

Curso de Programação Estatística

Andressa Cerqueira, Rafael Izbicki, Thiago Ramos

2024-10-09

Table of contents

Curso de Programação Estatística	7
Objetivo Geral	7
1 Breve Introdução ao Python	8
2 Breve Introdução ao R	9
2.1 Variáveis Numéricas	9
2.2 Variáveis Lógicas	11
2.3 Caracteres/Strings	13
2.4 Vetores e Sequências	14
2.5 Funções	17
2.6 Laços	19
2.7 If/Else	21
2.8 Listas	22
2.9 Estatística Descritiva	23
2.10 O operador %>% (pipe)	27
2.11 O pacote dplyr	29
2.12 O pacote ggplot2	35
2.13 O pacote tidyr	44
2.14 Leitura de Arquivos com readr	48
2.15 Exercícios	50
3 Geração de Números Aleatórios e Aplicação em Estatística	52
3.1 Objetivo da Aula	52
3.2 Conteúdo Teórico	52
3.3 Exemplo de Problema	52
4 R	53
5 Python	55
5.1 Exemplo: Simulação de um Jogo de Dados com Dado “Viciado”	56
6 R	58
7 Python	60
7.1 Pergunta: precisamos da função np.random.choice/set.seed?	61

8 R	62
9 Python	66
10 Números Pseudoaleatórios	69
10.1 O que é um número pseudoaleatório?	69
10.2 Geração de Números Pseudoaleatórios com o Gerador Linear Congruente (LCG)	69
11 R	71
12 Python	72
12.1 Por que o Gerador Linear Congruente Funciona?	72
13 R	75
14 Python	77
15 R	80
16 Python	82
16.1 Gerando números uniformes com uma moeda	83
17 R	84
18 Python	86
19 Técnica da Inversão para Variáveis Discretas	88
19.1 Geração de Variáveis Aleatórias Discretas Genéricas	88
20 R	89
21 Python	91
21.1 Inversa da CDF	92
21.2 Geração de Variáveis Aleatórias com Distribuição Geométrica	92
22 R	95
23 Python	98
23.1 Geração de Variáveis Aleatórias com Distribuição Poisson	100
24 R	102
25 Python	105
25.1 Exercícios	107

26 Técnica da Inversão para Variáveis Contínuas	109
26.1 Função Inversa	109
27 R	110
28 Python	112
28.1 Método da Inversão	114
28.2 Exemplo 1	114
29 R	115
30 Python	117
31 R	119
32 Python	121
32.1 Simulação de transformações de variáveis aleatórias	122
33 R	123
34 Python	125
35 R	127
36 Python	129
36.1 Exercícios	130
37 Método da Rejeição para Variáveis Discretas	132
38 Método da Rejeição para Variáveis Contínuas	133
38.1 Algoritmo	133
39 R	134
40 Python	136
40.1 Exemplo	138
41 R	139
42 Python	141
42.1 Resultados Teóricos	142
42.2 Exercícios	145
43 Transformações e Misturas	146
44 R	147

45 Python	148
45.1 Exercícios	148
46 Método de Box-Muller	149
47 R	150
48 Python	151
48.1 Exercícios	151
49 Método de Monte Carlo	152
49.1 Exemplo 1: Estimativa de uma Integral	153
50 R	154
51 Python	155
52 R	156
53 Python	158
53.1 Exemplo 2: Aproximando uma Probabilidade	159
54 R	160
55 Python	161
55.1 Exemplo 3: Aproximando o valor de π	161
56 R	162
57 Python	163
58 R	166
59 Python	168
59.1 Intervalos de Confiança para Estimativas de Monte Carlo	169
60 R	171
61 Python	172
61.1 Exemplo 4: Problema das Figurinhas	173
62 R	174
63 Python	175
64 R	176

65 Python	177
66 R	179
67 Python	180
68 R	181
69 Python	182
70 R	183
71 Python	184
71.1 Exercícios	184
References	186

Curso de Programação Estatística

Objetivo Geral

Este curso visa explorar o impacto das representações numéricas nos resultados de algoritmos de análise estatística. O foco será na programação, visualização e preparação de dados, além de discutir tópicos importantes como aleatoriedade, pseudoaleatoriedade, erros de truncamento e arredondamento, entre outros. O curso inclui ainda uma introdução à inferência por simulação estocástica, utilizando métodos como Monte Carlo e integrações numéricas. O material do curso foi amplamente baseado nas discussões apresentadas em Ross (2006).

Este livro apresenta códigos tanto em [R](#) quanto em [Python](#)

Autores: [Andressa Cerqueira](#), [Rafael Izbicki](#), [Thiago Rodrigo Ramos](#)

1 Breve Introdução ao Python

2 Breve Introdução ao R

Aqui faremos uma breve introdução ao R. Para uma introdução mais delhada, veja <https://curso-r.com/material/> e <https://r4ds.had.co.nz/>.

2.1 Variáveis Numéricas

Aqui vão alguns exemplos para começarmos a entender como o R funciona. Inicialmente, veremos como podemos usar o R como calculadora.

```
2 * 7
```

```
[1] 14
```

```
0 / 0
```

```
[1] NaN
```

```
10^1000
```

```
[1] Inf
```

```
log(1)
```

```
[1] 0
```

```
exp(1)
```

```
[1] 2.718282
```

Para entender o que uma função faz, você pode digitar o símbolo de interrogação seguido do nome da função, por exemplo:

```
?exp
```

O help contém as seguintes informações:

- **Description:** Resumo geral sobre o uso da função.
- **Usage:** Mostra como a função deve ser utilizada e quais argumentos podem ser especificados.
- **Arguments:** Explica o que é cada um dos argumentos.
- **Details:** Explica alguns detalhes sobre o uso e aplicação da função.
- **Value:** Mostra o que sai no output após usar a função (os resultados).
- **Note:** Notas sobre a função.
- **Authors:** Lista os autores da função.
- **References:** Referências para os métodos usados.
- **See also:** Outras funções relacionadas que podem ser consultadas.
- **Examples:** Exemplos do uso da função.

2.1.1 Armazenando resultados em variáveis

Podemos também armazenar resultados de contas em variáveis. Por exemplo:

```
x = 2 + 3
x
```

```
[1] 5
```

```
y = 2 * x
y
```

```
[1] 10
```

```
print(y)
```

```
[1] 10
```

Números grandes são impressos usando notação científica:

```
y = 2 * 1010
print(y)
```

```
[1] 2e+10
```

Para listar quais variáveis estão declaradas no ambiente, podemos usar:

```
ls()
```

```
[1] "x" "y"
```

Para remover uma variável:

```
rm(x)
ls()
```

```
[1] "y"
```

Para remover todas as variáveis existentes:

```
rm(list = ls()) # Essa função apaga todas as variáveis existentes
gc()           # Essa função libera a memória utilizada
```

```
      used (Mb) gc trigger (Mb) max used (Mb)
Ncells 628609 33.6   1279765 68.4   1279765 68.4
Vcells 1185295  9.1    8388608 64.0   1951069 14.9
```

2.2 Variáveis Lógicas

Alguns operadores lógicos definidos no R são mostrados na tabela abaixo:

#	Operador	Descrição
1	<code>x < y</code>	x menor que y?
2	<code>x <= y</code>	x menor ou igual a y?
3	<code>x > y</code>	x maior que y?
4	<code>x >= y</code>	x maior ou igual a y?
5	<code>x == y</code>	x igual a y?
6	<code>x != y</code>	x diferente de y?
7	<code>!x</code>	Negativa de x
8	<code>x y</code>	x ou y são verdadeiros?

#	Operador	Descrição
9	<code>x & y</code>	x e y são verdadeiros?
10	<code>xor(x, y)</code>	Apenas um dos dois é verdadeiro?

2.2.1 Exemplos de uso de operadores lógicos

```
1 == 3
```

```
[1] FALSE
```

```
1 == 1
```

```
[1] TRUE
```

```
1 <= 3
```

```
[1] TRUE
```

```
x = 1 > 3
print(x)
```

```
[1] FALSE
```

```
x = 1; y = 3
x < y
```

```
[1] TRUE
```

```
(x == 1) & (y == 3)
```

```
[1] TRUE
```

```
(x == 1) & (y != 3)
```

```
[1] FALSE
```

```
!TRUE
```

```
[1] FALSE
```

```
x = TRUE; y = FALSE  
x | y
```

```
[1] TRUE
```

O número um tem o valor verdadeiro, e o número zero tem o valor falso:

```
x = 1; y = 0  
x | y
```

```
[1] TRUE
```

Note que ao fazermos contas com variáveis lógicas, elas são transformadas em zero e um:

```
(TRUE | FALSE) + 2
```

```
[1] 3
```

2.3 Caracteres/Strings

Para declarar cadeias de caracteres, usamos aspas no R:

```
x = "Rafael"  
y = "Izbicki"
```

Várias operações podem ser feitas com esses objetos. Um exemplo é a função `paste`:

```
paste(x, y)
```

```
[1] "Rafael Izbicki"
```

```
paste(x, y, sep = "+")
```

```
[1] "Rafael+Izbicki"
```

2.4 Vetores e Sequências

Usualmente declaramos vetores usando o operador `c` (concatenação):

```
x = c(2, 5, 7, 1)
x
```

```
[1] 2 5 7 1
```

Para acessar seus componentes, fazemos:

```
x[2]
```

```
[1] 5
```

```
# [1] 5
```

```
x = x[c(2, 3)]
x[2:3]
```

```
[1] 7 NA
```

```
x[-c(2, 3)]
```

```
[1] 5
```

As operações `x[c(2, 3)]` e `x[-c(2, 3)]` são chamadas de *subsetting*; subsetting é a seleção de um subconjunto de um objeto. Veremos mais à frente outras maneiras de fazermos isso.

Podemos mudar alguns dos valores deste vetor usando:

```
x[2:3] = 0
x
```

```
[1] 5 0 0
```

Operações aritméticas podem ser facilmente feitas para cada elemento de um vetor. Alguns exemplos:

```
x = x - 1
x
```

```
[1] 4 -1 -1
```

```
x = 2 * x
x
```

```
[1] 8 -2 -2
```

Uma função útil para criar vetores é a `seq`:

```
x = seq(from = 1, to = 100, by = 2)
x
```

```
[1] 1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49
[26] 51 53 55 57 59 61 63 65 67 69 71 73 75 77 79 81 83 85 87 89 91 93 95 97 99
```

```
y = seq(from = 1, to = 100, length = 5)
y
```

```
[1] 1.00 25.75 50.50 75.25 100.00
```

Podemos também usar operadores lógicos em vetores:

```
x > 5
```

```
[1] FALSE FALSE FALSE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
[13] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
[25] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
[37] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
[49] TRUE TRUE
```

Note que `x > 5` é um vetor lógico.

Podemos usar as operações lógicas em vetores lógicos:

```
x = c(TRUE, TRUE, FALSE, FALSE)
y = c(TRUE, FALSE, TRUE, FALSE)

x | y
```

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

```
x & y
```

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

Operadores lógicos também podem ser usados para fazer *subsetting*:

```
x[x > 5] = 0
x
```

```
[1] 1 1 0 0
```

```
x[!(x <= 0)] = -1
x
```

```
[1] -1 -1 0 0
```

Podemos ter vetores de caracteres:

```
x = c("a", "c", "fgh")
x[-c(2, 3)]
```

```
[1] "a"
```

```
paste(x, "add")
```

```
[1] "a add" "c add" "fgh add"
```



```
paste(x, "add", collapse = "+")
```

```
[1] "a add+c add+fgh add"
```

Algumas funções adicionais úteis:

```
rep(2, 5)
```

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

```
rep(c(1, 2), 5)
```

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

```
rep(c(1, 2), each = 5)
```

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

```
sort(c(5, 10, 0, 20))
```

```
[1] 0 5 10 20
```

```
order(c(5, 10, 0, 20))
```

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

2.5 Funções

Para declarar funções em R, fazemos:

```
minhaFuncao <- function(argumento1, argumento2) {  
  # corpo da função  
}
```

Por exemplo:

```
potencia <- function(x, y) {  
  return(x^y)  
}  
potencia(2, 3)
```

```
[1] 8
```

```
# [1] 8
```

Note que `x` e `y` podem ser vetores:

```
potencia(x = c(1, 2, 3), y = c(0, 1, 2))
```

```
[1] 1 2 9
```

Você pode inverter a ordem dos argumentos, desde que eles sejam nomeados:

```
potencia(y = c(0, 1, 2), x = c(1, 2, 3))
```

```
[1] 1 2 9
```

2.5.1 Argumentos com valores default

Os argumentos podem ter valores *default*. Por exemplo:

```
potencia <- function(x, y = rep(1, length(x))) {  
  return(x^y)  
}
```

Neste caso, se o argumento `y` não for fornecido, ele será um vetor de uns do tamanho de `x`:

```
potencia(2)
```

```
[1] 2
```

```
potencia(2, 3)
```

```
[1] 8
```

2.6 Laços

Para calcular o fatorial de um número n , podemos usar um laço `while`:

```
n <- 4
i <- 1
nFatorial <- 1
while(i <= n) {
  nFatorial <- nFatorial * i
  i <- i + 1
}
nFatorial
```

```
[1] 24
```

Ou podemos usar um laço `for`:

```
n <- 4
nFatorial <- 1
for(i in 1:n) {
  nFatorial <- nFatorial * i
}
nFatorial
```

```
[1] 24
```

Lembre-se de que laços podem ser lentos no R, especialmente se o tamanho do objeto não for previamente declarado. Vamos comparar o tempo de execução de diferentes abordagens usando a função `system.time`. Para isso, vamos calcular a soma acumulada de um vetor.

```
x <- 1:100000
```

1. Laço sem declaração de tamanho:

```

aux <- function(x) {
  somaAcumulada <- NULL
  somaAcumulada[1] <- x[1]
  for(i in 2:length(x)) {
    somaAcumulada[i] <- somaAcumulada[i - 1] + x[i]
  }
}
system.time(aux(x))[1]

```

```

user.self
0.032

```

2. Laço com declaração de tamanho:

```

aux <- function(x) {
  somaAcumulada <- rep(NA, length(x))
  somaAcumulada[1] <- x[1]
  for(i in 2:length(x)) {
    somaAcumulada[i] <- somaAcumulada[i - 1] + x[i]
  }
}
system.time(aux(x))[1]

```

```

user.self
0.014

```

3. Função nativa do R:

```

aux <- function(x) {
  cumsum(x)
}
system.time(aux(x))[1]

```

```

user.self
0.001

```

Note que, em geral, funções nativas do R são muito mais rápidas.

2.7 If/Else

O R permite o uso de condicionais `if` e `else` para controlar o fluxo de execução. A sintaxe básica é:

```
if (condição1) {  
  # código se a condição1 for verdadeira  
} else if (condição2) {  
  # código se a condição1 for falsa e condição2 verdadeira  
} else {  
  # código se todas as condições forem falsas  
}
```

Exemplo:

```
testaX <- function(x) {  
  if (!is.numeric(x)) {  
    print("x não é um número")  
  } else if (x > 0) {  
    print("x é positivo")  
  } else if (x < 0) {  
    print("x é negativo")  
  } else {  
    print("x é nulo")  
  }  
}  
testaX(5)
```

```
[1] "x é positivo"
```

```
testaX(-5)
```

```
[1] "x é negativo"
```

```
testaX("5")
```

```
[1] "x não é um número"
```

2.8 Listas

Listas são como vetores, mas podem conter componentes de diferentes tipos (inclusive outras listas).

```
minhaLista <- list("um", TRUE, 3, c("q", "u", "a", "t", "r", "o"), "cinco")
minhaLista[[3]]
```

```
[1] 3
```

Você também pode atribuir nomes aos elementos de uma lista:

```
minhaLista <- list(primeiro = "um", segundo = TRUE, terceiro = 3,
                  quarto = c("q", "u", "a", "t", "r", "o"), quinto = "cinco")
minhaLista$quinto
```

```
[1] "cinco"
```

Listas são frequentemente usadas para retornar várias saídas de uma função. Por exemplo, ao ajustar um modelo de regressão linear, os resultados são armazenados em uma lista:

```
x <- rnorm(100)
y <- 1 + 2*x + rnorm(length(x), 0.1)
ajuste <- lm(y ~ x)
typeof(ajuste)
```

```
[1] "list"
```

Os componentes da lista podem ser acessados com:

```
names(ajuste)
```

```
[1] "coefficients" "residuals"    "effects"      "rank"
[5] "fitted.values" "assign"       "qr"          "df.residual"
[9] "xlevels"      "call"        "terms"       "model"
```

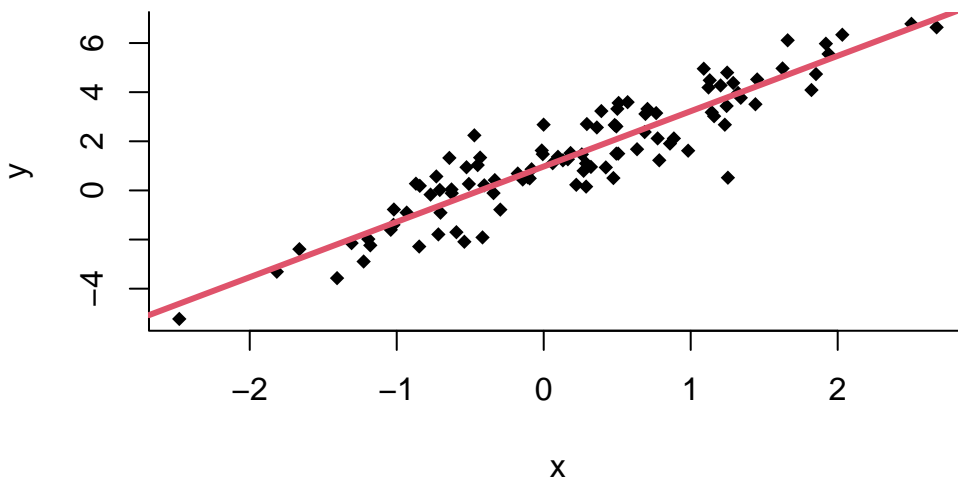
Para ver os coeficientes ajustados:

```
ajuste$coefficients
```

```
(Intercept)          x  
  0.9724057    2.2527105
```

Para adicionar a reta estimada ao gráfico:

```
plot(x, y, pch = 18, bty = "n")  
abline(ajuste, col = 2, lwd = 3)
```



Você também pode inicializar uma lista vazia:

```
minhaLista <- list()
```

2.9 Estatística Descritiva

O R é uma ferramenta poderosa para realizar análises descritivas. Vamos explorar alguns recursos para calcular estatísticas de variáveis quantitativas e qualitativas.

2.9.1 Variáveis quantitativas

```
x <- c(30, 1, 20, 5, 20, 60)  
y <- c(12, 0, 2, 15, 22, 20)  
  
mean(x)
```

```
[1] 22.66667
```

```
var(x)
```

```
[1] 448.6667
```

```
min(x)
```

```
[1] 1
```

```
which.min(x)
```

```
[1] 2
```

```
max(x)
```

```
[1] 60
```

```
which.max(x)
```

```
[1] 6
```

```
sort(x)
```

```
[1] 1 5 20 20 30 60
```

```
order(x)
```

```
[1] 2 4 3 5 1 6
```

```
median(x)
```

```
[1] 20
```



```
cor(x, y)
```

```
[1] 0.5229211
```

```
which(x > 10)
```

```
[1] 1 3 5 6
```

```
mean(y[x > 10])
```

```
[1] 14
```

2.9.2 Variáveis qualitativas

```
x <- c("S", "R", "P", "P", "Q", "P")  
y <- c("a", "a", "b", "a", "c", "d")  
  
x[y %in% c("a", "c")]
```

```
[1] "S" "R" "P" "Q"
```

```
table(x)
```

```
x  
P Q R S  
3 1 1 1
```

```
table(x, y)
```

```
      y  
x  a b c d  
P  1 1 0 1  
Q  0 0 1 0  
R  1 0 0 0  
S  1 0 0 0
```

2.9.3 Subconjuntos

Podemos utilizar subsetting para trabalhar com subconjuntos de interesse. Por exemplo, suponha que `dados` é um banco de dados com informações de alunos e queremos selecionar os alunos com idade menor que 20, altura maior que 1,70 e que estejam no primeiro ou segundo ano.

```
dados[dados$idade < 20 & dados$altura > 1.70 & dados$ano %in% c("Primeiro", "Segundo"), ]
```

2.9.4 Funções Apply

As funções da família `apply` (como `apply`, `lapply`, `sapply`) são úteis para aplicar uma função a um conjunto de dados. Vamos criar um banco de dados artificial para exemplificar:

```
dados <- data.frame(altura = rnorm(100, 1.7, 0.2), salario = rnorm(100, 3000, 500))
dados$peso <- dados$altura * 35 + rnorm(100, 0, 10)
```

2.9.4.1 `apply`

Para calcular a média de cada coluna de `dados`, usamos:

```
apply(dados, 2, mean)
```

	altura	salario	peso
	1.677187	3052.303104	59.438122

Aqui, o argumento 2 indica que a operação deve ser aplicada a cada coluna. Para calcular a soma de cada linha:

```
apply(dados, 1, sum)
```

[1]	2062.346	3667.433	3077.190	3836.503	3658.016	3688.631	2296.333	3420.136
[9]	2751.138	3835.294	3521.357	2836.284	2847.069	2151.939	2168.341	2949.562
[17]	2023.807	2868.123	3596.276	3313.307	2138.991	3361.501	3233.059	3149.487
[25]	3296.655	2449.965	3148.872	2734.500	3754.310	3127.992	1611.616	3647.683
[33]	4017.822	3919.512	3382.521	3531.782	2606.824	3111.279	2717.251	2659.249
[41]	3845.139	3311.428	3156.635	3893.421	3254.953	2901.505	3371.004	2590.344
[49]	2112.866	3735.882	3406.237	3707.758	3374.567	2960.588	3537.058	2355.337
[57]	2465.351	2790.355	3980.847	3437.992	2101.955	2542.670	2978.760	2278.723

```
[65] 2828.053 4031.376 2518.753 3650.600 2892.416 3480.280 2703.380 3613.652
[73] 3133.596 2535.971 3732.412 2841.356 3924.315 3036.076 2784.834 3775.178
[81] 3543.083 2557.632 3367.170 3309.441 3595.662 3648.089 3112.886 2395.571
[89] 2936.567 3936.432 3273.067 2666.768 3294.457 3724.719 3790.321 3285.652
[97] 2265.502 2749.922 2259.781 3885.541
```

2.9.4.2 Funções anônimas

Usando funções anônimas, podemos calcular várias estatísticas descritivas ao mesmo tempo:

```
estatisticas <- apply(dados, 2, function(x) {
  listaResultados <- list()
  listaResultados$media <- mean(x)
  listaResultados$mediaAparada <- mean(x, trim = 0.1)
  listaResultados$mediana <- median(x)
  listaResultados$maximo <- max(x)
  return(listaResultados)
})

estatisticas$altura
```

```
$media
[1] 1.677187
```

```
$mediaAparada
[1] 1.666687
```

```
$mediana
[1] 1.664309
```

```
$maximo
[1] 2.298977
```

2.10 O operador %>% (pipe)

O operador pipe foi uma das grandes revoluções recentes do R, tornando o código mais legível. Esse operador está definido no pacote **magrittr**, e vários outros pacotes, como o **dplyr**, fazem uso dele (veja a próxima seção).

A ideia é simples: o operador %>% usa o resultado do lado esquerdo como o primeiro argumento da função do lado direito. Um exemplo:

```
library(magrittr)
x <- c(1, 2, 3, 4)
x %>% sum() %>% sqrt()
```

```
[1] 3.162278
```

Isso é equivalente a:

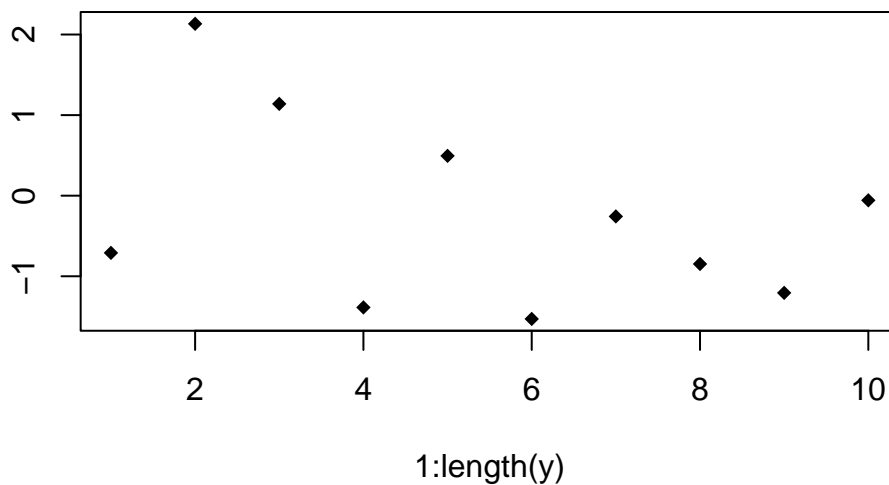
```
sqrt(sum(x))
```

```
[1] 3.162278
```

Mas a leitura com pipe é mais clara. No RStudio, você pode inserir o pipe com o atalho **ctrl + shift + m**.

O pipe envia o valor à esquerda apenas para o primeiro argumento da função à direita. Para utilizar o valor da esquerda em outro argumento, utilize o `"."`:

```
y <- rnorm(10)
y %>% plot(x = 1:length(y), y = ., pch = 18)
```



2.11 O pacote dplyr

O pacote `dplyr` é muito útil para manipulação eficiente de data frames. Vamos demonstrar alguns exemplos usando o conjunto de dados `hflights`.

```
library(dplyr)
library(hflights)
data(hflights)
```

Primeiro, converta os dados para o formato `tbl_df`, uma variação mais amigável do `data.frame`:

```
flights <- tbl_df(hflights)
```

2.11.1 Filter

A função `filter` retorna todas as linhas que satisfazem uma condição. Por exemplo, para mostrar todos os voos do dia 1º de janeiro:

```
flights %>% filter(Month == 1, DayofMonth == 1)
```

```
# A tibble: 552 x 21
```

	Year	Month	DayofMonth	DayOfWeek	DepTime	ArrTime	UniqueCarrier	FlightNum
	<int>	<int>	<int>	<int>	<int>	<int>	<chr>	<int>
1	2011	1	1	6	1400	1500	AA	428
2	2011	1	1	6	728	840	AA	460
3	2011	1	1	6	1631	1736	AA	1121
4	2011	1	1	6	1756	2112	AA	1294
5	2011	1	1	6	1012	1347	AA	1700
6	2011	1	1	6	1211	1325	AA	1820
7	2011	1	1	6	557	906	AA	1994
8	2011	1	1	6	1824	2106	AS	731
9	2011	1	1	6	654	1124	B6	620
10	2011	1	1	6	1639	2110	B6	622

```
# i 542 more rows
```

```
# i 13 more variables: TailNum <chr>, ActualElapsedTime <int>, AirTime <int>,  
#   ArrDelay <int>, DepDelay <int>, Origin <chr>, Dest <chr>, Distance <int>,  
#   TaxiIn <int>, TaxiOut <int>, Cancelled <int>, CancellationCode <chr>,  
#   Diverted <int>
```

Para condições alternativas, podemos usar o operador | (ou) ou %in%:

```
flights %>% filter(UniqueCarrier %in% c("AA", "UA"))
```

```
# A tibble: 5,316 x 21
```

	Year	Month	DayofMonth	DayOfWeek	DepTime	ArrTime	UniqueCarrier	FlightNum
	<int>	<int>	<int>	<int>	<int>	<int>	<chr>	<int>
1	2011	1	1	6	1400	1500	AA	428
2	2011	1	2	7	1401	1501	AA	428
3	2011	1	3	1	1352	1502	AA	428
4	2011	1	4	2	1403	1513	AA	428
5	2011	1	5	3	1405	1507	AA	428
6	2011	1	6	4	1359	1503	AA	428
7	2011	1	7	5	1359	1509	AA	428
8	2011	1	8	6	1355	1454	AA	428
9	2011	1	9	7	1443	1554	AA	428
10	2011	1	10	1	1443	1553	AA	428

```
# i 5,306 more rows
```

```
# i 13 more variables: TailNum <chr>, ActualElapsedTime <int>, AirTime <int>,  
#   ArrDelay <int>, DepDelay <int>, Origin <chr>, Dest <chr>, Distance <int>,  
#   TaxiIn <int>, TaxiOut <int>, Cancelled <int>, CancellationCode <chr>,  
#   Diverted <int>
```

2.11.2 Select

A função `select` seleciona colunas específicas de um data frame. Para selecionar as colunas `DepTime`, `ArrTime`, e `FlightNum`:

```
flights %>% select(DepTime, ArrTime, FlightNum)
```

```
# A tibble: 227,496 x 3
```

	DepTime	ArrTime	FlightNum
	<int>	<int>	<int>
1	1400	1500	428
2	1401	1501	428
3	1352	1502	428
4	1403	1513	428
5	1405	1507	428
6	1359	1503	428
7	1359	1509	428
8	1355	1454	428

```

  9      1443      1554      428
10      1443      1553      428
# i 227,486 more rows

```

Para selecionar todas as colunas que contêm “Taxi” ou “Delay”:

```
flights %>% select(contains("Taxi"), contains("Delay"))
```

```

# A tibble: 227,496 x 4
  TaxiIn TaxiOut ArrDelay DepDelay
  <int>   <int>   <int>   <int>
1      7      13     -10        0
2      6       9      -9        1
3      5      17      -8       -8
4      9      22       3        3
5      9       9      -3        5
6      6      13      -7       -1
7     12      15      -1       -1
8      7      12     -16       -5
9      8      22      44       43
10     6      19      43       43
# i 227,486 more rows

```

2.11.3 Arrange

A função `arrange` ordena o data frame de acordo com algum critério. Para ordenar pelo atraso de partida (`DepDelay`):

```
flights %>% arrange(DepDelay)
```

```

# A tibble: 227,496 x 21
  Year Month DayOfMonth DayOfWeek DepTime ArrTime UniqueCarrier FlightNum
  <int> <int>    <int>    <int>   <int>   <int> <chr>          <int>
1  2011    12        24         6   1112    1314 00           5440
2  2011     2        14         1   1917    2027 MQ           3328
3  2011     4        10         7   2101    2206 XE           2669
4  2011     8         3         3   1741    1810 XE           2603
5  2011     1        18         2   1542    1936 CO           1688
6  2011    10         4         2   1438    1813 EV           5412
7  2011     1       26         3   2248    2343 XE           2450

```

```

 8 2011      3          8          2    953    1156 CO          1882
 9 2011      3         18          5   2103    2156 XE          2261
10 2011      4          3          7   1048    1307 MQ          3796
# i 227,486 more rows
# i 13 more variables: TailNum <chr>, ActualElapsedTime <int>, AirTime <int>,
#   ArrDelay <int>, DepDelay <int>, Origin <chr>, Dest <chr>, Distance <int>,
#   TaxiIn <int>, TaxiOut <int>, Cancelled <int>, CancellationCode <chr>,
#   Diverted <int>

```

Para ordem decrescente:

```
flights %>% arrange(desc(DepDelay))
```

```

# A tibble: 227,496 x 21
   Year Month DayOfMonth DayOfWeek DepTime ArrTime UniqueCarrier FlightNum
   <int> <int>      <int>      <int>   <int>   <int>   <chr>          <int>
1  2011      8          1          1    156     452 CO              1
2  2011     12         12          1    650     808 AA             1740
3  2011     11          8          2    721     948 MQ             3786
4  2011      6         21          2   2334     124 UA             855
5  2011      6          9          4   2029    2243 MQ            3859
6  2011      5         20          5    858    1027 MQ            3328
7  2011      1         20          4    635     807 CO              59
8  2011      6         22          3    908    1040 CO             595
9  2011     10         25          2   2310     149 DL            1215
10 2011     12         13          2    706     824 MQ            3328
# i 227,486 more rows
# i 13 more variables: TailNum <chr>, ActualElapsedTime <int>, AirTime <int>,
#   ArrDelay <int>, DepDelay <int>, Origin <chr>, Dest <chr>, Distance <int>,
#   TaxiIn <int>, TaxiOut <int>, Cancelled <int>, CancellationCode <chr>,
#   Diverted <int>

```

2.11.4 Mutate

A função `mutate` cria novas variáveis. Para criar a variável `Speed` (velocidade) e adicioná-la ao banco:

```
flights <- flights %>% mutate(Speed = Distance / AirTime * 60)
```


2.11.5 Summarise e Group_by

A função `summarise` calcula estatísticas resumo. Para calcular o atraso médio de chegada por destino:

```
flights %>% group_by(Dest) %>% summarise(avg_delay = mean(ArrDelay, na.rm = TRUE))
```

```
# A tibble: 116 x 2
  Dest   avg_delay
  <chr>   <dbl>
1 ABQ     7.23
2 AEX     5.84
3 AGS      4
4 AMA     6.84
5 ANC    26.1
6 ASE     6.79
7 ATL     8.23
8 AUS     7.45
9 AVL     9.97
10 BFL    -13.2
# i 106 more rows
```

A função `summarise_each` aplica uma função a várias colunas ao mesmo tempo. Para calcular a média de `Cancelled` e `Diverted` por companhia aérea:

```
flights %>% group_by(UniqueCarrier) %>% summarise_each(funs(mean), Cancelled, Diverted)
```

```
# A tibble: 15 x 3
  UniqueCarrier Cancelled Diverted
  <chr>          <dbl>   <dbl>
1 AA            0.0185  0.00185
2 AS            0      0.00274
3 B6            0.0259  0.00576
4 CO            0.00678 0.00263
5 DL            0.0159  0.00303
6 EV            0.0345  0.00318
7 F9            0.00716  0
8 FL            0.00982 0.00327
9 MQ            0.0290  0.00194
10 OO           0.0139  0.00349
11 UA           0.0164  0.00241
```

12	US	0.0113	0.00147
13	WN	0.0155	0.00229
14	XE	0.0155	0.00345
15	YV	0.0127	0

Também podemos aplicar várias funções a uma mesma coluna:

```
flights %>% group_by(UniqueCarrier) %>% summarise_each(funs(mean, var), Cancelled, Diverted)
```

A tibble: 15 x 5

	UniqueCarrier	Cancelled_mean	Diverted_mean	Cancelled_var	Diverted_var
	<chr>	<dbl>	<dbl>	<dbl>	<dbl>
1	AA	0.0185	0.00185	0.0182	0.00185
2	AS	0	0.00274	0	0.00274
3	B6	0.0259	0.00576	0.0253	0.00573
4	CO	0.00678	0.00263	0.00674	0.00262
5	DL	0.0159	0.00303	0.0157	0.00302
6	EV	0.0345	0.00318	0.0333	0.00317
7	F9	0.00716	0	0.00712	0
8	FL	0.00982	0.00327	0.00973	0.00326
9	MQ	0.0290	0.00194	0.0282	0.00193
10	OO	0.0139	0.00349	0.0138	0.00347
11	UA	0.0164	0.00241	0.0161	0.00241
12	US	0.0113	0.00147	0.0111	0.00147
13	WN	0.0155	0.00229	0.0153	0.00229
14	XE	0.0155	0.00345	0.0153	0.00344
15	YV	0.0127	0	0.0127	0

Além disso, podemos usar o operador "." para aplicar funções com múltiplos argumentos:

```
flights %>% group_by(UniqueCarrier) %>% summarise_each(funs(min(., na.rm = TRUE), max(., na.rm = TRUE)), Cancelled, Diverted)
```

A tibble: 15 x 5

	UniqueCarrier	ArrDelay_min	DepDelay_min	ArrDelay_max	DepDelay_max
	<chr>	<int>	<int>	<int>	<int>
1	AA	-39	-15	978	970
2	AS	-43	-15	183	172
3	B6	-44	-14	335	310
4	CO	-55	-18	957	981
5	DL	-32	-17	701	730
6	EV	-40	-18	469	479

7	F9	-24	-15	277	275
8	FL	-30	-14	500	507
9	MQ	-38	-23	918	931
10	OO	-57	-33	380	360
11	UA	-47	-11	861	869
12	US	-42	-17	433	425
13	WN	-44	-10	499	548
14	XE	-70	-19	634	628
15	YV	-32	-11	72	54

Por fim, podemos realizar agrupamentos múltiplos:

```
flights %>% group_by(UniqueCarrier, DayOfWeek) %>% summarise_each(funs(min(., na.rm = TRUE)),
```

```
# A tibble: 105 x 6
# Groups:   UniqueCarrier [15]
  UniqueCarrier DayOfWeek ArrDelay_min DepDelay_min ArrDelay_max DepDelay_max
  <chr>          <int>      <int>      <int>      <int>      <int>
1 AA            1        -30        -14        978        970
2 AA            2        -30        -12        265        286
3 AA            3        -33        -15        179        168
4 AA            4        -38        -15        663        653
5 AA            5        -28        -13        255        234
6 AA            6        -39        -13        685        677
7 AA            7        -34        -15        507        525
8 AS            1        -30        -12        183        172
9 AS            2        -34        -12         92         68
10 AS           3        -29        -12        123        138
# i 95 more rows
```

Aqui está a seção sobre o **pacote ggplot2** e **tidyr** atualizada, mantendo a clareza e estrutura para fins didáticos:

2.12 O pacote ggplot2

Nota: Esta seção foi adaptada de curso-r.com.

O **ggplot2** é um pacote do R voltado para a criação de gráficos estatísticos. Ele é baseado na Gramática dos Gráficos (Grammar of Graphics) de Leland Wilkinson. Segundo essa gramática, um gráfico estatístico é um mapeamento dos dados por meio de atributos estéticos (cores, formas, tamanho) de formas geométricas (pontos, linhas, barras).

```
library(ggplot2)
```

2.12.1 Construindo gráficos

Vamos discutir os aspectos básicos para a construção de gráficos com o `ggplot2`, utilizando o conjunto de dados `mtcars`. Para visualizar as primeiras linhas:

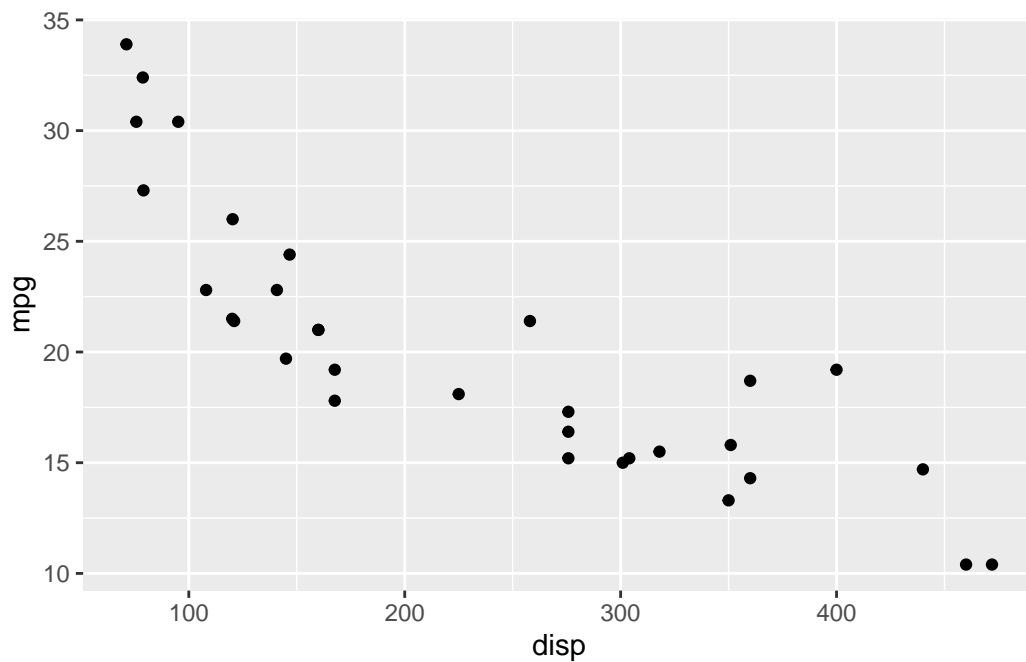
```
head(mtcars)
```

	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

2.12.2 As camadas de um gráfico

Os gráficos no `ggplot2` são construídos camada por camada. A primeira camada é criada com a função `ggplot`. Um exemplo de gráfico simples:

```
ggplot(data = mtcars) +  
  geom_point(aes(x = disp, y = mpg))
```

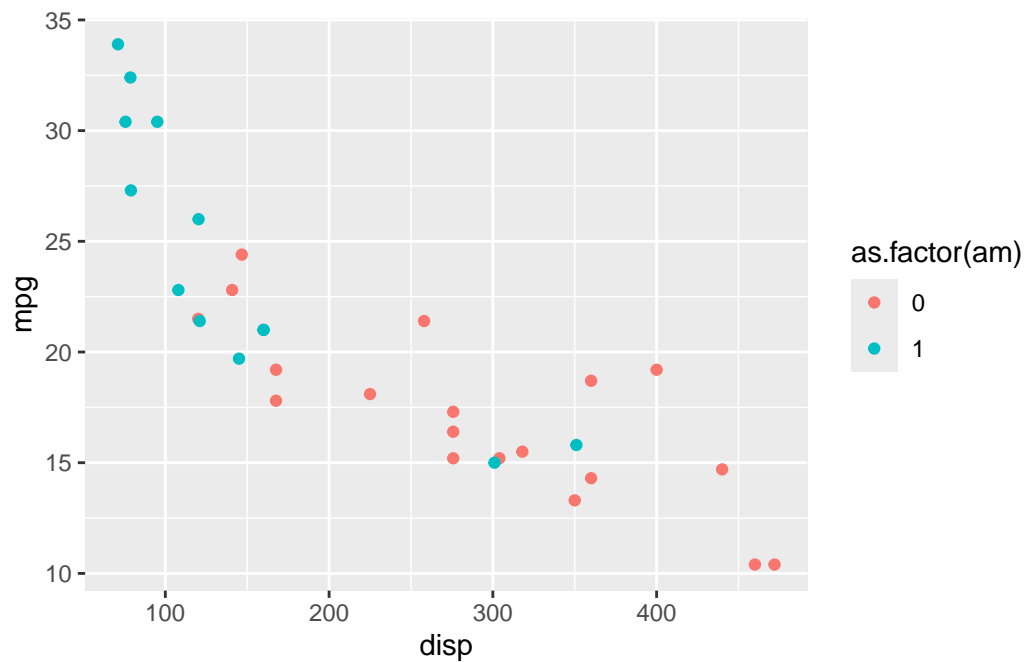


A função `aes` define o mapeamento entre as variáveis e os aspectos visuais. Neste caso, estamos criando um gráfico de dispersão (com `geom_point`) entre `disp` (cilindradas) e `mpg` (milhas por galão).

2.12.3 Aesthetics

Podemos mapear variáveis a diferentes aspectos estéticos, como cor e tamanho. Por exemplo, para mapear a variável `am` (transmissão) para a cor dos pontos:

```
ggplot(data = mtcars) +  
  geom_point(aes(x = disp, y = mpg, colour = as.factor(am)))
```



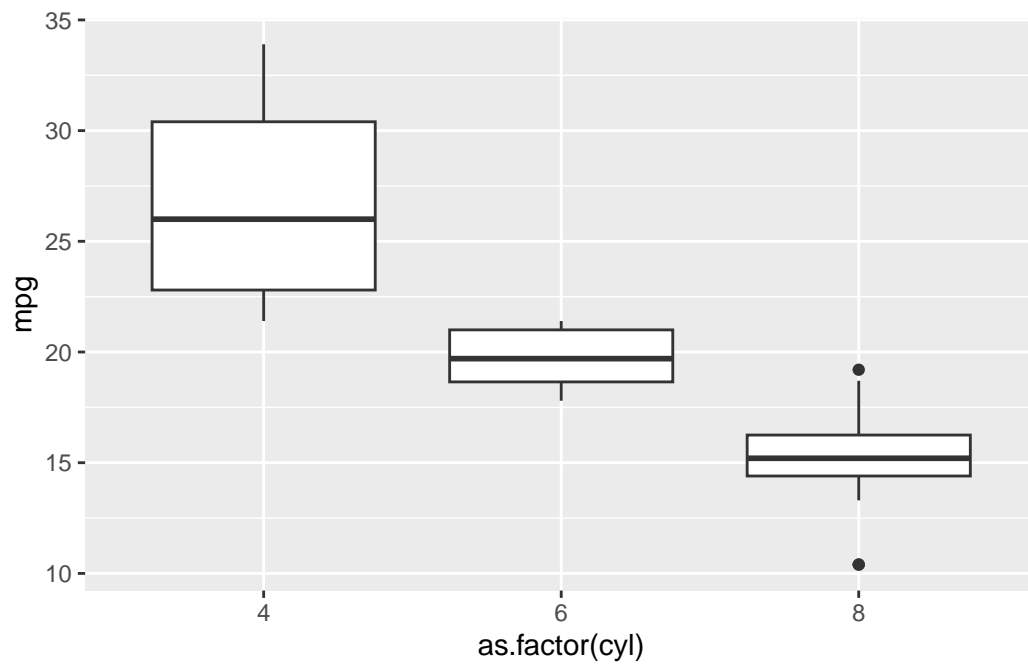
2.12.4 Geoms

Os `geoms` definem as formas geométricas usadas para a visualização dos dados. Além de `geom_point`, podemos usar:

- `geom_line` para linhas
- `geom_boxplot` para boxplots
- `geom_histogram` para histogramas

Exemplo de um boxplot:

```
ggplot(mtcars) +  
  geom_boxplot(aes(x = as.factor(cyl), y = mpg))
```

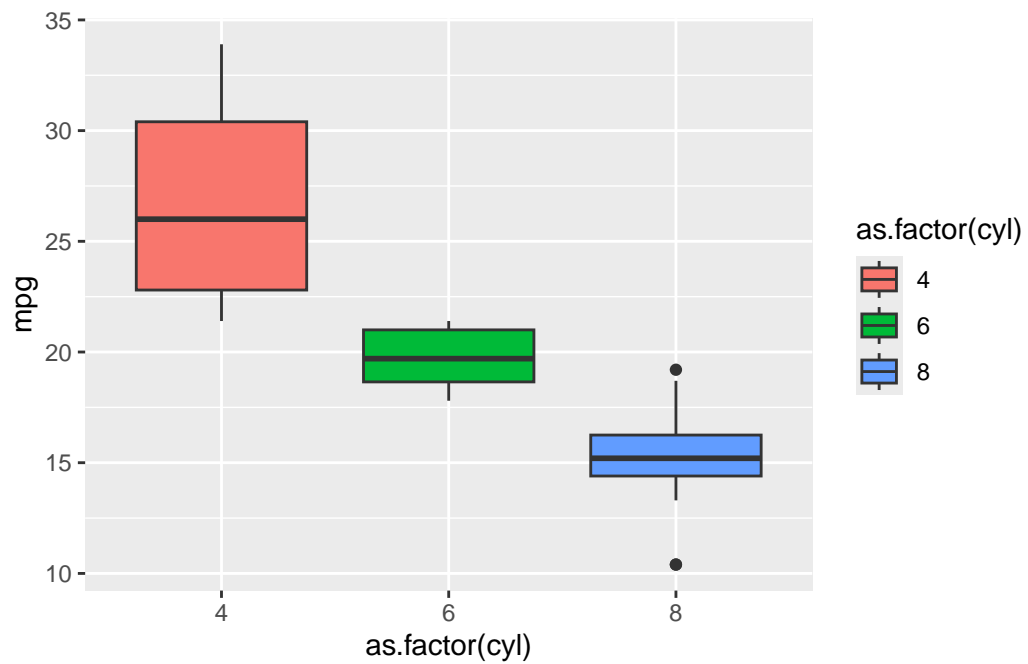


2.12.5 Personalizando os gráficos

2.12.5.1 Cores

Para mudar as cores do gráfico, podemos usar o argumento `colour` ou `fill`:

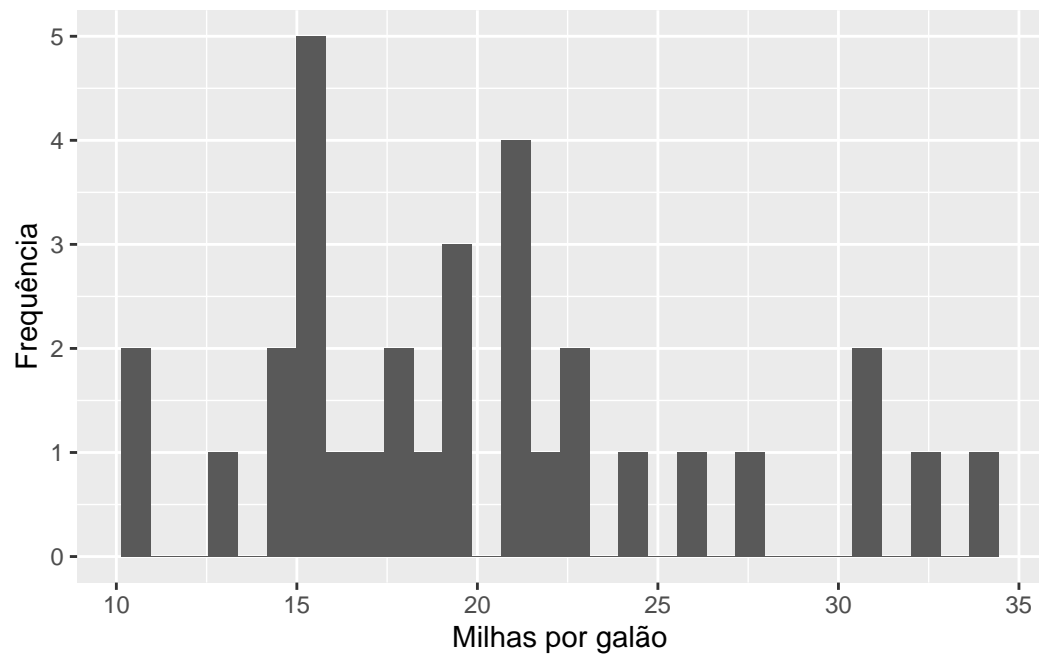
```
ggplot(mtcars) +  
  geom_boxplot(aes(x = as.factor(cyl), y = mpg, fill = as.factor(cyl)))
```



2.12.5.2 Eixos

Para alterar os rótulos dos eixos:

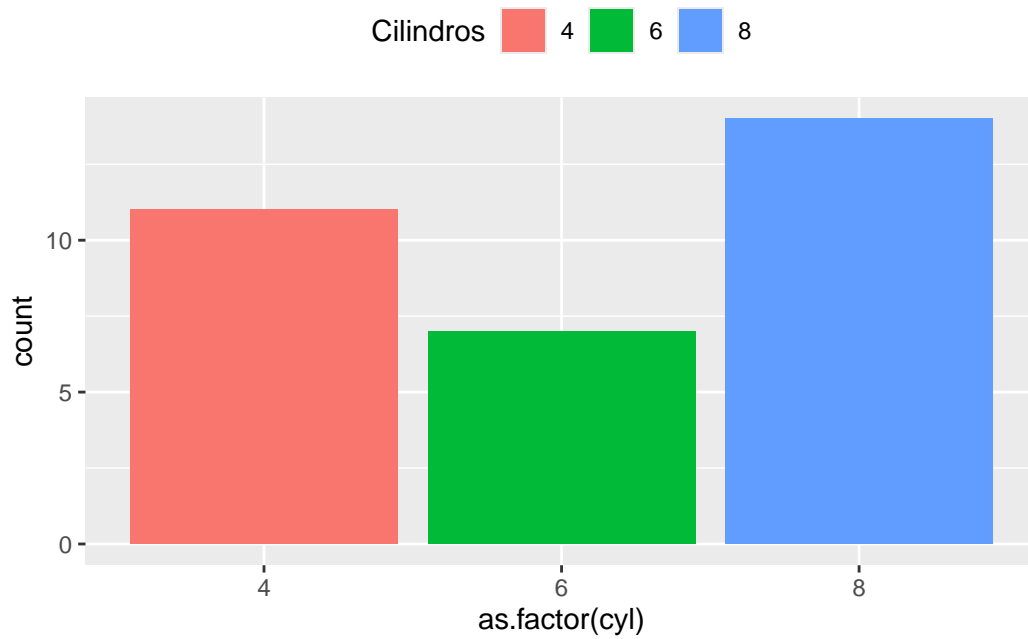
```
ggplot(mtcars) +  
  geom_histogram(aes(x = mpg)) +  
  xlab("Milhas por galão") +  
  ylab("Frequência")
```

2.12.5.3 Legendas

Podemos personalizar ou remover legendas facilmente:

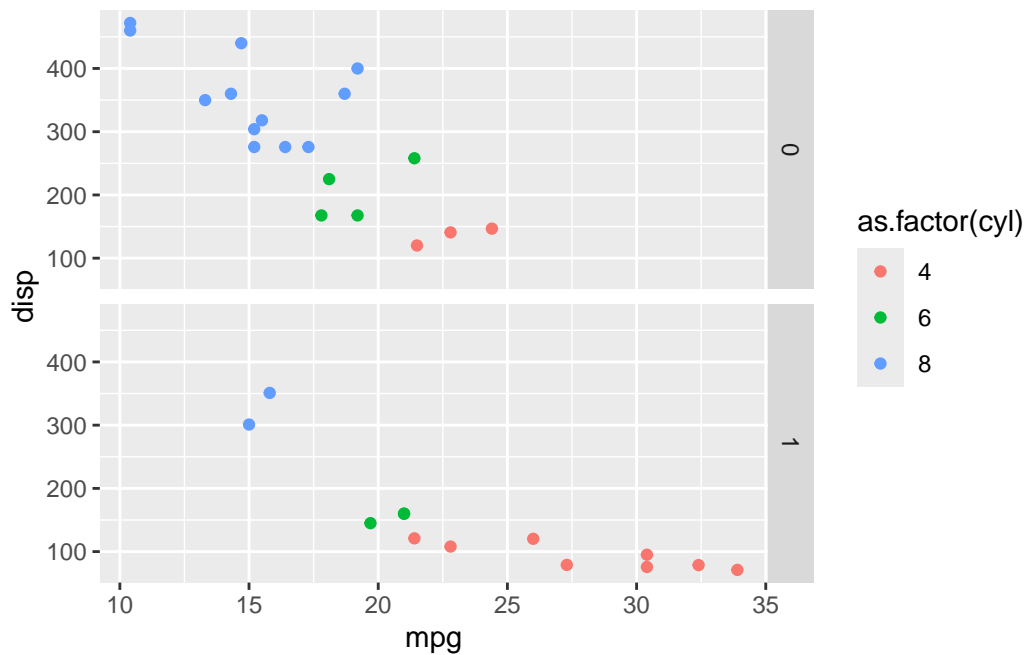
```
ggplot(mtcars) +  
  geom_bar(aes(x = as.factor(cyl), fill = as.factor(cyl))) +  
  labs(fill = "Cilindros") +  
  theme(legend.position = "top")
```



2.12.6 Facets

O `facet_grid` permite dividir o gráfico em subgráficos com base em uma variável:

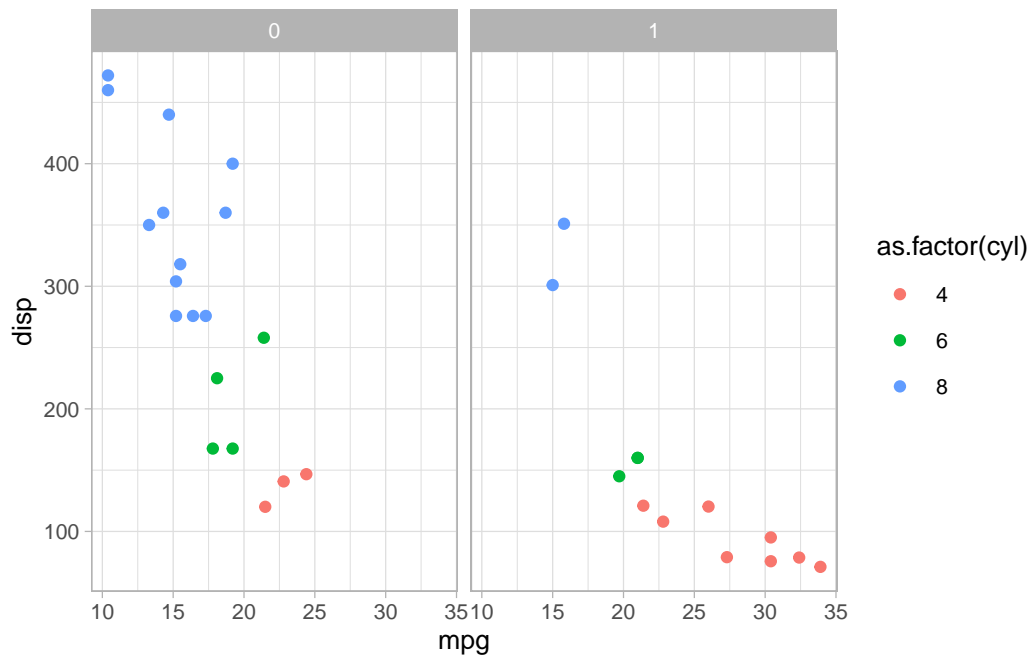
```
ggplot(mtcars) +  
  geom_point(aes(x = mpg, y = disp, colour = as.factor(cyl))) +  
  facet_grid(am ~ .)
```



2.12.7 Temas

Podemos mudar o tema de um gráfico com a função `theme_set`:

```
theme_set(theme_light(base_size = 10))
ggplot(mtcars) +
  geom_point(aes(x = mpg, y = disp, colour = as.factor(cyl))) +
  facet_grid(. ~ am)
```



2.13 O pacote tidyr

O pacote `tidyr` facilita a transformação e organização de dados no R. Para converter os dados de formato “wide” (largo) para “long” (longo), podemos utilizar a função `pivot_longer`, que substitui a antiga função `gather`. Considere os dados a seguir:

```
dados.originais <- data.frame(
  paciente = c("João", "Marcos", "Antonio"),
  pressao.antes = c(67, 80, 64),
  pressao.durante = c(54, 70, 60),
  pressao.depois = c(56, 90, 50)
)
dados.originais
```

	paciente	pressao.antes	pressao.durante	pressao.depois
1	João	67	54	56
2	Marcos	80	70	90
3	Antonio	64	60	50

Para reorganizar esses dados em um formato “long”, usando `pivot_longer`:

```
library(tidyr)
dados.novo.formato <- dados.originais %>%
  pivot_longer(cols = pressao.antes:pressao.depois,
               names_to = "instante",
               values_to = "pressao.arterial")
dados.novo.formato
```

```
# A tibble: 9 x 3
  paciente instante      pressao.arterial
  <chr>      <chr>          <dbl>
1 João     pressao.antes          67
2 João     pressao.durante        54
3 João     pressao.depois         56
4 Marcos   pressao.antes          80
5 Marcos   pressao.durante        70
6 Marcos   pressao.depois         90
7 Antonio  pressao.antes          64
8 Antonio  pressao.durante        60
9 Antonio  pressao.depois         50
```

A função `pivot_wider` é usada para converter dados de formato “long” (longo) para “wide” (largo), espalhando uma variável por várias colunas. Vamos começar com um conjunto de dados em formato “long” e então reorganizá-los para o formato “wide”.

Considere o conjunto de dados `dados.novo.formato` que organizamos anteriormente:

```
dados.novo.formato <- data.frame(
  paciente = c("João", "Marcos", "Antonio", "João", "Marcos", "Antonio", "João", "Marcos", "Antonio"),
  instante = c("pressao.antes", "pressao.antes", "pressao.antes", "pressao.durante", "pressao.durante", "pressao.durante", "pressao.depois", "pressao.depois", "pressao.depois"),
  pressao.arterial = c(67, 80, 64, 54, 70, 60, 56, 90, 50)
)
dados.novo.formato
```

```
  paciente      instante pressao.arterial
1   João  pressao.antes          67
2  Marcos  pressao.antes          80
3 Antonio  pressao.antes          64
4   João  pressao.durante          54
5  Marcos  pressao.durante          70
6 Antonio  pressao.durante          60
7   João  pressao.depois          56
```

8	Marcos	pressao.depois	90
9	Antonio	pressao.depois	50

Agora, vamos usar `pivot_wider` para transformar esses dados de volta ao formato “wide”:

```
dados.wide <- dados.novo.formato %>%
  pivot_wider(names_from = instante, values_from = pressao.arterial)
dados.wide
```

```
# A tibble: 3 x 4
  paciente pressao.antes pressao.durante pressao.depois
  <chr>          <dbl>          <dbl>          <dbl>
1 João           67            54            56
2 Marcos          80            70            90
3 Antonio         64            60            50
```

Temos os seguintes argumentos: - `names_from` especifica a coluna cujos valores serão usados para criar novos nomes de colunas (neste caso, a variável `instante`). - `values_from` especifica a coluna cujos valores serão preenchidos nas novas colunas criadas (neste caso, a variável `pressao.arterial`).

Aqui está um exemplo de como o pacote `tidyr` pode ser usado para manipular os dados antes de visualizá-los com gráficos no `ggplot2`.

2.13.1 Exemplo: Manipulação e Visualização com `pivot_longer` e `ggplot2`

O pacote `tidyr` é particularmente útil dentro do `ggplot2`. Vamos continuar com o exemplo dos dados de pressão arterial. Suponha que você queira visualizar as variações da pressão arterial dos pacientes em diferentes instantes.

Primeiro, vamos reorganizar os dados usando `pivot_longer` para colocar os dados no formato “long”, que é mais adequado para gráficos com o `ggplot2`:

```
library(tidyr)
library(ggplot2)

# Dados originais no formato "wide"
dados.originais <- data.frame(
  paciente = c("João", "Marcos", "Antonio"),
  pressao.antes = c(67, 80, 64),
  pressao.durante = c(54, 70, 60),
  pressao.depois = c(56, 90, 50)
```

```
)
dados.originais
```

	paciente	pressao.antes	pressao.durante	pressao.depois
1	João	67	54	56
2	Marcos	80	70	90
3	Antonio	64	60	50

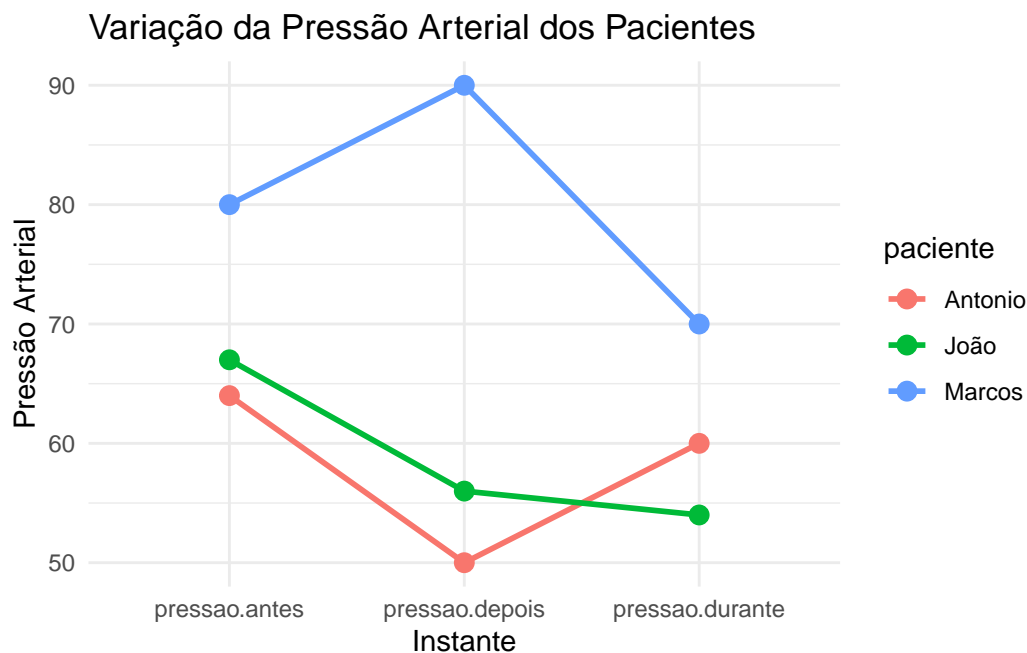
Para usá-los no ggplot, vamos os reorganizar no formato “long”:

```
dados.long <- dados.originais %>%
  pivot_longer(cols = pressao.antes:pressao.depois,
               names_to = "instante",
               values_to = "pressao.arterial")
dados.long
```

```
# A tibble: 9 x 3
  paciente instante      pressao.arterial
  <chr>      <chr>          <dbl>
1 João     pressao.antes         67
2 João     pressao.durante        54
3 João     pressao.depois         56
4 Marcos   pressao.antes         80
5 Marcos   pressao.durante        70
6 Marcos   pressao.depois         90
7 Antonio  pressao.antes         64
8 Antonio  pressao.durante        60
9 Antonio  pressao.depois         50
```

Agora, vamos criar um gráfico de linhas que mostra as mudanças na pressão arterial ao longo do tempo para cada paciente:

```
ggplot(dados.long, aes(x = instante, y = pressao.arterial, group = paciente, color = paciente)) +
  geom_line(size = 1) +
  geom_point(size = 3) +
  labs(title = "Variação da Pressão Arterial dos Pacientes",
       x = "Instante",
       y = "Pressão Arterial") +
  theme_minimal()
```



2.14 Leitura de Arquivos com readr

O `readr` é um pacote eficiente para leitura de arquivos de dados no R. Vamos explorar como utilizar o `readr` para ler e manipular arquivos CSV. Um arquivo CSV (Comma-Separated Values) é um formato comum para armazenar dados tabulares.

Para começar, carregue o pacote `readr`:

```
library(readr)
```

2.14.1 Exemplo: Leitura de um CSV com Dados Públicos

Vamos usar um banco de dados público disponível na internet. Um exemplo muito comum é o conjunto de dados de passageiros do Titanic, disponível em formato CSV. Para carregar diretamente da internet:

```
url <- "https://raw.githubusercontent.com/datasciencedojo/datasets/master/titanic.csv"
dados_titanic <- read_csv(url)
```

Aqui, estamos carregando os dados diretamente de um link. Após a leitura, podemos visualizar os primeiros registros:


```
head(dados_titanic)
```

```
# A tibble: 6 x 12
  PassengerId Survived Pclass Name      Sex      Age SibSp Parch Ticket  Fare Cabin
      <dbl>     <dbl>  <dbl> <chr>    <chr>  <dbl>  <dbl>  <dbl> <chr>   <dbl> <chr>
1           1         0      3 Braund~ male    22      1      0 A/5 2~  7.25 <NA>
2           2         1      1 Cuming~ fema~   38      1      0 PC 17~ 71.3  C85
3           3         1      3 Heikki~ fema~   26      0      0 STON/~  7.92 <NA>
4           4         1      1 Futrel~ fema~   35      1      0 113803 53.1  C123
5           5         0      3 Allen,~ male    35      0      0 373450  8.05 <NA>
6           6         0      3 Moran,~ male    NA      0      0 330877  8.46 <NA>
# i 1 more variable: Embarked <chr>
```

2.14.2 Manipulando os Dados Lidos

O `read_csv` detecta automaticamente os tipos de dados de cada coluna. No exemplo dos dados do Titanic, podemos realizar algumas análises rápidas, como visualizar a distribuição de sobreviventes:

```
table(dados_titanic$Survived)
```

```
0    1
549 342
```

2.14.3 Salvando o Conjunto de Dados

Se você quiser salvar o conjunto de dados em seu computador após manipulações, basta usar o `write_csv`:

```
write_csv(dados_titanic, "titanic_local.csv")
```

Aqui está o conjunto de exercícios no formato solicitado:

2.15 Exercícios

Exercício 1.

Execute as seguintes operações no R e interprete os resultados: $-5 + 9 - \frac{0}{0} - 10^5$

Utilize a função `log` para calcular o logaritmo natural de 10 e depois o logaritmo na base 10 de 1000.

Descubra a função que calcula a raiz quadrada de um número utilizando o símbolo de interrogação (?), e aplique-a ao número 144.

Exercício 2.

Armazene o resultado de 15×7 em uma variável chamada `resultado`. Use `resultado` para calcular o valor de $2 \times resultado$.

Liste todas as variáveis no ambiente do R usando a função `ls()`. Exclua a variável `resultado` e verifique novamente as variáveis presentes.

Exercício 3.

Crie um vetor `v` contendo os números 4, 8, 15, 16, 23, 42. Use subsetting para selecionar: - O segundo e o quarto elemento. - Todos os elementos exceto o terceiro. - Os elementos que são maiores que 10.

Multiplique cada elemento do vetor `v` por 2 e armazene o resultado em um novo vetor `v2`.

Exercício 4.

Verifique se 8 é maior que 5 e se 8 é igual a 10.

Crie dois vetores lógicos `a = c(TRUE, FALSE, TRUE)` e `b = c(FALSE, TRUE, FALSE)`. Aplique os operadores `&`, `|` e `xor()` nesses vetores.

Exercício 5.

Escreva uma função `soma_quadrados` que receba dois números como argumentos e retorne a soma de seus quadrados. Teste sua função com os números 3 e 4.

Modifique a função `soma_quadrados` para que o segundo argumento tenha um valor padrão de 2. Teste a função chamando-a com apenas um argumento.

Exercício 6.

Escreva um laço `for` que calcule a soma dos números de 1 a 100.

Crie um laço `while` que multiplique os números de 1 a 6 e retorne o resultado.

Exercício 7.

Crie duas variáveis com o seu primeiro e último nome. Use a função `paste()` para juntar as duas variáveis em uma frase que diga “Meu nome completo é [nome completo]”.

Altere o separador na função `paste()` para um traço - e junte novamente as variáveis.

Exercício 8.

Carregue o pacote `dplyr` e use o dataset `mtcars`. Selecione apenas as colunas `mpg`, `cyl`, e `hp`.

Filtre as observações onde `mpg` é maior que 20 e `hp` é menor que 150.

Ordene o dataset filtrado pela coluna `mpg` em ordem decrescente.

Exercício 9.

Usando o dataset `mtcars`, crie um gráfico de dispersão (`geom_point`) que mostre a relação entre `hp` (horsepower) e `mpg` (milhas por galão).

No gráfico anterior, mapeie a variável `cyl` para a cor dos pontos.

Exercício 10. Utilize a base pública de dados de voos da [nycflights13](#) para responder às perguntas abaixo.

1. Carregue o pacote `nycflights13` e explore o dataset `flights`. Visualize as primeiras 6 linhas da base.
2. Filtre todos os voos que decolaram no mês de junho e aterrissaram com atraso maior que 1 hora.
3. Agrupe os dados por companhia aérea (`carrier`) e calcule o atraso médio de chegada (`arr_delay`) para cada companhia.
4. Crie um gráfico de barras que mostre o atraso médio de chegada por companhia aérea.

3 Geração de Números Aleatórios e Aplicação em Estatística

3.1 Objetivo da Aula

Nesta aula, vamos introduzir o conceito de números pseudoaleatórios e como eles podem ser usados para resolver problemas estatísticos por meio de simulação. Vamos abordar a importância da aleatoriedade em estatísticas e em algoritmos de Monte Carlo.

3.2 Conteúdo Teórico

A geração de números aleatórios é essencial em várias áreas da estatística e da ciência de dados. Esses números são utilizados em simulações estocásticas, amostragem e para resolver problemas que envolvem incerteza. Contudo, em computadores, os números “aleatórios” gerados são na verdade pseudoaleatórios, pois seguem uma sequência previsível, gerada por um algoritmo determinístico.

Os números pseudoaleatórios são amplamente usados em algoritmos de Monte Carlo, que dependem da simulação repetida de processos aleatórios para estimar soluções para problemas matemáticos e estatísticos.

3.3 Exemplo de Problema

Vamos resolver o problema de estimar o valor de π (Pi) usando um método de Monte Carlo. A ideia é simular a área de um quarto de círculo inscrito em um quadrado. Gerando pontos aleatórios dentro do quadrado, podemos calcular a proporção desses pontos que também caem dentro do círculo e usar essa proporção para estimar o valor de π .

Como fazer isso?

4 R

```
# Definindo o número de pontos a serem gerados
n_pontos <- 1000

# Gerando pontos aleatórios (x, y) no intervalo [0, 1]
x <- runif(n_pontos, 0, 1)
y <- runif(n_pontos, 0, 1)

# Calculando a distância de cada ponto à origem
distancia <- sqrt(x^2 + y^2)

# Contando quantos pontos estão dentro do quarto de círculo (distância <= 1)
dentro_circulo <- distancia <= 1
pi_estimado <- 4 * sum(dentro_circulo) / n_pontos

# Exibindo o valor estimado de Pi
cat("Valor estimado de :", pi_estimado, "\n")
```

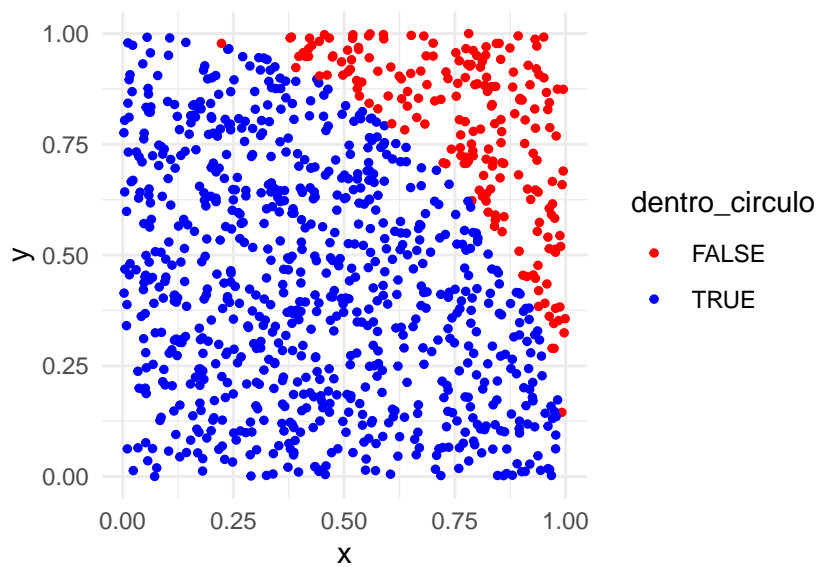
Valor estimado de : 3.184

```
# Visualizando a distribuição dos pontos
library(ggplot2)

dados <- data.frame(x = x, y = y, dentro_circulo = dentro_circulo)

ggplot(dados, aes(x = x, y = y, color = dentro_circulo)) +
  geom_point(size = 1) +
  scale_color_manual(values = c("red", "blue")) +
  ggtitle(paste0("Estimativa de usando Monte Carlo\nValor estimado: ", round(pi_estimado, 5))) +
  theme_minimal() +
  coord_equal() +
  labs(x = "x", y = "y")
```

Estimativa de π usando Monte Carlo
Valor estimado: 3.184



5 Python

```
import numpy as np
import matplotlib.pyplot as plt

# Definindo o número de pontos a serem gerados
n_pontos = 1000

# Gerando pontos aleatórios (x, y) no intervalo [0, 1]
x = np.random.uniform(0, 1, n_pontos)
y = np.random.uniform(0, 1, n_pontos)

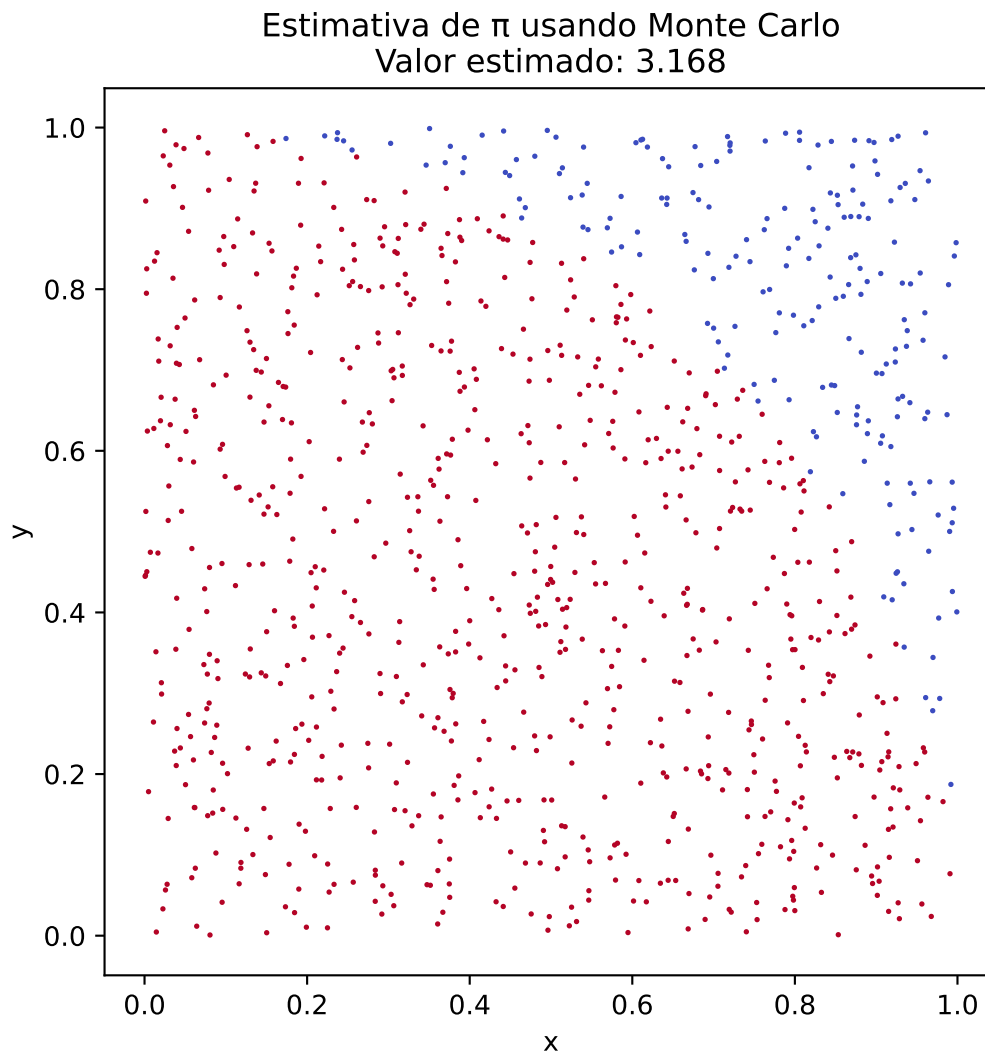
# Calculando a distância de cada ponto à origem
distancia = np.sqrt(x**2 + y**2)

# Contando quantos pontos estão dentro do quarto de círculo (distância <= 1)
dentro_circulo = distancia <= 1
pi_estimado = 4 * np.sum(dentro_circulo) / n_pontos

# Exibindo o valor estimado de Pi
print(f"Valor estimado de : {pi_estimado}")
```

Valor estimado de : 3.168

```
# Visualizando a distribuição dos pontos
plt.figure(figsize=(6,6))
plt.scatter(x, y, c=dentro_circulo, cmap='coolwarm', s=1)
plt.title(f'Estimativa de usando Monte Carlo\nValor estimado: {pi_estimado}')
plt.xlabel('x')
plt.ylabel('y')
plt.show()
```



5.1 Exemplo: Simulação de um Jogo de Dados com Dado “Viciado”

Imagine que estamos jogando um jogo em que o dado é “viciado” e não segue uma distribuição uniforme, ou seja, alguns números têm uma chance maior de serem sorteados. Por exemplo, o número 6 pode ter uma probabilidade maior, e os outros números, menores.

Isso nos permite mostrar como alterar a probabilidade de ocorrência de eventos em uma

distribuição discreta.

6 R

```
# Definindo as faces do dado e as probabilidades
faces <- 1:6
probabilidades <- c(0.05, 0.1, 0.15, 0.2, 0.25, 0.25) # Probabilidades associadas às faces

# Verificando que a soma das probabilidades é 1
cat("Soma das probabilidades:", sum(probabilidades), "\n")
```

Soma das probabilidades: 1

```
# Simulando 10000 lançamentos de um dado viciado
n_lancamentos <- 10000
set.seed(123) # Definindo seed para reprodutibilidade
resultados <- sample(faces, size = n_lancamentos, replace = TRUE, prob = probabilidades)

# Contando as frequências de cada face
frequencias <- table(resultados) / n_lancamentos

# Exibindo os resultados da simulação
cat("Frequências de cada face após", n_lancamentos, "lançamentos:\n")
```

Frequências de cada face após 10000 lançamentos:

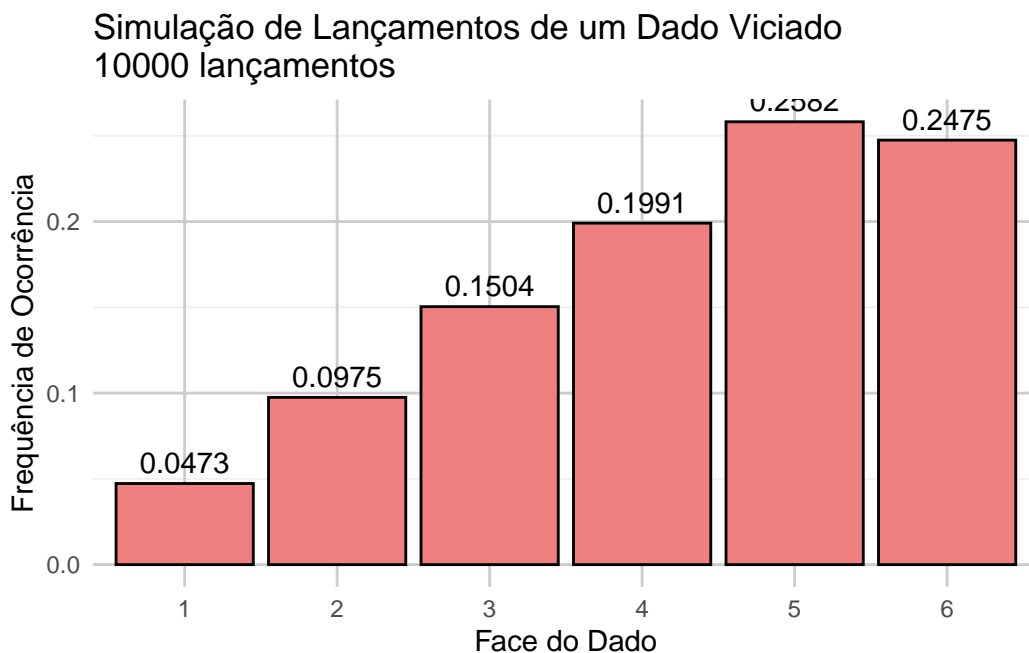
```
for (face in faces) {
  cat("Face", face, ":", frequencias[as.character(face)], "vezes\n")
}
```

```
Face 1 : 0.0473 vezes
Face 2 : 0.0975 vezes
Face 3 : 0.1504 vezes
Face 4 : 0.1991 vezes
Face 5 : 0.2582 vezes
Face 6 : 0.2475 vezes
```

```
# Visualizando os resultados em um gráfico de barras
library(ggplot2)

dados <- data.frame(faces = as.factor(faces), frequencias = as.numeric(frequencias))

ggplot(dados, aes(x = faces, y = frequencias)) +
  geom_bar(stat = "identity", fill = "lightcoral", color = "black") +
  ggtitle(paste0("Simulação de Lançamentos de um Dado Viciado\n", n_lancamentos, " lançamentos")) +
  xlab("Face do Dado") +
  ylab("Frequência de Ocorrência") +
  theme_minimal() +
  geom_text(aes(label = round(frequencias, 4)), vjust = -0.5) +
  theme(panel.grid.major = element_line(color = "grey80"))
```



7 Python

```
import numpy as np
import matplotlib.pyplot as plt

# Definindo as faces do dado e as probabilidades
faces = [1, 2, 3, 4, 5, 6]
probabilidades = [0.05, 0.1, 0.15, 0.2, 0.25, 0.25] # Probabilidades associadas às faces do

# Verificando que a soma das probabilidades é 1
print(f"Soma das probabilidades: {sum(probabilidades)}")
```

Soma das probabilidades: 1.0

```
# Simulando 10000 lançamentos de um dado viciado
n_lancamentos = 10000
resultados = np.random.choice(faces, size=n_lancamentos, p=probabilidades)

# Contando as frequências de cada face
frequencias = [np.sum(resultados == face) / n_lancamentos for face in faces]

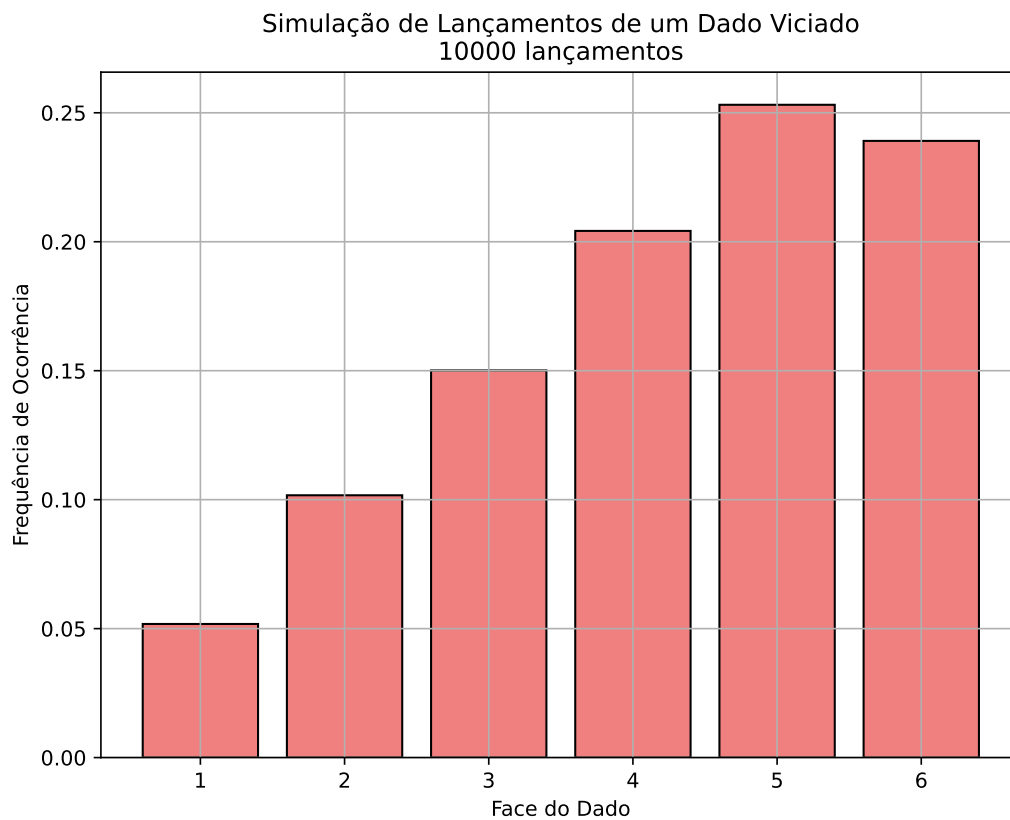
# Exibindo os resultados da simulação
print(f"Frequências de cada face após {n_lancamentos} lançamentos:")
```

Frequências de cada face após 10000 lançamentos:

```
for face, freq in zip(faces, frequencias):
    print(f"Face {face}: {freq} vezes")
```

Face 1: 0.0518 vezes
Face 2: 0.1017 vezes
Face 3: 0.1501 vezes
Face 4: 0.2042 vezes
Face 5: 0.2531 vezes
Face 6: 0.2391 vezes

```
# Visualizando os resultados em um gráfico de barras
plt.figure(figsize=(8,6))
plt.bar(faces, frequencias, color='lightcoral', edgecolor='black')
plt.title(f'Simulação de Lançamentos de um Dado Viciado\n{n_lancamentos} lançamentos')
plt.xlabel('Face do Dado')
plt.ylabel('Frequência de Ocorrência')
plt.grid(True)
plt.show()
```



7.1 Pergunta: precisamos da função `np.random.choice/set.seed`?

8 R

```
# Definindo as faces do dado e as probabilidades associadas (não uniformes)
faces <- 1:6
probabilidades <- c(0.05, 0.1, 0.15, 0.2, 0.25, 0.25) # Probabilidades associadas às faces do dado

# Função para gerar uma amostra baseada em intervalos de probabilidades
gerar_amostra_por_intervalos <- function(probabilidades, faces) {
  u <- runif(1) # Gerando um número aleatório uniforme
  limite_inferior <- 0 # Limite inferior do intervalo

  # Percorrendo as probabilidades e verificando em qual intervalo o número cai
  for (i in seq_along(probabilidades)) {
    limite_superior <- limite_inferior + probabilidades[i] # Definindo o limite superior do intervalo
    if (limite_inferior <= u && u < limite_superior) {
      return(faces[i]) # Retorna a face correspondente ao intervalo
    }
    limite_inferior <- limite_superior # Atualiza o limite inferior para o próximo intervalo
  }
}

# Simulando lançamentos do dado viciado utilizando a verificação dos intervalos
n_lancamentos <- 10000
set.seed(123) # Definindo seed para reprodutibilidade
resultados <- replicate(n_lancamentos, gerar_amostra_por_intervalos(probabilidades, faces))

# Contando as frequências de cada face
frequencias <- sapply(faces, function(face) sum(resultados == face) / n_lancamentos)

# Exibindo os resultados da simulação
cat("Frequências de cada face após", n_lancamentos, "lançamentos:\n")
```

Frequências de cada face após 10000 lançamentos:

```
for (i in seq_along(faces)) {
  cat("Face", faces[i], ":", frequencias[i], "vezes\n")
}
```

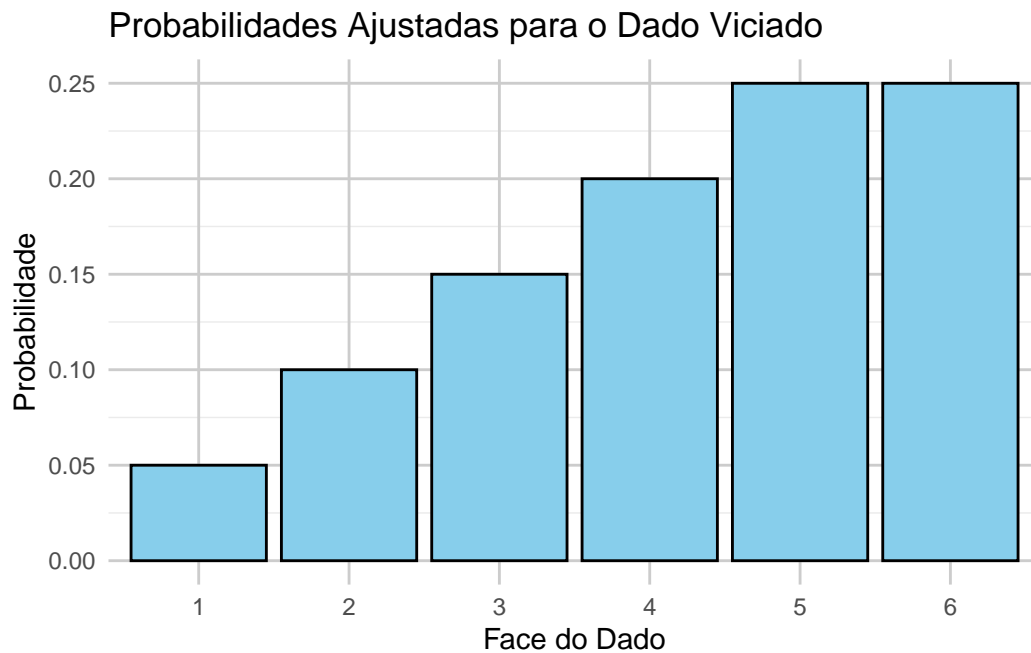
```
Face 1 : 0.0521 vezes
Face 2 : 0.0964 vezes
Face 3 : 0.1527 vezes
Face 4 : 0.2045 vezes
Face 5 : 0.2508 vezes
Face 6 : 0.2435 vezes
```

```
# Visualizando os resultados em gráficos
```

```
library(ggplot2)
```

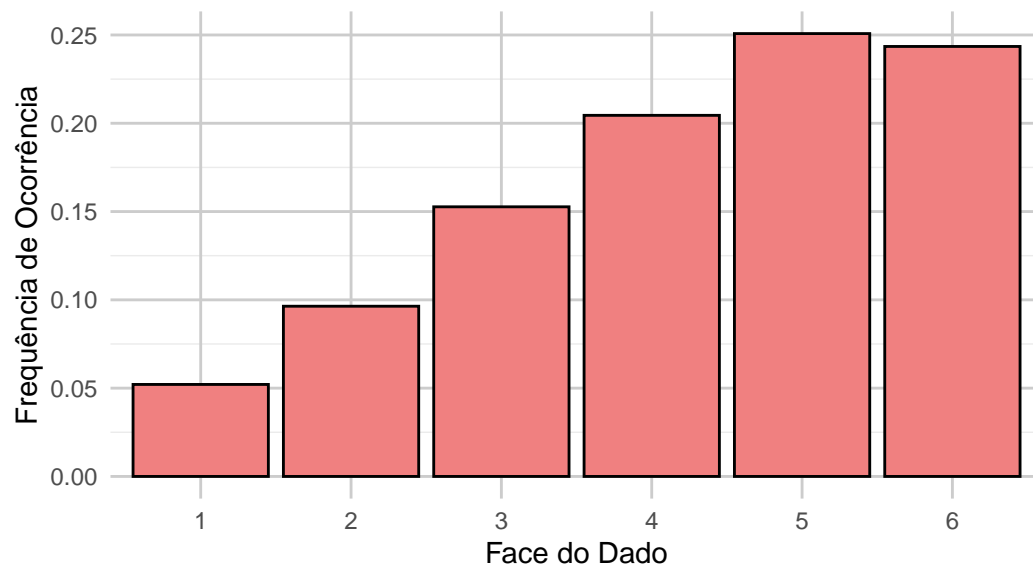
```
# Gráfico das probabilidades ajustadas
```

```
dados_probabilidades <- data.frame(faces = as.factor(faces), probabilidades = probabilidades)
ggplot(dados_probabilidades, aes(x = faces, y = probabilidades)) +
  geom_bar(stat = "identity", fill = "skyblue", color = "black") +
  ggtitle("Probabilidades Ajustadas para o Dado Viciado") +
  xlab("Face do Dado") +
  ylab("Probabilidade") +
  theme_minimal() +
  theme(panel.grid.major = element_line(color = "grey80"))
```



```
# Gráfico das frequências obtidas
dados_frequencias <- data.frame(faces = as.factor(faces), frequencias = frequencias)
ggplot(dados_frequencias, aes(x = faces, y = frequencias)) +
  geom_bar(stat = "identity", fill = "lightcoral", color = "black") +
  ggtitle(paste0("Simulação de Lançamentos de um Dado Viciado\n", n_lancamentos, " lançamentos")) +
  xlab("Face do Dado") +
  ylab("Frequência de Ocorrência") +
  theme_minimal() +
  theme(panel.grid.major = element_line(color = "grey80"))
```


Simulação de Lançamentos de um Dado Viciado
10000 lançamentos



9 Python

```
import numpy as np
import matplotlib.pyplot as plt

# Definindo as faces do dado e as probabilidades associadas (não uniformes)
faces = [1, 2, 3, 4, 5, 6]
probabilidades = [0.05, 0.1, 0.15, 0.2, 0.25, 0.25] # Probabilidades associadas às faces do

# Gerando um número aleatório e verificando em qual intervalo ele cai
def gerar_amostra_por_intervalos(probabilidades, faces):
    u = np.random.uniform(0, 1) # Gerando um número aleatório uniforme
    limite_inferior = 0 # Limite inferior do intervalo

    # Percorrendo as probabilidades e verificando em qual intervalo o número cai
    for i, p in enumerate(probabilidades):
        limite_superior = limite_inferior + p # Definindo o limite superior do intervalo
        if limite_inferior <= u < limite_superior:
            return faces[i] # Retorna a face correspondente ao intervalo
        limite_inferior = limite_superior # Atualiza o limite inferior para o próximo inter

# Simulando lançamentos do dado viciado utilizando a verificação dos intervalos
n_lancamentos = 10000
resultados = [gerar_amostra_por_intervalos(probabilidades, faces) for _ in range(n_lancamentos)]

# Contando as frequências de cada face
frequencias = [np.sum(np.array(resultados) == face) / n_lancamentos for face in faces]

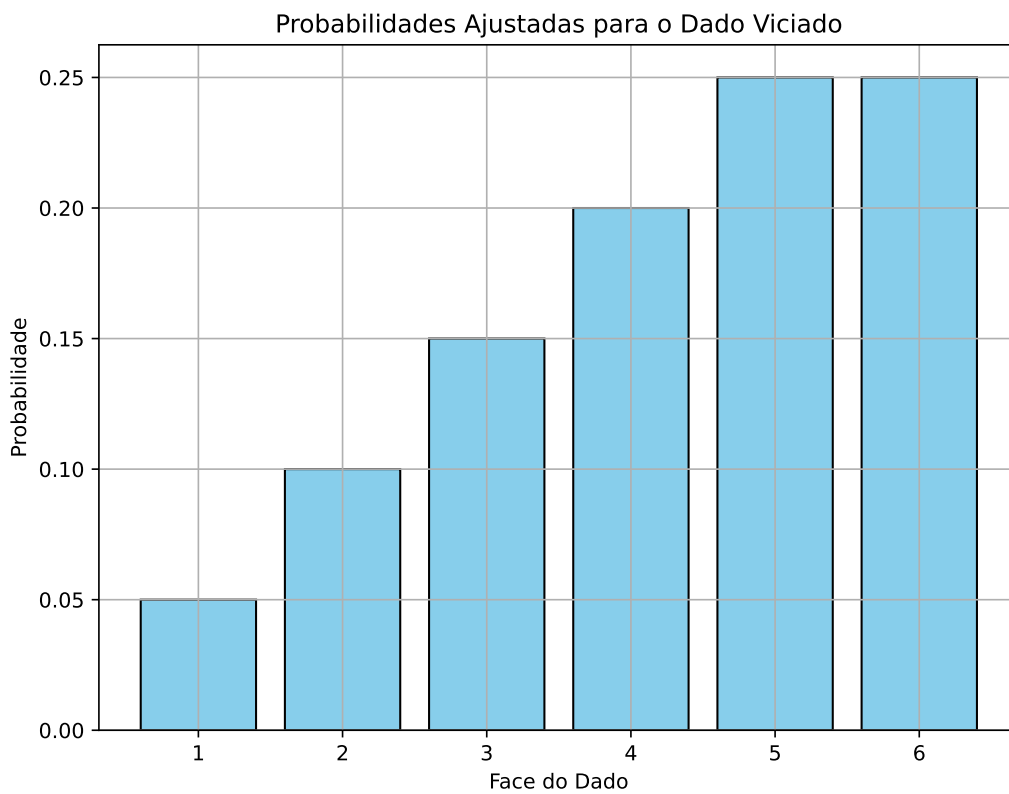
# Exibindo os resultados da simulação
print(f"Frequências de cada face após {n_lancamentos} lançamentos:")
```

Frequências de cada face após 10000 lançamentos:

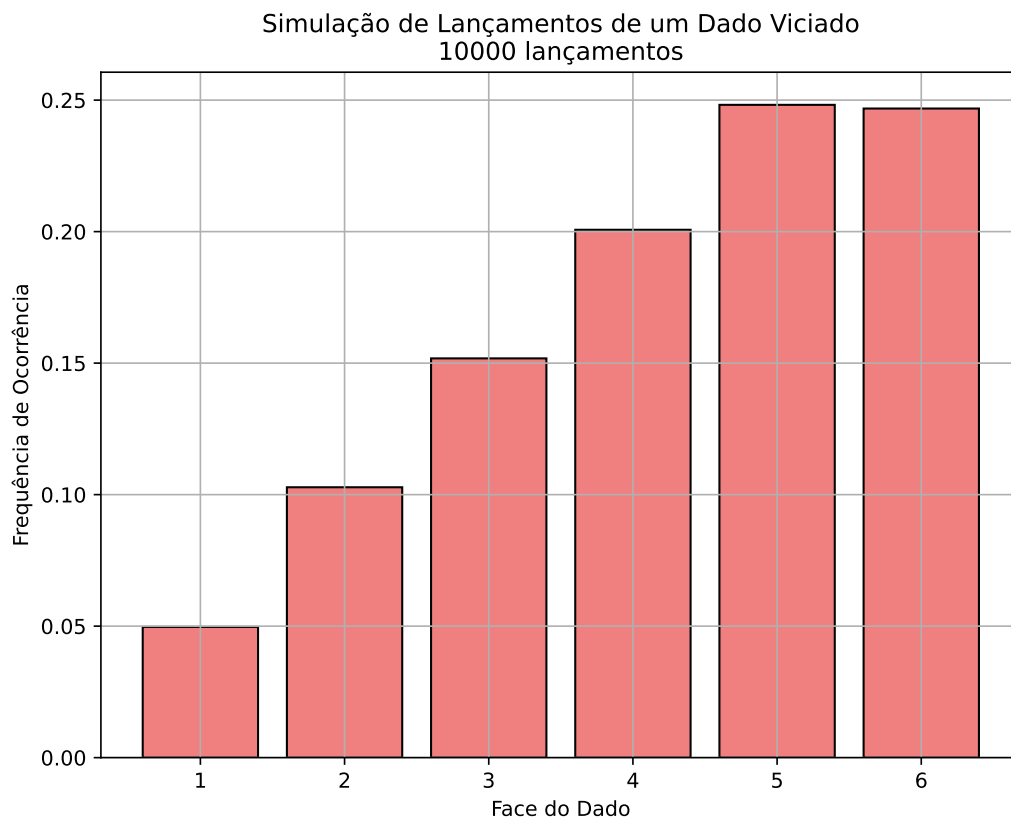
```
for face, freq in zip(faces, frequencias):
    print(f"Face {face}: {freq} vezes")
```

Face 1: 0.0497 vezes
Face 2: 0.1028 vezes
Face 3: 0.1518 vezes
Face 4: 0.2007 vezes
Face 5: 0.2482 vezes
Face 6: 0.2468 vezes

```
# Gráfico das probabilidades ajustadas
plt.figure(figsize=(8,6))
plt.bar(faces, probabilidades, color='skyblue', edgecolor='black')
plt.title('Probabilidades Ajustadas para o Dado Viciado')
plt.xlabel('Face do Dado')
plt.ylabel('Probabilidade')
plt.grid(True)
plt.show()
```



```
# Gráfico das frequências obtidas
plt.figure(figsize=(8,6))
plt.bar(faces, frequencias, color='lightcoral', edgecolor='black')
plt.title(f'Simulação de Lançamentos de um Dado Viciado\n{n_lancamentos} lançamentos')
plt.xlabel('Face do Dado')
plt.ylabel('Frequência de Ocorrência')
plt.grid(True)
plt.show()
```



Ou seja, se conseguimos simular uma distribuição uniforme, conseguimos simular uma distribuição discreta. Isso vale de forma mais geral?

10 Números Pseudoaleatórios

Números aleatórios têm muitas aplicações na computação, como em simulações, amostragem estatística, criptografia e jogos de azar. No entanto, os computadores, por serem sistemas determinísticos, não podem gerar números realmente aleatórios de forma autônoma. Em vez disso, utilizam algoritmos determinísticos que geram números que parecem aleatórios, e esses números são chamados de pseudoaleatórios.

10.1 O que é um número pseudoaleatório?

Um número pseudoaleatório é gerado a partir de uma fórmula matemática que, a partir de uma semente (um valor inicial), gera uma sequência de números que tem as propriedades desejadas de uma sequência aleatória. Essa sequência parece aleatória, mas se a mesma semente for usada, a sequência será a mesma.

10.2 Geração de Números Pseudoaleatórios com o Gerador Linear Congruente (LCG)

[Link para o wikipedia](#)

O **Gerador Linear Congruente** (LCG) é um dos métodos mais antigos e simples para gerar números pseudoaleatórios. Ele segue a fórmula:

$$X_{n+1} = (a \cdot X_n + c) \mod m$$

Onde: - X_n é o número atual (ou a semente inicial), - a é o multiplicador, - c é o incremento, - m é o módulo, ou seja, o intervalo dos números gerados.

A sequência gerada pelo LCG depende diretamente dos parâmetros a , c , m e da semente inicial X_0 . Um conjunto mal escolhido de parâmetros pode resultar em uma sequência com um período curto, o que compromete a aleatoriedade da sequência.

10.2.1 O que é a Função Módulo?

A **função módulo** (também conhecida como operação de resto) retorna o **resto da divisão** de um número por outro. Em termos matemáticos, para dois números inteiros a e b , a operação módulo é representada como:

$$r = a \mod b$$

Onde: - a é o dividendo, - b é o divisor, - r é o resto da divisão de a por b .

Por exemplo, se temos $a = 17$ e $b = 5$, a divisão de 17 por 5 dá 3 com um resto de 2, então:

$$17 \mod 5 = 2$$

No contexto do **Gerador Linear Congruente (LCG)**, a função módulo é usada para garantir que os números gerados fiquem dentro de um intervalo específico, geralmente entre 0 e $m - 1$, onde m é o módulo definido no algoritmo.

11 R

```
# Exemplo de uso da função módulo em R

# Definindo os valores
a <- 17
b <- 3

# Calculando o módulo de a por b
resto <- a %% b

# Exibindo o resultado
cat("O resultado de", a, "%%", b, "é:", resto, "\n")
```

O resultado de 17 %% 3 é: 2

12 Python

```
# Exemplo de uso da função módulo em Python

# Definindo os valores
a = 17
b = 3

# Calculando o módulo de a por b
resto = a % b

# Exibindo o resultado
print(f"O resultado de {a} % {b} é: {resto}")
```

O resultado de 17 % 3 é: 2

12.1 Por que o Gerador Linear Congruente Funciona?

O Gerador Linear Congruente (LCG) é um dos métodos mais simples e eficientes para gerar números pseudoaleatórios. Sua eficácia se baseia em um bom equilíbrio entre a escolha dos parâmetros (multiplicador a , incremento c , módulo m e semente inicial X_0) e as propriedades matemáticas que garantem uma sequência suficientemente “aleatória”. Para que o LCG funcione bem, os parâmetros precisam ser cuidadosamente selecionados para garantir que a sequência gerada tenha um período longo, seja bem distribuída e evite padrões repetitivos.

12.1.1 A Fórmula do LCG

A fórmula básica do LCG é:

$$X_{n+1} = (a \cdot X_n + c) \mod m$$

Onde: - X_n é o número gerado na n -ésima iteração, - a é o multiplicador, - c é o incremento, - m é o módulo, - X_0 é a semente inicial.

O número gerado em cada iteração é o **resto da divisão** de $(a \cdot X_n + c)$ por m . Essa operação garante que os números fiquem dentro do intervalo $[0, m - 1]$. A normalização posterior geralmente transforma esses números em valores no intervalo $[0, 1)$.

12.1.2 O Papel de m

O valor de m , conhecido como módulo, define o intervalo no qual os números gerados estarão contidos. Em muitos casos, m é escolhido como uma potência de 2 (por exemplo, $m = 2^{32}$ ou $m = 2^{64}$) porque cálculos modulares com potências de 2 são mais rápidos em hardware.

A escolha de m também influencia o **período máximo** da sequência. Se todos os parâmetros forem escolhidos corretamente, o LCG pode gerar uma sequência com o período máximo, que é m . Isso significa que a sequência não repetirá nenhum número até que m números tenham sido gerados.

12.1.3 O Papel de a , c e a Condição de Coprimos

Para garantir que o gerador tenha o período máximo (ou seja, m números diferentes antes de repetir a sequência), a escolha dos parâmetros a (multiplicador), c (incremento) e m (módulo) deve satisfazer as seguintes condições baseadas em **teorias de números**:

1. **O incremento c deve ser coprimo com m :**

- Dois números são **coprimos** se o maior divisor comum deles for 1, ou seja, $\gcd(c, m) = 1$. Isso garante que, ao somar c , todos os possíveis valores de X_n possam ser atingidos antes de repetir a sequência.
- Se c não for coprimo com m , a sequência gerada pode pular certos valores, resultando em um período mais curto do que o esperado.

2. **O valor de $a - 1$ deve ser divisível por todos os fatores primos de m :**

- Se m é uma potência de 2 (por exemplo, $m = 2^k$), a escolha de a deve ser tal que $(a - 1)$ seja divisível por 2 para garantir que o período seja maximizado.

3. **Se m for divisível por 4, então $(a - 1)$ também deve ser divisível por 4:**

- Isso é necessário para garantir que todos os resíduos modulares possíveis possam ser gerados, especialmente quando m é uma potência de 2.

12.1.4 Exemplo de uma Escolha Correta de Parâmetros

Um exemplo clássico de um bom conjunto de parâmetros é:

- $m = 2^{32}$ (módulo com 32 bits),
- $a = 1664525$ (multiplicador),
- $c = 1013904223$ (incremento),
- $X_0 = 42$ (semente inicial, que pode ser qualquer valor).

Esses parâmetros foram escolhidos para garantir que o LCG tenha um longo período e uma boa distribuição dos números gerados. O módulo $m = 2^{32}$ é uma potência de 2, o que torna as operações modulares mais rápidas, e os valores de a e c satisfazem as condições matemáticas para maximizar o período.

13 R

```
# Importando o pacote necessário
library(gmp)

# Parâmetros do exemplo
m <- 2^32
a <- 1664525
c <- 1013904223

# Verificando as condições
# 1. O incremento c deve ser coprimo com m
coprimo_c_m <- gcd(c, m) == 1

# 2. a - 1 deve ser divisível por todos os fatores primos de m
a_menos_1 <- a - 1

print(a_menos_1)
```

```
[1] 1664524
```

```
# Verificando se a - 1 é divisível por 2 (único fator primo de m = 2^32)
divisivel_por_2 <- (a_menos_1 %% 2 == 0)

# 3. Se m for divisível por 4, a - 1 também deve ser divisível por 4
divisivel_por_4 <- (a_menos_1 %% 4 == 0)

list(coprimo_c_m, divisivel_por_2, divisivel_por_4)
```

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

```
[[2]]
[1] TRUE
```

```
[[3]]  
[1] TRUE
```

14 Python

```
import math

# Parâmetros do exemplo
m = 2**32
a = 1664525
c = 1013904223

# Verificando as condições
# 1. O incremento c deve ser coprimo com m
coprimo_c_m = math.gcd(c, m) == 1

# 2. a - 1 deve ser divisível por todos os fatores primos de m
a_menos_1 = a - 1

print(a_menos_1)
```

1664524

```
# Verificando se a - 1 é divisível por 2 (único fator primo de m = 2^32)
divisivel_por_2 = (a_menos_1 % 2 == 0)

# 3. Se m for divisível por 4, a - 1 também deve ser divisível por 4
divisivel_por_4 = (a_menos_1 % 4 == 0)

coprimo_c_m, divisivel_por_2, divisivel_por_4
```

(True, True, True)

14.0.1 Por que o LCG Funciona Bem?

O LCG funciona porque: - **As operações modulares** garantem que os números gerados estejam dentro de um intervalo fixo e possam cobrir todo o espaço de possíveis valores de maneira

ordenada. - **A escolha adequada dos parâmetros** garante que a sequência tenha um longo período (o maior possível dado m), evita padrões repetitivos e assegura que a sequência seja **pseudoaleatória** o suficiente para muitas aplicações, como simulações e métodos de Monte Carlo.

No entanto, o LCG pode não ser adequado para todas as aplicações, especialmente em criptografia, onde a previsibilidade é um problema. Para a maioria dos usos científicos e de simulação, ele ainda é uma escolha eficiente e simples.

14.0.2 Efeito dos Parâmetros no Gerador Linear Congruente (LCG)

Os parâmetros no Gerador Linear Congruente (LCG) têm um impacto significativo sobre a qualidade e as propriedades da sequência de números pseudoaleatórios gerados. Os parâmetros principais são:

1. Multiplicador a :

- Esse parâmetro é essencial para garantir que a sequência de números gerados tenha um bom período (o número de valores distintos antes de a sequência começar a se repetir). Se o valor de a não for bem escolhido, o período da sequência pode ser curto e a qualidade dos números gerados diminui.
- Bons valores de a são cruciais para evitar padrões repetitivos ou ciclos curtos.

2. Incremento c :

- O incremento c adiciona um valor fixo à sequência e é um dos fatores que pode garantir que todos os valores no intervalo $[0, m)$ sejam atingidos em algum momento, desde que os outros parâmetros também sejam bem escolhidos.
- Quando $c = 0$, o gerador é chamado de **multiplicativo**. Nessa forma, o LCG pode ter um comportamento menos uniforme.

3. Módulo m :

- O módulo define o intervalo dos números gerados. Comumente, m é escolhido como uma potência de 2 (por exemplo, $m = 2^{32}$) para facilitar os cálculos modulares em hardware e software.
- O valor de m também determina o período máximo da sequência de números. Com um módulo de m , o período máximo teórico que o LCG pode ter é m , mas isso depende da escolha correta dos parâmetros a e c .

4. Semente X_0 :

- A semente é o valor inicial de X_0 usado pelo LCG para iniciar a sequência. Mudar a semente resultará em uma sequência diferente, mas com o mesmo período e comportamento determinado pelos outros parâmetros.

- A semente garante que o algoritmo possa ser **reproduzido**. Se dois programas utilizarem a mesma semente com os mesmos parâmetros, ambos produzirão a mesma sequência de números.

14.0.3 Impacto dos Parâmetros:

1. Período da Sequência:

- O período é a quantidade de números gerados antes que a sequência comece a se repetir. Para obter o período máximo, os parâmetros a , c , m e a semente X_0 precisam ser cuidadosamente escolhidos.
- Se os parâmetros não forem bons, o gerador pode produzir uma sequência com um ciclo muito curto ou, pior, um conjunto pequeno de valores.

2. Distribuição dos Números:

- Embora o LCG gere números no intervalo $[0, 1)$, o quão bem distribuídos esses números estão nesse intervalo depende dos parâmetros.
- Parâmetros mal escolhidos podem causar uma distribuição não uniforme, onde certos intervalos terão mais números gerados que outros, levando a um comportamento indesejável.

3. Padrões Repetitivos:

- Se os parâmetros forem mal escolhidos, podem surgir padrões repetitivos que comprometem a aleatoriedade dos números. Esses padrões tornam o LCG inadequado para algumas aplicações, como criptografia ou simulações que exigem alta qualidade de aleatoriedade.

Por essas razões, a escolha dos parâmetros a , c , m e da semente X_0 é crítica para garantir que o LCG produza números pseudoaleatórios de alta qualidade e com um período longo.

15 R

```
# Carregando o pacote ggplot2
library(ggplot2)

# Classe para o Gerador Congruente Linear
LinearCongruentialGenerator <- setRefClass(
  "LinearCongruentialGenerator",
  fields = list(a = "numeric", c = "numeric", m = "numeric", semente = "numeric"),
  methods = list(
    initialize = function(semente, a = 1103515245, c = 12345, m = 2^32) {
      .self$a <- a
      .self$c <- c
      .self$m <- m
      .self$semente <- semente
    },
    gerar = function() {
      # Atualizando a semente
      .self$semente <- (.self$a * .self$semente + .self$c) %% .self$m
      return(.self$semente / .self$m) # Normalizando para [0, 1)
    }
  )
)

# Inicializando o gerador com uma semente
lcg <- LinearCongruentialGenerator$new(semente = 5)

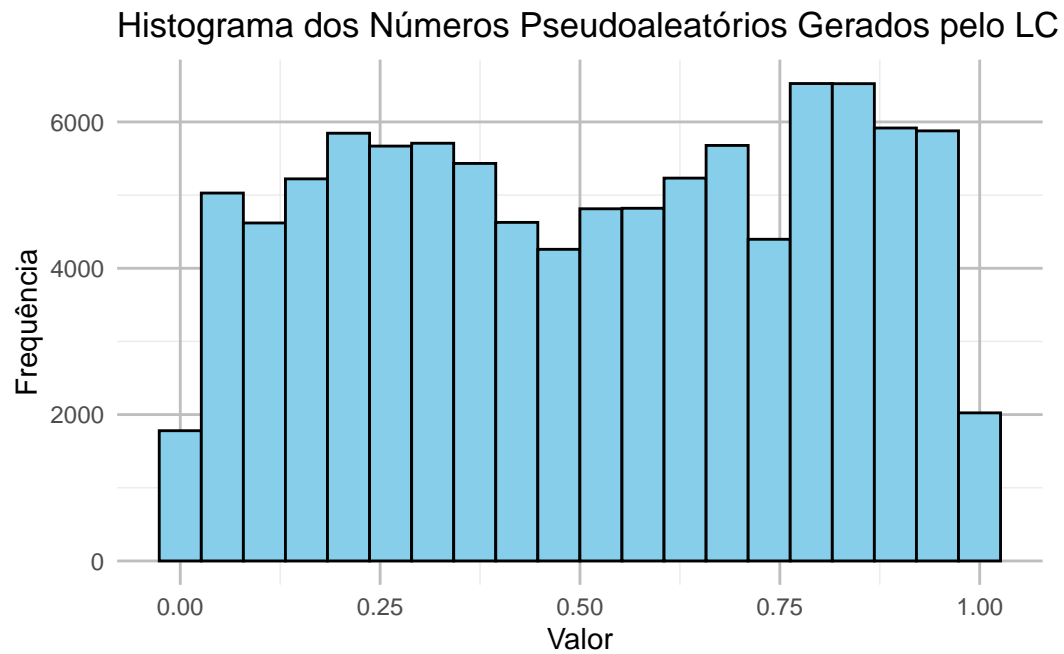
# Gerando 1000 números pseudoaleatórios
numeros_gerados <- sapply(1:100000, function(x) lcg$gerar())

# Convertendo para um data.frame
dados <- data.frame(numeros_gerados = numeros_gerados)

# Plotando o histograma dos números gerados
ggplot(dados, aes(x = numeros_gerados)) +
  geom_histogram(bins = 20, fill = 'skyblue', color = 'black') +
```



```
ggtitle('Histograma dos Números Pseudoaleatórios Gerados pelo LCG') +
xlab('Valor') +
ylab('Frequência') +
theme_minimal() +
theme(panel.grid.major = element_line(color = "grey"))
```



16 Python

```
import matplotlib.pyplot as plt

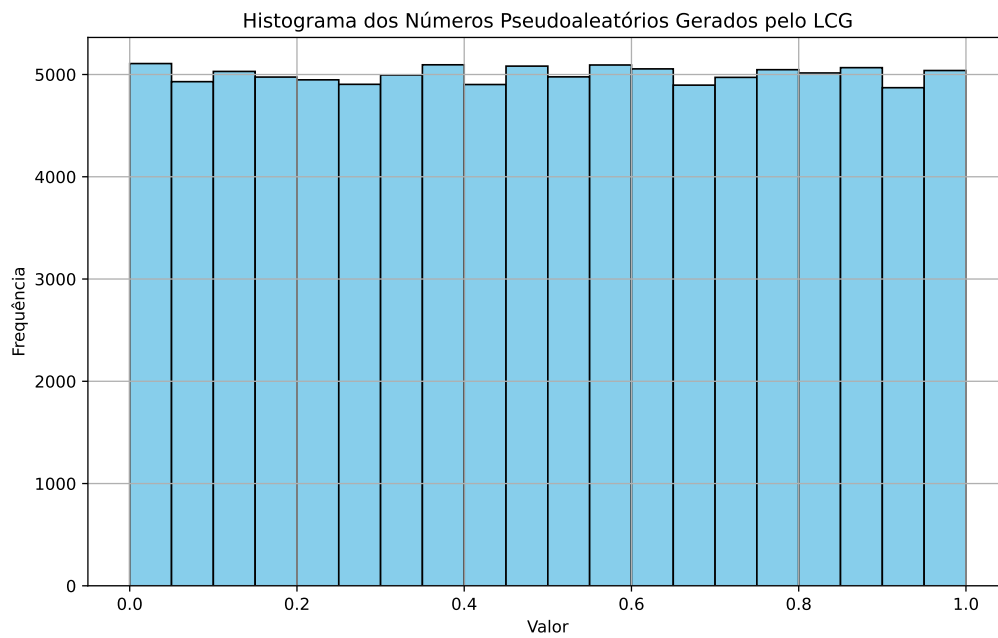
class LinearCongruentialGenerator:
    def __init__(self, semente, a=1103515245, c=12345, m=2**32):
        self.a = a
        self.c = c
        self.m = m
        self.semente = semente

    def gerar(self):
        # Atualizando a semente
        self.semente = (self.a * self.semente + self.c) % self.m
        return self.semente / self.m # Normalizando para [0, 1)

# Inicializando o gerador com uma semente
lcg = LinearCongruentialGenerator(semente=5)

# Gerando 1000 números pseudoaleatórios
numeros_gerados = [lcg.gerar() for _ in range(100000)]

# Plotando o histograma dos números gerados
plt.figure(figsize=(10, 6))
plt.hist(numeros_gerados, bins=20, color='skyblue', edgecolor='black')
plt.title('Histograma dos Números Pseudoaleatórios Gerados pelo LCG')
plt.xlabel('Valor')
plt.ylabel('Frequência')
plt.grid(True)
plt.show()
```



16.1 Gerando números uniformes com uma moeda

17 R

```
# Carregando pacotes necessários
library(ggplot2)

# Função para simular o lançamento de uma moeda justa
lancar_moeda <- function() {
  # Lançar moeda justa: 0 para coroa (K) e 1 para cara (C)
  sample(c(0, 1), 1)
}

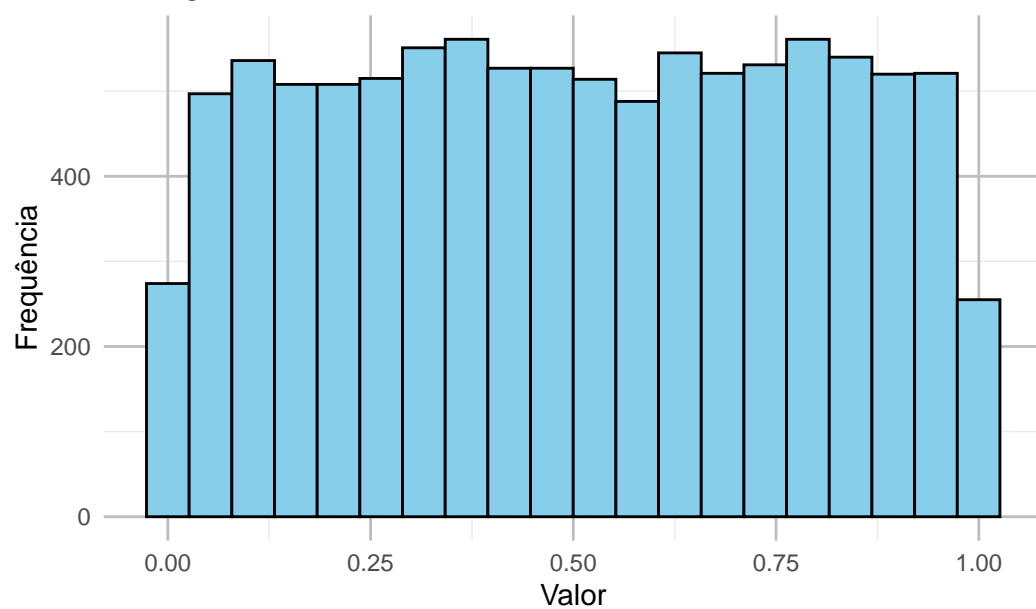
# Função para gerar um número uniformemente distribuído usando uma moeda
gerar_numero_uniforme <- function(n_bits = 32) {
  numero <- 0
  for (i in 1:n_bits) {
    bit <- lancar_moeda()
    # Atualizando o número, multiplicando pela base 2
    numero <- numero + bit * (2-(i))
  }
  return(numero)
}

# Gerando 10000 números uniformemente distribuídos
numeros_uniformes <- sapply(1:10000, function(x) gerar_numero_uniforme())

# Convertendo para um data.frame
dados <- data.frame(numeros_uniformes = numeros_uniformes)

# Plotando o histograma dos números gerados
ggplot(dados, aes(x = numeros_uniformes)) +
  geom_histogram(bins = 20, fill = 'skyblue', color = 'black') +
  ggtitle('Histograma de Números Uniformes Gerados Usando uma Moeda Justa') +
  xlab('Valor') +
  ylab('Frequência') +
  theme_minimal() +
  theme(panel.grid.major = element_line(color = "grey"))
```

Histograma de Números Uniformes Gerados Usando uma Mo



18 Python

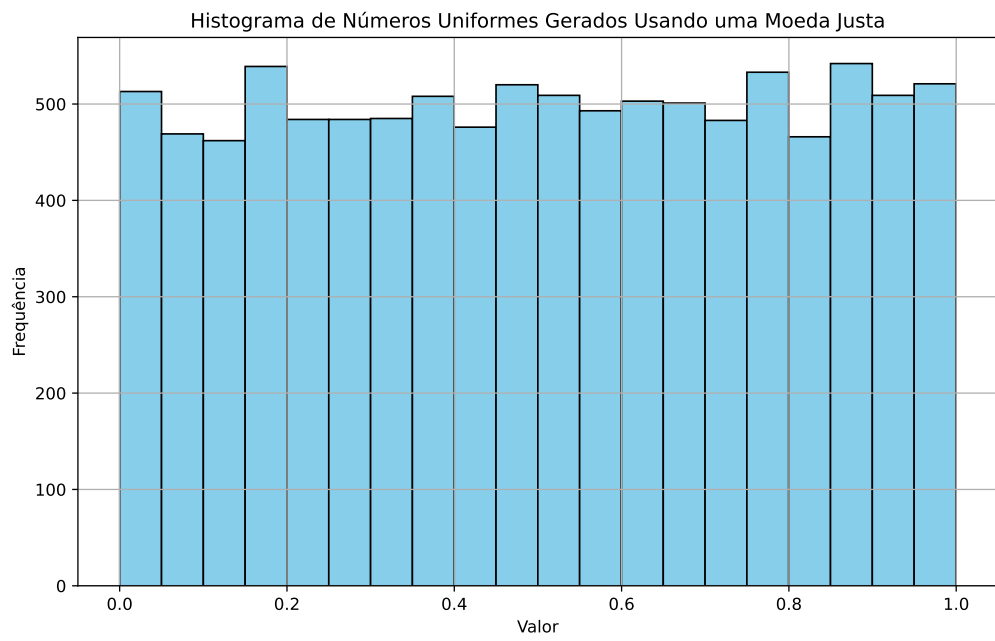
```
import random
import matplotlib.pyplot as plt

# Função para simular o lançamento de uma moeda justa
def lancar_moeda():
    # Lançar moeda justa: 0 para coroa (K) e 1 para cara (C)
    return random.choice([0, 1])

# Função para gerar um número uniformemente distribuído usando uma moeda
def gerar_numero_uniforme(n_bits=32):
    numero = 0
    for i in range(n_bits):
        bit = lancar_moeda()
        # Atualizando o número, multiplicando pela base 2
        numero += bit * (2 ** -(i + 1)) # Cada bit tem um peso de 2-(posição)
    return numero

# Gerando 1000 números uniformemente distribuídos
numeros_uniformes = [gerar_numero_uniforme() for _ in range(10000)]

# Plotando o histograma dos números gerados
plt.figure(figsize=(10, 6))
plt.hist(numeros_uniformes, bins=20, color='skyblue', edgecolor='black')
plt.title('Histograma de Números Uniformes Gerados Usando uma Moeda Justa')
plt.xlabel('Valor')
plt.ylabel('Frequência')
plt.grid(True)
plt.show()
```



19 Técnica da Inversão para Variáveis Discretas

A **técnica da inversão** é uma maneira poderosa de gerar variáveis aleatórias (v.a.) a partir de uma distribuição arbitrária, usando números aleatórios uniformes no intervalo $[0, 1)$. A ideia básica é usar a **função de distribuição acumulada (CDF)** para mapear um número uniforme gerado entre 0 e 1 para o valor correspondente da v.a. discreta.

Passos Gerais para Gerar Variáveis Aleatórias Discretas com a Técnica da Inversão:

1. **Calcular a CDF** da variável aleatória que se deseja gerar.
2. **Gerar um número aleatório uniforme** $u \in [0, 1)$.
3. Encontrar o valor da variável aleatória cuja CDF seja maior ou igual a u .
4. Retornar esse valor como a variável aleatória gerada.

19.1 Geração de Variáveis Aleatórias Discretas Genéricas

Dado um conjunto de probabilidades $p(x_i)$, onde x_i são os valores possíveis da variável aleatória discreta, e $p(x_i)$ são suas respectivas probabilidades, a CDF $F(x)$ é calculada como:

$$F(x_i) = \sum_{j=1}^i p(x_j)$$

O algoritmo para gerar uma variável aleatória discreta genérica é:

1. Gerar um número aleatório $u \in [0, 1)$.
2. Encontrar o menor valor x_i tal que $F(x_i) \geq u$.
3. Retornar x_i .

20 R

```
# Exemplo de valores e probabilidades de uma variável aleatória discreta
valores <- c(0, 1, 2, 3, 4, 5, 6)
probabilidades <- c(0, 0.1, 0.2, 0.3, 0.25, 0.15, 0)

# Calculando a CDF
cdf <- cumsum(probabilidades)

# Gerando um número aleatório uniforme
u <- runif(1)

# Encontrando o valor correspondente na CDF
valor_gerado <- NA
for (i in seq_along(valores)) {
  if (u < cdf[i]) {
    valor_gerado <- valores[i]
    break
  }
}

# Ajustando o gráfico para corrigir a visualização da CDF e garantir que os valores estejam corretos
library(ggplot2)

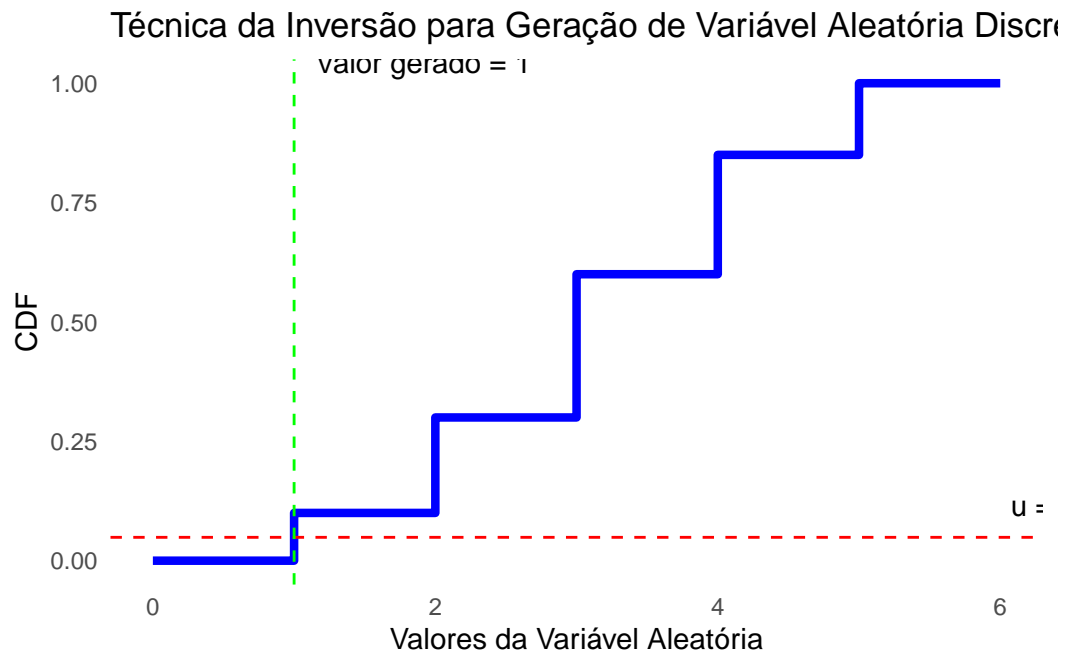
# Criando um dataframe para os valores e CDF
df <- data.frame(valores = valores, cdf = cdf)

# Gráfico da CDF com número aleatório e valor gerado
ggplot(df, aes(x = valores, y = cdf)) +
  geom_step(direction = "hv", color = "blue", size = 1.5) +
  geom_hline(yintercept = u, color = "red", linetype = "dashed") +
  geom_vline(xintercept = valor_gerado, color = "green", linetype = "dashed") +
  labs(title = "Técnica da Inversão para Geração de Variável Aleatória Discreta",
       x = "Valores da Variável Aleatória",
       y = "CDF") +
  annotate("text", x = max(valores), y = u, label = sprintf("u = %.2f", u), hjust = -0.1, vjust = "top")
```

```

annotate("text", x = valor_gerado, y = max(cdf), label = paste("Valor gerado =", valor_gerado),
theme_minimal() +
theme(panel.grid = element_blank())

```



21 Python

```
import numpy as np
import matplotlib.pyplot as plt

# Exemplo de valores e probabilidades de uma variável aleatória discreta
valores = [0, 1, 2, 3, 4, 5, 6]
probabilidades = [0, 0.1, 0.2, 0.3, 0.25, 0.15, 0]

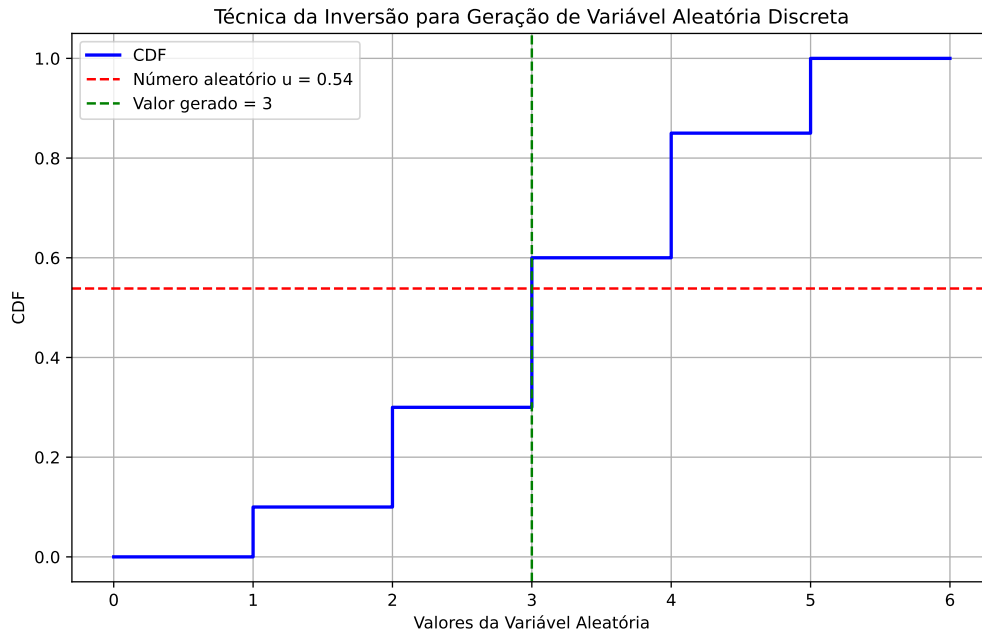
# Calculando a CDF
cdf = np.cumsum(probabilidades)

# Gerando um número aleatório uniforme
u = np.random.uniform(0, 1)

# Encontre o valor correspondente na CDF
valor_gerado = None
for i, valor in enumerate(valores):
    if u < cdf[i]:
        valor_gerado = valor
        break

# Ajustando o gráfico para corrigir a visualização da CDF e garantir que os valores estejam corretos
plt.figure(figsize=(10, 6))

# Ajustando o eixo x para que a CDF comece e termine corretamente
plt.step(valores, cdf, label='CDF', color='blue', linewidth=2, where='post')
plt.axhline(y=u, color='red', linestyle='--', label=f'Número aleatório u = {u:.2f}')
plt.axvline(x=valor_gerado, color='green', linestyle='--', label=f'Valor gerado = {valor_gerado}')
plt.title('Técnica da Inversão para Geração de Variável Aleatória Discreta')
plt.xlabel('Valores da Variável Aleatória')
plt.ylabel('CDF')
plt.legend()
plt.grid(True)
plt.show()
```



21.1 Inversa da CDF

A **inversa da CDF** (função de distribuição acumulada), também conhecida como a **função quantil** ou **função percentil**, é uma função utilizada para gerar variáveis aleatórias a partir de uma distribuição específica.

Definição: Seja $F(x)$ a função de distribuição acumulada (CDF) de uma variável aleatória X . A inversa da CDF, denotada por $F^{-1}(p)$, é definida como:

$$F^{-1}(p) = \inf\{x \in \mathbb{R} : F(x) \geq p\}, \quad \text{para } p \in [0, 1]$$

Em palavras: - A inversa da CDF $F^{-1}(p)$ mapeia um número p , que representa uma probabilidade acumulada, de volta ao valor x correspondente da variável aleatória X , tal que a probabilidade acumulada até x é igual a p . - Isso significa que, se $p = F(x)$, então $F^{-1}(p) = x$.

21.2 Geração de Variáveis Aleatórias com Distribuição Geométrica

A distribuição geométrica modela o número de tentativas até o primeiro sucesso em uma sequência de experimentos de Bernoulli. Se a probabilidade de sucesso em cada tentativa é p ,

a PMF é dada por:

21.2.1 Derivação da Inversa da CDF para a Distribuição Geométrica

A fórmula da **inversa da CDF** para a distribuição geométrica foi obtida a partir da definição da função de distribuição acumulada (CDF) da distribuição geométrica e da aplicação da técnica da inversão.

21.2.1.1 CDF da Distribuição Geométrica

A função de distribuição acumulada (CDF) da distribuição geométrica com parâmetro p (a probabilidade de sucesso) para o número de falhas $k - 1$ antes do primeiro sucesso é dada por:

$$F(k) = 1 - (1 - p)^k$$

Essa equação expressa a probabilidade acumulada de obter o primeiro sucesso em até k tentativas.

21.2.1.2 Inversa da CDF

Queremos encontrar a **inversa da CDF**, ou seja, a fórmula que, dado um valor u entre 0 e 1, nos permita calcular o valor k tal que $F(k) = u$.

Sabemos que:

$$u = F(k) = 1 - (1 - p)^k$$

Nosso objetivo é resolver essa equação para k . Vamos fazer isso passo a passo.

21.2.1.3 Isolando o termo com k

Começamos isolando o termo $(1 - p)^k$:

$$u = 1 - (1 - p)^k$$

Subtraindo 1 de ambos os lados:

$$u - 1 = -(1 - p)^k$$

Multiplicando ambos os lados por -1 :

$$1 - u = (1 - p)^k$$

21.2.1.4 Aplicando o Logaritmo

Agora aplicamos o logaritmo natural (\log base e) em ambos os lados para resolver k :

$$\log(1 - u) = \log((1 - p)^k)$$

Usando a propriedade dos logaritmos que permite trazer o expoente k para frente:

$$\log(1 - u) = k \cdot \log(1 - p)$$

21.2.1.5 Isolando k

Agora, isolamos k :

$$k = \frac{\log(1 - u)}{\log(1 - p)}$$

Como k precisa ser um número inteiro (já que a distribuição geométrica conta o número de tentativas), usamos a função de arredondamento “para cima” ($\lceil \cdot \rceil$), conhecida como a **função teto**:

$$k = \lceil \frac{\log(1 - u)}{\log(1 - p)} \rceil$$

21.2.2 Conclusão

Portanto, a fórmula da inversa da CDF da distribuição geométrica é:

$$k = \lceil \frac{\log(1 - u)}{\log(1 - p)} \rceil$$

Esta fórmula nos permite gerar variáveis aleatórias com distribuição geométrica a partir de um número aleatório uniforme $u \in [0, 1)$.

22 R

```
# Função para gerar a inversa da CDF para a distribuição geométrica
inversa_cdf_geometrica <- function(p, u) {
  # Usando a fórmula inversa da CDF geométrica:  $F^{-1}(u) = \text{ceil}(\log(1 - u) / \log(1 - p))$ 
  k <- ceiling(log(1 - u) / log(1 - p))
  return(as.integer(k))
}

# Parâmetro p da distribuição geométrica
p <- 0.5

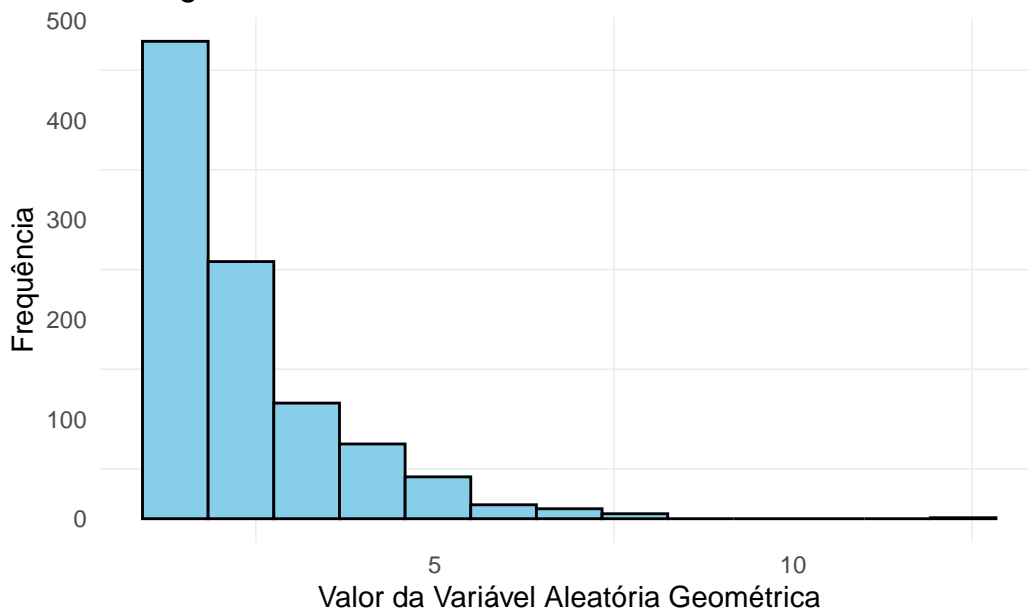
# Gerando 1000 números uniformemente distribuídos
uniformes <- runif(1000)

# Gerando a variável aleatória geométrica correspondente para cada número uniforme
geometricas <- sapply(uniformes, inversa_cdf_geometrica, p = p)

# Plotando um histograma das variáveis geométricas geradas
library(ggplot2)

df <- data.frame(geometricas = geometricas)
ggplot(df, aes(x = geometricas)) +
  geom_histogram(bins = max(geometricas) + 1, color = "black", fill = "skyblue", boundary = "none") +
  labs(title = "Histograma de Variáveis Aleatórias Geométricas Usando a Inversa da CDF",
       x = "Valor da Variável Aleatória Geométrica",
       y = "Frequência") +
  theme_minimal() +
  theme(panel.grid.major = element_blank())
```

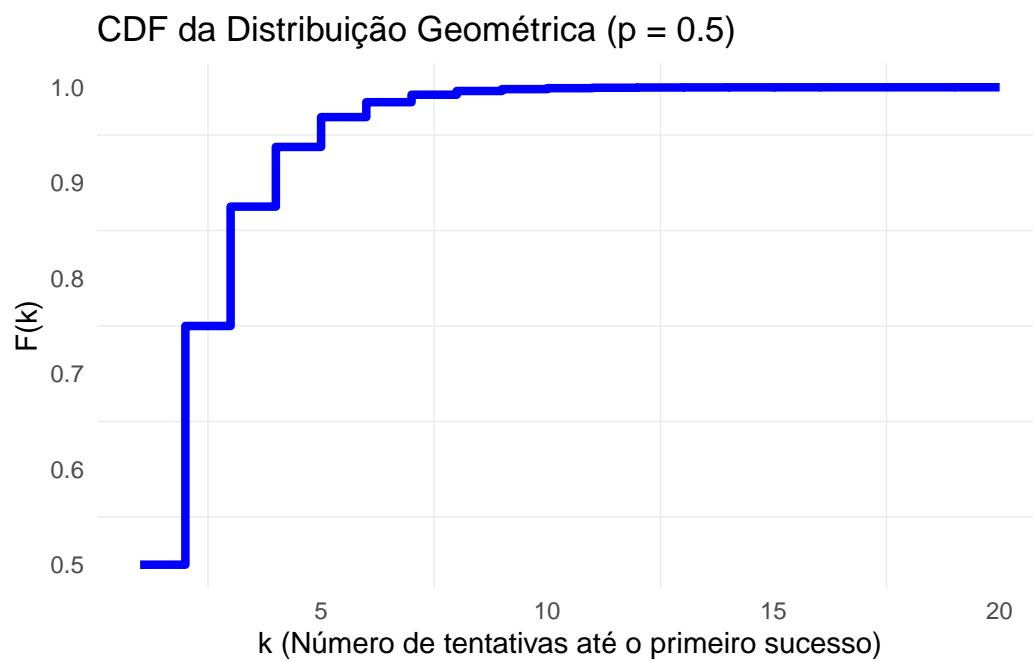
Histograma de Variáveis Aleatórias Geométricas Usando a Inv



```
# Função para calcular a CDF da distribuição geométrica
cdf_geometrica <- function(k, p) {
  return(1 - (1 - p)^k)
}

# Gerando valores de k para plotar a CDF
k_values <- 1:20
cdf_values <- sapply(k_values, cdf_geometrica, p = p)

# Plotando a CDF da distribuição geométrica
df_cdf <- data.frame(k_values = k_values, cdf_values = cdf_values)
ggplot(df_cdf, aes(x = k_values, y = cdf_values)) +
  geom_step(direction = "hv", color = "blue", size = 1.5) +
  labs(title = "CDF da Distribuição Geométrica (p = 0.5)",
       x = "k (Número de tentativas até o primeiro sucesso)",
       y = "F(k)") +
  theme_minimal() +
  theme(panel.grid.major = element_blank())
```

23 Python

```
import numpy as np
import matplotlib.pyplot as plt

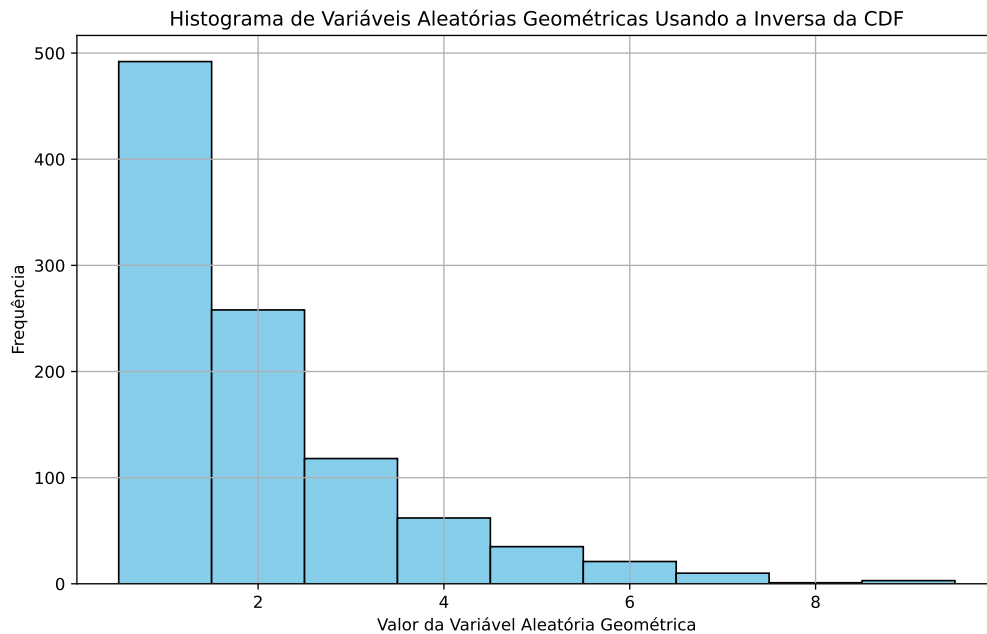
# Função para gerar a inversa da CDF para a distribuição geométrica
def inversa_cdf_geometrica(p, u):
    # Usando a fórmula inversa da CDF geométrica:  $F^{-1}(u) = \text{ceil}(\log(1 - u) / \log(1 - p))$ 
    k = np.ceil(np.log(1 - u) / np.log(1 - p))
    return int(k)

# Parâmetro p da distribuição geométrica
p = 0.5

# Gerando 100 números uniformemente distribuídos
uniformes = np.random.uniform(0, 1, 1000)

# Gerando a variável aleatória geométrica correspondente para cada número uniforme
geometricas = [inversa_cdf_geometrica(p, u) for u in uniformes]

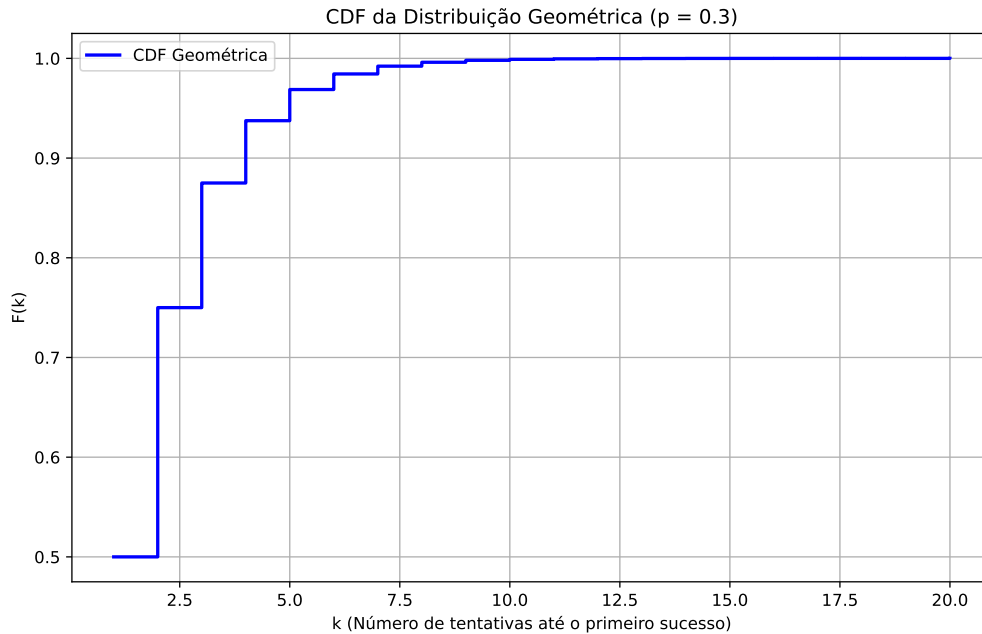
# Plotando um histograma das variáveis geométricas geradas
plt.figure(figsize=(10, 6))
plt.hist(geometricas, bins=range(1, max(geometricas) + 1), color='skyblue', edgecolor='black')
plt.title('Histograma de Variáveis Aleatórias Geométricas Usando a Inversa da CDF')
plt.xlabel('Valor da Variável Aleatória Geométrica')
plt.ylabel('Frequência')
plt.grid(True)
plt.show()
```



```
# Calculando a CDF para a distribuição geométrica
def cdf_geometrica(k, p):
    return 1 - (1 - p)**k

# Gerando valores de k para plotar a CDF
k_values = np.arange(1, 21)
cdf_values = [cdf_geometrica(k, p) for k in k_values]

# Plotando a CDF da distribuição geométrica
plt.figure(figsize=(10, 6))
plt.step(k_values, cdf_values, where='post', color='blue', label='CDF Geométrica', linewidth=2)
plt.title('CDF da Distribuição Geométrica (p = 0.3)')
plt.xlabel('k (Número de tentativas até o primeiro sucesso)')
plt.ylabel('F(k)')
plt.grid(True)
plt.legend()
plt.show()
```



23.1 Geração de Variáveis Aleatórias com Distribuição Poisson

A distribuição de Poisson é usada para modelar o número de eventos que ocorrem em um intervalo de tempo ou espaço fixo, onde os eventos ocorrem com uma taxa constante λ e de forma independente.

A função de probabilidade de massa (PMF) da distribuição de Poisson é dada por:

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad k = 0, 1, 2, \dots$$

23.1.1 Derivação da Inversa da CDF para a Distribuição Poisson Usando a Fórmula Recursiva

A **distribuição de Poisson** tem uma fórmula recursiva que pode ser usada para calcular as probabilidades de forma mais eficiente. Em vez de recalcular a probabilidade $P(X = k)$ a cada vez, podemos usar a seguinte relação recursiva:

$$P(X = k + 1) = \frac{\lambda}{k + 1} \cdot P(X = k)$$

Onde $P(X = 0) = e^{-\lambda}$.

Essa relação recursiva permite gerar variáveis aleatórias de Poisson sem precisar calcular fatoriais repetidamente, o que é mais eficiente para grandes valores de λ ou grandes números de eventos k .

23.1.2 Técnica da Inversão Usando a Fórmula Recursiva

Para gerar uma variável aleatória de Poisson usando a técnica da inversão e a fórmula recursiva, o processo é o seguinte:

1. Gerar um número aleatório uniforme $u \in [0, 1)$.
2. Calcular a CDF acumulada para valores de k , somando as probabilidades da PMF (usando a relação recursiva) até que a CDF acumulada seja maior ou igual a u .
3. O menor valor k tal que $F(k) \geq u$ será o número de eventos gerado pela distribuição Poisson.

23.1.3 Explicação:

1. **Fórmula Recursiva:** A relação recursiva $P(X = k + 1) = \frac{\lambda}{k+1} \cdot P(X = k)$ permite calcular as probabilidades de forma eficiente, atualizando a probabilidade de $P(X = k + 1)$ a partir de $P(X = k)$.
2. **Eficiência:** Esta técnica evita o cálculo repetido de fatoriais, tornando-a mais eficiente, especialmente quando se está gerando variáveis para grandes valores de λ ou quando se deseja calcular probabilidades para muitos eventos k .

Este método é amplamente utilizado por ser mais rápido e computacionalmente eficiente para a geração de variáveis aleatórias de Poisson.

24 R

```
# Função para gerar a inversa da CDF para a distribuição Poisson usando a técnica de inversão
inversa_cdf_poisson_recurativa <- function(lam, u) {
  k <- 0
  p <- exp(-lam) # P(X=0)
  F <- p # Iniciamos com a probabilidade P(X=0)

  # Continuamos somando até que F >= u
  while (u > F) {
    k <- k + 1
    p <- p * lam / k # Atualiza a probabilidade recursivamente para o próximo valor
    F <- F + p
  }

  return(k)
}

# Parâmetro lambda da distribuição Poisson
lam <- 3

# Gerando 1000 números uniformemente distribuídos
uniformes <- runif(1000)

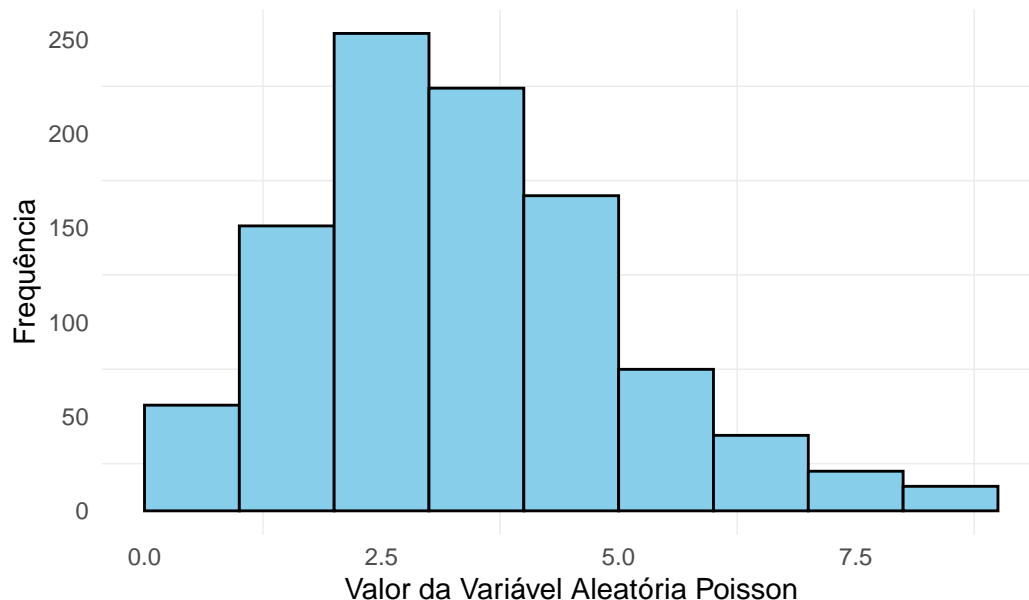
# Gerando a variável aleatória Poisson correspondente para cada número uniforme
poisson_vars <- sapply(uniformes, inversa_cdf_poisson_recurativa, lam = lam)

# Plotando o histograma das variáveis Poisson geradas
library(ggplot2)

df <- data.frame(poisson_vars = poisson_vars)
ggplot(df, aes(x = poisson_vars)) +
  geom_histogram(bins = max(poisson_vars) + 1, color = "black", fill = "skyblue", boundary =
  labs(title = "Histograma de Variáveis Aleatórias Poisson Usando a Fórmula Recursiva ( = 3)",
    x = "Valor da Variável Aleatória Poisson",
    y = "Frequência") +
```

```
theme_minimal() +
theme(panel.grid.major = element_blank())
```

Histograma de Variáveis Aleatórias Poisson Usando a Fórmula

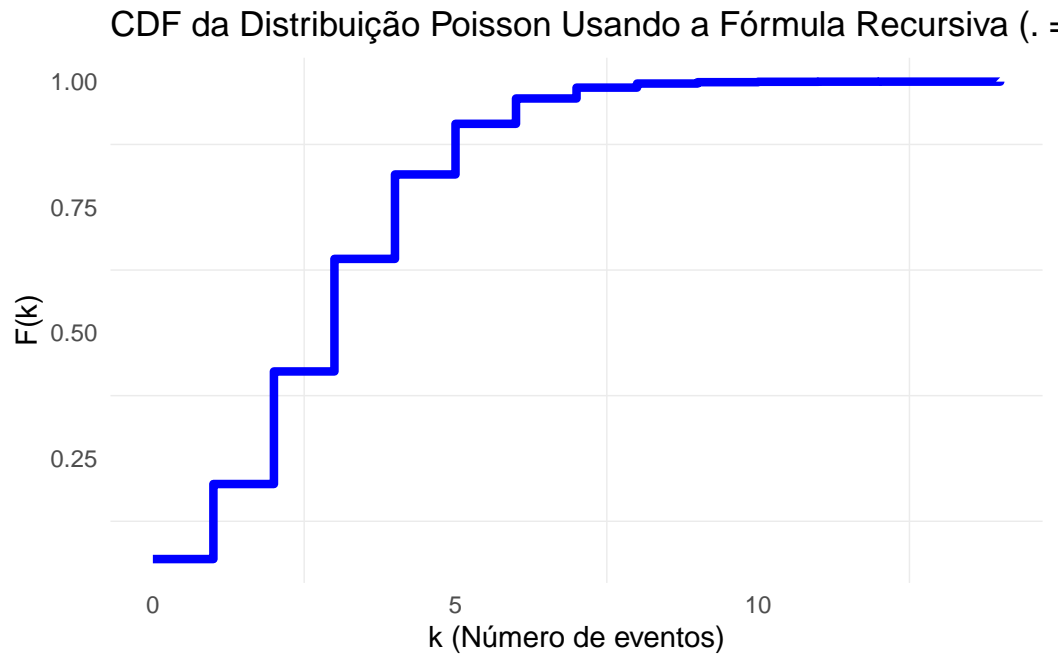


```
# Função para calcular a CDF da distribuição Poisson
cdf_poisson <- function(k, lam) {
  cdf <- 0
  p <- exp(-lam) # P(X=0)
  for (i in 0:k) {
    cdf <- cdf + p # Adiciona a probabilidade à CDF
    if (i < k) {
      p <- p * lam / (i + 1) # Atualiza a probabilidade recursivamente
    }
  }
  return(cdf)
}

# Gerando valores de k para a CDF
k_values <- 0:14
cdf_values <- sapply(k_values, cdf_poisson, lam = lam)

# Plotando a CDF da distribuição Poisson
df_cdf <- data.frame(k_values = k_values, cdf_values = cdf_values)
ggplot(df_cdf, aes(x = k_values, y = cdf_values)) +
```

```
geom_step(direction = "hv", color = "blue", size = 1.5) +
labs(title = "CDF da Distribuição Poisson Usando a Fórmula Recursiva (λ = 3)",
     x = "k (Número de eventos)",
     y = "F(k)") +
theme_minimal() +
theme(panel.grid.major = element_blank())
```



25 Python

```
import numpy as np
import matplotlib.pyplot as plt
import math

# Função para gerar a inversa da CDF para a distribuição Poisson usando a técnica de inversão
def inversa_cdf_poisson_recurativa(lam, u):
    k = 0
    p = math.exp(-lam) # P(X=0)
    F = p # Iniciamos com a probabilidade P(X=0)

    # Continuamos somando até que F >= u
    while u > F:
        k += 1
        p = p * lam / k # Atualiza a probabilidade recursivamente para o próximo valor
        F += p

    return k

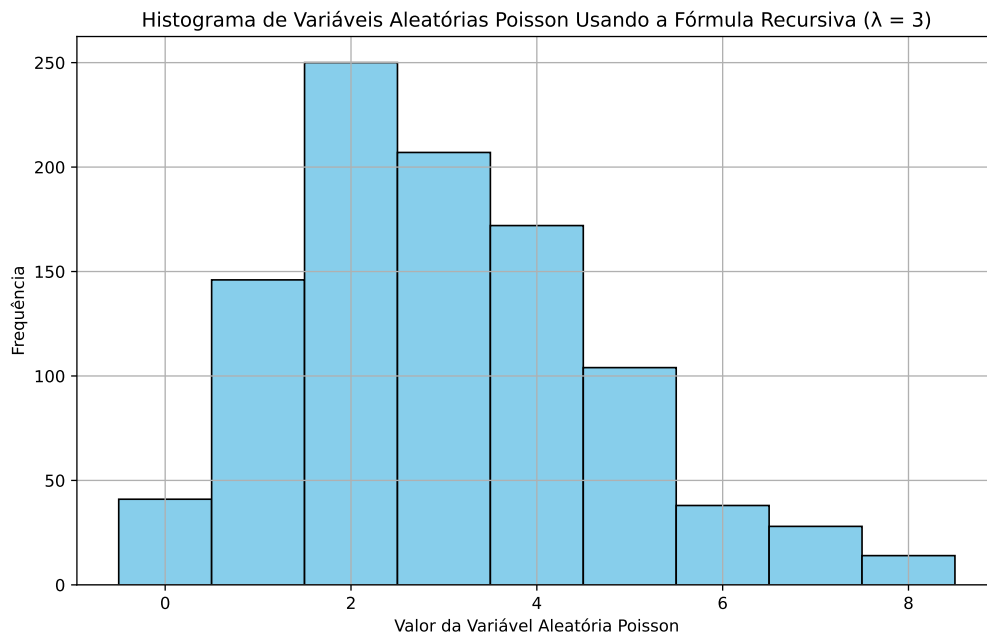
# Parâmetro lambda da distribuição Poisson
lam = 3

# Gerando 1000 números uniformemente distribuídos
uniformes = np.random.uniform(0, 1, 1000)

# Gerando a variável aleatória Poisson correspondente para cada número uniforme
poisson_vars = [inversa_cdf_poisson_recurativa(lam, u) for u in uniformes]

# Plotando o histograma das variáveis Poisson geradas
plt.figure(figsize=(10, 6))
plt.hist(poisson_vars, bins=range(0, max(poisson_vars) + 1), color='skyblue', edgecolor='black')
plt.title('Histograma de Variáveis Aleatórias Poisson Usando a Fórmula Recursiva (λ = 3)')
plt.xlabel('Valor da Variável Aleatória Poisson')
plt.ylabel('Frequência')
plt.grid(True)
```

```
plt.show()
```

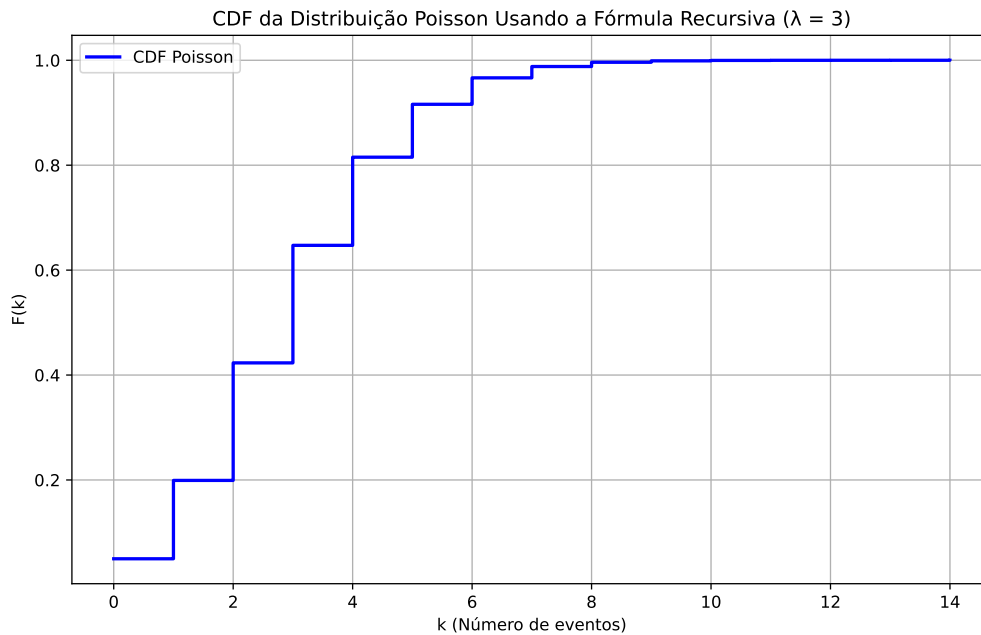


```
# Calculando a CDF da distribuição Poisson
def cdf_poisson(k, lam):
    cdf = 0
    p = math.exp(-lam) # P(X=0)
    for i in range(k+1):
        cdf += p # Adiciona a probabilidade à CDF
        if i < k: # Atualiza a probabilidade recursivamente
            p = p * lam / (i + 1)
    return cdf

# Gerando valores de k para a CDF
k_values = np.arange(0, 15)
cdf_values = [cdf_poisson(k, lam) for k in k_values]

# Plotando a CDF da distribuição Poisson
plt.figure(figsize=(10, 6))
plt.step(k_values, cdf_values, where='post', color='blue', label='CDF Poisson', linewidth=2)
plt.title('CDF da Distribuição Poisson Usando a Fórmula Recursiva (  $\lambda = 3$  )')
plt.xlabel('k (Número de eventos)')
```

```
plt.ylabel('F(k)')
plt.grid(True)
plt.legend()
plt.show()
```



25.1 Exercícios

Exercício 1. Seja X uma v.a. tal que $\mathbb{P}(X = 1) = 0.3$, $\mathbb{P}(X = 3) = 0.1$ e $\mathbb{P}(X = 4) = 0.6$.

- Escreva um pseudo-algoritmo para gerar um valor de X .
- Implemente uma função para gerar n valores de X .
- Compare a distribuição das frequências obtidas na amostra simulada com as probabilidades reais.

Exercício 2. Considere X uma v.a. tal que

$$\mathbb{P}(X = i) = \alpha \mathbb{P}(X_1 = i) + (1 - \alpha) \mathbb{P}(X_2 = i), \quad i = 0, 1, \dots$$

onde $0 \leq \alpha \leq 1$ e X_1, X_2 são v.a. discretas.

A distribuição de X é chamada de distribuição de mistura. Podemos escrever

$$X = \begin{cases} X_1, & \text{com probabilidade } \alpha \\ X_2, & \text{com probabilidade } 1 - \alpha \end{cases}$$

Pseudo-Algoritmo:

- (1) Seja $U \sim Unif(0, 1)$
- (2) Se $U \leq \alpha$, gere um valor da distribuição de X_1 . Senão, gere um valor da distribuição de X_2

Com base no pseudo-algoritmo implemente um algoritmo para gerar uma amostra de tamanho n da distribuição mistura de uma Poisson e de uma Geométrica com base nas funções implementadas nos Exercícios (2) e (3).

26 Técnica da Inversão para Variáveis Contínuas

Os valores que uma variável aleatória X pode assumir são chamados de **suporte** da distribuição de X .

Variáveis Aleatórias Contínuas são variáveis aleatórias que têm suporte em um conjunto não enumerável de valores, como intervalos na reta real, \mathbb{R} , ou $(0, \infty)$, por exemplo.

Uma variável aleatória X contínua tem função de distribuição acumulada (f.d.a.) puder ser expressa como

$$P(X \leq a) = F(a) = \int_{-\infty}^a f(x)dx, \quad \forall a \in \mathbb{R},$$

em que $f : \mathbb{R} \rightarrow [0, \infty)$ é uma função integrável, chamada de função densidade de probabilidade.

26.1 Função Inversa

Sabemos que $F : \mathbb{R} \rightarrow [0, 1]$ é estritamente crescente quando X é contínua, e, portanto, podemos definir sua função inversa $F^{-1} : [0, 1] \rightarrow \mathbb{R}$. A seguinte figura ilustra F e sua inversa.

27 R

```
# Carregando as bibliotecas necessárias
library(ggplot2)

# Definindo a função de distribuição acumulada F(x) - função logística
F <- function(x) {
  return(1 / (1 + exp(-x)))
}

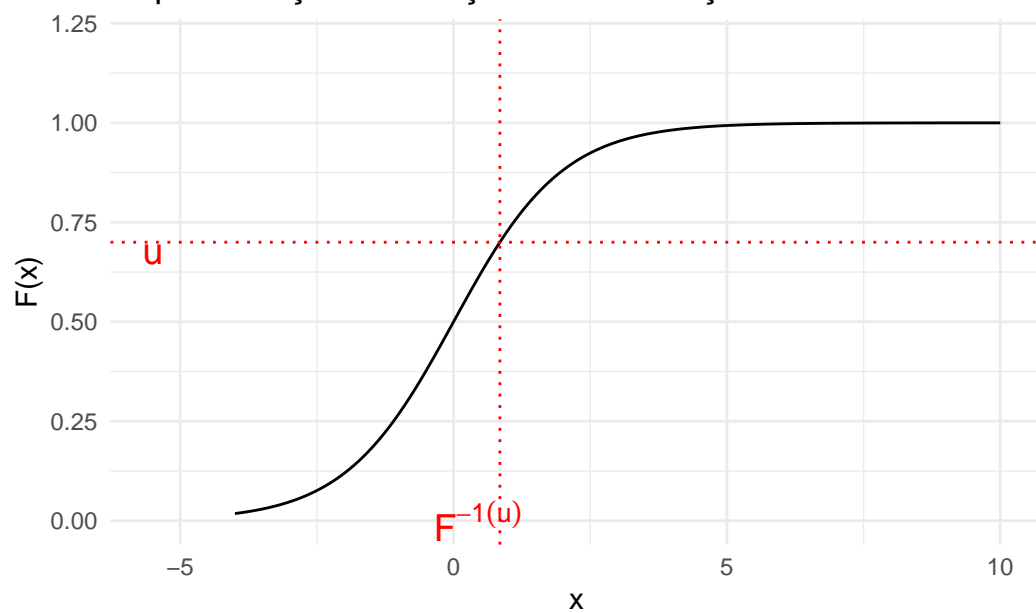
# Definindo a inversa da função de distribuição acumulada F_inv(u)
F_inv <- function(u) {
  return(-log((1 / u) - 1))
}

# Gerando valores de x e u
x <- seq(-4, 10, length.out = 400)
u <- seq(0.01, 0.99, length.out = 400)

# Definindo o valor de U para plotar as linhas
u_value <- 0.7
x_value <- F_inv(u_value)

# Criando o gráfico
ggplot(data = data.frame(x = x, F_x = F(x))) +
  geom_line(aes(x = x, y = F_x), color = "black") +
  geom_hline(yintercept = u_value, linetype = "dotted", color = "red") +
  geom_vline(xintercept = x_value, linetype = "dotted", color = "red") +
  annotate("text", x = x_value - 0.4, y = -0, label = expression(F^{-1}(u)), color = "red", size = 5) +
  annotate("text", x = -5.5, y = u_value - 0.02, label = "u", color = "red", size = 5) +
  labs(title = "Representação da Função de Distribuição Acumulada e sua Inversa",
       x = "x", y = "F(x)") +
  ylim(0, 1.2) +
  theme_minimal()
```

Representação da Função de Distribuição Acumulada e sua Inversa



28 Python

```
import numpy as np
import matplotlib.pyplot as plt

# Definindo a função de distribuição acumulada F(x)
def F(x):
    return 1 / (1 + np.exp(-x)) # Função logística como exemplo de F(x)

# Definindo a inversa da função de distribuição acumulada F_inv(u)
def F_inv(u):
    return -np.log((1 / u) - 1)

# Gerando valores de x e u para plotar
x = np.linspace(-4, 10, 400)
u = np.linspace(0.01, 0.99, 400) # U entre 0 e 1 (evitando extremos para evitar erros na in

# Plotando a função de distribuição acumulada F(x) com truncamento do eixo y no zero
plt.figure(figsize=(8, 6))
plt.plot(x, F(x), color="black")

# Adicionando linhas pontilhadas para representar U e F_inv(U)
u_value = 0.7 # Exemplo de valor de U
x_value = F_inv(u_value)

plt.hlines(u_value, min(x), x_value, linestyles='dotted', colors='red')
plt.vlines(x_value, 0, u_value, linestyles='dotted', colors='red')

# Etiquetas
plt.text(x_value-0.4, -0.05, r"$F^{-1}(u)$", fontsize=12, color='red')
plt.text(-5.5, u_value - 0.02, r"$u$", fontsize=14, color='red')

# Rótulos e estilo do gráfico
plt.title(r'Representação da Função de Distribuição Acumulada e sua Inversa', fontsize=14)
plt.xlabel(r'$x$', fontsize=12)
plt.ylabel(r'$F(x)$', fontsize=12)
```



```
plt.ylim(0, 1.2)
```

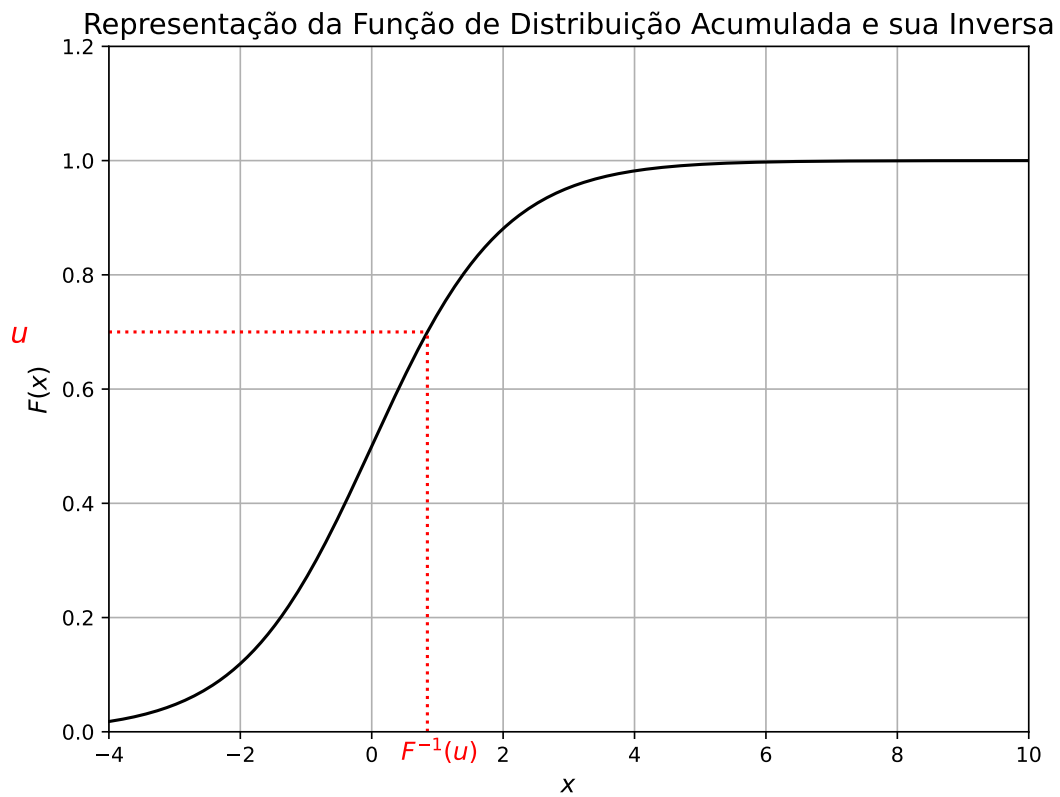
```
(0.0, 1.2)
```

```
plt.xlim(-4, 10)
```

```
(-4.0, 10.0)
```

```
plt.grid(True)
```

```
# Exibir o gráfico  
plt.show()
```



28.1 Método da Inversão

Uma maneira de gerar valores de uma variável aleatória contínua X , é o **método da inversão**, que é originado da seguinte proposição:

Proposição: Seja $U \sim \text{Unif}(0, 1)$. Para qualquer variável aleatória contínua com função de distribuição acumulada F , a variável:

$$X = F^{-1}(U)$$

tem distribuição F .

Prova:

$$\mathbb{P}(X \leq x) = \mathbb{P}(F^{-1}(U) \leq x) = \mathbb{P}(F(F^{-1}(U)) \leq F(x)) = \mathbb{P}(U \leq F(x)) = F(x).$$

Assim, o método da inversão consiste em:

1. Gerar $U \sim \text{Unif}(0, 1)$.
2. Calcular $X = F^{-1}(U)$.

28.2 Exemplo 1

Seja X uma v.a. com:

$$F(x) = x^n, \quad \text{para } 0 < x < 1.$$

A função inversa é:

$$u = F(x) = x^n \implies x = u^{1/n}.$$

Portanto, o pseudo-algoritmo para gerar X a partir do método da inversão é:

1. Gere $U \sim \text{Unif}(0, 1)$.
2. Calcule $X = U^{1/n}$.

29 R

```
# Carregar a biblioteca ggplot2
library(ggplot2)

# Definir o parâmetro n da distribuição  $F(x) = x^n$ 
n <- 3

# Gerar 1000 valores U de uma distribuição uniforme (0,1)
U <- runif(1000, min = 0, max = 1)

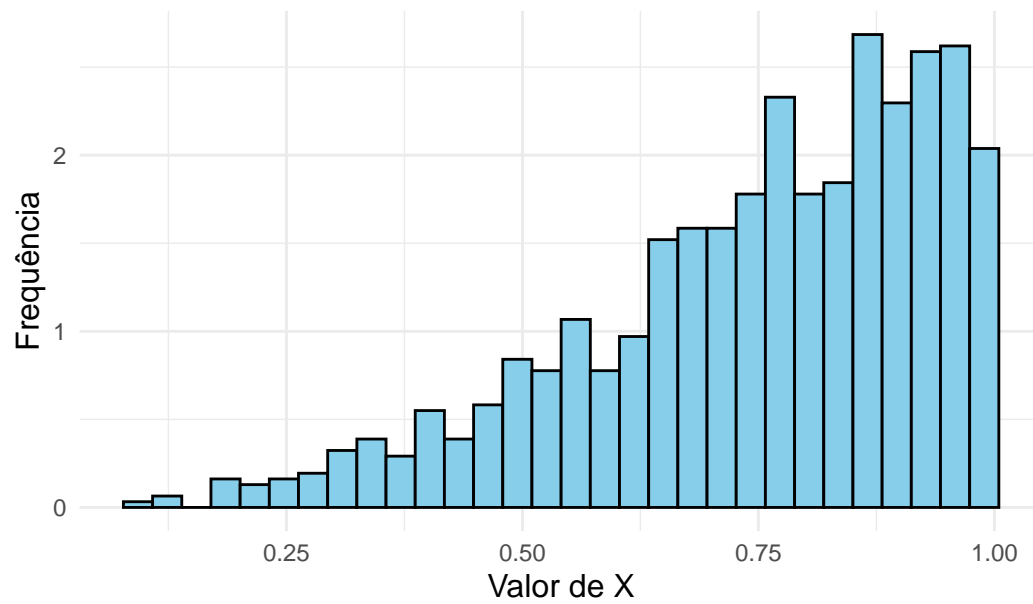
# Calcular  $X = U^{(1/n)}$ 
X <- U^(1/n)

# Criar um dataframe para o ggplot2
data <- data.frame(X = X)

# Plotar o histograma usando ggplot2
p <- ggplot(data, aes(x = X)) +
  geom_histogram(aes(y = ..density..), bins = 30, fill = 'skyblue', color = 'black') +
  labs(title = 'Histograma de variáveis geradas pela inversão:  $F(x) = x^n$ ,  $n = 3$ ',
       x = 'Valor de X', y = 'Frequência') +
  theme_minimal() +
  theme(plot.title = element_text(size = 14),
        axis.title.x = element_text(size = 12),
        axis.title.y = element_text(size = 12))

# Exibir o gráfico
print(p)
```

Histograma de variáveis geradas pela inversão: $F(x) = x^n$,



30 Python

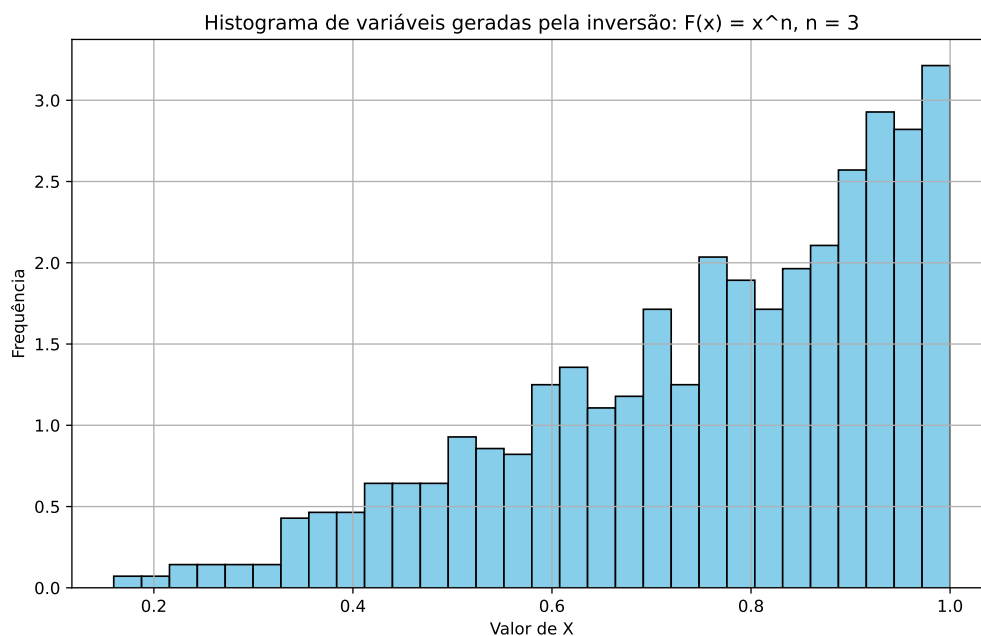
```
import numpy as np
import matplotlib.pyplot as plt

# Parâmetro n da distribuição  $F(x) = x^n$ 
n = 3

# Gerando 1000 valores U de uma distribuição uniforme (0,1)
U = np.random.uniform(0, 1, 1000)

# Calculando  $X = U^{1/n}$ 
X = U**(1/n)

# Plotando um histograma dos valores gerados
plt.figure(figsize=(10, 6))
plt.hist(X, bins=30, color='skyblue', edgecolor='black', density=True)
plt.title('Histograma de variáveis geradas pela inversão:  $F(x) = x^n$ ,  $n = 3$ ')
plt.xlabel('Valor de X')
plt.ylabel('Frequência')
plt.grid(True)
plt.show()
```



30.0.1 Exemplo 2

Seja $X \sim \text{Exp}(\lambda)$, com:

$$F(x) = 1 - e^{-\lambda x}, \quad \text{para } x > 0.$$

A função inversa é:

$$u = F(x) = 1 - e^{-\lambda x} \implies x = -\frac{\log(1-u)}{\lambda}.$$

Um pseudo-algoritmo para gerar X é, portanto,:

1. Gere $U \sim \text{Unif}(0, 1)$.
2. Calcule $X = -\frac{\log(1-U)}{\lambda}$.

31 R

```
# Carregar a biblioteca ggplot2
library(ggplot2)

# Definir o parâmetro lambda da distribuição exponencial
lambda <- 2

# Gerar 1000 valores U de uma distribuição uniforme (0,1)
U <- runif(1000, min = 0, max = 1)

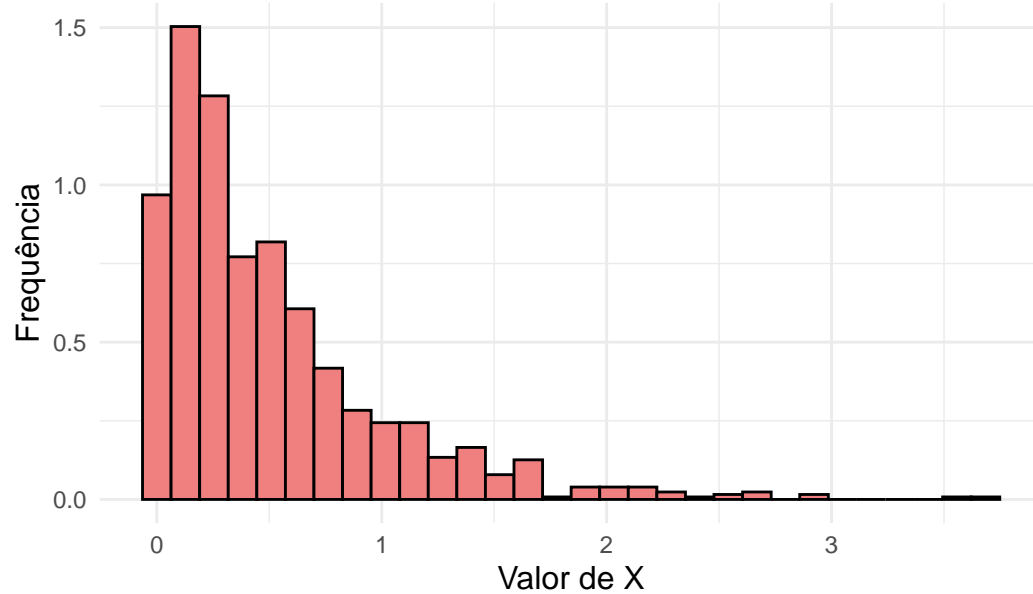
# Calcular X usando a inversa da CDF da distribuição exponencial
X <- -log(1 - U) / lambda

# Criar um dataframe para o ggplot2
data <- data.frame(X = X)

# Plotar o histograma usando ggplot2
p <- ggplot(data, aes(x = X)) +
  geom_histogram(aes(y = ..density..), bins = 30, fill = 'lightcoral', color = 'black') +
  labs(title = 'Histograma de variáveis geradas pela inversão: Distribuição Exponencial',
       x = 'Valor de X', y = 'Frequência') +
  theme_minimal() +
  theme(plot.title = element_text(size = 14),
        axis.title.x = element_text(size = 12),
        axis.title.y = element_text(size = 12))

# Exibir o gráfico
print(p)
```

Histograma de variáveis geradas pela inversão: Distribuição



32 Python

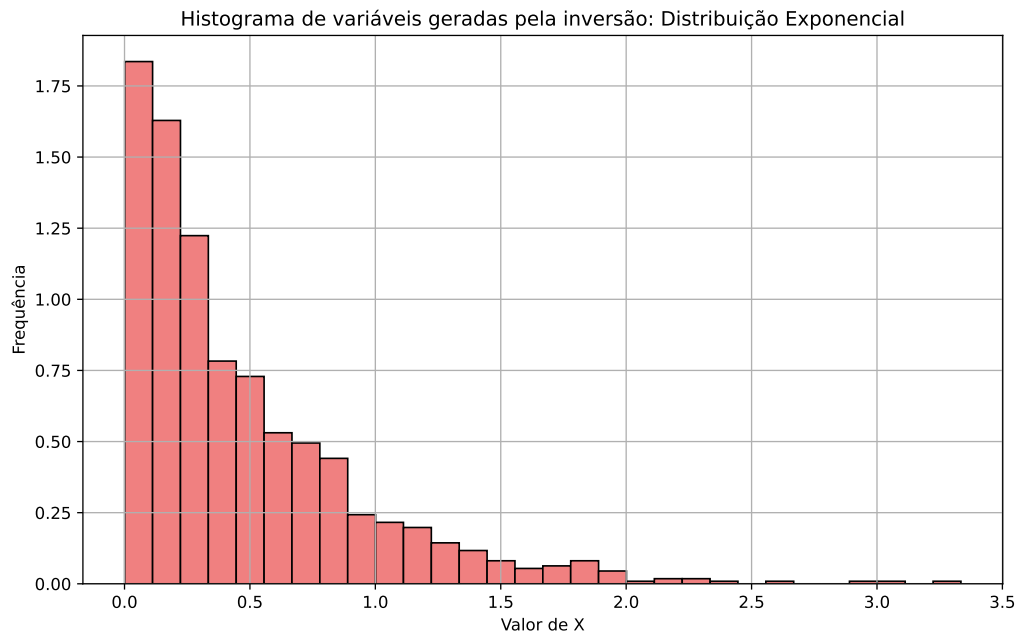
```
import numpy as np
import matplotlib.pyplot as plt

# Parâmetro lambda da distribuição exponencial
lamdb = 2

# Gerando 1000 valores U de uma distribuição uniforme (0,1)
U = np.random.uniform(0, 1, 1000)

# Calculando X usando a inversa da CDF da exponencial
X = -np.log(1 - U) / lamdb

# Plotando um histograma dos valores gerados
plt.figure(figsize=(10, 6))
plt.hist(X, bins=30, color='lightcoral', edgecolor='black', density=True)
plt.title('Histograma de variáveis geradas pela inversão: Distribuição Exponencial')
plt.xlabel('Valor de X')
plt.ylabel('Frequência')
plt.grid(True)
plt.show()
```



32.1 Simulação de transformações de variáveis aleatórias

Agora que já sabemos uma maneira de simular uma variável aleatória X , descreveremos como gerar valores de uma transformação dessa variável, ou seja, $g(X)$. Para isso, basta aplicar a função de transformação g diretamente aos valores simulados de X . Veremos a seguinte alguns exemplos disso em funcionamento.

32.1.1 Exemplo 1: Simulando $Y \sim Unif(1, 2)$

Para gerar valores de $Y \sim Unif(1, 2)$, usamos o fato de que Y é uma simples transformação de $U \sim Unif(0, 1)$. A relação é:

$$Y = U + 1.$$

Assim, podemos usar o seguinte pseudo-algoritmo para gerar Y a partir de U :

1. Gere $U \sim Unif(0, 1)$.
2. Calcule $Y = U + 1$.

33 R

```
# Carregar biblioteca ggplot2
library(ggplot2)

# Gerar 1000 valores U de uma distribuição uniforme (0,1)
U <- runif(1000, min = 0, max = 1)

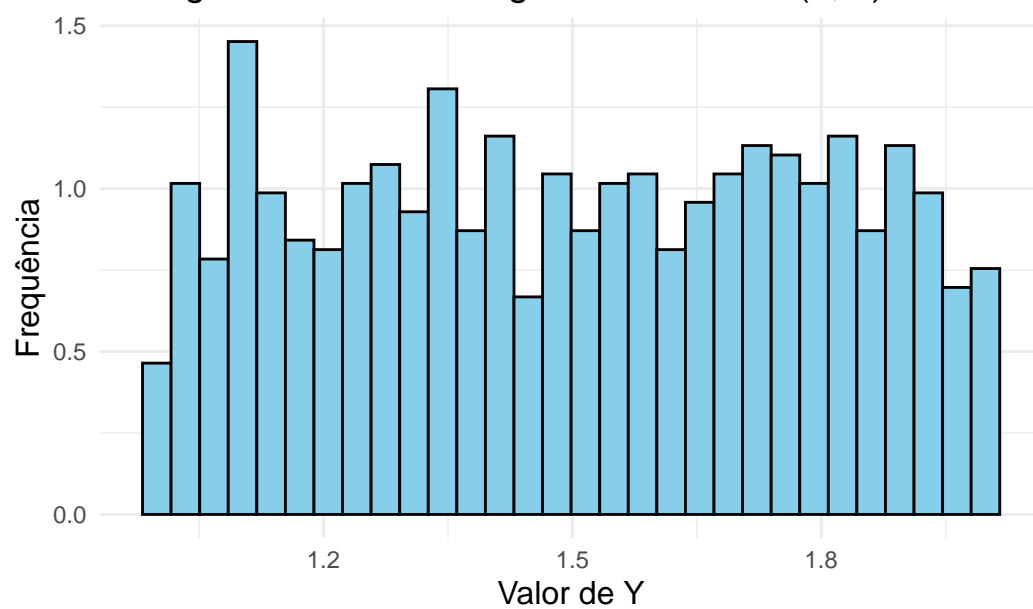
# Calcular Y = U + 1 para ter Y ~ Unif(1, 2)
Y <- U + 1

# Criar um dataframe para o ggplot2
data <- data.frame(Y = Y)

# Plotar o histograma usando ggplot2
p <- ggplot(data, aes(x = Y)) +
  geom_histogram(aes(y = ..density..), bins = 30, fill = 'skyblue', color = 'black') +
  labs(title = 'Histograma de variáveis geradas: Y ~ Unif(1, 2)',
       x = 'Valor de Y', y = 'Frequência') +
  theme_minimal() +
  theme(plot.title = element_text(size = 14),
        axis.title.x = element_text(size = 12),
        axis.title.y = element_text(size = 12))

# Exibir o gráfico
print(p)
```

Histograma de variáveis geradas: $Y \sim \text{Unif}(1, 2)$



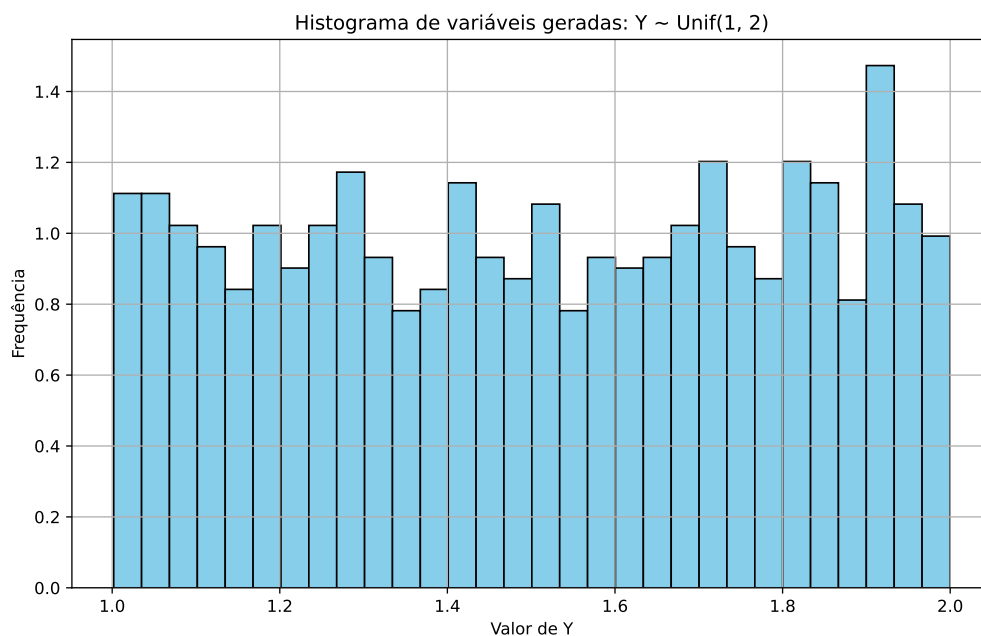
34 Python

```
import numpy as np
import matplotlib.pyplot as plt

# Gerando 1000 valores U de uma distribuição uniforme (0,1)
U = np.random.uniform(0, 1, 1000)

# Calculando Y = U + 1 para ter Y ~ Unif(1, 2)
Y = U + 1

# Plotando um histograma dos valores gerados
plt.figure(figsize=(10, 6))
plt.hist(Y, bins=30, color='skyblue', edgecolor='black', density=True)
plt.title('Histograma de variáveis geradas: Y ~ Unif(1, 2)')
plt.xlabel('Valor de Y')
plt.ylabel('Frequência')
plt.grid(True)
plt.show()
```



34.0.1 Exemplo 2: Simulando $Y \sim \text{Gamma}(n, \lambda)$

Para gerar valores de $Y \sim \Gamma(n, \lambda)$, usamos o fato de que a soma de que Y pode ser representado como

$$Y = \sum_{i=1}^n X_i,$$

em que cada $X_i \sim \text{Exp}(\lambda)$, e X_i 's são independentes.

Assim, podemos gerar Y da seguinte forma:

1. Gere $U_1, \dots, U_n \sim \text{Unif}(0, 1)$ independentemente.
2. Calcule $X_i = -\frac{\log(1-U_i)}{\lambda}$ para $i = 1, \dots, n$.
3. Calcule $Y = X_1 + X_2 + \dots + X_n$.

35 R

```
# Carregar biblioteca ggplot2
library(ggplot2)

# Definir parâmetros
n <- 5 # número de somas
lambda <- 2 # parâmetro da distribuição exponencial

# Gerar 1000 valores U para cada uma das n somas
U <- matrix(runif(1000 * n, min = 0, max = 1), ncol = n)

# Calcular  $X_i = -\log(1 - U_i) / \lambda$  para cada  $U_i$ 
X <- -log(1 - U) / lambda

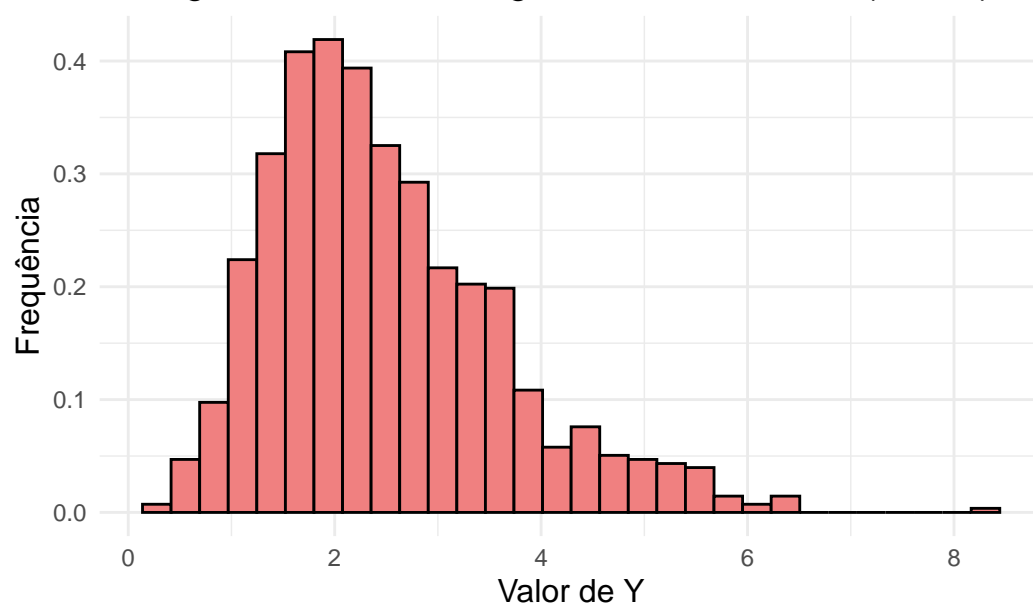
# Somar os valores de X para obter  $Y \sim \text{Gamma}(n, \lambda)$ 
Y <- rowSums(X)

# Criar um dataframe para o ggplot2
data <- data.frame(Y = Y)

# Plotar o histograma usando ggplot2
p <- ggplot(data, aes(x = Y)) +
  geom_histogram(aes(y = ..density..), bins = 30, fill = 'lightcoral', color = 'black') +
  labs(title = paste('Histograma de variáveis geradas: Y ~ Gamma(', n, ', ', lambda, ')'),
       x = 'Valor de Y', y = 'Frequência') +
  theme_minimal() +
  theme(plot.title = element_text(size = 14),
        axis.title.x = element_text(size = 12),
        axis.title.y = element_text(size = 12))

# Exibir o gráfico
print(p)
```

Histograma de variáveis geradas: $Y \sim \text{Gamma}(5, 2)$



36 Python

```
import numpy as np
import matplotlib.pyplot as plt

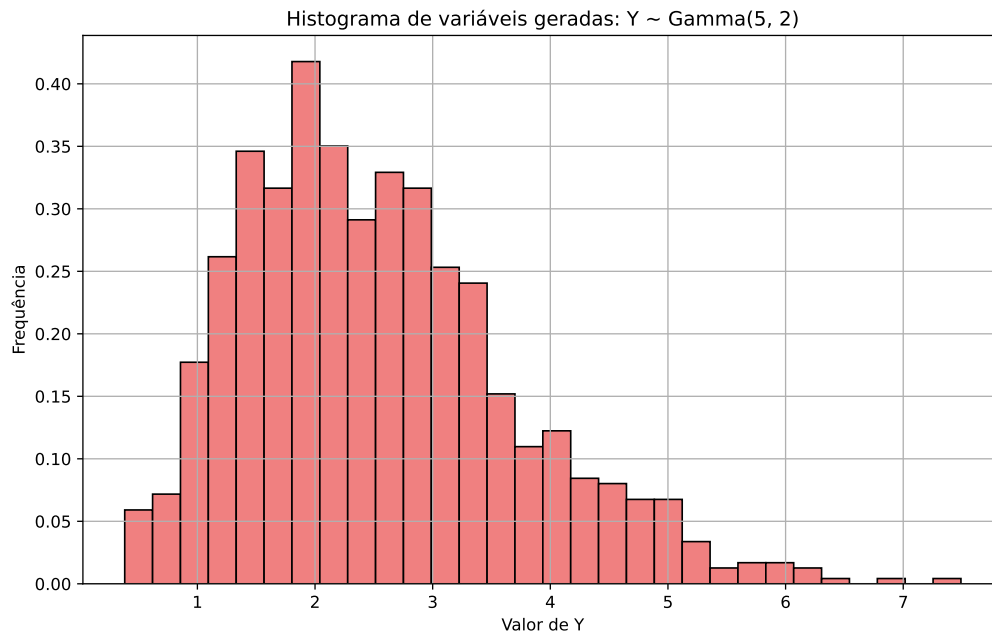
# Definindo parâmetros
n = 5 # número de somas
lamdb = 2 # parâmetro da distribuição exponencial

# Gerando 1000 valores U para cada uma das n somas
U = np.random.uniform(0, 1, (1000, n))

# Calculando  $X_i = -\log(1 - U_i) / \lambda$  para cada  $U_i$ 
X = -np.log(1 - U) / lamdb

# Somando os valores de X para obter  $Y \sim \text{Gamma}(n, \lambda)$ 
Y = np.sum(X, axis=1)

# Plotando um histograma dos valores gerados
plt.figure(figsize=(10, 6))
plt.hist(Y, bins=30, color='lightcoral', edgecolor='black', density=True)
plt.title(f'Histograma de variáveis geradas:  $Y \sim \text{Gamma}(\{n\}, \{\lambda\})$ ')
plt.xlabel('Valor de Y')
plt.ylabel('Frequência')
plt.grid(True)
plt.show()
```



36.1 Exercícios

Exercício 1.

Utilizando o método da inversão, simule $U \sim U(1, 3)$.

Exercício 2.

- Implemente uma função para gerar uma amostra de tamanho n da distribuição Exponencial de parâmetro λ .
- Compare a distribuição empírica dos valores simulados com a densidade da Exponencial $f(x) = \lambda e^{-\lambda x}, x > 0$.

Exercício 3.

- Implemente uma função para gerar uma amostra de tamanho n da distribuição $Gama(a, b)$, para a sendo um valor inteiro.
- Compare a distribuição empírica dos valores simulados com a densidade da Gama $f(x) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx}, x > 0$.

Exercício 4. Seja X uma v.a. com função densidade dada por:

$$f(x) = \frac{1}{8}x, \quad 0 < x < 4.$$

- a) Escreva um pseudo-algoritmo para simular um único valor da variável X pelo método da inversão.
- b) Compare a distribuição empírica dos valores simulados com a densidade de X .

37 Método da Rejeição para Variáveis Discretas

Fill in

38 Método da Rejeição para Variáveis Contínuas

A amostragem por rejeição é uma técnica útil para gerar variáveis aleatórias a partir de uma distribuição complexa, utilizando uma distribuição proposta mais simples. A ideia básica é simular de uma distribuição fácil de amostrar e, em seguida, rejeitar ou aceitar essas amostras com base na distribuição alvo.

38.1 Algoritmo

O método assume que conhecemos uma distribuição $g(x)$, fácil de simular, que cubra o suporte da distribuição alvo $f(x)$. Essa distribuição é chamada de **distribuição proposta**. Além disso ele assume que existe c tal que $\frac{f(x)}{g(x)} \leq c$ para todo x , e que conseguimos calcular c . O método de aceitação e rejeição segue os seguintes passos:

1. Gere um valor Y da distribuição proposta $g(x)$.
2. Gere um valor $U \sim \text{Uniform}(0, 1)$.
3. Aceite Y como uma amostra de $f(x)$ se $U \leq \frac{f(Y)}{c \cdot g(Y)}$, caso contrário, rejeite Y e repita o processo.

A figura a seguir ilustra esse processo quando g é uma distribuição uniforme entre 0 e 1.

39 R

```
library(ggplot2)

# Função densidade alvo f(x)
f <- function(x) {
  20 * x * (1 - x)^3
}

# Função densidade proposta g(x) (distribuição uniforme)
g <- function(x) {
  rep(1, length(x)) # g(x) é uniforme no intervalo (0, 1)
}

# Constante c (máximo de f(x))
opt_result <- optimize(f = function(x) -f(x), interval = c(0, 1))
c <- f(opt_result$minimum)

# Gerar valores de x
x <- seq(0, 1, length.out = 1000)

# Valor de Y para aceitar/rejeitar (escolhido aleatoriamente)
Y <- 0.6

# DataFrame para o gráfico
df <- data.frame(x = x, fx = f(x), gx = c * g(x))

# Plotando as funções f(x) e a linha horizontal c*g(x)
ggplot(df, aes(x = x)) +
  geom_line(aes(y = fx, color = "f(x)"), size = 1) +
  geom_hline(aes(yintercept = c * g(x), color = "c * g(x)"), size = 1.5) +

  # Destacar a área de aceitação e rejeição
  geom_text(aes(x = Y + 0.025, y = 0.05, label = "Y"), color = "black", size = 5) +

  geom_segment(aes(x = Y, xend = Y, y = 0, yend = f(Y)), color = "green", size = 1.5) +
```

```

geom_segment(aes(x = Y, xend = Y, y = f(Y), yend = c * g(Y)), color = "red", size = 1.5) +

# Mover rótulos de Aceitar/Rejeitar para a esquerda da linha Y
geom_text(aes(x = Y + 0.06, y = f(Y) / 2, label = "Aceitar x"), color = "green", size = 4)
geom_text(aes(x = Y + 0.06, y = f(Y) + (c * g(Y) - f(Y)) / 2, label = "Rejeitar x"), color = "red", size = 4)

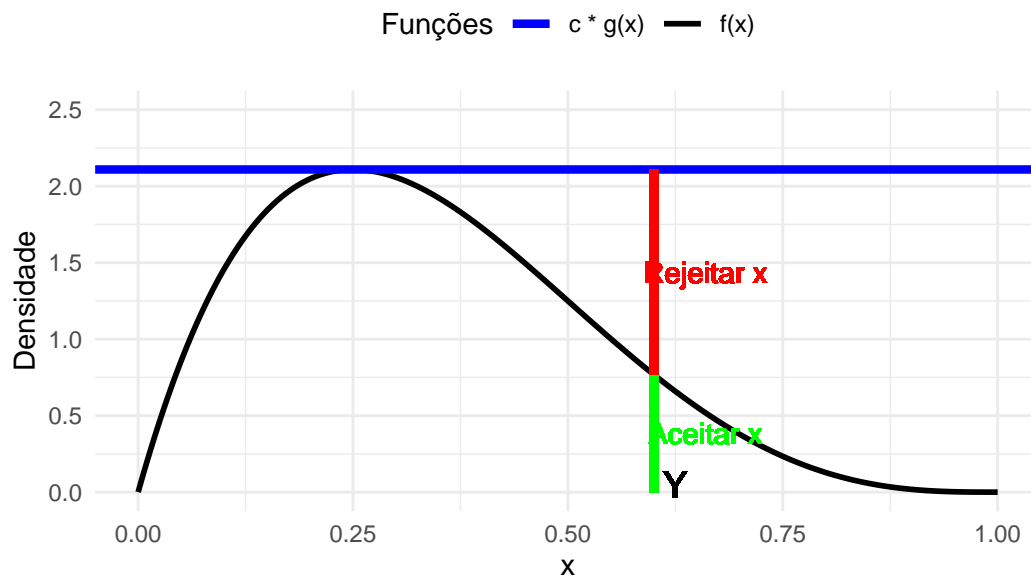
# Limites e rótulos
xlim(0, 1) +
ylim(0, 2.5) +
labs(x = "x", y = "Densidade", title = "Amostragem por Rejeição para g uniforme") +

# Legenda
scale_color_manual(values = c("f(x)" = "black", "c * g(x)" = "blue"), name = "Funções") +

theme_minimal() +
theme(legend.position = "top")

```

Amostragem por Rejeição para g uniforme



40 Python

```
import numpy as np
import matplotlib.pyplot as plt
from scipy.optimize import minimize_scalar

# Função densidade alvo f(x)
def f(x):
    return 20 * x * (1 - x)**3

# Função densidade proposta g(x) (distribuição uniforme)
def g(x):
    return 1 # g(x) é uniforme no intervalo (0, 1)

# Constante c
result = minimize_scalar(lambda x: -f(x), bounds=(0, 1), method='bounded')
c = f(result.x)

# Gerar valores de x
x = np.linspace(0, 1, 1000)

# Valor de Y para aceitar/rejeitar (escolhido aleatoriamente)
Y = 0.6
U = 0.8
accept_threshold = f(Y) / (c * g(Y))

# Plotando as funções f(x) e a linha horizontal c*g(x)
# Ajustando o código para truncar o eixo y no zero
plt.plot(x, f(x), label='f(x)', color='black')
plt.hlines(c * g(x), 0, 1, color='blue', label='c*g(x)', linewidth=2) # Linha horizontal c*g(x)

# Destacar a área de aceitação e rejeição
plt.vlines(Y, 0, f(Y), colors='green', label='Aceitar $x$', linewidth=2)
plt.vlines(Y, f(Y), c * g(Y), colors='red', label='Rejeitar $x$', linewidth=2)

# Adicionar rótulo de Y na posição da linha vertical
```



```
plt.text(Y+0.025, 0.05, 'Y', color='black', fontsize=12, horizontalalignment='center')

# Adicionar rótulos de Aceitar/Rejeitar
plt.text(Y + 0.01, f(Y) / 2, 'Aceitar x', color='green', fontsize=10, verticalalignment='center')
plt.text(Y + 0.01, f(Y) + (c * g(Y) - f(Y)) / 2, 'Rejeitar x', color='red', fontsize=10, verticalalignment='center')

# Limites e rótulos
plt.xlim(0, 1)
```

(0.0, 1.0)

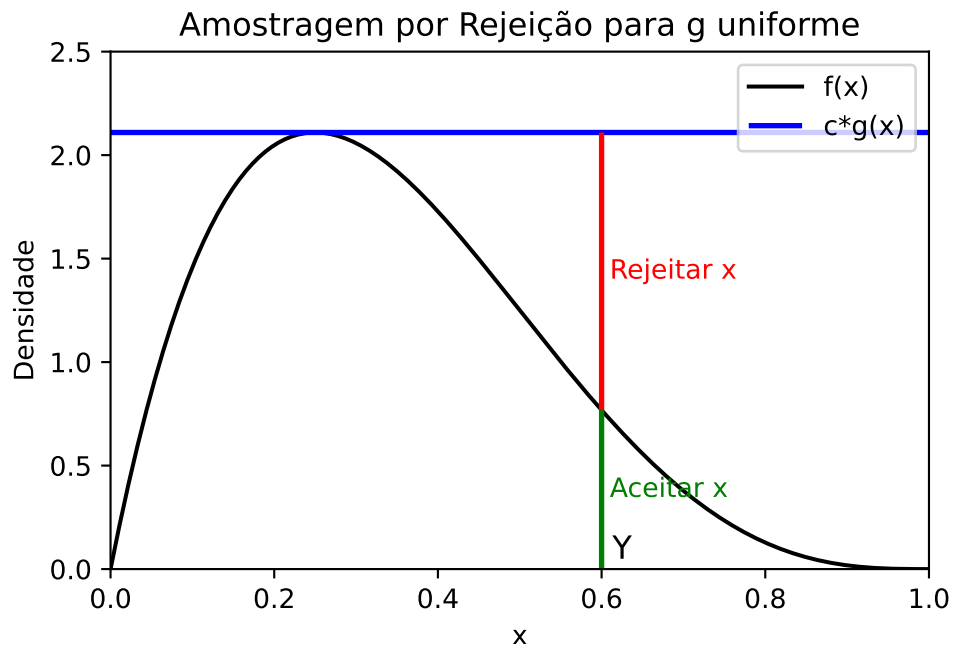
```
plt.ylim(0, 2.5) # Truncando o eixo y no zero
```

(0.0, 2.5)

```
plt.xlabel('x')
plt.ylabel('Densidade')

# Adicionar a legenda para f(x) e c*g(x)
plt.legend(['f(x)', 'c*g(x)'])

# Mostrar gráfico
plt.title('Amostragem por Rejeição para g uniforme')
plt.show()
```



40.1 Exemplo

Vamos demonstrar a amostragem por rejeição gerando amostras de uma distribuição alvo $f(x)$, utilizando a distribuição uniforme como distribuição proposta.

A função densidade alvo que utilizaremos é:

$$f(x) = 20x(1-x)^3 \quad \text{para } 0 < x < 1$$

E a distribuição proposta será $g(x) = 1$ para $0 < x < 1$, que é uma distribuição uniforme.

Implementamos o seguinte algoritmo:

41 R

```
library(ggplot2)

# Função densidade alvo f(x)
f <- function(x) {
  20 * x * (1 - x)^3
}

# Amostragem por rejeição
amostragem_rejeicao <- function(f, c, n_amstras) {
  amostras <- numeric(0)
  while (length(amostras) < n_amstras) {
    # Gere Y ~ g(x) (uniforme entre 0 e 1)
    Y <- runif(1, 0, 1)

    # Gere U ~ Uniforme(0, 1)
    U <- runif(1, 0, 1)

    # Verifica se aceitamos Y
    if (U <= f(Y) / (c)) {
      amostras <- c(amostras, Y)
    }
  }
  return(amostras)
}

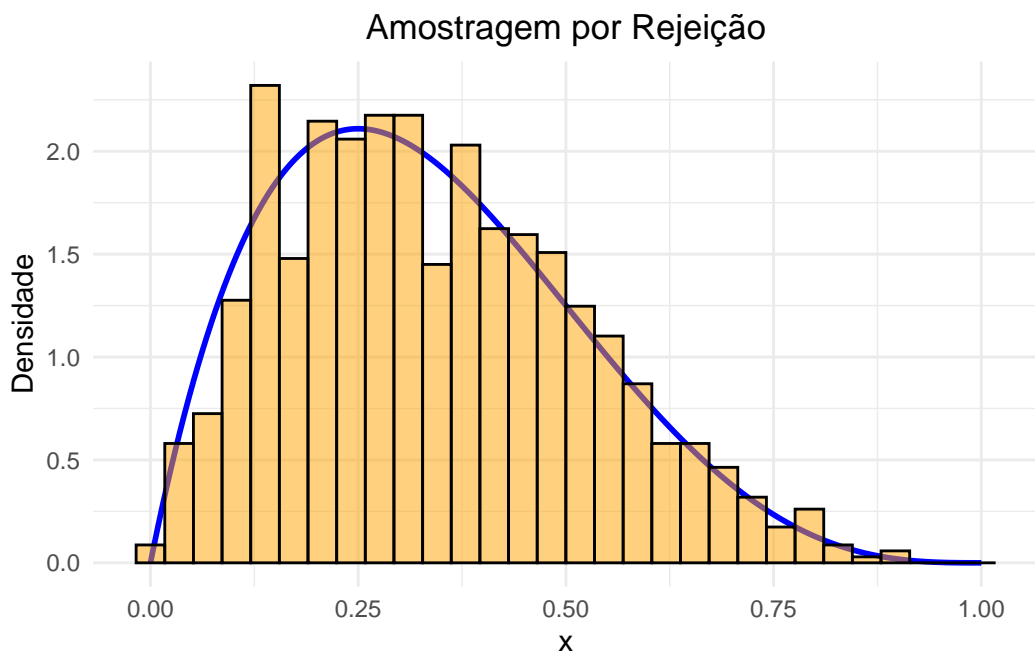
# A constante c é o limite superior de f(x)
c <- 135 / 64

# Gere 1000 amostras usando o método de aceitação e rejeição
set.seed(123)
amostras <- amostragem_rejeicao(f, c, 1000)

# Criação do data frame para o ggplot2
x_vals <- seq(0, 1, length.out = 1000)
```

```
f_vals <- f(x_vals)
df <- data.frame(x = x_vals, y = f_vals)

# Gráfico com ggplot2
ggplot() +
  geom_line(data = df, aes(x = x, y = y), color = 'blue', size = 1, linetype = 'solid') +
  geom_histogram(aes(x = amostras, y = ..density..), bins = 30, fill = 'orange', alpha = 0.5) +
  ggtitle("Amostragem por Rejeição") +
  xlab("x") +
  ylab("Densidade") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))
```



42 Python

```
import numpy as np
import matplotlib.pyplot as plt

# Função densidade alvo f(x)
def f(x):
    return 20 * x * (1 - x)**3

# Função densidade proposta g(x) (distribuição uniforme)
def g(x):
    return 1 # g(x) é uniforme no intervalo (0, 1)

# Amostragem por rejeição
def amostragem_rejeicao(f, g, c, n_amstras):
    amostras = []
    while len(amostras) < n_amstras:
        # Gere Y ~ g(x)
        Y = np.random.uniform(0, 1)

        # Gere U ~ Uniforme(0, 1)
        U = np.random.uniform(0, 1)

        # Verifica se aceitamos Y
        if U <= f(Y) / (c * g(Y)):
            amostras.append(Y)

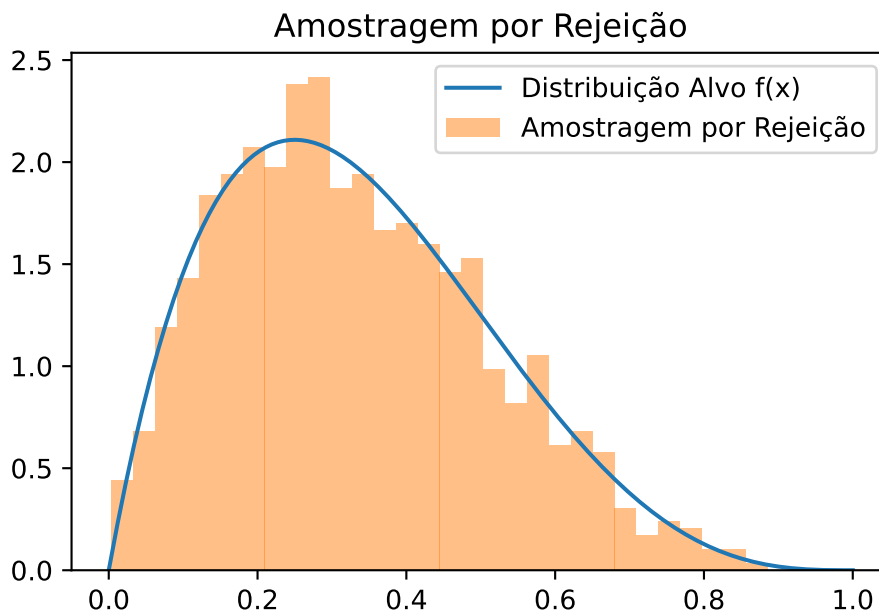
    return np.array(amostras)

# A constante c é o limite superior de f(x)/g(x)
c = 135 / 64 # Pré-calculado

# Gere 1000 amostras usando o método de aceitação e rejeição
amostras = amostragem_rejeicao(f, g, c, 1000)

# Plotando os resultados
```

```
x = np.linspace(0, 1, 1000)
plt.plot(x, f(x), label='Distribuição Alvo f(x)')
plt.hist(amostras, bins=30, density=True, alpha=0.5, label='Amostragem por Rejeição')
plt.legend()
plt.title("Amostragem por Rejeição")
plt.show()
```



42.1 Resultados Teóricos

42.1.1 Correção do Método de Aceitação e Rejeição

Teorema:

A variável aleatória Y gerada pelo método da rejeição tem densidade $f(x)$.

Prova:

Considere a função de distribuição acumulada (CDF) de X , a variável gerada pelo método:

$$P(X \leq x) = P(Y \leq x \mid \text{aceitar } Y)$$

Podemos escrever essa probabilidade como:

$$P(X \leq x) = \frac{P(Y \leq x, U \leq \frac{f(Y)}{c \cdot g(Y)})}{P(U \leq \frac{f(Y)}{c \cdot g(Y)})}$$

Usando a probabilidade condicional, o numerador pode ser expresso como:

$$P(Y \leq x, U \leq \frac{f(Y)}{c \cdot g(Y)}) = \int_{-\infty}^x \int_0^{\frac{f(y)}{c \cdot g(y)}} g(y) du dy$$

Resolvendo a integral em u :

$$P(Y \leq x, U \leq \frac{f(Y)}{c \cdot g(Y)}) = \int_{-\infty}^x \frac{f(y)}{c \cdot g(y)} g(y) dy = \frac{1}{c} \int_{-\infty}^x f(y) dy$$

De maneira semelhante, o denominador é dado por:

$$P\left(U \leq \frac{f(Y)}{c \cdot g(Y)}\right) = \int_{-\infty}^{\infty} \frac{f(y)}{c \cdot g(y)} g(y) dy = \frac{1}{c} \int_{-\infty}^{\infty} f(y) dy$$

Como $\int_{-\infty}^{\infty} f(y) dy = 1$ (pois $f(x)$ é uma função densidade), o denominador resulta em $\frac{1}{c}$. Substituindo essas expressões na equação original, obtemos:

$$P(X \leq x) = \frac{\frac{1}{c} \int_{-\infty}^x f(y) dy}{\frac{1}{c}} = \int_{-\infty}^x f(y) dy$$

Portanto, a variável X gerada pelo método de aceitação e rejeição tem função de distribuição acumulada $\int_{-\infty}^x f(y) dy$, o que implica que X tem densidade $f(x)$. Assim, o método gera amostras corretamente distribuídas de acordo com $f(x)$, como desejado.

42.1.2 Eficiência computacional

Teorema:

A variável aleatória gerada pelo método de aceitação e rejeição tem função de densidade $f(x)$. O número de passos que o algoritmo necessita para gerar X tem distribuição geométrica com média c , onde c é a constante de normalização que define o limite superior da razão $\frac{f(x)}{g(x)}$.

Prova:

Seja Y a variável aleatória gerada pela distribuição proposta $g(x)$, e $U \sim \text{Unif}(0, 1)$ uma variável aleatória uniforme. O método de aceitação e rejeição aceita Y como amostra de $f(x)$ se $U \leq \frac{f(Y)}{cg(Y)}$, caso contrário, o valor é rejeitado e o processo é repetido.

A probabilidade de aceitar uma amostra Y , dado que $Y \leq x$, é dada por:

$$P(Y \leq x \text{ e aceitar}) = P\left(Y \leq x, U \leq \frac{f(Y)}{cg(Y)}\right)$$

Podemos reescrever essa probabilidade como uma integral:

$$P(Y \leq x \text{ e aceitar}) = \int_{-\infty}^x \int_0^{\frac{f(y)}{cg(y)}} g(y) du dy$$

Resolvendo a integral em u , temos:

$$P(Y \leq x \text{ e aceitar}) = \int_{-\infty}^x \frac{f(y)}{cg(y)} g(y) dy = \frac{1}{c} \int_{-\infty}^x f(y) dy$$

A probabilidade de aceitar qualquer valor Y é dada por:

$$P(\text{aceitar}) = P\left(U \leq \frac{f(Y)}{cg(Y)}\right) = \int_{-\infty}^{\infty} \int_0^{\frac{f(y)}{cg(y)}} g(y) du dy$$

Simplificando:

$$P(\text{aceitar}) = \int_{-\infty}^{\infty} \frac{f(y)}{cg(y)} g(y) dy = \frac{1}{c} \int_{-\infty}^{\infty} f(y) dy = \frac{1}{c}$$

Assim, a cada passo, a probabilidade de aceitar um valor é $\frac{1}{c}$, o que implica que o número de passos necessários para aceitar uma amostra tem distribuição geométrica com média c .

Por fim, sabemos que a função de distribuição acumulada da variável X , dado que ela foi aceita, é:

$$P(X \leq x) = P(Y \leq x \mid \text{aceitou}) = \frac{P(Y \leq x, \text{aceitou})}{P(\text{aceitar})}$$

Substituindo as expressões:

$$P(X \leq x) = \frac{\frac{1}{c} \int_{-\infty}^x f(y) dy}{\frac{1}{c}} = \int_{-\infty}^x f(y) dy$$

Portanto, a variável X gerada pelo método de aceitação e rejeição tem função de densidade $f(x)$, como desejado.

42.2 Exercícios

Exercício 1. Seja $X \sim \text{Gama}(3/2, 1)$ com função densidade dada por

$$f(x) = \frac{1}{\Gamma(3/2)} x^{1/2} e^{-x}, \text{ para } x > 0.$$

- a) Escreva um pseudo-algoritmo para simular um valor da distribuição de X usando o método da aceitação e rejeição usando como distribuição proposta a distribuição exponencial de parâmetro $2/3$.

Observação: Note que $\mathbb{E}[X] = 3/2$ e $\mathbb{E}[Y] = 3/2$, por isso o parâmetro da distribuição exponencial proposta foi tomado como $2/3$.

- b) Crie uma função para gerar um valor de X .
- c) Qual foi o número de passos necessários para gerar um valor de X ? Compare com o valor real.
- d) Crie uma função para gerar uma amostra de tamanho n de X .
- e) Qual foi o número médio de passos necessários para gerar a amostra de tamanho n ?
- f) Compare a distribuição empírica dos valores simulados com a distribuição real de X .

43 Transformações e Misturas

44 R

45 Python

45.1 Exercícios

Exercício 1.

46 Método de Box-Muller

47 R

48 Python

48.1 Exercícios

Exercício 1.

49 Método de Monte Carlo

Os **Métodos de Monte Carlo** (MMC) são uma técnica poderosa que utiliza a simulação de valores aleatórios para calcular aproximações de valores esperados e integrais que não podem ser resolvidos analiticamente. Neste contexto, aprendemos a utilizar variáveis aleatórias simuladas para realizar o cálculo aproximado de integrais.

Seja X uma v.a. discreta com função de probabilidade $p(x)$ ou uma v.a. contínua com função densidade $f(x)$. O Método de Monte Carlo propõe um estimador para

$$\theta = \mathbb{E}[g(X)] = \begin{cases} \int_{-\infty}^{\infty} g(x)f(x) dx & \text{se } X \text{ é contínua} \\ \sum_x g(x)p(x) & \text{se } X \text{ é discreta} \end{cases}$$

Especificamente, ele funciona da seguinte forma. Seja X_1, \dots, X_n uma sequência de variáveis aleatórias i.i.d. com a mesma distribuição que X . Então:

$$\hat{\theta}_n = \frac{1}{n} \sum_{i=1}^n g(X_i)$$

é o estimador de Monte Carlo para θ .

Pseudo-Algoritmo para aproximar θ :

1. Gere X_1, \dots, X_n valores independentes da v.a. X .
2. Calcule:

$$\hat{\theta}_n = \frac{1}{n} \sum_{i=1}^n g(X_i)$$

Porque Funciona?

O método de Monte Carlo funciona devido à Lei Forte dos Grandes Números, que garante que a média amostral convergirá quase certamente para a esperança verdadeira conforme o número de amostras tende ao infinito:

$$\hat{\theta}_n = \frac{1}{n} \sum_{i=1}^n g(X_i) \longrightarrow \theta \text{ (quase certamente)}$$

Assim, quando n cresce, a estimativa de Monte Carlo converge para o verdadeiro valor de θ .

Importante: Freqüentemente n é denotado por B na literatura de Monte Carlo, de forma a não haver confusão com o tamanho de amostras não geradas no computador.

49.1 Exemplo 1: Estimativa de uma Integral

Queremos obter uma estimativa para

$$\theta = \int_0^1 e^{-x} dx$$

Para isso, basta observar que se $U \sim Unif(0,1)$, então $\theta = \mathbb{E}[e^{-U}]$. De fato, se $U \in Unif(0,1)$,

$$\theta = \int_0^1 e^{-x} dx = \int_0^1 e^{-u} f(u) du.$$

Assim, o estimador desta integral via MMC pode ser obtido da seguinte forma:

50 R

```
n <- 100
u <- runif(n, min = 0, max = 1)
theta_hat <- mean(exp(-u))

valor_real <- 1 - exp(-1)
theta_hat
```

```
[1] 0.6251406
```

```
valor_real
```

```
[1] 0.6321206
```

51 Python

```
import numpy as np
n = 100
u = np.random.uniform(0, 1, n)
theta_hat = np.mean(np.exp(-u))

valor_real = 1 - np.exp(-1)
theta_hat, valor_real
```

```
(np.float64(0.6280312864908527), np.float64(0.6321205588285577))
```

Vamos verificar como o valor de n influencia na aproximação:

52 R

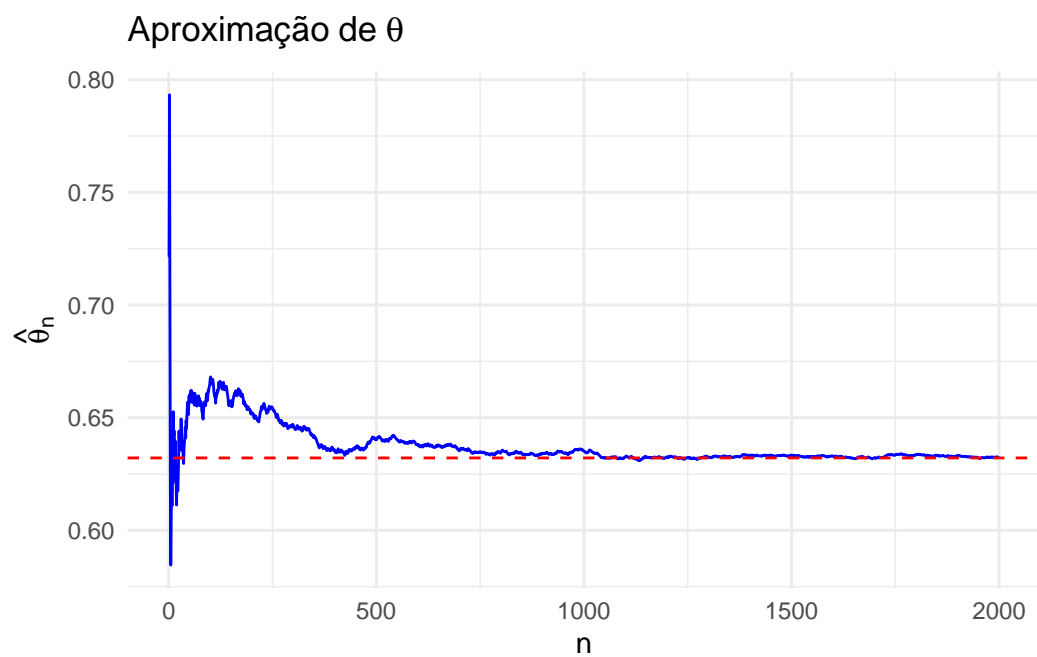
```
library(ggplot2)

set.seed(58)
N <- 1:2000
theta_hat <- numeric(length(N))
u <- numeric(length(N)) # Pré-alocação do vetor

# Preenchendo o vetor u e calculando theta_hat
for (i in seq_along(N)) {
  u[i] <- runif(1)
  theta_hat[i] <- mean(exp(-u[1:i]))
}

# Criando o dataframe para o ggplot
df <- data.frame(N = N, theta_hat = theta_hat)

# Plotar o gráfico usando ggplot2
ggplot(df, aes(x = N, y = theta_hat)) +
  geom_line(color = "blue") +
  geom_hline(yintercept = 1 - exp(-1), color = "red", linetype = "dashed") +
  labs(x = "n", y = expression(hat(theta)[n]), title = expression(Aproximação de theta)) +
  theme_minimal()
```



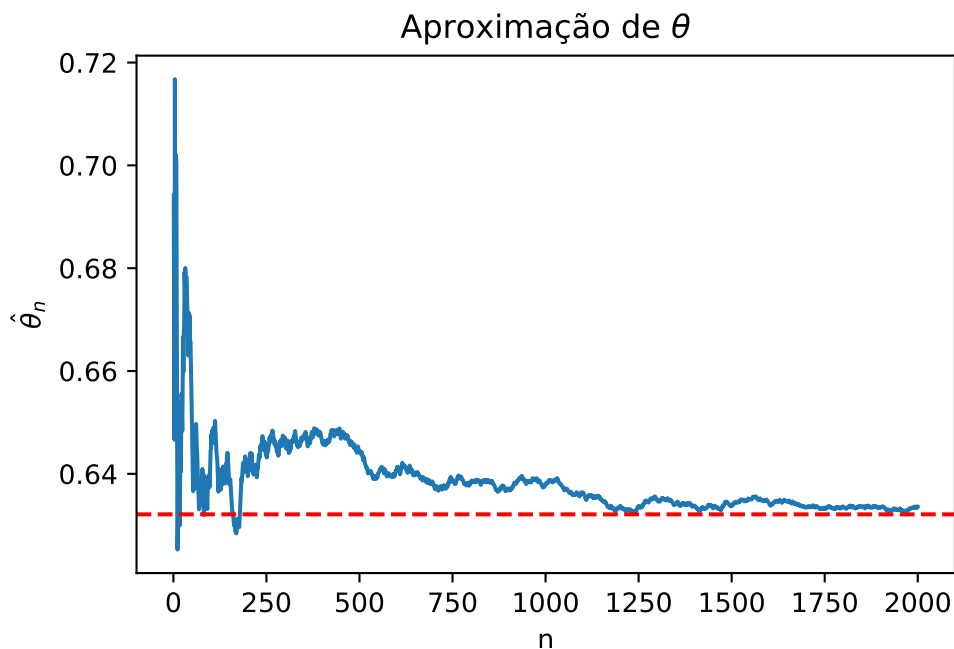
53 Python

```
import numpy as np
import matplotlib.pyplot as plt

np.random.seed(58)
N = np.arange(1, 2001, 1)
theta_hat = np.zeros(len(N))
u = np.array([])

for i in range(len(N)):
    u = np.append(u, np.random.uniform(0, 1))
    theta_hat[i] = np.mean(np.exp(-u))

plt.plot(N, theta_hat)
plt.axhline(y=1 - np.exp(-1), color='red', linestyle='--')
plt.xlabel('n')
plt.ylabel(r'$\hat{\theta}_n$')
plt.title(r'Aproximação de $\theta$')
plt.show()
```



Esse gráfico mostra a evolução da estimativa à medida que o número de variáveis aleatórias n aumenta, comparando com o valor real da integral.

53.1 Exemplo 2: Aproximando uma Probabilidade

Seja $X \sim \text{Gama}(2, 3)$. Queremos aproximar o valor de $\mathbb{P}(X \geq 0.4)$ usando o método de Monte Carlo. Note que

$$\theta := \mathbb{P}(X \geq 0.4) = \int g(x)f(x)dx,$$

em que $g(x) = I(x \geq 0.4)$ e $f(x)$ é a densidade da $\text{Gama}(2, 3)$. Assim, podemos aproximar a probabilidade gerada via MC.

Neste caso, o algoritmo corresponde a gerar $X_i \sim \text{Gama}(2, 3)$, definir e definir uma variável indicadora Y_i que vale 1 quando $X_i \geq 0.4$ e 0 caso contrário. A estimativa de Monte Carlo é dada pela média dos Y_i 's:

54 R

```
set.seed(58)
library(ggplot2)

N <- 50000
x <- rgamma(N, shape = 2, rate = 3) # Usando rate = 1/scale
y <- as.integer(x >= 0.4)

valor_aproximado <- mean(y)
valor_real <- 1 - pgamma(0.4, shape = 2, rate = 3)

valor_aproximado
```

```
[1] 0.66282
```

```
valor_real
```

```
[1] 0.6626273
```


55 Python

```
import numpy as np
from scipy.stats import gamma

N = 50000
x = gamma.rvs(2, scale=1/3, size=N)
y = (x >= 0.4).astype(int)

valor_aproximado = np.mean(y)
valor_real = 1 - gamma.cdf(0.4, 2, scale=1/3)

valor_aproximado, valor_real
```

```
(np.float64(0.66098), np.float64(0.6626272662068446))
```

55.1 Exemplo 3: Aproximando o valor de π

Neste exemplo, queremos aproximar o valor de π utilizando o método de Monte Carlo. A ideia é gerar pontos aleatórios em um quadrado e contar quantos caem dentro de um círculo inscrito no quadrado. Vamos seguir o raciocínio a partir da geometria básica.

55.1.0.1 Geometria

- Considere um quadrado com lado 2 centrado na origem, ou seja, o quadrado vai de $(-1, -1)$ até $(1, 1)$.
- Dentro deste quadrado, inscreva um círculo de raio 1, também centrado na origem.
- A área do quadrado é 4 (já que $2 \times 2 = 4$) e a área do círculo é $\pi \cdot r^2 = \pi \cdot 1^2 = \pi$.

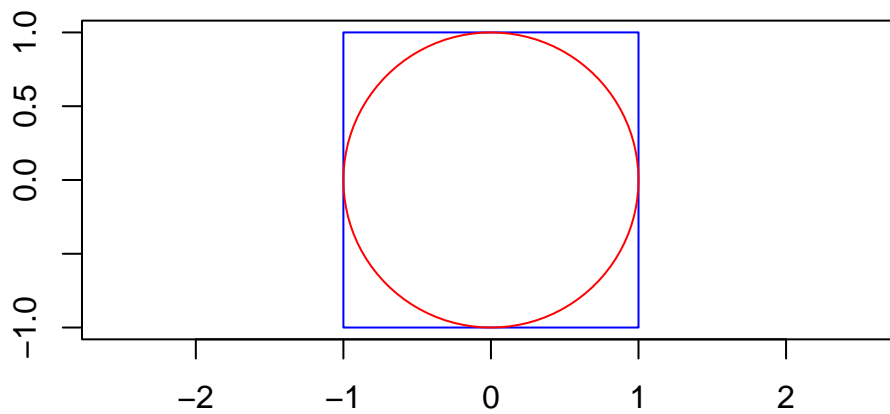
A razão entre a área do círculo e a área do quadrado é dada por:

$$\frac{\text{Área do círculo}}{\text{Área do quadrado}} = \frac{\pi}{4}$$

56 R

```
require(plotrix)
require(grid)

plot(c(-1, 1), c(-1,1), type = "n", asp=1,xlab="",yla="")
rect( -1, -1, 1, 1,border="blue")
draw.circle( 0, 0, 1 ,border="red")
```



57 Python

```
import matplotlib.pyplot as plt
from matplotlib.patches import Rectangle, Circle

# Configura o gráfico
fig, ax = plt.subplots()
ax.set_aspect('equal') # Define o aspecto como 1:1 (quadrado)
ax.set_xlim(-1, 1)
```

(-1.0, 1.0)

```
ax.set_ylim(-1, 1)
```

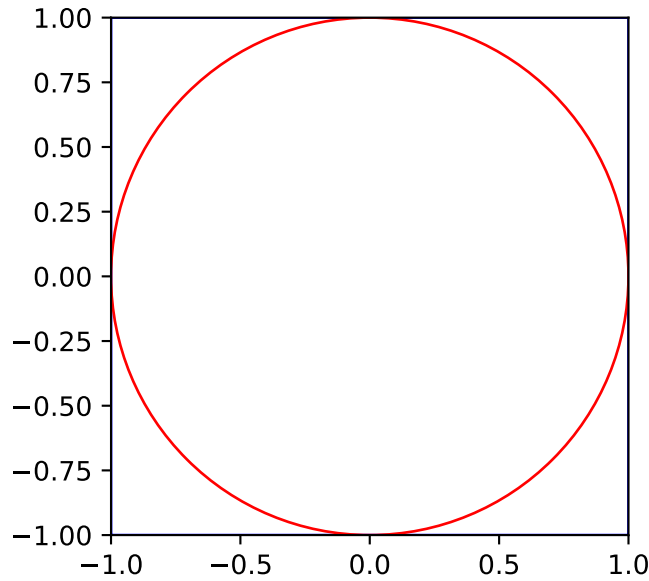
(-1.0, 1.0)

```
ax.set_xlabel('')
ax.set_ylabel('')

# Desenha o retângulo
rect = Rectangle((-1, -1), 2, 2, edgecolor='blue', facecolor='none')
ax.add_patch(rect)

# Desenha o círculo
circle = Circle((0, 0), 1, edgecolor='red', facecolor='none')
ax.add_patch(circle)

# Mostra o gráfico
plt.show()
```



Para estimar π usando Monte Carlo, procedemos da seguinte forma:

1. Geramos pontos aleatórios (x, y) no quadrado $[-1, 1] \times [-1, 1]$.
2. Verificamos se cada ponto está dentro do círculo, o que ocorre se $x^2 + y^2 \leq 1$.
3. A fração de pontos que caem dentro do círculo aproxima a razão $\frac{\pi}{4}$.
4. Multiplicamos essa fração por 4 para obter uma estimativa de π .

57.0.0.1 Matemática do Estimador

Formalmente, se $(X, Y) \sim \text{Unif}(-1, 1) \times \text{Unif}(-1, 1)$ e $g(x, y) = I((x, y) \text{ está no círculo})$, temos que

$$\theta := \int g(x, y) f(x, y) dx dy = \frac{\text{Área do círculo}}{\text{Área do quadrado}} = \frac{\pi}{4}$$

Assim, se (X_i, Y_i) é um ponto gerado uniformemente dentro do quadrado, o estimador de Monte Carlo para $\frac{\pi}{4}$ é dado por:

$$\hat{\theta}_n = \frac{1}{n} \sum_{i=1}^n g(x_i, y_i) = \frac{1}{n} \sum_{i=1}^n z_i,$$

em que $z_i = 1$ se o ponto i está dentro do círculo (i.e., se $x_i^2 + y_i^2 \leq 1$) e $z_i = 0$ caso contrário.

Multiplicando por 4, obtemos a estimativa de π :

$$\hat{\pi}_n = 4 \cdot \hat{\theta}_n = 4 \cdot \frac{1}{n} \sum_{i=1}^n z_i$$

De acordo com a **Lei Forte dos Grandes Números**, sabemos que:

$$\hat{\pi}_n \longrightarrow \pi \quad (\text{quase certamente, quando } n \rightarrow \infty).$$

Ou seja, à medida que o número de pontos simulados n aumenta, a estimativa $\hat{\pi}_n$ convergirá para o valor verdadeiro de π .

Implementação

Os trechos de código fornecidos em R e Python simulam esse processo, gerando n pontos e calculando a aproximação de π com base nos pontos que caem dentro do círculo. Além disso, são gerados gráficos que mostram como a estimativa de π melhora conforme o número de simulações aumenta, evidenciando a convergência mencionada.

Agora que todos os detalhes matemáticos do exemplo estão claros, o código para simulação pode ser executado para observar a aproximação prática de π .

58 R

```
set.seed(459)
library(ggplot2)

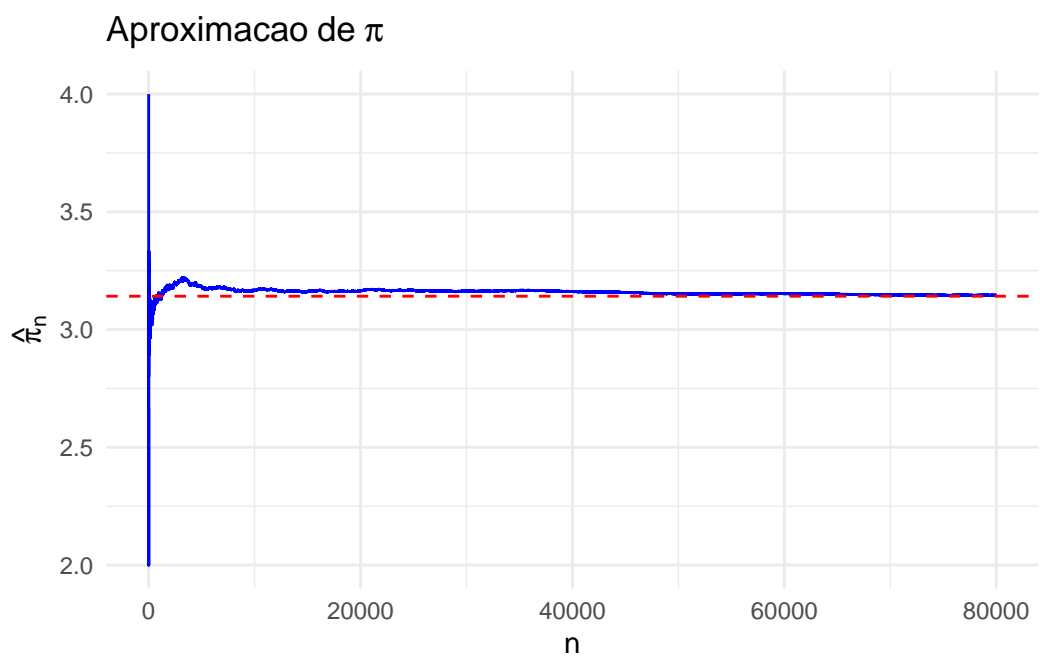
N <- 1:80000
z <- numeric(length(N))

# Loop para gerar os pontos e verificar se estão dentro do círculo
for (i in seq_along(N)) {
  x <- 2 * runif(1) - 1
  y <- 2 * runif(1) - 1
  z[i] <- (x^2 + y^2 <= 1)
}

# Cálculo da estimativa de pi
theta_hat <- cumsum(z) / N
pi_hat <- theta_hat * 4

# Criando o dataframe para o ggplot
df <- data.frame(N = N, pi_hat = pi_hat)

# Plotar o gráfico usando ggplot2
ggplot(df, aes(x = N, y = pi_hat)) +
  geom_line(color = "blue") +
  geom_hline(yintercept = pi, color = "red", linetype = "dashed") +
  labs(x = "n", y = expression(hat(pi)[n]), title = expression(Aproximacao-de~pi)) +
  theme_minimal()
```



59 Python

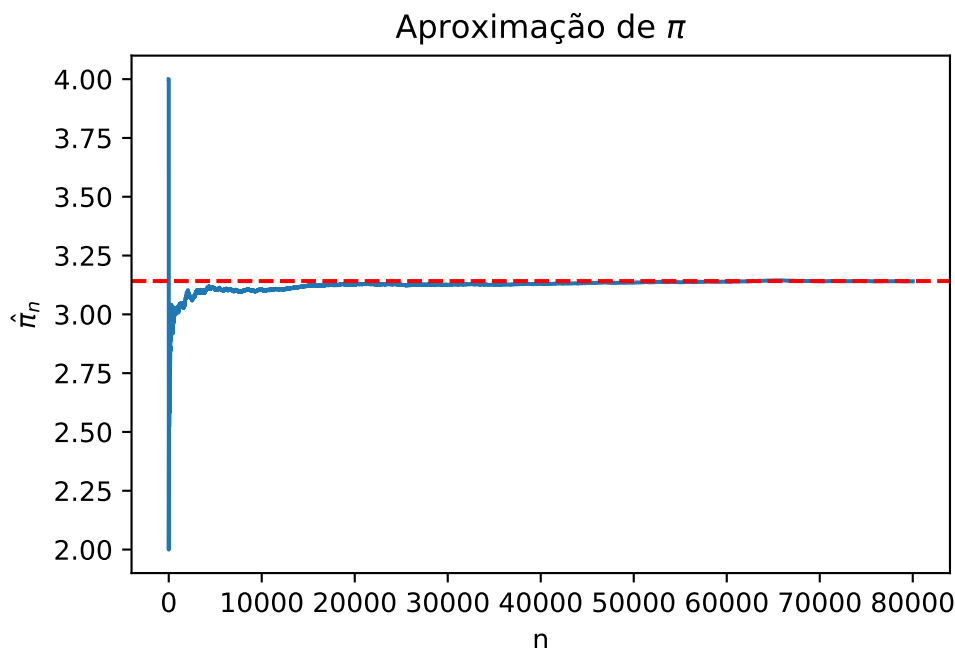
```
import numpy as np
import matplotlib.pyplot as plt

np.random.seed(459)
N = np.arange(1, 80001, 1)
z = np.zeros(len(N))

for i in range(len(N)):
    x = 2 * np.random.uniform(0, 1) - 1
    y = 2 * np.random.uniform(0, 1) - 1
    z[i] = (x**2 + y**2 <= 1)

theta_hat = np.cumsum(z) / N
pi_hat = theta_hat * 4

plt.plot(N, pi_hat)
plt.axhline(y=np.pi, color='red', linestyle='--')
plt.xlabel('n')
plt.ylabel(r'$\hat{\pi}_n$')
plt.title(r'Aproximação de $\pi$')
plt.show()
```

Esse gráfico mostra como a estimativa de π melhora conforme o número de pontos simulados n aumenta.

59.1 Intervalos de Confiança para Estimativas de Monte Carlo

O Método de Monte Carlo possibilita quantificar a incerteza em nossas estimativas. Para isso, podemos utilizar **intervalos de confiança**.

Suponha que temos uma sequência de amostras X_1, X_2, \dots, X_n de uma variável aleatória X , e calculamos a média amostral $\hat{\theta}_n = \frac{1}{n} \sum_{i=1}^n g(X_i)$. Pelo **Teorema Central do Limite**, sabemos que, para n suficientemente grande, a média amostral $\hat{\theta}_n$ é aproximadamente normalmente distribuída:

$$\hat{\theta}_n \sim N\left(\theta, \frac{\sigma^2}{n}\right)$$

onde $\sigma^2 = \text{Var}[g(X)]$ é a variância de $g(X)$. Essa aproximação permite a construção de um intervalo de confiança para θ .

O intervalo de confiança aproximado para θ com nível de confiança $(1 - \alpha) \times 100\%$ é dado por:

$$\hat{\theta}_n \pm z_{\alpha/2} \frac{\hat{\sigma}}{\sqrt{n}}$$

onde $z_{\alpha/2}$ é o quantil da distribuição normal padrão associado à probabilidade $\alpha/2$ e $\hat{\sigma}$ é a estimativa da variância de $g(X)$, obtida a partir das amostras.

Pseudo-Algoritmo para construir um intervalo de confiança:

1. Gere X_1, \dots, X_n amostras i.i.d. da v.a. X .
2. Calcule $\hat{\theta}_n = \frac{1}{n} \sum_{i=1}^n g(X_i)$.
3. Estime a variância de $g(X)$ como:

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (g(X_i) - \hat{\theta}_n)^2$$

4. Calcule o intervalo de confiança:

$$\hat{\theta}_n \pm z_{\alpha/2} \frac{\hat{\sigma}}{\sqrt{n}}$$

59.1.1 Exemplo: Intervalo de Confiança para a Estimativa de π

Neste exemplo, construímos um intervalo de confiança para a estimativa de π usando o método de Monte Carlo, seguindo os passos acima.

60 R

```
set.seed(0)
N <- 10000 # número de amostras
z <- numeric(N)

# Loop para gerar os pontos e verificar se estão dentro do círculo
for (i in 1:N) {
  x <- 2 * runif(1) - 1
  y <- 2 * runif(1) - 1
  z[i] <- (x^2 + y^2 <= 1)
}

theta_hat <- mean(z)
sigma_hat <- sqrt(var(z) / N)
alpha <- 0.05 # Nível de significância
z_alpha2 <- qnorm(1 - alpha / 2)

# Intervalo de confiança
IC <- c(theta_hat - z_alpha2 * sigma_hat, theta_hat + z_alpha2 * sigma_hat) * 4

theta_hat * 4 # Estimativa de pi
```

```
[1] 3.1308
```

```
IC # Intervalo de confiança
```

```
[1] 3.098466 3.163134
```

61 Python

```
import numpy as np
from scipy.stats import norm

np.random.seed(0)
N = 10000 # número de amostras
z = np.zeros(N)

# Loop para gerar os pontos e verificar se estão dentro do círculo
for i in range(N):
    x = 2 * np.random.uniform(0, 1) - 1
    y = 2 * np.random.uniform(0, 1) - 1
    z[i] = (x**2 + y**2 <= 1)

theta_hat = np.mean(z)
sigma_hat = np.sqrt(np.var(z) / N)
alpha = 0.05 # Nível de significância
z_alpha2 = norm.ppf(1 - alpha / 2)

# Intervalo de confiança
IC = (theta_hat - z_alpha2 * sigma_hat, theta_hat + z_alpha2 * sigma_hat)
IC = [i * 4 for i in IC]

theta_hat * 4 # Estimativa de pi
```

```
np.float64(3.1228)
```

```
IC # Intervalo de confiança
```

```
[np.float64(3.090360848359833), np.float64(3.1552391516401666)]
```

Neste caso, o intervalo de confiança construído em torno da estimativa $\hat{\pi}_n$ oferece uma ideia de quão próxima nossa estimativa está do valor verdadeiro de π , com uma certa confiança.

61.1 Exemplo 4: Problema das Figurinhas

Um colecionador está juntando figurinhas para completar um álbum da Copa. Qual é a probabilidade de que, ao comprar n pacotes, pelo menos um deles contenha duas ou mais figurinhas iguais? Durante o curso de probabilidade, você aprenderá como calcular essa probabilidade. Outra abordagem possível é o uso do método de Monte Carlo. Nesse caso, simulamos a compra de n pacotes, cada um com 5 figurinhas, de um total de 640 figurinhas disponíveis no álbum. Repetimos essa simulação $B = 1000$ vezes e contamos quantas vezes ocorreu pelo menos uma repetição de figurinhas em algum dos pacotes. A proporção de simulações com repetições é a estimativa de Monte Carlo da probabilidade que estamos interessados em calcular.

62 R

```
nFigurinhas <- 640
B <- 1000
nPacotes <- 1:50
aoMenosUmaRepetidaNoMesmoPacote <- matrix(NA, length(nPacotes), B)
for(ii in 1:length(nPacotes)) {
  for(jj in 1:B) {
    figurinhas <- sample(1:nFigurinhas, nPacotes[ii] * 5, replace = TRUE)
    figurinhas <- matrix(figurinhas, nPacotes[ii], 5)
    aoMenosUmaRepetidaNoMesmoPacote[ii,jj] <- sum(apply(figurinhas, 1, function(xx) length(u
  })
}

probCoincidencia <- apply(aoMenosUmaRepetidaNoMesmoPacote, 1, mean)
```

63 Python

```
import numpy as np
import random

nFigurinhas = 640
B = 1000
nPacotes = np.arange(1, 51) # Intervalo de 1 a 50 (Python é exclusivo no final)
aoMenosUmaRepetidaNoMesmoPacote = np.empty((len(nPacotes), B))

for ii in range(len(nPacotes)):
    for jj in range(B):
        figurinhas = [random.choices(range(1, nFigurinhas + 1), k=5) for _ in range(nPacotes)]
        figurinhas = np.array(figurinhas)
        aoMenosUmaRepetidaNoMesmoPacote[ii, jj] = np.sum([len(set(pacote)) != len(pacote) for _ in range(nPacotes)])

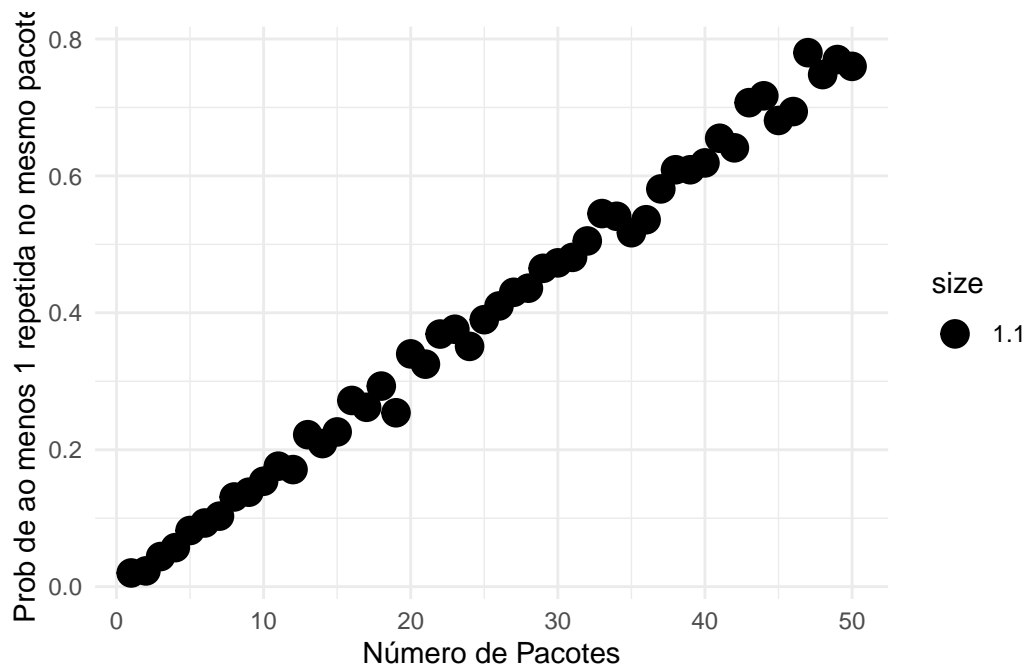
probCoincidencia = np.mean(aoMenosUmaRepetidaNoMesmoPacote, axis=1)
```

Agora podemos visualizar os resultados obtidos por meio do gráfico a seguir:

64 R

```
dados <- data.frame(probCoincidencia = probCoincidencia, nPacotes = nPacotes)

library(ggplot2)
ggplot(dados, aes(x = nPacotes, y = probCoincidencia)) +
  geom_point(aes(size = 1.1)) +
  xlab("Número de Pacotes") +
  ylab("Prob de ao menos 1 repetida no mesmo pacote") +
  theme(legend.position = "none")+
  theme_minimal()
```

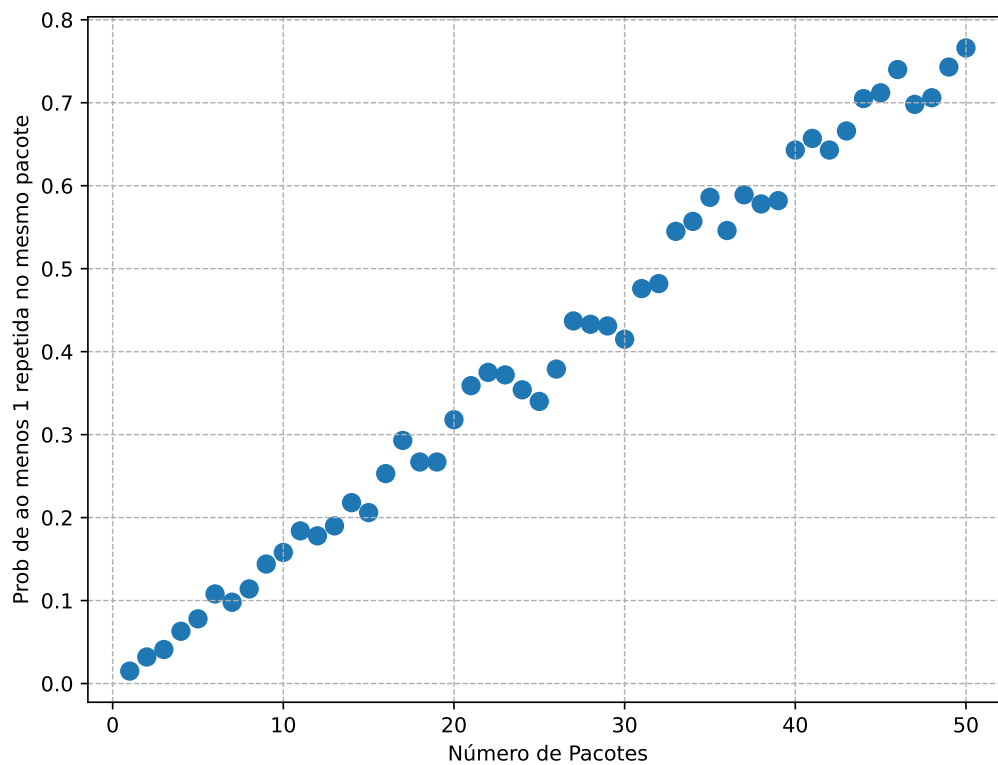


65 Python

```
import pandas as pd
import matplotlib.pyplot as plt

# Criação do DataFrame
dados = pd.DataFrame({'nPacotes': nPacotes, 'probCoincidencia': probCoincidencia})

# Criação do gráfico
plt.figure(figsize=(8, 6))
plt.scatter(dados['nPacotes'], dados['probCoincidencia'], s=70) # 's' ajusta o tamanho dos p
plt.xlabel('Número de Pacotes')
plt.ylabel('Prob de ao menos 1 repetida no mesmo pacote')
plt.grid(True, which='both', linestyle='--', linewidth=0.7)
plt.gca().set_facecolor('white') # Para seguir o tema minimalista
plt.show()
```



Simulação para completar o álbum

Uma segunda pergunta que também podemos responder via simulação é: quantos pacotes necessitamos em média para completar o álbum? A solução estimada pode ser obtida com o código abaixo:

66 R

```
numeroPacotes <- rep(NA, B)
for(jj in 1:B) {
  albumCompleto <- FALSE
  nPacotes <- 0
  figurinhas <- NULL
  while(!albumCompleto) {
    figurinhas <- c(figurinhas, sample(1:nFigurinhas, 5, replace = TRUE))
    if(length(unique(figurinhas)) == nFigurinhas) {
      albumCompleto <- TRUE
    }
    nPacotes <- nPacotes + 1
  }
  numeroPacotes[jj] <- nPacotes
}
```

67 Python

```
import numpy as np
import random

nFigurinhas = 640
B = 1000
numeroPacotes = np.empty(B)

for jj in range(B):
    albumCompleto = False
    nPacotes = 0
    figurinhas = []

    while not albumCompleto:
        figurinhas.extend(random.choices(range(1, nFigurinhas + 1), k=5))
        if len(set(figurinhas)) == nFigurinhas:
            albumCompleto = True
            nPacotes += 1

    numeroPacotes[jj] = nPacotes
```

O número médio de pacotes necessários para completar o álbum é:

68 R

```
mean(numeroPacotes)
```

```
[1] 901.629
```

69 Python

```
mean_numeroPacotes = np.mean(numeroPacotes)
mean_numeroPacotes
```

```
np.float64(905.341)
```

Com essa simulação, também podemos estimar as seguintes probabilidades:

70 R

```
cat("Probabilidade de precisar de mais de 800 pacotes: ", mean(numeroPacotes > 800) * 100, "%")
```

Probabilidade de precisar de mais de 800 pacotes: 69.3%

```
cat("Probabilidade de precisar de mais de 1000 pacotes: ", mean(numeroPacotes > 1000) * 100, "%")
```

Probabilidade de precisar de mais de 1000 pacotes: 23.2%

71 Python

```
prob_mais_800 = np.mean(numeroPacotes > 800) * 100
prob_mais_1000 = np.mean(numeroPacotes > 1000) * 100

print(f"Probabilidade de precisar de mais de 800 pacotes: {prob_mais_800}%")
```

Probabilidade de precisar de mais de 800 pacotes: 71.0%

```
print(f"Probabilidade de precisar de mais de 1000 pacotes: {prob_mais_1000}%")
```

Probabilidade de precisar de mais de 1000 pacotes: 24.099999999999998%

71.1 Exercícios

Exercício 1.

Utilize o método de Monte Carlo para aproximar a integral

$$I = \int_0^{10} \sin(x^2) dx$$

Forneça um intervalo de confiança para avaliar a precisão da estimativa.

Exercício 1.

Utilize o método de Monte Carlo para aproximar a integral

$$I = \int_1^{\infty} \frac{1}{x^3} dx$$

Forneça um intervalo de confiança para avaliar a precisão da estimativa.

Exercício 2.

Seja $X \sim \text{Exp}(\lambda)$ com $\lambda = 2$. Estime a esperança $\mathbb{E}[X^2]$ utilizando o método de Monte Carlo. Forneça um intervalo de confiança para avaliar a precisão da estimativa. Compare com o valor real dessa quantidade.

Exercício 3.

Considere que X tem distribuição Normal(0,1) e Y tem distribuição Gamma(1,1). Estime $P(X \times Y > 3)$. Forneça um intervalo de confiança para avaliar a precisão da estimativa.

References

Ross, Sheldon M. 2006. *Simulation, Fourth Edition*. USA: Academic Press, Inc.