CAR00110

Aprendizado em Linguagem R: Estatística Descritiva I
Apostila I

Prof. Lucas Helal, MMSc, PhD

2023-08-16

Programando em R

1. Sintaxe Básica

Como é costumeiro, começaremos aprendendo a linguagem R como qualquer outra linguagem de programação: por meio de um programa *"Hello World"*. Com isso, iremos aprender que você pode programar em R tanto no **prompt de comando** - ou **console**, quanto em um **script** ou mesmo em documento dinâmico (que veremos mais a frente).

1.1. O prompt de comando/console

Uma vez com o *setup* do seu R Studio (ou qualquer outro ambiente de desenvolvimento) configurado, a tela de `Console` aparecerá como uma das abas dos seus quadrantes que ficam à esquerda da tela - esta é a forma nativa carregada pelo R Studio, mas você pode reconfigurar a posição como quiser. De forma fácil, você simplesmente digita o comando direto no `prompt/console`.
Você verá algo como:

```
R version 4.3.0 (2023-04-21) -- Already Tomorrow
Copyright (C) 2023 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin20 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

>
```

Se você digitar e apertar `enter` diretamente:

```
> 2 + 2
[1] 4
```

Ou...

```
> myString <- "Hello, World!"
> print(myString)

[1] "Hello, World!"
```

Onde $[n]$ sempre sinalizará um **output de resultado**.

Na sentença #1, você atribuiu a palavra (`string`) **Hello World** ao objeto `myString`.

Na sentença #2, você aplicou uma função que indica à IDE, em linguagem R, que você está dando o comando para que a IDE/Linguagem R leia o que está contido dentro do objeto previamente criado - a função `print` por meio do comando `print()`.

1.2. O R Script

Nem sempre você quererá rodar códigos e não tê-los em mãos para utilizar outra vez, ou mesmo otimiza-lo. Por conta disso, é muito comum que em qualquer linguagem de programação ou pacote estatístico, que o código seja escrito em um documento na IDE que se está usando ou mesmo utilizando editores de texto simples (bloco de notas, html, etc.).

No R, a gente possui como primeiro recurso o R Script, que é um documento em que você escreve todas as linhas de código antes de visualizar o resultado dos códigos.

- Caso você queira escrever algo que não faça parte do código, utilize o símbolo `#` antes de começar a escrever. Assim o R Studio não entenderá o que está escrito como código
- Números e palavras são entendidos diferentemente em linguagem R. Para números, não há a necessidade de nenhum recurso adicional. Para palavras, você deve utilizar as aspas `"nonono"` - dessa forma, o R Studio não compreenderá a palavra como comentário, e sim como parte do código que você está escrevendo
- A linguagem R possui a característica marcante de ser orientada à objetos. Ou seja, nós digitamos uma parte de código e transformamos em um objeto, que pode ser re-transformado em outro objeto à medida que você for acumulando informações. Por exemplo: `1+2` deve ser **idealmente** escrito como `a <- 1+2`. Para pedir o resultado, em uma nova linha de comando, escreva a letra `a` (no caso desse exemplo) e rode o comanmdo.

1.3. Tipos de variáveis/dados em linguagem R

Em R, temos uma variedade de tipos de dados/variáveis ou classes de objetos. Aqui listo os mais frequentes para uso, assumindo que você irá aprender ainda muito mais à medida que continua estudando linguagem R.
Para fins didáticos, dividirei os tipos de variáveis de acordo com a classificação estatística.

PS: aqui vai uma pequena lista com símbolos mais utilizados em Estatística (com os comandos para documento Markdown).

x - x/Observação. **Comando** –> `$x$`

x_1 - O primeiro valor observado. **Comando** –> `$x_1$`

$\overline{x}$ - A média amostral. **Comando** –> `$\overline x$`

μ - A média populacional (parâmetro). **Comando** –> `$\mu$`

$\hat{p}$ - A proporção em uma amostra. **Comando** –> `$\hat p$`

$\hat{P}$ - A proporção em uma população (parâmetro). **Comando** –> `$\hat p$`

n, N - Tamanhos amostrais e populacionais, respectivamente. **Comando** –> `$n,N$`

s - Desvio padrão amostral. **Comando** –> `$s$`

s^2 - Variância amostral. **Comando** –> `$s^2$`

σ - Desvio padrão populacional (parâmetro). **Comando** –> `$\sigma$`

σ^2 - Variância populacional (parâmetro). **Comando** –> `$\sigma^2$`

$\mid y \mid$ - O valor absoluto de uma variável (módulo). **Comando** –> `$|\ y\ |$`

k - Denota *múltiplas aparições* para qualquer variável. **Comando** –> `$k$`

Variáveis Numéricas ou Quantitativas

Em Matemática e Estatística, uma variável numérica ou quantitativa é aquela que pode ser definida como **contínua** ou **discreta**, de forma que as variáveis contínuas podem ser definidas como:

Contínuas

São aquelas variáveis que pertecem ao conjunto dos números reais e que podem ser expandidas por números decimais de forma indefinida. Em outras palavras, variáveis contínuas **medem** quantidades em uma dimensão, como a **distância** entre dois pontos; a **temperatura** ambiente.

$$\mathbb{R} \ \exists \iff [A] \ onde: \ \forall \ a_n \ \therefore \ há \ pelo \ menos \ um \ -\infty \ \leq \ a \ \leq +\infty \tag{1.0}$$

Ou:

$$\mathbb{R} \text{ é } o\ conjunto\ dos\ números\ reais\ existe\ se,\ e\ somente\ se,\ existir\ um\ conjunto\ qualquer\ \mathbb{A}\ ... \quad (1.1)$$

Complete a sentença...

E, daqui, é importante saber que:

$$X \ \text{é} \ \textit{contínua} \ \ e \ \in \ \mathbb{R} \quad (1.3)$$

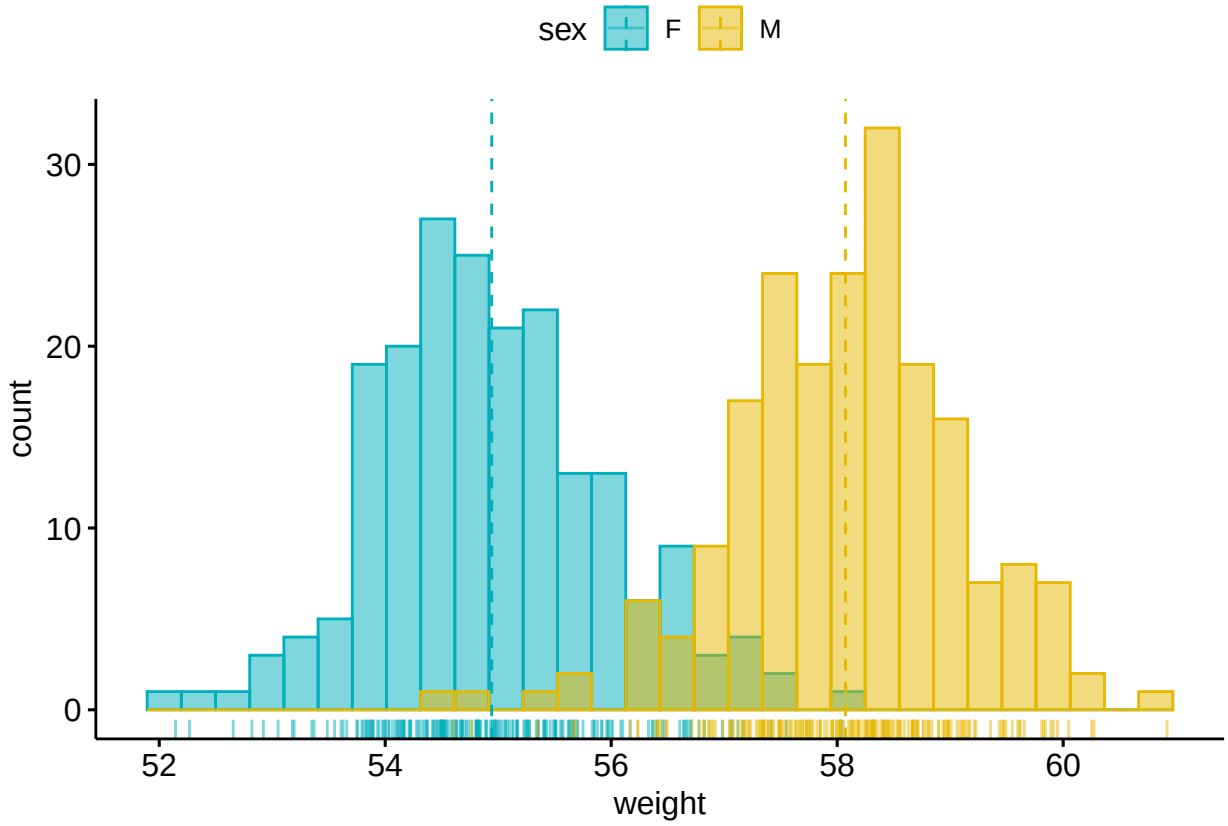
Variáveis contínuas pertencem ao conjunto dos números reais, mas nem todos os elementos do conjunto dos números reais são variáveis contínuas.

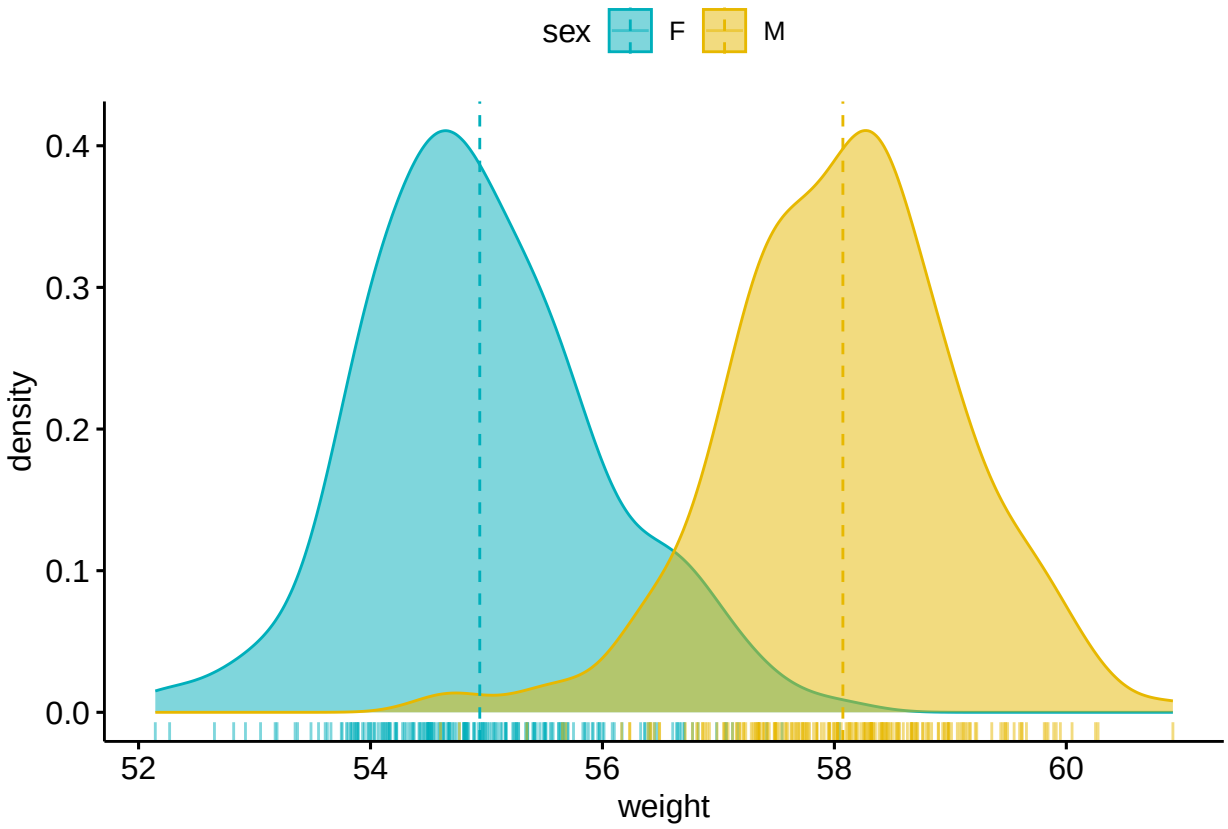
Em R, variáveis contínuas são chamadas de `numéricas`; ou, mais precisamente, `numeric`.

$$A = \{10.7,\ 92, -43.223,\ \pi,\ \sqrt{2}7,\ \ \sin\theta\} \quad (1.4)$$

Exemplos de fenônemos que podem ser explicados por variáveis contínuas:
- O peso corporal

- A altura humana
- Os valores de troponina ultrassensível
- O diâmetro de uma artéria coronária descendente anterior
- A renda familiar...


sex
F
M
30
20
10
0
count
52
54
56
58
60
weight



Quer construir VOCÊ uma distribuição normal em linguagem R?

1.4 Exercícios

Em todos os exercícios você deverá utilizar boas práticas em programação (ex: comentar o código, identificar corretamente os objetos, não omitir etapas que parecem óbvias etc.).

Estes deverão ser realizados em R Markdown.

https://livro.curso-r.com/9-2-r-markdown.html

1.4.1. Operações matemáticas

- a) Você deve realizar 8 operações matemáticas básicas, sendo 2 para soma, 2 para subtração, 2 para multiplicação e 2 para divisão. Para cada tipo de operação, uma delas deve ser feita com números e a outra com objetos

- b) Escreva uma operação em modo função de primeiro grau (ex: f(x) = ax + b). Escreva-a no .Rmd em linguagem LaTeX (lista completa de símbolos: https://linorg.usp.br/CTAN/info/symbols/comprehensive/symbols-a4.pdf) e identifique o que cada termo da equação quer dizer em Estatística, com um exemplo voltado à saúde.
- c) Crie um vetor, armazenado em um objeto qualquer, de palavras simples e compostas
- d) Crie um vetor, armazenando em um objeto qualquer, que misture palavras e números contínuos

1.4.2. Linguagem R e Estatística Descritiva

- e) Realize a plotagem de um gráfico hipotético da distribuição normal. Abaixo há instruções detalhadas da estrutura do código.

```
-  Defina os atributos do banco de dados fictício que você irá criar. Você vai precisar:

  -  Conjunto de dados observados (eixo X)
  -  Distribuição de probabilidades deste conjunto, com base na Normal teórica
  -  Utilização correta da função que gerará a distribuição desejada.


-  Aqui vai um esqueleto pra você:


  -  X: <nome_do_objeto_1> <- seq(min,max, length=n) - Este comando indica
  que você definirá, para os valores de X, um banco fictício, no formato sequência
  de valores. Você deve colocar um valor mínimo, um valor máximo, e no atributo length,
  indicar a quantidade de números no vetor


  -  Y: <nome_do_objeto_2> <- dnorm(<objeto_que_representa_X>) - Com este comando,
  você cria um novo banco, que é referente às probabilidades de cada valor X ser observado.
  Como falado em aula, um gráfico de distribuição de observações possui em seu eixo X
  os valores possíveis e observados, e no eixo Y, a probabilidade acumulada deste
  valor ocorrer. O comando indicado significa: você criará uma distribuição normal
  dos valores de $X$, utilizando a função dnorm, e indicando, dentro dela, o objeto
  previamente criado que representa o vetor $X$


  - <nome_do_objeto_3> <- plot(x, y, xlab="legenda do eixo X", ylab="legenda no eixo Y",
  main="Título" ) - função plot gera o gráfico com base nos atributos colocados


  -  Quarto passo --> agora é com você! :)


#  Caso tenha dúvidas ou algum comando não rode, utilize, no console,
a função help("<nome_da_função>") ou digite ?<nome_da_função>.
```

```
#  Por vezes o que pode travar um código é o uso de letras maiúsculas; acentos;
a ausência de parênteses para fechar cada bloco de código; a ausência de aspas
nas palavras; espaços não-solicitados.
```

- f) Indique, por favor, a estimativa da média, da mediana, da probabilidade máxima encontrada e do desvio-padrão amostral. Neste momento não é necessário utilizar nenhuma função para retornar tais valores - tente identificar visualmente e aponte no documento os locais de referência utilizados

1.4.3. Pré-exposição a conteúdo futuro

- g) Aqui a ideia é que você tente por si mesmo/a identificar conceitos fundamentais que aprenderemos na próxima aula. Use e abuse da literatura de apoio, assim como da minha disponibilidade e do nosso fórum via Discord. :)

- Traga o conceito formal e o conceito prático de uma variável discreta, com referências

- Tente escrever a notação formal da mesma tal qual foi feito neste documento. Use a lógica :)
- Identifique o nome computacional deste tipo de variável. É universal
- Crie um vetor que contenha somente variáveis discretas, armazene-o em um objeto e leia-o
- Peça formalmente em linguagem R para que o R Studio teste se trata-se de uma variável discreta mesmo ou não (TRUE ou FALSE)
- Identifique o conceito de `array` em linguagem R
- Identifique o ceonceito de `fator` em linguagem R
- Identifique o conceito de `data frame` em linguagem R
- Identifique o conceito de `pipe` em linguagem R, traga sua simbologia e exemplifique como uma sintaxe com e sem `pipe` é escrita
- Identifique o conceito de `packages` em R e tente listar os 10 pacotes essenciais para o estudo da Estatística e da Epidemiologia que não vêm nativos quando fazemos o download da linguagem R. Com uma breve busca na internet, em fóruns, é fácil de encontrar.
- Demonstre como se `instala` um pacote e como se `lê` um pacote utilizando o ambiente R Studio.
