

CAR00110 - Aula 01

Linguagem R: Sintaxe Básica

Prof. Lucas Helal, MMSc, PhD

2023-08-12

1. A representação simbólica da linguagem humana.

Notações e símbolos, de forma escrita, são corriqueiramente usados para substituir expressões da convivência humana em diversas circunstâncias do nosso dia a dia.

Por exemplo, suponhamos que você tenha realizado anotações ao acordar pela manhã:

E então você rasurou o texto com marcações...

Está claro que você quis comunicar alguma coisa, não? Quer seja para uma outra pessoa que possa ler; quer seja pra você mesmo... Com símbolos comuns, fáceis de interpretar, tanto pra mim quanto pra você... e provavelmente para a maioria das pessoas ao redor do mundo.

Porém, é provável que nem sempre tenha sido assim. Mesmo depois do advento do conceito de idioma e também de escrita. Nem sempre foi assim para a humanidade, nem sempre foi assim para mim ou para você, que precisamos de tempo, após aprender a ler e escrever. Como tantas outras coisas - mas, depois de um tempo ficou fácil e universal; deixou de ser assustador. E aqui está por que ressalto isto.

Na ciência, o uso de notações e outras formas de comunicar, cada vez mais inovadoras, é abundante. Algumas vezes universal, algumas vezes de significado particular. Ponto é que todo cientista que consegue entender notações de sua área, assim como linguagens não-humanas, hoje, nem sempre conseguia.

Portanto, ao discutirmos sintaxe e notação científica particulares da Estatística e da programação (em nosso caso particular, a linguagem R), é natural que precisemos esperar a curva de aprendizado se fazer valer; e fundamental não se assustar.

1.1 Taxonomia Aplicada à Ciência

1.1.1 Fundamentos

Metafísica: ramo da Filosofia que se ocupa da natureza fundamental da realidade

Epistemologia: ramo da Filosofia que se ocupa do conhecimento científico; por vezes chamada de **Filosofia da Ciência**

Conhecimento: ato de conhecer, sendo uma intersecção de crenças e da verdade

Crença: estado imaginário em que se assume algo ser verdade ou provável

Verdade: propriedade de estar em acordo com o fato real ou a realidade; não é sinônimo de científico e vice-versa; é independente da crença ou do conhecimento prévio

Ontologia: ramo da Filosofia que atribui o conceito de realidade às coisas sem necessariamente se apropriar de fatos, onde a prova lógica é suficiente

Ciência: *stricto sensu*, a maneira de adquirir conhecimento com base no método científico; maneira de se apropriar da verdade por meio do método científico; só é declarativa por meio daquilo que é empírico

Empírico: tipo de conhecimento gerado por meio de testes e experimentos; fundamental para a existência do conceito de Ciência

Método Científico: metodologia de sistema próprio criada como uma das alternativas possíveis para a apropriação da realidade

1.1.2 Método Científico

Conceitos

Teoria: são modelos explicativos bem fundamentados para descrever eventos que existem na natureza, envolvendo fatos, lei, hipóteses, evidências que resultaram de experimentos baseados no método científico

Teorema: proposição demonstrável por axiomas e postulados

Lei: regra que descreve um fenômeno ocorrente com certa regularidade, provada empiricamente

Corolário: consequências de um teorema; é uma afirmação; deduzida de uma verdade já demonstrada

Princípio: proposição que se põe no início de uma dedução, assumida como não pertencente ao que vai ser deduzido, e, provisoriamente, admitida como inquestionável

Postulado: proposição que foi admitida como verdade sem necessidade de demonstração, mas que o caráter não-empírico não é aparente

Axioma: proposição que foi admitida como verdade sem necessidade de demonstração, mas que é visivelmente de caráter não-empírico

Indução: extrapolação de um determinado atributo observado em casos particulares para o todo

Dedução: caracterização de um atributo a casos particulares por meio do que foi observado no todo

Inerte: atributo próprio de uma teoria, de ser imune à argumentos de autoridade, hierarquias ou daquilo que não é capaz de refutá-la que não seja empiricamente

Particularidades

Observação: qualquer coisa que se modifique em um experimento científico

Hipótese nula: uma afirmação temporária a se tentar refutar; uma teoria; ponto de partida que existe no mundo das ideias; local de referência para se testar uma hipótese

Hipótese alternativa: aquilo que o pesquisador imagina ser verossímil de encontrar em um dado experimento; é resultado da pergunta de pesquisa; afirmação temporária como resposta à pergunta de pesquisa, anterior ao experimento

Dado: conjunto de significado que comporá uma ou mais variáveis (inclusive exposições e desfechos)

Variável: é um aglutinado de dados que compõe parte do experimento; um conceito operacionalizado que variará experimento a experimento

Exposição: um tipo de variável à qual atribuímos a capacidade de modificar outra variável; o fator causal/associativo candidato em uma dada hipótese pode ser chamado de exposição

Desfecho: um tipo de variável à qual atribuímos a capacidade de ser modificada por outra variável; o fator que indicará evidência ou não de que a hipótese em estudo foi tensionada

Argumentos Clássicos

Modus ponens

Também conhecido como afirmação do antecedente; regra de inferência simples denotada por $\frac{p \implies q, p}{\therefore q}$, isto é: “ p implica q ; p é verdade; portanto, q deve ser verdade.”

Modus tollens:

“*O modo que nega por negação.*” Nome formal de uma regra de inferência indireta, denotado por $\frac{p \implies q, \neg q}{\therefore \neg p}$, isto é: “ p implica q ; q é falso; portanto, p deve ser falso.”

Silogismo hipotético:

Conhecido como inferência por consequência, denotado por: $\frac{p \implies q \wedge q \implies r}{\therefore p \implies r}$, isto é: “ p implica q ; q implica r ; portanto, p implicar r deve ser verdadeiro.”

Simbologia Lógica

\exists : *existe pelo menos um...*

$\exists!$: *existe um e apenas um...*

\nexists : *não existe...*

\forall : *para todo...*

\neg : *negação*

\vee : *ou*

\wedge : *e*

\implies : *implica*

\Rightarrow : *o que está implicado está à direita*

\Leftarrow : *é implicado por...; somente se*

\Leftarrow : *o que está implicado está à esquerda*

\Leftrightarrow : *se, e somente se...*

\Leftrightarrow : *equivale a...*

\equiv : *equivale a...*

\therefore : *então...*

1.2 Taxonomia Aplicada à Estatística

Exemplo

p_1 : A quantidade de vezes que o valor de uma variável aparece.

p_2 : f_j

p_3 : $p_1 = p_2$, onde:

p_1 é a linguagem humana e p_2 a notação usual para simbolizar, de forma genérica, que uma dada variável e , de valor f , apareceu ou aparecerá j vezes.

A isto chamamos de conversão, que pode ser congruente ou não-congruente - a primeira, depende de três critérios:

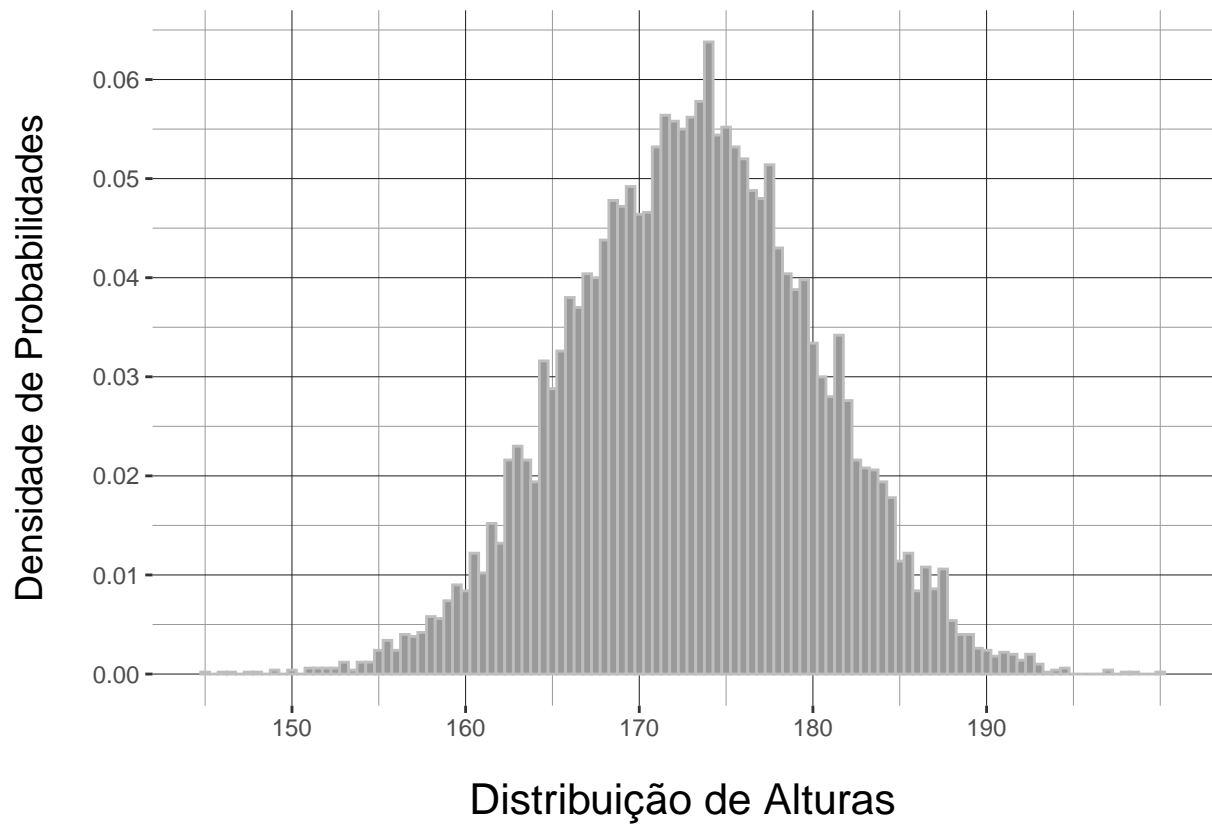
- a) correspondência semântica entre os elementos;
- b) a desambiguidade semântica dos elementos;
- c) e uma única ordem possível de disposição dos elementos.

Na premissa **P3**, caso o primeiro termo estivesse *em maiúsculo*; ou o segundo termo *sobrescrito* em vez de *subscrito*; a premissa **P3** não seria verdadeira, pois haveria violação da premissa **P2**, por não atender a convenção.

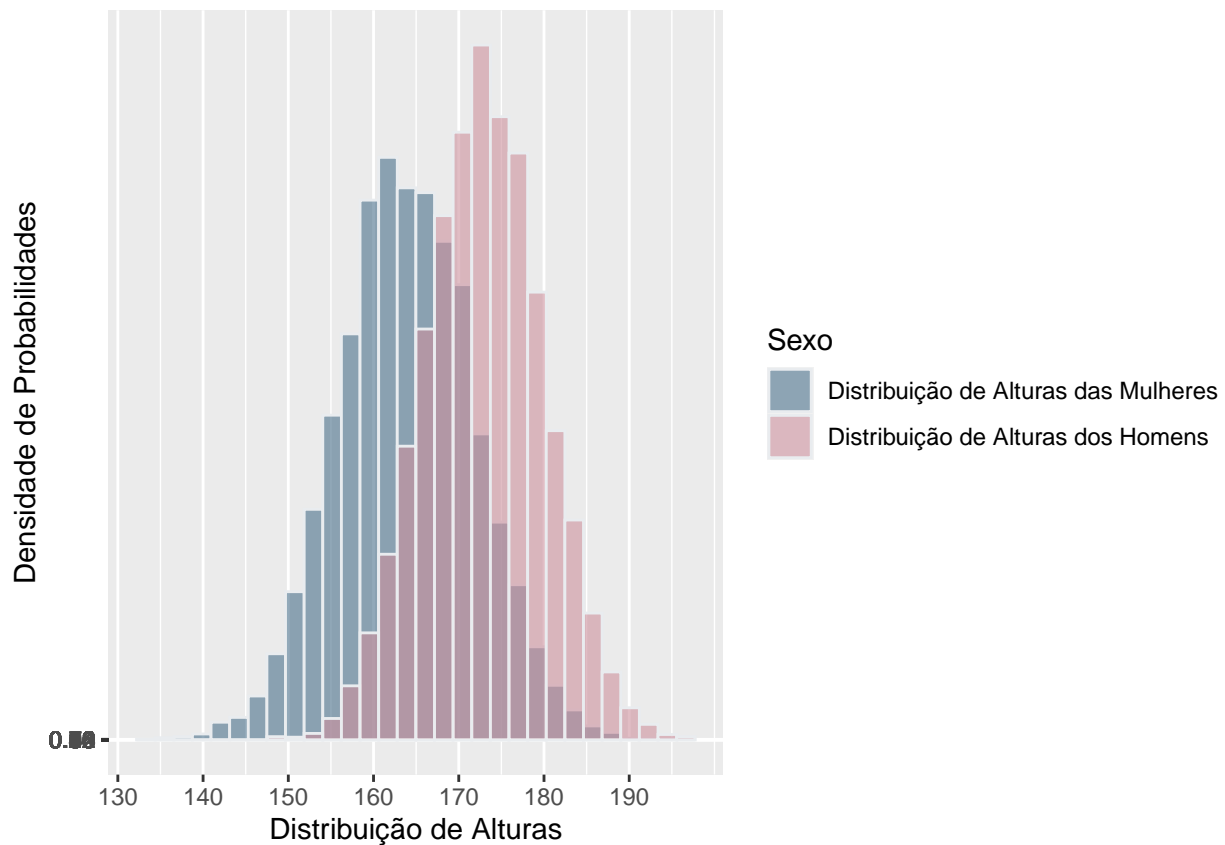
Vamos aprender Estatística em Linguagem R

O que você enxerga nesse gráfico?

##	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
## Mazda RX4	21.0	6	160	110	3.90	2.620	16.46	0	1	4	4
## Mazda RX4 Wag	21.0	6	160	110	3.90	2.875	17.02	0	1	4	4
## Datsun 710	22.8	4	108	93	3.85	2.320	18.61	1	1	4	1
## Hornet 4 Drive	21.4	6	258	110	3.08	3.215	19.44	1	0	3	1
## Hornet Sportabout	18.7	8	360	175	3.15	3.440	17.02	0	0	3	2
## Valiant	18.1	6	225	105	2.76	3.460	20.22	1	0	3	1



E neste gráfico?



A se a gente amostrar médias de mesmo tamanho K vezes?

```
sum(sample(1:6, 2, replace = T))
```

```
## [1] 9
```

Aqui estão todas as combinações possíveis de um conjunto $S = 1, 2, 3, 4, 5, 6$ tomado 2 a 2 $\therefore A = K_2$.

(1, 1)(1, 2)(1, 3)(1, 4)(1, 5)(1, 6)
 (2, 1)(2, 2)(2, 3)(2, 4)(2, 5)(2, 6)
 (3, 1)(3, 2)(3, 3)(3, 4)(3, 5)(3, 6)
 (4, 1)(4, 2)(4, 3)(4, 4)(4, 5)(4, 6)
 (5, 1)(5, 2)(5, 3)(5, 4)(5, 5)(5, 6)
 (6, 1)(6, 2)(6, 3)(6, 4)(6, 5)(6, 6)

O que dá as seguintes probabilidades:

$$P(S) = \begin{cases} 1/36, & S = 2 \\ 2/36, & S = 3 \\ 3/36, & S = 4 \\ 4/36, & S = 5 \\ 5/36, & S = 6 \\ 6/36, & S = 7 \\ 5/36, & S = 8 \\ 4/36, & S = 9 \\ 3/36, & S = 10 \\ 2/36, & S = 11 \\ 1/36, & S = 12 \end{cases} \quad (1)$$

Simulação 1:

```
# Vetor para desfechos
S <- 2:12

# Vetor para probabilidades
PS <- c(1:6, 5:1) / 36

# Esperança/Parâmetro de S
ES <- sum(S * PS)

# Variância de S
VarS <- sum((S - ES)^2 * PS)
VarS
```

```
## [1] 5.833333
```

Então:

```
S
```

```
## [1] 2 3 4 5 6 7 8 9 10 11 12
```

```
PS
```

```
## [1] 0.02777778 0.05555556 0.08333333 0.11111111 0.13888889 0.16666667
## [7] 0.13888889 0.11111111 0.08333333 0.05555556 0.02777778
```

```
ES
```

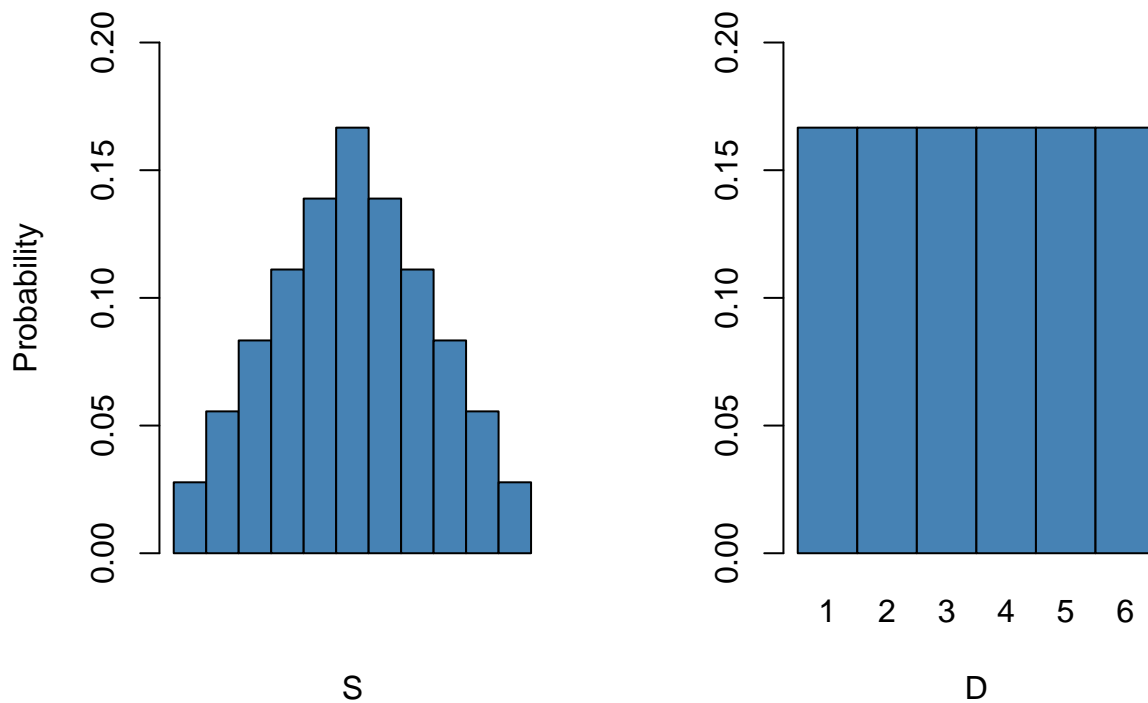
```
## [1] 7
```

```
VarS
```

```
## [1] 5.833333
```

Cria-se um mini data set:

Somatório de Duas Rodadas de um jogo que acontece se eu jogo apenas um



O Teorema do Limite Central

A demonstração matemática da média amostral (\hat{x})

$$Y = \frac{1}{n} \sum_{i=1}^n Y_i = \frac{1}{n} (Y_1 + Y_2 + \dots + Y_n).$$

A demonstração da estimativa da média e variância populacionais (parâmetros):

$$E(Y) = E\left(\frac{1}{n} \sum_{i=1}^n Y_i\right) = \frac{1}{n} E\left(\sum_{i=1}^n Y_i\right) = \frac{1}{n} \sum_{i=1}^n E(Y_i) = \frac{1}{n} \cdot n \cdot \mu_Y = \mu_Y$$

$$\begin{aligned}
\text{Var}(Y) &= \text{Var}\left(\frac{1}{n} \sum_{i=1}^n Y_i\right) \\
&= \frac{1}{n^2} \sum_{i=1}^n \text{Var}(Y_i) + \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1, j \neq i}^n \text{cov}(Y_i, Y_j) \\
&= \frac{\sigma_Y^2}{n} \\
&= \sigma_Y^2/n.
\end{aligned}$$

$$\sigma_Y = \frac{\sigma_Y}{\sqrt{n}}.$$

A distribuição amostral de médias se $X \sim N$

$$Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2/n)$$

Amostragem

Aqui, plotaremos a distribuição amostral de médias

```
# Check se 'sample.avg' é um vetor
is.vector(sample.avg)
```

```
## [1] TRUE
```

```
# Primeiras entradas no console
head(sample.avg)
```

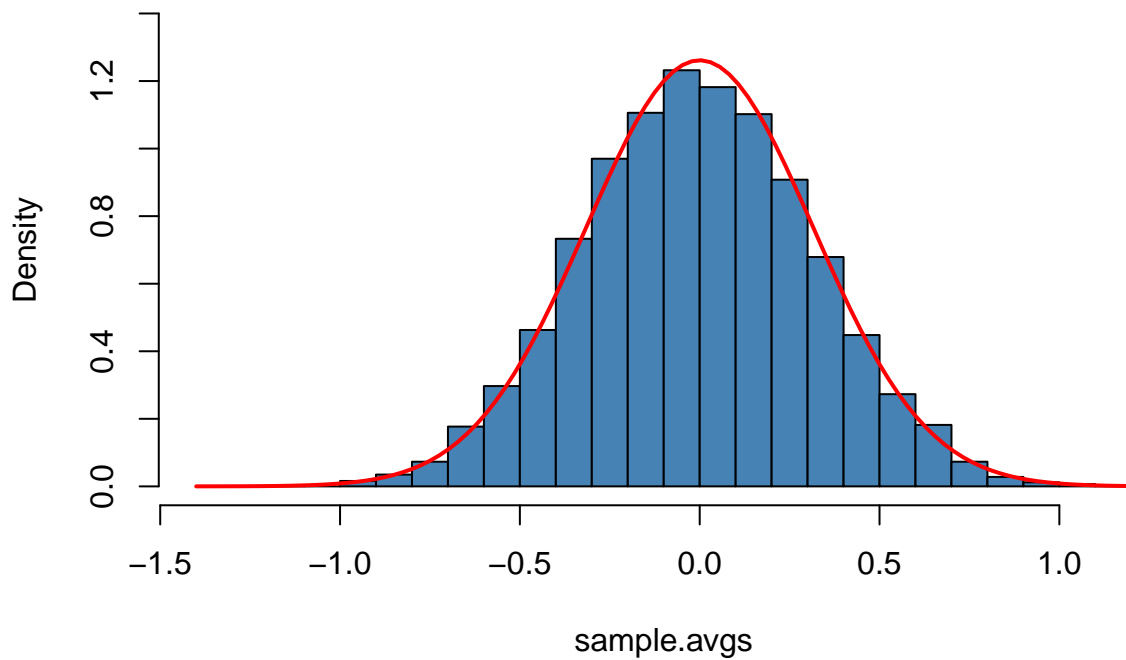
```
## [1] 0.02564534 -0.04577475 -0.14469432 -0.26804360 -0.08673719 0.40976494
```

Histograma de Densidades

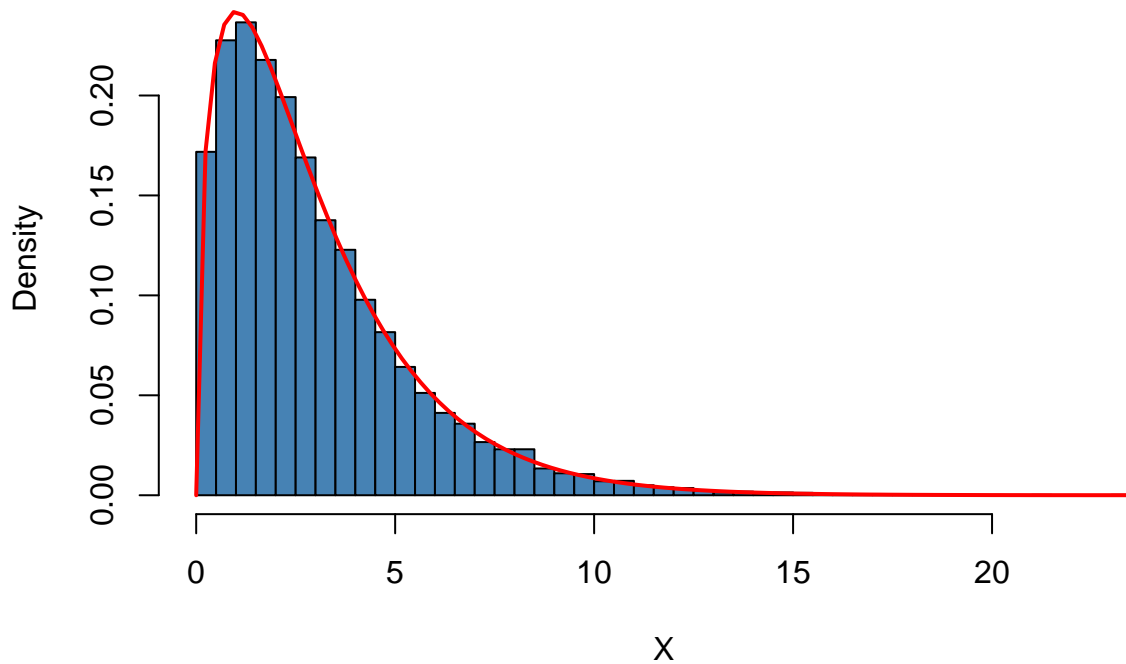
```
# Plot do histograma
hist(sample.avg,      # chama a função
      ylim = c(0, 1.4), # limite superior e inferior vertical
      col = "steelblue", # cor
      freq = F,         # atributo de reposição
      breaks = 20)      # número de colunas
```

```
# Distribuição de probabilidades (teórica) por cima do histograma
curve(dnorm(x, sd = 1/sqrt(n)),      # chama a função `curve` e a função densidade de prob `dnorm`
      col = "red",                  # cor da linha
      lwd = "2",
      add = T)
```

Histogram of sample.avgs



Agora se, modificarmos o espaço amostral e os graus de liberdade



E aqui demonstramos visualmente o Teorema do Limite Central

