Introdução ao R

Teoria das Probabilidades
Parte II

Denise Manfredini
PPGEco/UFSC

27 de março de 2019

Introdução

Objetivo

Relembrar conceitos básicos da teoria das probabilidades.

O que vamos ver no capítulo?

- Gerar números aleatórios contínuos;
- Média e variância de variáveis contínuas; e
- Computar distribuições de variáveis contínuas.

Bibliografia

Hanck, Arnold, Gerber, Schmelzer (2018). Introduction to Econometrics with R. GitHub/bookdown. ▸ Link

Adicional

Farnsworth (2014). Econometrics in R ▸ Link

Variáveis Contínuas

Assumem valores em uma escala contínua na reta real.
Ex.: Peso de um objeto.

Distribuição de Probabilidade de V.A. Contínuas

Como é possível ter um continum de possíveis valores para a variável aleatória contínua, não podemos mais usar o conceito de distribuição de probabilidades.

- Distribuição de probabilidade dá lugar para a **função densidade de probabilidade** (FDP); e
- A **função de distribuição acumulada** (FDA) agora identifica qual a probabilidade da v.a. assumir um número maior ou menor do que um valor particular.

Isso implica que temos que aprender a usar funções e integrais no R!

Funções com o R

Escrevendo a função: $x^2 + 2x = 0$

```
f <- function(x){ x^2 + 2*x }

f(3)
[1] 15
```

Integrais com o R

Integrando a função: $\frac{1}{x^2}$ de 1 até $+\infty$

```
z <- function(x){ 1/x^2 }

integral_z <- integrate(z,
                        lower = 1,
                        upper = Inf)$value

[1] 1
```

Funções com o R

Exercício 2

Escreva a função: $w(x) = e^{-x}$

Dica: *e* é a função *exp*()

Exercício 3

Ache o valor da integral: $\int_0^{+\infty} e^{-x} dx$

Funções com o R

Exercício 2

Escreva a função: $w(x) = e^{-x}$

Dica: *e* é a função *exp*()

```
w <- function(x){ exp(-x) }
```

Exercício 3

Ache o valor da integral: $\int_0^{+\infty} e^{-x} dx$

```
integrate(w,lower = 0, upper = Inf)$value
 [1] 1
```

Probabilidade para uma V.A. Contínua

Função Densidade de Probabilidade

Seja $f_Y(y)$ a função densidade de probabilidade de Y. A probabilidade de que Y esteja entre a e b, com $a < b$, é:

$$P(a \leq Y \leq b) = \int_a^b f_Y(y)\mathrm{d}y$$

Propriedades:

$F(b) = \Pr(X \leq x) = \int_{-\infty}^{x} f(t)dt$
$f(x) \geq 0$ para qualquer x
$\int_{-\infty}^{\infty} f(t)dt = 1$

Esperança

Esperança

$$E(Y) = \mu_Y = \int y f_Y(y) \mathrm{d}y$$

No R:

```
  # defina a FDP
 f <- function(x)x/4*exp(-x^2/8)

 # defina a função ex
ex <- function(x)x*f(x)

# calcule o valor esperado de X
expected_value <- integrate(ex, 0, Inf)$value
```

Variância

Variância

$$\text{Var}(Y) = \sigma_Y^2 = \int (y - \mu_Y)^2 f_Y(y) \mathrm{d}$$

$$\text{Var}(Y) = E\left(\mathrm{Y}^2\right) - E(Y)^2, \text{ em que} E\left(Y^2\right) = \int_0^\infty y^2 f_y(y) \mathrm{d}y$$

No R:

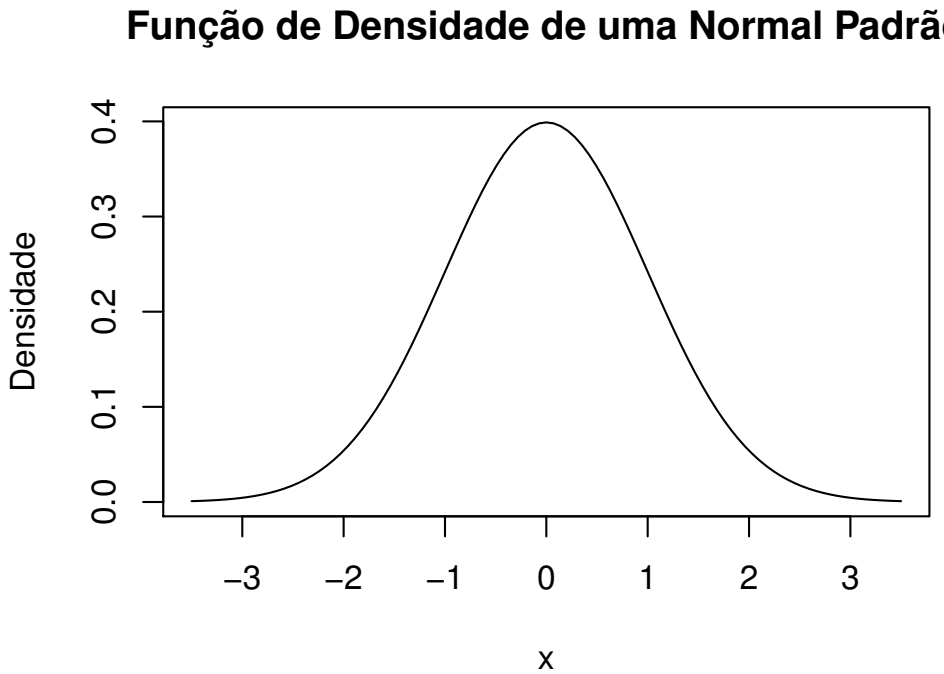
```
# defina a função ex2
ex2 <- function(x)x^2*f(x)

# calcule a variância de X
variance <- integrate(ex2, 0, Inf)$value - expected_value^2
```

Distribuição Normal

FDP da Distribuição Normal

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp -(x-\mu)^2 / \left(2\sigma^2\right)$$



Distribuição Normal no R

FDP da distribuição Normal no R.

- $x <- dnorm(x, mean, sd)$

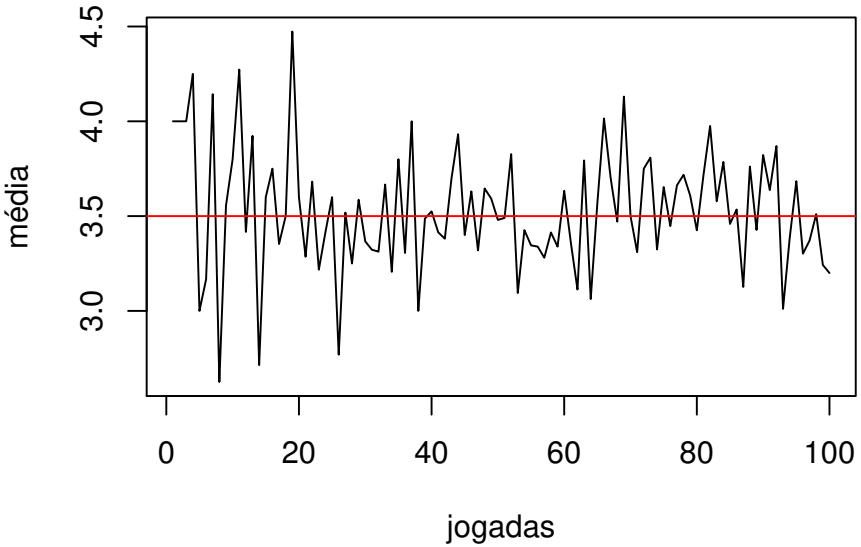
Função de probabilidade acumulada da Normal.

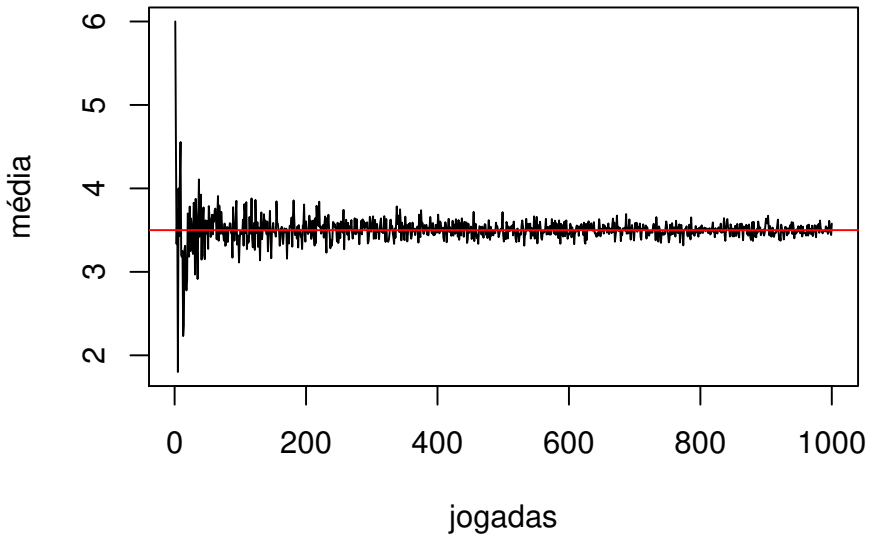
- $x <- pnorm(q, mean, sd)$

Função que simula variáveis de uma distribuição Normal.

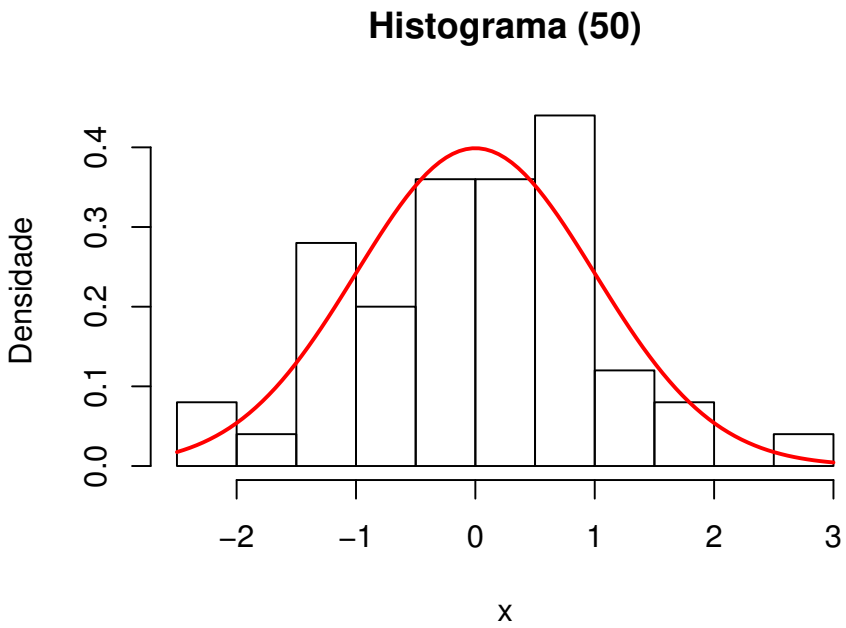
- $x <- rnorm(n, mean, sd)$

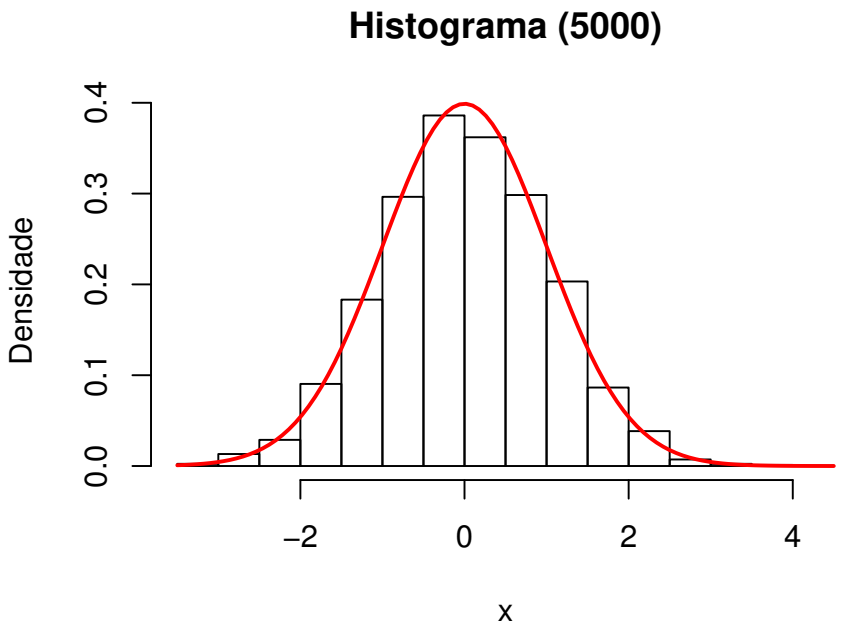
Lei dos Grandes Números





Teorema do Limite Central





Intervalos de Confiança para a Média

Bibliografia:

UC Business Analytics R Programming Guide ▸ Link

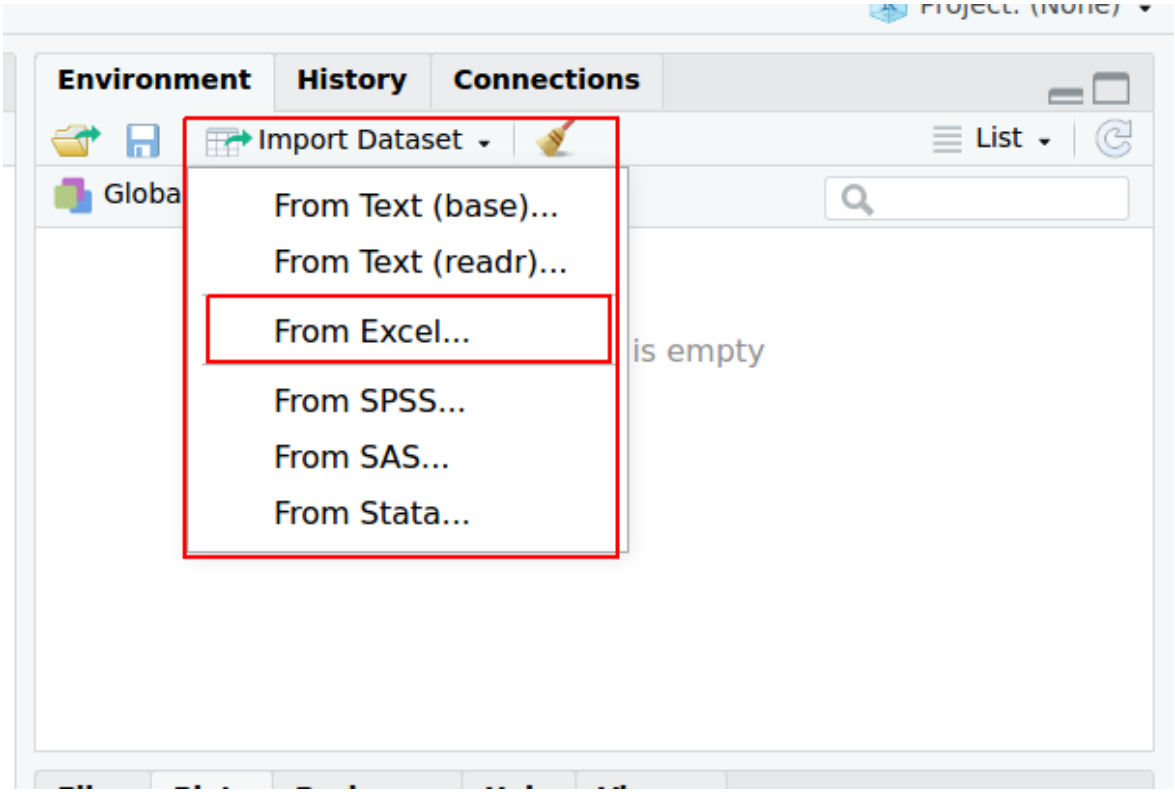
Intervalo de Confiança

- Qual o preço médio de venda dos imóveis em Windsor, Canadá [1]?

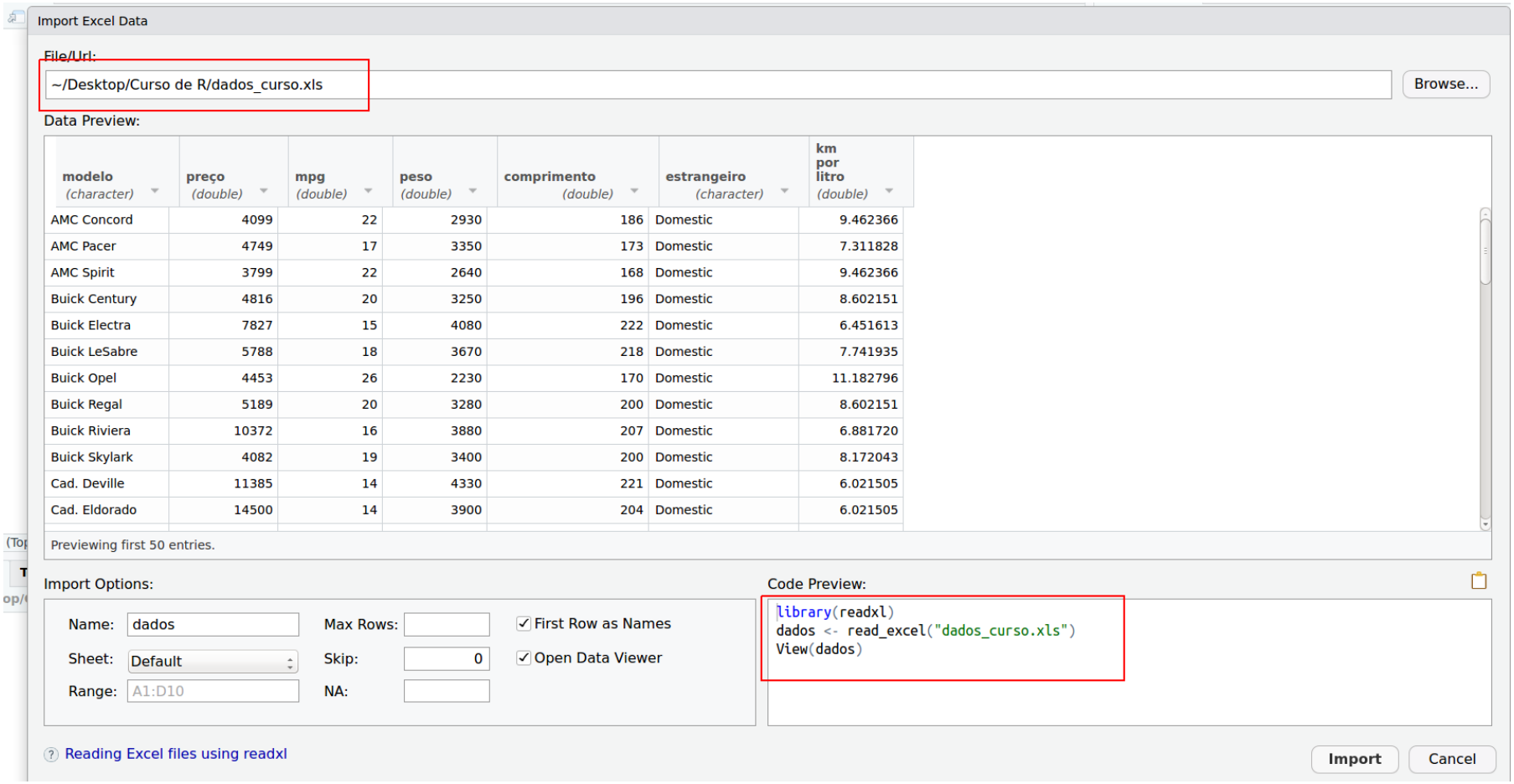
[1] baseada no artigo de Anglin e Gençay (1996)

Intervalo de Confiança

- O primeiro passo na análise de dados é **carregar os dados**.
- O R aceita diversos formatos de dados (.xlsx, .csv, .txt, .mat, etc.)

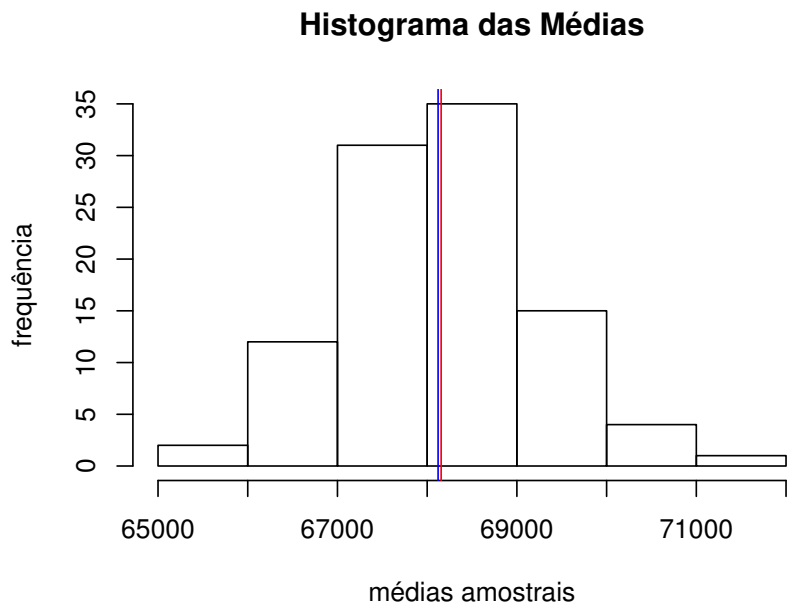


Intervalo de Confiança



Intervalo de Confiança

Estimativa Pontual ± Margem de Erro



Intervalo de Confiança

Se a população tem como processo gerador uma distribuição normal ou a amostral é grande o suficiente (lembre do Teorema do Limite Central), a equação abaixo é um IC confiável:

$$\bar{x} \pm t_{\alpha/2} \frac{S}{\sqrt{n}}$$

em que:

- $\bar{x}$ é a estimativa pontual;
- $t_{\alpha/2}$ distribuição-t (Student);
- s é o desvio-padrão da amostra; e
- $\sqrt{n}$ é a raiz quadrada do tamanho da amostra.

Intervalo de Confiança

```
# tira amostras dos dados originais
set.seed(123)
windsor_sample <- sample_frac(windsor_pop, .5)

# Parâmetros para o IC
x <- windsor_sample$price
xbar <- mean(x)                          # média
multi <- qt(.975, df = length(x) - 1)    # t-Student
sigma <- sd(x)                           # desvio padrão
denom <- sqrt(length(x))                 # raiz quadrada de n
```

Intervalo de Confiança

$$\bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$$

Calculando o IC

```
# calcula o erro padrão
se <- multi * (sigma / denom)

# intervalo inferior e superior
xbar + c(-se, se)
```

```
[1] 65288.43 71502.40
```

Intervalo de Confiança

Outra forma de calcular o IC é utilizando a função t.test.

t.test(x, ...) - Teste t de Student

```
# nomeia a função t.test
 ttest <- t.test(windsor_sample$price, conf.level = 0.95)

# mostra apenas o resultado do IC
 ttest$conf.int
```

Testes de Hipótese

Bibliografia:

Hypothesis Testing in R using Hollywood Movies Dataset. ▸ Link

UC Business Analytics R Programming Guide. ▸ Link

Análise das notas dos filmes no IMDB

Hipótese Nula

Ho: Não existe diferença significativa na média das notas dos filmes lançados em 2015.

Hipótese Alternativa

H1: Existe diferença significativa na média das notas dos filmes lançados em 2015.

Teste de Hipótese

Quando o desvio-padrão da população não é conhecido, o teste-t pode ser usado para os testes de hipóteses:

$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}$$

Intervalo de Confiança

```
# define os dados
sample <- year2015$Metascore
pop <- movies$Metascore

 sample_mean <- mean(sample)   # média da amostra

 pop_mean <- mean(pop)   # média da população

 n <- length(sample)  # tamanho da amostra

 var <- var(sample)  # variância da amostra
```

Teste de Hipótese

$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}$$

Calculando o teste de hipótese

```
# calcula o teste-t
tstatistic <- (sample_mean - pop_mean) / (sqrt(var/(n)))
```

```
[1]-1.20705
```

Teste de Hipótese

Outra forma de calcular o valor do teste de hipótese é utilizando a função t.test.

t.test(x, ...) - Teste t de Student

```
# nomeia a função t.test
t.test(year2015$Metascore, mu = mean(movies$Metascore))


length(year2015$Metascore) -1 # grau de liberdade

qt(0.975,108) # valor tabelado da distribuição t (superior)
qt(0.025,108) # valor tabelado da distribuição t (inferior)
```

Exercícios

Probabilidade para uma V.A. Contínua no R

Considere uma variável aleatória X com FDP igual a:

$$f_X(x) = \frac{x}{4} e^{-x^2/8}, \quad x \geq 0$$

Exercício 4

Escreva a função da FDP no R.

Exercício 5

Confira se a função é uma FDP.

Dica: Para ser FDP: $X \geq 0$ **E** a área total tem que ser igual a um.

Probabilidade para uma V.A. Contínua no R

Exercício 4

Escreva a função da FDP no R.

```
FDP <- function(x){x/4*exp(-x^2/8)}
```

Exercício 5

Confira se a função é uma FDP.

Dica: Para ser FDP 1) $X \geq 0$ e a área total tem que ser igual a um.

```
integrate(FDP, 0, Inf)$value
[1] 1
```

Esperança e Variância no R

Considere uma variável aleatória X com FDP igual a:

$$f_X(x) = \frac{3}{x^4}, \quad x > 1$$

Exercício 6

Encontre a esperança e variância.

Esperança e Variância no R

Exercício 6

Encontre a esperança e variância.

```
    # defina a FDP
f <- function(x) 3 / x^4

  # calcule E(X)
g <- function(x) x * f(x)
EX <- integrate(g,
                lower = 1,
                upper = Inf)$value
```

```
  # calcule Var(X)
h <- function(x) x^2 * f(x)
VarX <- integrate(h,
                  lower = 1,
                  upper = Inf)$value
                  - EX^2
```

EX [Resultado: 1.5]

VarX [Resultado: 0.75]

Distribuição Normal no R

Exercício 7

Seja $Z \sim N(0, 1)$.
Calcule valor da densidade da normal padrão em $c = 3$

Distribuição Normal no R

Exercício 7

Seja $Z \sim N(0, 1)$.
Calcule valor da densidade da normal padrão em $c = 3$

```
# usando a função
    dnorm(3)

    [1] 0.004431848
```

```
# escrevendo a FDP da normal

normal <- function(x) 1 /
          sqrt(2*pi*1)*
          exp(-(x-0)^2/(2*1^2))

    normal(3)
    [1] 0.004431848
```

Distribuição Normal no R

Exercício 8

Seja $Y \sim N(2, 12)$.
Gere 5 números aleatórios dessa distribuição.

Distribuição Normal no R

Exercício 8

Seja $Y \sim N(2, 12)$.
Gere 5 números aleatórios dessa distribuição.

```
set.seed(1)
rnorm(5, 2, 12)
```

```
[1] -5.517446 4.203720 -8.027543 21.143370 5.954093
```

Distribuição Normal no R

Exercício 9

Seja $Z \sim N(0,1)$.**Calcule** $P(|Z| \leq 1.64)$. Dica:

$P(|Z| \leq z) = P(-z \leq Z \leq z)$

Distribuição Normal no R

Exercício 9

Seja $Z \sim N(0,1)$.**Calcule** $P(|Z| \leq 1.64)$. Dica:

$P(|Z| \leq z) = P(-z \leq Z \leq z)$

```
pnorm(1.64)-pnorm(-1.64)
[1] 0.8989948
```

Intervalo de Confiança no R

Exercício 10

Baseado na amostra dos imóveis (windsor_sample) e em um intervalo de 95% de confiança. Qual o preço médio dos imóveis com 3 quartos da mostra de Windsor, Canadá?

Intervalo de Confiança no R

Exercício 10

Baseado na amostra dos imóveis (windsor_sample) e em um intervalo de 95% de confiança. Qual o preço médio dos imóveis com 3 quartos da mostra de Windsor, Canadá?

```
windsor_3B <- subset(windsor_sample, windsor_sample$bed=='3')
ttest_3B <- t.test(windsor_3B$price, conf.level = 0.95)
ttest_3B$conf.in
```

```
[1] 66170.35 74041.38
```

Teste de Hipótese no R

Exercício 11

Existe diferença significativa na média das notas dos filmes com menos de 120 minutos?

Teste de Hipótese no R

Exercício 11

Existe diferença significativa na média das notas dos filmes com menos de 120 minutos?

```
# define os dados
sample <- time120$Metascore
pop <- movies$Metascore

# usando a função t-test
tstatistic_f <- t.test(time120$Metascore, mu = mean(movies$Metascore),
                       conf.level = 0.95)
tstatistic_f$statistic
```

Teste de Hipótese no R

```
# Comparando com o valor tabelado
df<-length(time120$Metascore) -1 # grau de liberdade
qt(0.975,df) # valor tabelado da distribuição t (superior)
qt(0.025,df) # valor tabelado da distribuição t (inferior)

# se rejeita a hipótese nula |estat.calculada| > valor.tabelado_superior
abs(tstatistic_f$statistic) > qt(0.975,df) # Rejeito H0?

|-2.619625| > +1.965669
[1] TRUE
```

Denise Manfredini
Doutoranda em Economia
Universidade Federal de Santa Catarina

manfredini.denise@gmail.com