é uma linguagem para computação estatística, enquanto **RStudio** é um ambiente de desenvolvimento integrado (IDE) que facilita o trabalho. Ou seja, o R é quem faz o trabalho pesado e o RStudio é uma das várias maneiras de se usar o R com menos esforço.

Processo de instalação: 1. Baixe o R em: https://cran.r-project.org/ 2. Baixe o RStudio em: https://posit.co/products/open-source/rstudio/?sid=1 3. Siga os passos de instalação de cada um deles em suas próprias páginas.

1.1.2 Sintaxe Básica

Imagine que você vai ler um artigo. Você imprime esse documento e inicia sua leitura, mas começa sentir sono e resolve parar e ir tomar um café. Quando você retorna para continuar sua leitura, seu artigo sumiu! Pior que isso, você também esqueceu onde havia parado de ler!

Então você precisa novamente imprimir o documento e iniciar sua leitura novamente. Agora imagine que isso acontece a cada vez que para de ler e se distancia do seu documento. Seria um sofrimento ler qualquer artigo, uma vez que sempre seria necessário imprimir e ler o documento de uma vez.

O mesmo ocorre na análise de dados e computação em geral, nós queremos ter uma forma de ler ou registrar um dado e depois poder retornar a usá-lo sem grandes problemas. Para isso usamos variáveis (que não é a mesma variável da estatística). Então, no R, uma variável representa um nome associado a um dado gravado na memória. Por ser somente um nome, não há restrições para o que ele nomeia (o tipo de dado), somente não se aceita que ele seja um nome feio, que usa caractéres proibidos (numeros no início, “$”“,”“, “,”).

```
# | label: creating_variables

# para criar comentários comece a linha com #
# esses comentários são desconsiderados pelo interpretador do R (não processados)
```

```r
# forma de atribuição mais comum
x <- 10
# forma menos comum
y = 20
# Atribuição reversa - para aqueles que vivem no Upside Down
30 -> z

# R é case-sensitive (diferencia maiusculas de minusculas)
Var1 <- 5
var1 <- 10

# Boas práticas para nomear variável
# Nunca:
# - Começar com números: 2var (incorreto)
# - Usar espaços ou carácteres especiais: minha variavel (incorreto)
# - Usar palavras reservadas: mean, if, for

# Use nomes descritivos, quem lê seu código não sabe o que você pensou
p_valor_teste_t <- 0.032
ic_95_inferior <- 12.3
ic_95_superior <- 18.7
```

1.1.3 Tipos de Dados

Numeric (double/integer):

```r
# | label: data_types_numeric

# Números reais (padrão)
altura <- 1.75
peso <- 68.5
# "numeric"
class(altura)
```

```
[1] "numeric"
```

```r
class(1)
```

```
[1] "numeric"
```

```r
# Inteiros (com L)
n_amostras <- 100L
# "integer"
class(n_amostras)
```

```
[1] "integer"
```

Character (texto):

```r
# | label: data_types_character

tratamento <- "vacina"
```

```r
# c() é um vetor (agrupamento de dados atômicos)
grupo <- c("controle", "tratado", "placebo")
# "character"
class(tratamento)
```

```
[1] "character"
```

Logical (booleano, verdadeiro/falso):

```r
# | label: data_types_logical

significativo <- TRUE
hipotese_nula <- FALSE
p_valor <- 0.01
# operações lógicas: Retorna TRUE ou FALSE
p_valor < 0.05
```

```
[1] TRUE
```

```r
class(hipotese_nula)
```

```
[1] "logical"
```

1.1.4 Vetores e Operações Básicas

Os vetores são a estrutura fundamental do R. **Tudo é vetor em sua essência**!

```r
# | label: data_types_vector

# Como criar vetores
dados <- c(23, 45, 12, 67, 34)
sequencia <- 1:10
seq_regular <- seq(0, 1, by=0.1)
repeticao <- rep(c(0,1), times=5)

# vetor nomeado
idades <- c(fulano=21, cicrano=43)
names(idades)
```

```
[1] "fulano"  "cicrano"
```

```r
# Operações vetorizadas são realizadas elemento por elemento
x <- c( 1,  2,  3,  4,  5)
y <- c(10, 20, 30, 40, 50)

x + y
```

```
[1] 11 22 33 44 55
```

```r
x * 2
```

```
[1]  2  4  6  8 10
```

```r
x^2
```

```
[1]  1  4  9 16 25
```

```r
sqrt(x)
```

```
[1] 1.000000 1.414214 1.732051 2.000000 2.236068
```

```r
# Operações em vetores de tamanho diferente
# Cuidado! porque ocorre reciclagem do menor vetor
c(1, 2, 3) + c(10, 20)
```

```
Warning in c(1, 2, 3) + c(10, 20): comprimento do objeto maior não é múltiplo
do comprimento do objeto menor
```

```
[1] 11 22 13
```

1.1.5 Funções Básicas

Funções possuem um padrão nome_da_funcao(argumento1, argumento2, ...). Ela é um bloco de código com uma finalidade específica, que abstrai a complexidade de como é feito algo para quem a usa. Então, por exemplo, se uso uma função media(x), eu não preciso saber o "como" e somente o que ela faz (calcula a média de um grupo de elementos em x).

Funções também são úteis quando repetimos um bloco de código em vários momentos de uma análise, pois, podemos definir uma função para executar esse bloco de código uma vez e depois só executá-la (princípio DRY).

```r
# | label: functions_basic

dados <- c(23, 45, 12, 67, 34, 28, 51)

# funções do dia-a-dia
sum(dados)        # Soma
```

```
[1] 260
```

```r
mean(dados)       # Média aritmética
```

```
[1] 37.14286
```

```r
median(dados)     # Mediana
```

```
[1] 34
```

```r
var(dados)        # Variância amostral (n-1)
```

```
[1] 345.1429
```

```r
sd(dados)         # Desvio padrão
```

```
[1] 18.57802
```

```r
min(dados)        # Mínimo
```

```
[1] 12
```

```r
max(dados)        # Máximo
```

```
[1] 67
```

```r
range(dados)      # Min e Max
```

```
[1] 12 67
```

```
quantile(dados)   # Quartis
```

```
  0%  25%  50%  75% 100%
12.0 25.5 34.0 48.0 67.0
```

```
summary(dados)    # Resumo estatístico
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  12.00   25.50   34.00   37.14   48.00   67.00
```

```
# Outras funções úteis
# Tamanho do vetor
length(dados)
```

```
[1] 7
```

```
# Ordenação
sort(dados, decreasing = FALSE)
```

```
[1] 12 23 28 34 45 51 67
```

```
# Valores únicos
unique(dados)
```

```
[1] 23 45 12 67 34 28 51
```

```
# Tabela de frequências
dados <- c(rep("a", 2), rep("b", 4), rep("c", 8), rep("d", 1))
table(dados)
```

```
dados
a b c d
2 4 8 1
```

```
prop.table(table(dados))
```

```
dados
         a          b          c          d
0.13333333 0.26666667 0.53333333 0.06666667
```

1.1.6 Operadores Aritméticos e Lógicos

Operadores Aritméticos:

```
# | label: arithmetics_operations

# Básicos
10 + 5    # Adição
```

```
[1] 15
```

```
10 - 5    # Subtração
```

```
[1] 5
```

```r
10 * 5    # Multiplicação
```

```
[1] 50
```

```r
10 / 5    # Divisão
```

```
[1] 2
```

```r
10 ^ 2    # Potenciação
```

```
[1] 100
```

```r
10 ** 2   # Potenciação
```

```
[1] 100
```

```r
10 %% 3   # Módulo (resto): 1
```

```
[1] 1
```

```r
10 %/% 3  # Divisão inteira: 3
```

```
[1] 3
```

```r
amostra <- sample(0:200, 1e6, replace = TRUE)
media <- sum(amostra) / length(amostra)
variancia <- sum((amostra - mean(amostra))^2) / (length(amostra) - 1)
```

Operadores Lógicos:

```r
# | label: boolean_operations

# Comparação
5 > 3      # maior
```

```
[1] TRUE
```

```r
5 < 3      # menor
```

```
[1] FALSE
```

```r
5 >= 3     # maior ou igual
```

```
[1] TRUE
```

```r
5 <= 3     # menor ou igual
```

```
[1] FALSE
```

```r
5 == 3     # igual
```

```
[1] FALSE
```

```r
5 != 3     # diferente
```

```
[1] TRUE
```

```r
"a" == "b"
```

```
[1] FALSE
```

```r
# Operadores booleanos
p <- TRUE
q <- FALSE
# operação   conectivo
!p           # NEGAÇÃO
```

```
[1] FALSE
```

```r
p & q        # E - conjunção (só é V em VV)
```

```
[1] FALSE
```

```r
p | q        # OU - disjunção inclusiva (só é F em FF)
```

```
[1] TRUE
```

```r
xor(p, q)    # OU OU - disjunção exclusiva (é F sempre que iguais - VV, FF)
```

```
[1] TRUE
```

```r
!p | q       # equivalente à condicional
```

```
[1] FALSE
```

```r
(!p | q) & (!q | p) # equivalente à bicondicional
```

```
[1] FALSE
```

```r
10 < 12 & 12 > 5
```

```
[1] TRUE
```

```r
# short circuit evaluation
10 > 12 && nao_existo
```

```
[1] FALSE
```

```r
10 < 12 || nao_existo
```

```
[1] TRUE
```

```r
# Vetorização
idades <- c(18, 25, 30, 17, 22)
idades >= 18
```

```
[1]  TRUE  TRUE  TRUE FALSE  TRUE
```

1.2 Estruturas de dados

1.2.1 Matrizes

Estruturas bidimensionais com elementos do **mesmo tipo**.

```r
# | label: matrices

# Criação de matrizes
matriz1 <- matrix(1:12, nrow=3, ncol=4)
matriz2 <- matrix(1:12, nrow=3, ncol=4, byrow=TRUE)
```

```r
# Matriz de correlação
dados <- matrix(rnorm(100), ncol=5)
cor_matrix <- cor(dados)

# Operações matriciais
A <- matrix(c(1,2,3,4), nrow=2)
B <- matrix(c(5,6,7,8), nrow=2)

A + B          # Soma elemento por elemento
```

```
     [,1] [,2]
[1,]    6   10
[2,]    8   12
```

```r
A * B          # Multiplicação elemento por elemento
```

```
     [,1] [,2]
[1,]    5   21
[2,]   12   32
```

```r
A %*% B        # Multiplicação matricial verdadeira
```

```
     [,1] [,2]
[1,]   23   31
[2,]   34   46
```

```r
t(A)           # Transposta
```

```
     [,1] [,2]
[1,]    1    2
[2,]    3    4
```

```r
solve(A)       # Inversa (se existir)
```

```
     [,1] [,2]
[1,]   -2  1.5
[2,]    1 -0.5
```

```r
det(A)         # Determinante
```

```
[1] -2
```

```r
# Dimensões
dim(A)
```

```
[1] 2 2
```

```r
nrow(A)
```

```
[1] 2
```

```r
ncol(A)
```

```
[1] 2
```

1.2.2 Data Frames

São basicamente tabelas com colunas de **variados tipos** em que cada linha representa um registro (planilha do excel). É o principal tipo de dado com que trabalhamos na prática. Todas as colunas no `data.frame` devem apresentar o mesmo tamanho.

```
# | label: data_frames

# Criação
df <- data.frame(
  id = 1:5,
  tratamento = c("A", "B", "A", "B", "A"),
  peso_inicial = c(65.2, 70.1, 68.5, 72.3, 66.8),
  peso_final = c(68.1, 71.5, 71.2, 73.8, 69.5),
  melhorou = c(TRUE, TRUE, TRUE, TRUE, NA)
)

# Estrutura e resumo
str(df)         # Estrutura do data.frame
```

```
'data.frame':   5 obs. of  5 variables:
 $ id          : int  1 2 3 4 5
 $ tratamento  : chr  "A" "B" "A" "B" ...
 $ peso_inicial: num  65.2 70.1 68.5 72.3 66.8
 $ peso_final  : num  68.1 71.5 71.2 73.8 69.5
 $ melhorou    : logi  TRUE TRUE TRUE TRUE NA
```

```
summary(df)     # Resumo estatístico simples de cada coluna
```

```
       id     tratamento         peso_inicial     peso_final    melhorou
 Min.   :1   Length:5           Min.   :65.20   Min.   :68.10   Mode:logical
 1st Qu.:2   Class :character   1st Qu.:66.80   1st Qu.:69.50   TRUE:4
 Median :3   Mode  :character   Median :68.50   Median :71.20   NA's:1
 Mean   :3                      Mean   :68.58   Mean   :70.82
 3rd Qu.:4                      3rd Qu.:70.10   3rd Qu.:71.50
 Max.   :5                      Max.   :72.30   Max.   :73.80
```

```
head(df)        # Primeiras x linhas
```

```
  id tratamento peso_inicial peso_final melhorou
1  1          A         65.2       68.1     TRUE
2  2          B         70.1       71.5     TRUE
3  3          A         68.5       71.2     TRUE
4  4          B         72.3       73.8     TRUE
5  5          A         66.8       69.5       NA
```

```
tail(df)        # Últimas x linhas
```

```
  id tratamento peso_inicial peso_final melhorou
1  1          A         65.2       68.1     TRUE
2  2          B         70.1       71.5     TRUE
3  3          A         68.5       71.2     TRUE
4  4          B         72.3       73.8     TRUE
5  5          A         66.8       69.5       NA
```

```r
# Acessando colunas
df$peso_inicial
```

```
[1] 65.2 70.1 68.5 72.3 66.8
```

```r
df[["peso_inicial"]]
```

```
[1] 65.2 70.1 68.5 72.3 66.8
```

```r
df[, "peso_inicial"]
```

```
[1] 65.2 70.1 68.5 72.3 66.8
```

```r
df[, 3]
```

```
[1] 65.2 70.1 68.5 72.3 66.8
```

```r
df[c(1, 3), 3]
```

```
[1] 65.2 68.5
```

```r
# Criando nova variável no data.frame
df$ganho_peso <- df$peso_final - df$peso_inicial
df$ganho_percentual <- (df$ganho_peso / df$peso_inicial) * 100

head(df)
```

```
  id tratamento peso_inicial peso_final melhorou ganho_peso ganho_percentual
1  1          A         65.2       68.1     TRUE        2.9         4.447853
2  2          B         70.1       71.5     TRUE        1.4         1.997147
3  3          A         68.5       71.2     TRUE        2.7         3.941606
4  4          B         72.3       73.8     TRUE        1.5         2.074689
5  5          A         66.8       69.5       NA        2.7         4.041916
```

1.2.3 Listas

Estruturas mais flexíveis - podem conter elementos de **tipos e tamanhos diferentes**.

```r
# | label: lists

# Lista com resultados de uma análise estatística
resultado_teste <- list(
  nome_teste = "Teste t de Student",
  estatistica_t = 2.453,
  graus_liberdade = 48,
  p_valor = 0.018,
  intervalo_confianca = c(1.23, 5.67),
  dados_originais = df,
  matriz_cov = matrix(rnorm(9), 3, 3)
)

# Acessando elementos
resultado_teste$p_valor
```

```
[1] 0.018
```

```r
resultado_teste[["p_valor"]]
```

```
[1] 0.018
```

```r
resultado_teste[[4]]
```

```
[1] 0.018
```

```r
teste_t <- t.test(df$peso_final, df$peso_inicial, paired=TRUE)
str(teste_t)
```

```
List of 10
 $ statistic  : Named num 6.89
  ..- attr(*, "names")= chr "t"
 $ parameter  : Named num 4
  ..- attr(*, "names")= chr "df"
 $ p.value    : num 0.00232
 $ conf.int   : num [1:2] 1.34 3.14
  ..- attr(*, "conf.level")= num 0.95
 $ estimate   : Named num 2.24
  ..- attr(*, "names")= chr "mean difference"
 $ null.value : Named num 0
  ..- attr(*, "names")= chr "mean difference"
 $ stderr     : num 0.325
 $ alternative: chr "two.sided"
 $ method     : chr "Paired t-test"
 $ data.name  : chr "df$peso_final and df$peso_inicial"
 - attr(*, "class")= chr "htest"
```

```r
names(teste_t)
```

```
 [1] "statistic"   "parameter"   "p.value"     "conf.int"    "estimate"
 [6] "null.value"  "stderr"      "alternative" "method"      "data.name"
```

1.2.4 Fatores

Variáveis categóricas com níveis fixos - essencial para modelos estatísticos.

```r
# | label: factors

# Criação de fatores
sexo <- factor(c("M", "F", "F", "M", "F"), levels = c("M", "F"))
levels(sexo)
```

```
[1] "M" "F"
```

```r
# Recodificação
levels(sexo) <- c("Masculino", "Feminino")

# Fator ordenado
educacao <- factor(
  c("Médio", "Superior", "Fundamental", "Superior", "Médio"),
  levels = c("Fundamental", "Médio", "Superior"),
  ordered = TRUE
```

```
)
as.integer((educacao))
```

```
[1] 2 3 1 3 2
```

```
# GLMs e ANOVA tratam fatores como dummies ou como variáveis discretas
df$tratamento[5] <- "C"
df$grupo <- df$tratamento
lm(peso_final ~ grupo, data=df)
```

```

Call:
lm(formula = peso_final ~ grupo, data = df)

Coefficients:
(Intercept)       grupoB       grupoC
      69.65         3.00        -0.15
```

```
df$grupo <- factor(df$tratamento)
lm(peso_final ~ grupo, data=df)
```

```

Call:
lm(formula = peso_final ~ grupo, data = df)

Coefficients:
(Intercept)       grupoB       grupoC
      69.65         3.00        -0.15
```

```
df$grupo <- factor(df$tratamento, ordered = TRUE)
lm(peso_final ~ grupo, data=df)
```

```

Call:
lm(formula = peso_final ~ grupo, data = df)

Coefficients:
(Intercept)      grupo.L      grupo.Q
    70.6000      -0.1061      -2.5107
```

1.2.5 Arrays

Generalizações de matrizes para **múltiplas dimensões**.

```
# | label: arrays

# Array de 3 dimensões (exemplo: medidas x indivíduos x tempo)
n_pacientes <- 5
n_tempos <- 3
peso_inicial <- rnorm(n_pacientes, mean = 70, sd = 5)
altura_inicial <- rnorm(n_pacientes, mean = 170, sd = 10)
idade_inicial <- rpois(n_pacientes, lambda = 23)
```

```r
# matriz vazia
peso_tempo <- matrix(nrow = n_pacientes, ncol = n_tempos)
altura_tempo <- matrix(nrow = n_pacientes, ncol = n_tempos)
idade_tempo <- matrix(nrow = n_pacientes, ncol = n_tempos)

for(i in 1:n_pacientes) {
  peso_tempo[i, ] <- peso_inicial[i] + cumsum(c(0, rnorm(n_tempos-1, mean=0.5, sd=1)))
  altura_tempo[i, ] <- altura_inicial[i] + rnorm(n_tempos, mean=0, sd=0.5)
  idade_tempo[i, ] <- idade_inicial[i] + c(0, 0.25, 0.5)
}

imc_tempo <- peso_tempo / (altura_tempo/100)^2

medidas_tempo <- array(
  c(t(peso_tempo), t(altura_tempo), t(imc_tempo), t(idade_tempo)),
  dim = c(n_pacientes, n_tempos, 4),
  dimnames = list(
    paste("Paciente", 1:5),
    c("Mês_0", "Mês_3", "Mês_6"),
    c("Peso", "Altura", "IMC", "Idade")
  )
)

medidas_tempo[1, , ] # paciente 1
```

```
          Peso   Altura      IMC Idade
Mês_0 74.68949 168.3604 26.34993 28.00
Mês_3 63.96022 176.3027 20.57747 20.50
Mês_6 71.70416 169.8894 24.84344 17.25
```

```r
medidas_tempo[, , 4] # idade
```

```
           Mês_0 Mês_3 Mês_6
Paciente 1 28.00 20.50 17.25
Paciente 2 28.25 20.00 17.50
Paciente 3 28.50 20.25 29.00
Paciente 4 20.00 20.50 29.25
Paciente 5 20.25 17.00 29.50
```

```r
medidas_tempo[, 2, ] # mes 2
```

```
              Peso   Altura      IMC Idade
Paciente 1 63.96022 176.3027 20.57747 20.50
Paciente 2 71.84751 149.6320 32.08947 20.00
Paciente 3 72.47510 150.5019 31.99669 20.25
Paciente 4 73.17602 150.0540 32.49926 20.50
Paciente 5 71.54084 170.5523 24.59454 17.00
```

```r
n <- 1000
dados_epi <- data.frame(
  Sexo = sample(c("M", "F"), n, replace = TRUE),
  Idade = sample(c("0-20", "21-40", "41-60", "60+"), n, replace = TRUE),
```

```
  Exposicao = sample(c("Sim", "Não"), n, replace = TRUE, prob = c(0.3, 0.7)),
  Doenca = sample(c("Presente", "Ausente"), n, replace = TRUE, prob = c(0.1, 0.9))
)
tabela_4d <- table(dados_epi)
print(dim(tabela_4d))  # 2 x 4 x 2 x 2
```

```
[1] 2 4 2 2
```

```
# Análise de odds ratio estratificado
for(sexo in c("M", "F")) {
  for(idade in unique(dados_epi$Idade)) {
    subtabela <- tabela_4d[sexo, idade, , ]
    if(all(subtabela > 0)) {
      #|                | desfecho: Sim  | desfecho: Não |
      #|----------------|----------------|---------------|
      #| preditor: Sim  | A              | B             |
      #| preditor: Não  | C              | D             |
      # OR = a*d/b*c

      OR <- (subtabela[2,2] * subtabela[1,1]) /
            (subtabela[2,1] * subtabela[1,2])
      cat(sprintf("OR para %s, %s: %.2f\n", sexo, idade, OR))
    }
  }
}
```

```
OR para M, 60+: 0.25
OR para M, 0-20: 2.33
OR para M, 21-40: 11.59
OR para M, 41-60: 0.70
OR para F, 60+: 0.79
OR para F, 0-20: 1.43
OR para F, 21-40: 0.97
OR para F, 41-60: 1.78
```

1.2.6 Indexação e Seleção de Dados

```
# | label: indexing

# VETORES
x <- c(10, 20, 30, 40, 50)
x[2]
```

```
[1] 20
```

```
x[c(1,3,5)]
```

```
[1] 10 30 50
```

```
x[-2]
```

```
[1] 10 30 40 50
```

```r
x[x > 25]
```

```
[1] 30 40 50
```

```r
# MATRIZES [linha, coluna]
mat <- matrix(1:12, nrow=3)
mat[2, 3]
```

```
[1] 8
```

```r
mat[2, ]
```

```
[1]  2  5  8 11
```

```r
mat[, 3]
```

```
[1] 7 8 9
```

```r
mat[1:2, 3:4]
```

```
     [,1] [,2]
[1,]    7   10
[2,]    8   11
```

```r
# DATA FRAMES
df[2, 3]
```

```
[1] 70.1
```

```r
df[2, ]
```

```
  id tratamento peso_inicial peso_final melhorou ganho_peso ganho_percentual
2  2          B         70.1       71.5     TRUE        1.4         1.997147
  grupo
2     B
```

```r
df[, "peso_inicial"]
```

```
[1] 65.2 70.1 68.5 72.3 66.8
```

```r
df$peso_inicial
```

```
[1] 65.2 70.1 68.5 72.3 66.8
```

```r
df[df$tratamento == "A", ]
```

```
  id tratamento peso_inicial peso_final melhorou ganho_peso ganho_percentual
1  1          A         65.2       68.1     TRUE        2.9         4.447853
3  3          A         68.5       71.2     TRUE        2.7         3.941606
  grupo
1     A
3     A
```

```r
df[df$ganho_peso > 2 & df$tratamento == "A", c("id", "ganho_peso")]
```

```
  id ganho_peso
1  1        2.9
3  3        2.7
```

```r
# LISTAS
lista <- list(a=1:5, b=matrix(1:4,2), c="texto")
lista[[1]]
```

```
[1] 1 2 3 4 5
```

```r
lista$a
```

```
[1] 1 2 3 4 5
```

```r
lista[["b"]]
```

```
     [,1] [,2]
[1,]    1    3
[2,]    2    4
```

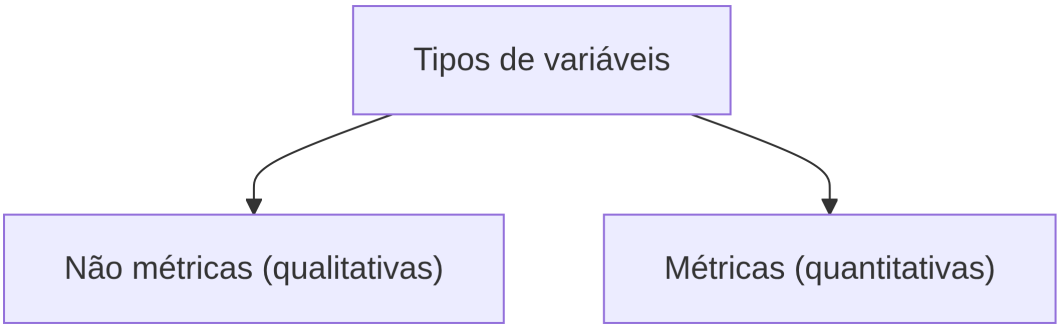
Chapter 2

Tipos de variáveis e escalas de mensuração e precisão

2.1 Tipos de variáveis

Variável é uma característica em estudo em uma população (ou amostra) que pode ser **medida, contada ou categorizada**. O tipo influencia na escolha de estatísticas descritivas, gráficos e métodos de análise que serão adotados para analisar os dados. Assim, vamos começar nosso estudo pela classificação das variáveis relativa ao fato de serem mensuradas ou categorizadas, bem como as escalas de mensuração.

Em geral, variáveis são classificadas como **métricas e não métricas**. Variáveis não métricas ou qualitativas são características registradas dentro de um número finito de categorias, em que o indivíduo possui ou não determinada categoria em análise (ex. sexo). Variáveis métricas ou quantitivas são contadas ou mensuradas (nesse caso não há ausência de determinada característica, embora ele possa apresentar 0 dela).



Tipos de variáveis.

2.1.1 Variáveis Não Métricas (Qualitativas)

Representam características, atributos ou qualidades que não podem ser medidas numericamente de forma intrínseca. Normalmente, são apresentadas na análise em tabelas de frequência, sem medidas de posição, dispersão ou forma. Uma medida que pode ser associada é a moda, valor ou categoria mais frequente.

Características principais:

- Representam categorias ou classes
- Não admitem operações aritméticas significativas
- Podem ser codificadas numericamente, mas os números são apenas rótulos

Exemplos: A) Raça de animais (Holandesa, Jersey, Angus); B) Diagnóstico clínico (positivo, negativo); C) Coloração de pelagem (preto, branco, malhado).

2.1.2 Variáveis Métricas (Quantitativas)

Representam quantidades mensuráveis ou que são contadas. Permitem maior gama de possibilidades gráficas (gráficos de linhas, dispersão, histograma, boxplot...), medidas de posição, dispersão e forma.

Podem ser discretas, quando assumem um conjunto finito ou enumerável de valores, ou contínuas, quando os valores estão em um intervalo de números reais (em tese, assumem infinitos valores dentro do intervalo).

Características principais:

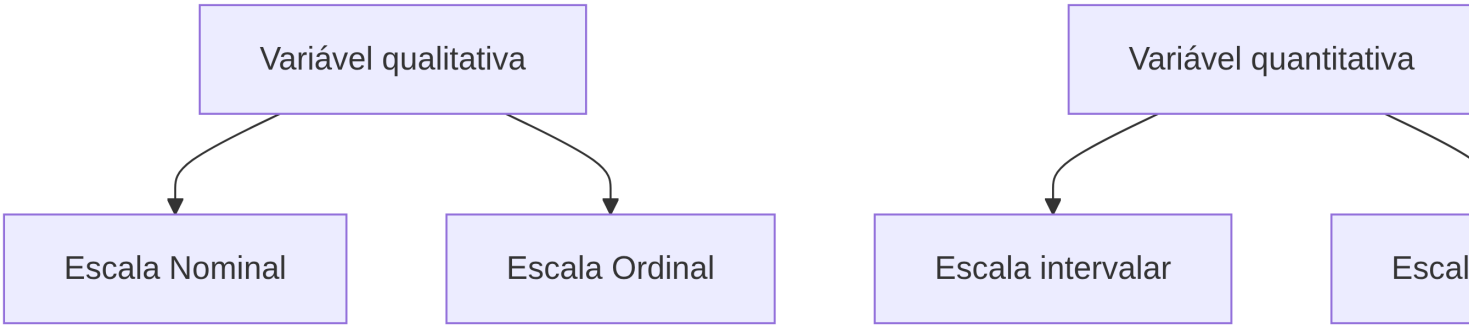
- Expressam magnitude ou quantidade
- Permitem operações aritméticas (soma, média, etc.)
- Possuem unidade de medida (kg, cm, °C, etc.)

Exemplos: A) Peso corporal (kg); B) Contagem de leucócitos (células/ L); C) Produção de leite (litros/dia); D) número de filhos.

2.2 Escalas de mensuração (Stevens, 1946)

Mensuração é o ato de atribuir números ou rótulos a um indíviduo de acordo com regras específicas para representar quantidades ou qualidades de um atributo (a variável em estudo). A escala é o conjunto de possíveis valores que o atributo poderá assumir, considerando a regra de mensuração.

Existem várias classificações de tipos de escala, mas vamos usar a de Stevens (1946), pela simplicidade e amplo uso.



Tipos de escalas.

2.2.1 Escala Nominal (Variáveis Não Métricas)

É a mais básica das escalas, onde números ou símbolos servem apenas para identificar e classificar. Não estabelece relações de orgem ou grandeza. Possibilita contar o número de elementos em cada categoria e testar hipóteses sobre a distribuição de unidades dentro das categorias, mas medidas de posição, dispersão e forma não fazem muito sentido para ela.

Exemplos:

```
Tipo sanguíneo: A, B, AB, O
Sexo: M=Macho, F=Fêmea
Diagnóstico de LVC: 0=Negativo, 1=Positivo
Espécie animal: 1=Bovino, 2=Suíno, 3=Equino
País de Origem: Brasil, Bolívia, Peru, Canadá, ...
```

Propriedades matemáticas:

- Equivalência ($=$, $\neq$)
- Não há ordenação intrínseca

Ferramentas ou medidas para caracterização dos dados:

- Tabelas de frequência
- Gráficos de barras ou setores
- Contagem de frequências, moda

Análise estatística possível:

- Teste qui-quadrado (χ^2)
- Coeficiente de contingência (CC)
- V de Cramér

2.2.2 Escala Ordinal (Variáveis Não Métricas)

Ainda classifica o atributo em classes ou categorias, mas existe uma relação inerente de ordem entre as diferentes categorias, mesmo que as distâncias entre categorias não sejam necessariamente iguais. Por exemplo, a idade dos indivíduos sendo mensurada nas faixas etárias jovem (15 a 21 anos), adulto (22 a 65 anos), idoso (mais que 65 anos). Embora, exista uma ordem entre as faixas (idoso $>$ adulto $>$ jovem, na idade), o intervalo entre cada classe não é igual, ou seja, não faz sentido eu pensar em algo como "1 idoso $=$ 3 jovem anos".

Exemplos:

```
Escore de condição corporal: 1 (magro) < 2 < 3 < 4 < 5 (obeso)
Grau de claudicação: 0 (normal) < 1 (leve) < 2 (moderado) < 3 (severo)
Escolaridade: fundamental < médio < superior
Escala Linkert: 1 (Total discordância) < 2 < 3 (Neutro) < 4 < 5 (Total cordância)
```

Propriedades matemáticas:

- Relações de ordem e equivalência ($<$, $>$, $=$)
- Intervalos entre categorias não são uniformes

Ferramentas ou medidas para caracterização dos dados:

- Tabelas de frequência
- Gráficos de barras ou setores
- Contagem de frequências, moda, mediana, percentis, correlação de Spearman

Análise estatística possível:

- Teste de Mann-Whitney U
- Teste de Kruskal-Wallis
- Correlação de postos de Spearman (ρ de Spearman)
- Correlação de postos de Kendall (τ de Kendall)

2.2.3 Escala Intervalar (Variável Quantitativa)

Além de haver ordenação, possui intervalos iguais entre valores, ou seja, unidade de medida constante. Porém, o zero da escala é arbitrário (não representa ausência absoluta). Assim, conseguimos calcular a diferença entre dois valores mensurados, mas não podemos inferir que um seja múltiplo do outro.

Exemplos:

```
Temperatura em Celsius: 0°C é arbitrário
Calendário: ano 0 é convenção
Escores padronizados (z-score): média=0 por construção
```

Propriedades matemáticas:

- Relações de ordem e equivalência ($<$, $>$, $=$)
- Permite inferir diferença
- Razões não têm significado
- Se X está em escala intervalar, então:
  - $Y = a + bX$ também está em escala intervalar
  - A razão X/X não é invariante sob transformação linear

```
# peso em quilos ou gramas
x1 = 10 # 10 kg
x2 = 100 # 100 kg

x2 / x1
```

```
[1] 10
```

```
y1 = x1 * 1000 # 10000 g
y2 = x2 * 1000 # 100000 g

y2 / y1
```

```
[1] 10
```

```
# temperatura em celsius ou fahrenheit (1.8 * C + 32)
x1 = 10 # 10 C
x2 = 20 # 20 C

x2 / x1
```

```
[1] 2
```

```
y1 = 32 + (1.8 * x1) # 50 F
y2 = 32 + (1.8 * x2) # 68 F

y2 / y1
```

```
[1] 1.36
```

```
x3 = 30
y3 = 32 + (1.8 * x3)

(x3 - x1) / (x3 - x2) == (y3 - y1) / (y3 - y2)
```

```
[1] TRUE
```

Ferramentas ou medidas para caracterização dos dados:

- Gráficos de linhas, dispersão, histograma, boxplot
- Média, desvio-padrão, mediana, percentis, correlação (com os devidos cuidados)

```
x <- rnorm(1000, mean=283.15, sd=5.15) # aproximadamente 20°C
mean(x)
```

```
[1] 283.1564
```

```
sd(x)
```

```
[1] 5.196768
```

```
sd(x) / mean(x) * 100
```

```
[1] 1.8353
```

```
y = x - 273.15
mean(y)
```

```
[1] 10.00635
```

```
sd(y)
```

```
[1] 5.196768
```

```
sd(y) / mean(y) * 100
```

```
[1] 51.93468
```

```
y = x*2
mean(y)
```

```
[1] 566.3127
```

```
sd(y)
```

```
[1] 10.39354
```

```
sd(y) / mean(y) * 100
```

```
[1] 1.8353
```

```
require(tidyverse)
require(gridExtra)

temperatures <- tibble(kelvin = x, celsius = y)

k_plot <- ggplot(data = temperatures, aes(x=kelvin, y = ..density..)) +
  geom_histogram(fill = "cornsilk", colour = "grey60", size = .2) +
  geom_density() +
  labs(
    x = "Kelvin",
    y = "Densidade"
  ) +
  theme_bw()
```

```
c_plot <- ggplot(data = temperatures, aes(x=celsius, y = ..density..)) +
  geom_histogram(fill = "cornsilk", colour = "grey60", size = .2) +
  geom_density() +
  labs(
    x = "Celsius",
    y = "Densidade"
  ) +
  theme_bw()

marrangeGrob(
    list(k_plot, c_plot),
    ncol = 2, nrow = 1,
    top = "Distribuição da variável original e sua transformada na forma Y = aX + b"
)
```

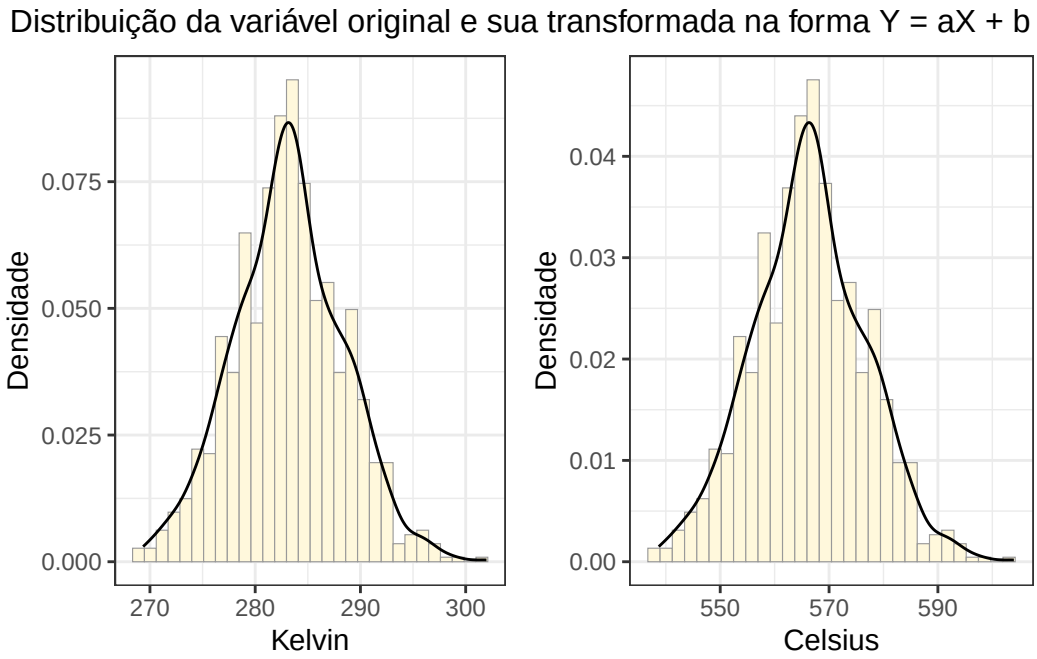


Figura 2.1: Distribuição de uma variável na escala razão e sua transformada nas escalar intervalar.

Análise estatística possível:

- Testes paramétricos ou não-paramétricos
- Coeficiente de correlação de Pearson (ρ de Pearson)

2.2.4 Escala de Razão (Variável Quantitativa)

Possui todas as propriedades da escala intervalar, mais um zero absoluto significativo.

Exemplos:

`Peso: 0 kg = ausência de massa`

```
Contagem: 0 células = nenhuma célula
Concentração: 0 mg/dL = ausência da substância
Tempo de reação: 0 segundos = instantâneo
```

Propriedades matemáticas:

- Zero representa ausência completa
- Relações de ordem e equivalência (<, >, =)
- Permite inferir diferença e razão
- Todas as operações aritméticas são válidas (+, -, *, /)
- Operações permitidas: média, desvio padrão, correlação de Pearson

Ferramentas ou medidas para caracterização dos dados:

- Gráficos de linhas, dispersão, histograma, boxplot
- Média, desvio-padrão, coeficiente de variação, mediana, percentis, correlação

Análise estatística possível:

- Testes paramétricos ou não-paramétricos
- Coeficiente de correlação de Pearson (ρ de Pearson)

Tabela resumo do efeito da transformação linear nas propriedades das variáveis quantitativas

Propriedade	Transformação Linear Positiva (Y=a+bX, b>0 e a>0)	Transformação de Proporção (Y=bX, b>0 e a=0)
Ordem (X $<$ X)	Invariante	Invariante
Diferenças	Não invariante (multiplicadas por b)	Não invariante
Razões de diferenças	Invariante	Invariante
Razões diretas (X /X)	Não invariante	Invariante
Correlação	Invariante	Invariante
Média	Não invariante	Não invariante
CV	Não invariante	Invariante

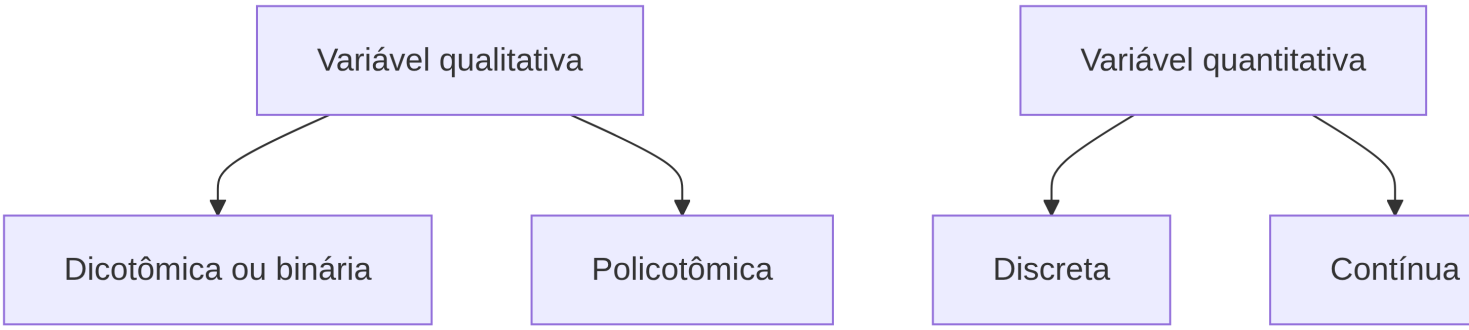
Invariante significa que determinada propriedade permanece inalterada após a transformação.

Em termos matemáticos:

- Dada uma propriedade $P(X)$ de uma variável X
- e uma transformação T de X em uma variável Y: $X \to Y = T(X)$
- P é invariante sob T se: $P(X) = P(Y)$

2.3 Número de categorias e pecisão

Variáveis qualitativas podem ser classificadas em dicotômicas quando apresentam apenas duas categorias ou policotômicas quando possuem mais que duas categorias. Já as variáveis quantitativas podem ser classificadas conforme a escala de precisão em discretas ou contínuas.



Tipos de escalas.

2.3.1 Variável Dicotômica ou Binária (Dummy)

Variável com exatamente duas categorias mutuamente exclusivas. Nas variáveis dummy, uma categoria indica a presença de determinada característica e a outra a ausência.

Representação matemática:

```
X = {0, 1} ou X = {Sim, Não}
P(X=1) = p, P(X=0) = 1-p
```

Propriedades estatísticas:

- $E[X] = p$
- $Var(X) = p(1-p)$
- Distribuição: $X \sim Bernoulli(p)$

Exemplos:

- Presença/ausência de doença
- Vacinado/não vacinado
- Tratamento/controle

2.3.2 Variável Policotômica

Variável categórica com três ou mais categorias.

Tipos:

- Nominal policotômica: sem ordem (raça, região)
- Ordinal policotômica: com ordem (estágio da doença: I, II, III)

Uma variável qualitativa com n categorias pode ser representada por $n-1$ variáveis dummies (dicotômicas).

Estágio da doença	D1	D2
Estágio I	0	0
Estágio II	0	1
Estágio III	1	0

2.3.3 Variável Quantitativa Discreta

Assume valores em um conjunto enumerável (geralmente inteiros), geralmente finito ("infinito somente nos extremos").

Características:

- Resultado de contagem
- Valores isolados na reta numérica
- Função de probabilidade: $P(X = x)$
  - $E[X] = \mu_x = \sum_{i=1}^{n} x_i P(X = x_i), i = 1, 2, ..., n$

Distribuições comuns:

- $Poisson(\lambda)$: contagem de eventos raros
- $Binomial(n, p)$: número de sucessos em n tentativas
- Binomial Negativa, $BinNeg(n, p)$: número de falhas até o fim do experimento

Exemplos:

```
Número de crias: 0, 1, 2, 3, ...
Contagem de parasitas: valores inteiros  0
Número de tratamentos: 1, 2, 3, ...
```

2.3.4 Variável Quantitativa Contínua

Pode assumir qualquer valor em um intervalo dos reais (infinitos valores dentro do intervalo).

Características:

- Resultado de medição
- Infinitos valores possíveis em qualquer intervalo
- Função densidade de probabilidade: $f(x)$
  - $f(x) > 0$
  - $\int f(x)dx = 1$
  - a probabilidade de X assumir valores dentro de um intervalo $[a, b]$ é justamente a área sob a curva de $f(x)$ limitada pelos pontos a e b, logo, $P(a \leq X \leq b) = \int_a^b f(x)dx$
  - $E[X] = \int_{-\infty}^{+\infty} xf(x)dx$

Distribuições comuns:

- $Normal(\mu, \sigma^2)$: muitos fenômenos naturais
- $Exponencial(\lambda)$: tempo entre eventos
- $Gamma(\alpha, \beta)$: tempos de espera, concentrações

Exemplos:

```
Peso corporal: qualquer valor positivo real
Glicemia: valores contínuos em mg/dL
Produção de leite: litros com precisão decimal
```

2.3.5 Implicações práticas do tipo de variável na análise estatística

Testes de hipótese

- Nominal: Qui-quadrado, Teste Exato de Fisher, Coeficiente de associação
- Ordinal: Mann-Whitney, Kruskal-Wallis, Coeficiente de correlação de Spearman

- Quantitativa: Correlação de Pearson, Teste t, ANOVA (se normal), Correlação de Spearman, Mann-Whitney, Kruskal-Wallis (se não normal)

Modelagem Estatística

- Binária: Regressão Logística
- Policotômica: Regressão Multinomial
- Ordinal: Regressão Ordinal (Proportional Odds)
- Contagem: Regressão Poisson/Binomial Negativa
- Contínua: Regressão Linear (se normal)

Cuidados na interpretação de suas variáveis em estudo

1. **Não tratar ordinal como quantitativa**: A média de escores 1, 2, 3 possui algum significado? E o desvio padrão dos estágios 1, 2, 3 e 4 de determinada neoplasia?
2. **Verificar escala antes de calcular razões**: 20°C não é "duas vezes mais quente" que 10°C
3. **Considerar a precisão da medida**: Uma balança com precisão de 1kg torna a variável peso, que é inerentemente contínua, em uma variável discreta na prática.