Introdução ao R

MATF14 - Estatística Econômica I

Rodney Fonseca

30/10/2024

Introdução

Software R

R é uma linguagem e um ambiente para computação estatística e para preparação de gráficos

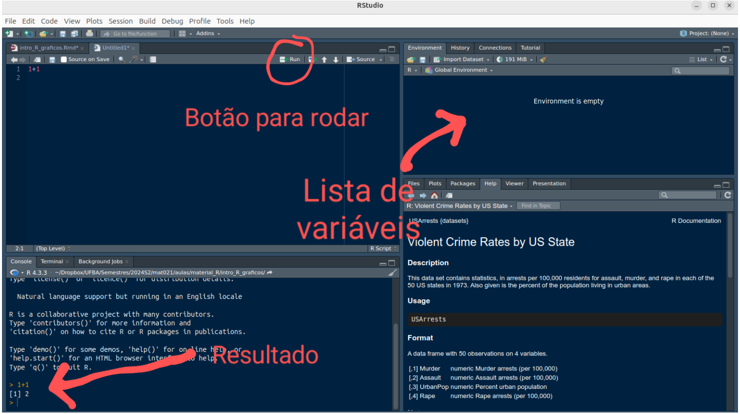
Vantagens

- software gratuito
- métodos para simulação probabilística e cálculo de probabilidades
- diversos tipos de técnicas para análise de dados
- ferramentas gráficas

Referências

- Livro Frery, Cribari-Neto (2011) *Elementos de Estatística Computacional Usando Plataformas de Software Livre/Gratuito*
- Página do curso de *Ciência de dados* da profa. Carolina Mota e prof. Gilberto Sassi do DEst-UFBA: https://ufba.netlify.app/paginas/catalogo
- Página do curso *Estatística Computacional com R* do prof. Paulo Justiniano (UFPR) e equipe: http://cursos.leg.ufpr.br/ecr/index.html

Ambiente do RStudio



Comandos básicos

Operações básicas

Adição

```
1+1
```

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

Subtração

```
5 - 3
```

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

Multiplicação

```
2 * 3
```

```
## [1] 6
```

Divisão

```
10/5
```

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

Operações básicas

Potência

```r
2^4
```

```
## [1] 16
```

Logaritmo (natural)

```r
log(10)
```

```
## [1] 2.302585
```

Exponencial

```r
exp(3)
```

```
## [1] 20.08554
```

Operadores lógicos

Maior/menor que

```r
5 > 3
```

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

```r
3>=3
```

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

```r
3>3
```

```
## [1] FALSE
```

Operadores lógicos

Igual

```r
5 == 3
```

```
## [1] FALSE
```

```r
3 == 3
```

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

Diferente

```r
4 != 2
```

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

```r
4 != 4
```

```
## [1] FALSE
```

Tipos de dados em R

Número real

```r
class(0.5)
```

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

Valor lógico

```r
class(TRUE)
```

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

Caractere

```r
class('UFBA')
```

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

Variável

- É como uma caixa nomeada. Você pode trocar o conteúdo da caixa, mas o nome permanece o mesmo.

Usamos os símbolos <- ou = para atribuir valores a uma variável

```
x <- 2
```

Para ver o valor da variável, basta rodar o seu nome

```
x
```

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

Variável

Podemos fazer operações com a variável

```
5 * x
```

```
## [1] 10
```

podemos trocar o seu valor

```
x <- 20
x
```

```
## [1] 20
```

Podemos atribuir o valor de operações à outras variáveis

```
idade <- 2 * x
idade
```

```
## [1] 40
```

Funções

- Em matemática, funções recebem um ou mais argumentos e retornam um ou mais valores. Funções em R são similares.

Função para criar uma sequência de 9 números

```r
seq(from = 1, to = 9)
```

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

Fatorial de n

```r
factorial(5)
```

```
## [1] 120
```

Funções

- Em matemática, funções recebem um ou mais argumentos e retornam um ou mais valores. Funções em R são similares.

Função para criar uma sequência de 9 números

```r
seq(from = 1, to = 9)
```

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

Fatorial de n

```r
factorial(5)
```

```
## [1] 120
```

Coeficiente binomial (combinação)

```r
choose(5,2)
```

```
## [1] 10
```

Vetores

- Vetores são listas indexadas de *variáveis do mesmo tipo*. É como um armário de gavetas numeradas como 1, 2, 3, . . . Você pode mudar o conteúdo da gaveta 5 sem alterar o conteúdo da gaveta 1.

Vetores podem ser criados com a fórmula `c(x1, x2, ...)`

```r
meu_vetor <- c(1, 2, 5, 8, 1, 3)
```

Vetores

- Vetores são listas indexadas de *variáveis do mesmo tipo*. É como um armário de gavetas numeradas como 1, 2, 3, . . . Você pode mudar o conteúdo da gaveta 5 sem alterar o conteúdo da gaveta 1.

Vetores podem ser criados com a fórmula `c(x1, x2, ...)`

```r
meu_vetor <- c(1, 2, 5, 8, 1, 3)
```

Valores de elementos de um vetor podem ser vistos assim

```r
meu_vetor[3]
```

```
## [1] 5
```

Vetores

- Vetores são listas indexadas de *variáveis do mesmo tipo*. É como um armário de gavetas numeradas como 1, 2, 3, ... Você pode mudar o conteúdo da gaveta 5 sem alterar o conteúdo da gaveta 1.

Vetores podem ser criados com a fórmula `c(x1, x2, ...)`

```
meu_vetor <- c(1, 2, 5, 8, 1, 3)
```

Valores de elementos de um vetor podem ser vistos assim

```
meu_vetor[3]
```

```
## [1] 5
```

Podemos alterar elementos do vetor

```
meu_vetor[3] <- 1
meu_vetor
```

```
## [1] 1 2 1 8 1 3
```

Vetores

Soma dos elementos de um vetor

```r
sum(meu_vetor)
```

```
## [1] 16
```

Produto dos elementos de um vetor

```r
prod(meu_vetor)
```

```
## [1] 48
```

Aplicando uma função aos elementos

```r
log(meu_vetor)
```

```
## [1] 0.0000000 0.6931472 0.0000000 2.0794415 0.0000000 1
```

Variáveis aleatórias no software R

Funções para variáveis aleatórias

- R conta com funções para simulação e cálculo de probabilidades de diversas variáveis aleatórias: Bernoulli, binomial, geométrica, Poisson, normal, etc.
- Exemplos de funções: gerador de valores aleatórios, função de probabilidade, função de distribuição acumulada, etc.

Funções para variáveis aleatórias

- R conta com funções para simulação e cálculo de probabilidades de diversas variáveis aleatórias: Bernoulli, binomial, geométrica, Poisson, normal, etc.
- Exemplos de funções: gerador de valores aleatórios, função de probabilidade, função de distribuição acumulada, etc.
- Vamos explorar algumas destas funções para as distribuições Bernoulli e binomial

Distribuição Bernoulli

- Relembrando: a distribuição Bernoulli é usada para modelar experimentos com dois resultados possíveis, *sucesso* ou *fracasso*, em que a probabilidade de sucesso é $p \in [0, 1]$
- Se $X \sim \textit{Bernoulli}(p)$, então $X \in \{0, 1\}$ e X assume valor 1 com probabilidade $p = P(X = 1)$.
- Temos ainda que $E(X) = p$ e $var(X) = p(1 - p)$

Distribuição Bernoulli

- Relembrando: a distribuição Bernoulli é usada para modelar experimentos com dois resultados possíveis, *sucesso* ou *fracasso*, em que a probabilidade de sucesso é $p \in [0, 1]$
- Se $X \sim Bernoulli(p)$, então $X \in \{0, 1\}$ e X assume valor 1 com probabilidade $p = P(X = 1)$.
- Temos ainda que $E(X) = p$ e $var(X) = p(1 - p)$
- Note que também podemos dizer que $X \sim Bin(1, p)$

Distribuição Bernoulli

- Podemos usar a função *rbinom(n_amostras, 1, p)* para gerar valores de $X \sim Bernoulli(p)$, em que *n_amostras* é o número de realizações simuladas de X

Distribuição Bernoulli

- Podemos usar a função *rbinom(n_amostras, 1, p)* para gerar valores de $X \sim Bernoulli(p)$, em que *n_amostras* é o número de realizações simuladas de X
- Ex.: Gerando 10 valores de $X \sim Bernoulli(1/4)$:

```
amostra_bern <- rbinom(10, 1, 1/4)
amostra_bern
```

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

Distribuição Bernoulli

- Podemos montar tabelas para checar a frequência dos valores gerados

Tabela de frequências absolutas

```
table(amostra_bern)
```

```
## amostra_bern
## 0 1
## 9 1
```

Tabela de frequências relativas

```
table(amostra_bern)/10
```

```
## amostra_bern
##   0   1
## 0.9 0.1
```

Distribuição Bernoulli

- Calculando a média amostral dos valores

```r
mean(amostra_bern)
```

```
## [1] 0.1
```

- Calculando a variância amostral dos valores

```r
var(amostra_bern)
```

```
## [1] 0.1
```

Distribuição Bernoulli

- Calculando a média amostral dos valores

```r
mean(amostra_bern)
```

```
## [1] 0.1
```

- Calculando a variância amostral dos valores

```r
var(amostra_bern)
```

```
## [1] 0.1
```

Observação: as funções *mean* e *var* não fornecem os valores teóricos, somente estimativas com base na amostra gerada

Distribuição Bernoulli

- Para calcular probabilidades, podemos usar a função **dbinom(k, 1, p)**. O exemplo abaixo fornece $P(X = 1)$:

```
dbinom(1, 1, 1/4)
```

```
## [1] 0.25
```

Distribuição Bernoulli

- Para calcular probabilidades, podemos usar a função **dbinom(k, 1, p)**. O exemplo abaixo fornece $P(X = 1)$:

```r
dbinom(1, 1, 1/4)
```

```
## [1] 0.25
```

- A função de distribuição acumulada $F(t) = P(X \leq t)$ é **pbinom(t, 1, p)**. O exemplo abaixo fornece $P(X \leq 1/2)$:

```r
pbinom(1/2, 1, 1/4)
```

```
## [1] 0.75
```

Distribuição Bernoulli

- Vamos considerar um exemplo com 1000 valores simulados

```
amostra_maior_bern <- rbinom(1000, 1, 1/4)
table(amostra_maior_bern)/1000
```

```
## amostra_maior_bern
##     0     1
## 0.759 0.241
```

- As frequências ficam mais próximas das probabilidades teóricas

Distribuição Bernoulli

- Vamos considerar um exemplo com 1000 valores simulados

```
amostra_maior_bern <- rbinom(1000, 1, 1/4)
table(amostra_maior_bern)/1000
```

```
## amostra_maior_bern
##     0     1
## 0.759 0.241
```

- As frequências ficam mais próximas das probabilidades teóricas
- A média teórica é $E(x) = p = 0.25$, enquanto a amostral foi

```
mean(amostra_maior_bern)
```

```
## [1] 0.241
```

- Quanto maior a amostra de valores gerados for, mais próxima a média amostral será da média teórica (populacional).

Distribuição Binomial

- Relembrando: a distribuição binomial conta o número de sucessos em n repetições de experimentos de Bernoulli com probabilidade p de sucesso
- Se $X \sim Bin(n, p)$, então $X \in \{0, 1, \ldots, n\}$ e X assume valor $k \in \{0, 1, \ldots, n\}$ com probabilidade

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

- Temos ainda que $E(X) = np$ e $var(X) = np(1 - p)$

Distribuição Binomial

- A função *rbinom(n_amostras, n, p)* para gerar valores de $X \sim Bin(n, p)$, em que *n_amostras* é o número de realizações simuladas de X, enquanto n e p são os parâmetros da distribuição binomial

Distribuição Binomial

- A função *rbinom(n_amostras, n, p)* para gerar valores de $X \sim Bin(n, p)$, em que *n_amostras* é o número de realizações simuladas de X, enquanto n e p são os parâmetros da distribuição binomial
- Ex.: Gerando 10 valores de $X \sim Bin(5, 1/2)$:

```
amostra_bin <- rbinom(10, 5, 1/2)
amostra_bin
```

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

Tabela de frequência dos valores gerados

```
table(amostra_bin)
```

```
## amostra_bin
## 1 2 3 4
## 1 2 5 2
```

Distribuição Binomial

- $X \sim Bin(n, p)$, podemos calcular a probabilidade $P(X = k)$ usando a função **dbinom(k, n, p)**.
- No exemplo abaixo, consideramos que $X \sim Bin(5, 1/2)$ e calculamos $P(X = 3)$:

```r
dbinom(3, 5, 1/2)
```

```
## [1] 0.3125
```

- Usando a fórmula teórica $P(X = 3) = \binom{5}{3}\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right)^{5-3}$, tal cálculo seria

```r
choose(5, 3) * ( (1/2)^3 ) * ( (1/2)^(5 - 3) )
```

```
## [1] 0.3125
```

Distribuição Binomial

- Vamos gerar uma amostra grande de $X \sim Bin(5, 1/2)$:

```
tam_amostra <- 1000000
vx <- rbinom(tam_amostra, 5, 1/2)
table(vx)/tam_amostra
```

```
## vx
##        0        1        2        3        4        5
## 0.031465 0.156483 0.312334 0.312410 0.156142 0.031166
```

- A média amostral dessa amostra é

```
mean(vx)
```

```
## [1] 2.498779
```

enquanto a média teórica é $E(X) = np = 5 * (1/2) = 2.5$.