# Hands-on R - Testes estatísticos básicos

*Rodrigo Heldt - rodrigoheldt@gmail.com*

*IMED - HOPEAD - Nov 13, 2017*

## Testes estatísticos básicos

### 0.1 Carregar pacotes

```
#install.packages("corrplot")
library(corrplot)
```

### 0.2 Importando a base de dados de um arquivo csv

```
base <- read.csv("C:/Users/Rodrigo/Dropbox/Hands-on R Workshop/Parte 2 - Pre-processamento dos dados/Da
                 sep = ",")

base$Educacao <- as.factor(base$Educacao)
```

### 1. Testes de diferença de médias

#### 1.1 Teste t-student

**Teste t-student: utilizado quando tem-se dois grupos (ex.: homens e mulheres) e uma variável intervalar ou de razão (ex.: altura) para a qual se quer se há ou não diferença entre as médias dos dois grupos.**

**Teste de hipótese:**

**H0 = não há diferença entre as médias dos dois grupos**

**H1 = há diferença entre as médias dos dois grupos**

##### 1.1.1. Exemplo 1: SalarioPorHora X Mulher/Homem

```
## Contar quantas observacoes de cada grupo existem usando a funcao 'table'
table(base$Mulher)
```

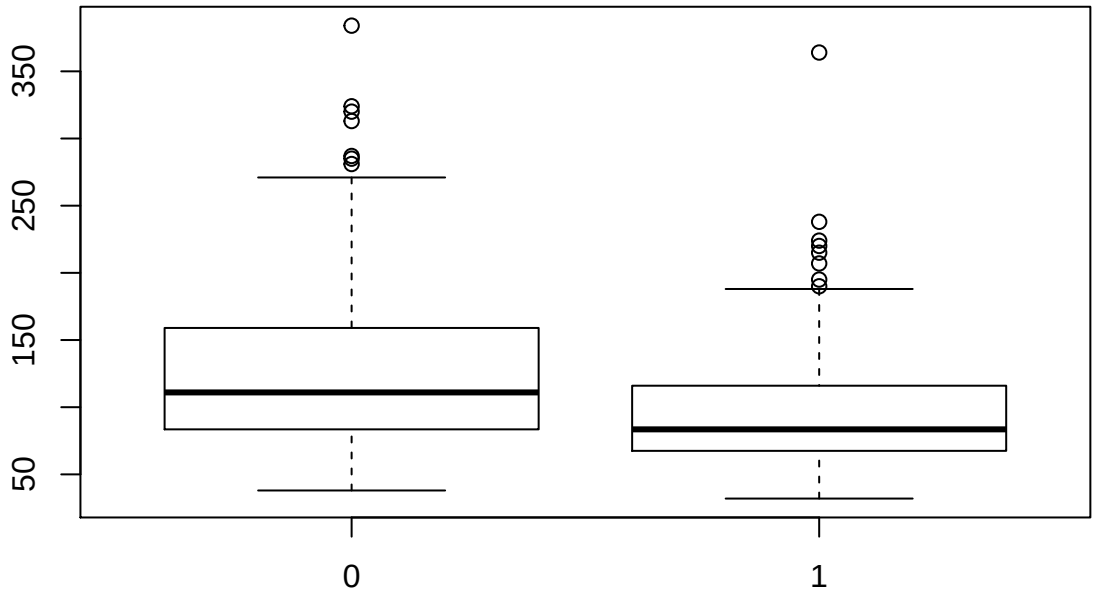
```
##
##   0   1
## 316 184
```

```
## Verificar se ha diferenca significativa entre as medias de salario dos grupos mulher e homem.
# p-value = 6.394e-09
# p-value menor do que 0.05, entao ha diferenca significativa entre as medias de salario dos
# grupos Mulher e Homem.
```

```r
t.test(base$SalarioPorHora ~ base$Mulher)
```

```
##
##  Welch Two Sample t-test
##
## data:  base$SalarioPorHora by base$Mulher
## t = 5.9204, df = 448.12, p-value = 6.394e-09
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  18.57644 37.03721
## sample estimates:
## mean in group 0 mean in group 1
##       125.13291        97.32609
```

```r
## Fazer um boxplot para visualizar as medidas descritivas entre os dois grupos
boxplot(base$SalarioPorHora ~ base$Mulher)
```

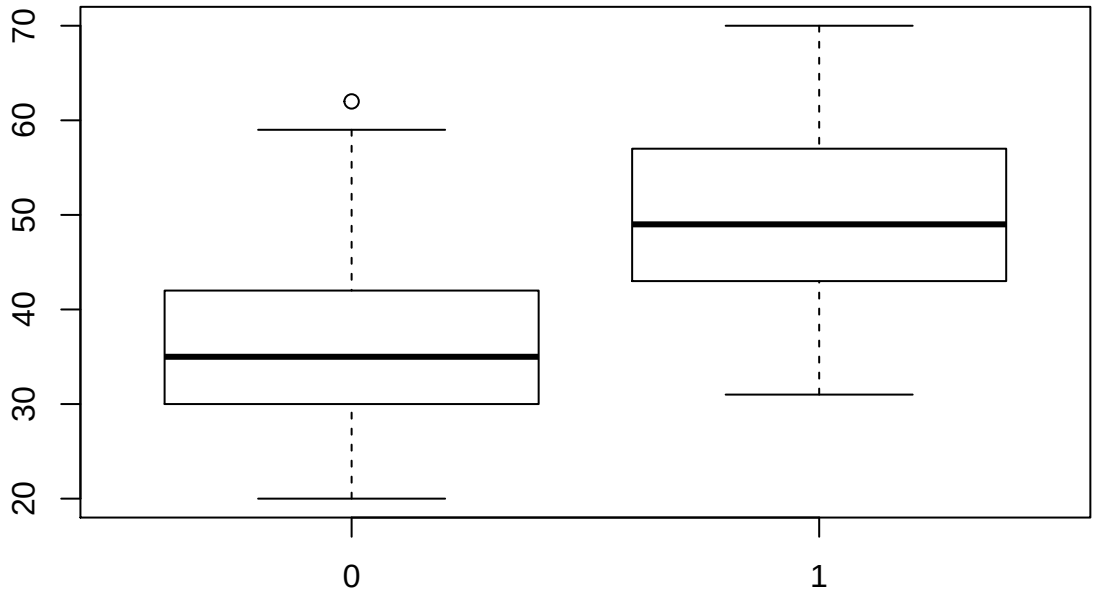



### 1.1.2. Exemplo 1: Idade X PartTime/NonPartTime job

```r
## Verificar se ha diferenca significativa entre as medias de idade dos grupos part-time e nao
## part-time job.
# p-value = 2.2e-16
# p-value menor do que 0.05, entao ha diferenca significativa entre as medias de idade dos
# grupos part-time e nao part-time job.
t.test(base$Idade ~ base$PartTime)
```

```
##
##  Welch Two Sample t-test
##
## data:  base$Idade by base$PartTime
## t = -15.201, df = 239.2, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -15.71699 -12.11072
## sample estimates:
## mean in group 0 mean in group 1
##        36.00281        49.91667
```

```r
# Fazer um boxplot para visualizar as medidas descritivas entre os dois grupos
boxplot(base$Idade ~ base$PartTime)
```




### 1.2 Análise de Variância (ANOVA)

**ANOVA: Esse teste e utilizado principalmente para medir a diferença entre as médias quando tem-se 3 ou mais grupos analisados de forma conjunta.**

**Teste de hipótese:**

**H0 = não há diferença entre as médias dos dois grupos**

**H1 = há diferença entre as médias dos dois grupos**

### 1.2.1. Exemplo 1: SalarioPorHora X Nível de educação

```
## Calucar as medias de SalarioPorHora por nivel de educacao
by(data = base$SalarioPorHora, INDICES = base$Educacao, FUN = mean)
```

```
## base$Educacao: 1
## [1] 89.32653
## ------------------------------------------------------------
## base$Educacao: 2
## [1] 101.403
## ------------------------------------------------------------
## base$Educacao: 3
## [1] 130.9048
## ------------------------------------------------------------
## base$Educacao: 4
## [1] 193.9846
```

```
## Verificar se ha diferenca significativa entre as medias de SalarioPorHora dos grupos
## entre os grupos por nivel de educacao.

# p-value = 2e-16
# p-value menor do que 0.05, entao ha diferenca significativa entre as medias
# niveis de educacao 1, 2, 3 e 4.

anova <- aov(base$SalarioPorHora ~ base$Educacao)
summary(anova)
```

```
##                Df Sum Sq Mean Sq F value Pr(>F)
## base$Educacao   3 586026  195342     104 <2e-16 ***
## Residuals     496 932027    1879
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Fazer teste post hoc para verificar entre quais grupos ha diferenca de medias
# Os resultados evidenciam que apenas a diferenca de médias entre os níveis 2 e 1 de educação
# nao e significativa
TukeyHSD(anova)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = base$SalarioPorHora ~ base$Educacao)
##
## $`base$Educacao`
##          diff        lwr       upr     p adj
## 2-1  12.07645 -0.4489782  24.60189 0.0634360
## 3-1  41.57823 28.0644538  55.09201 0.0000000
## 4-1 104.65808 88.6642818 120.65189 0.0000000
## 3-2  29.50178 14.9381983  44.06536 0.0000016
## 4-2  92.58163 75.6914743 109.47179 0.0000000
## 4-3  63.07985 45.4442902  80.71542 0.0000000
```

## 2 Análise de Correlação

### 2.1. Exemplo1: SalarioPorHora X Idade

```
# Analise de Correlacao utilizando a funcao 'cor.test()'
# cor=0.50 / p-value = 2.2e-16
# p-value < 0.05, entao significante
cor.test(base$SalarioPorHora, base$Idade)
```

```
##
##  Pearson's product-moment correlation
##
## data:  base$SalarioPorHora and base$Idade
## t = 13.044, df = 498, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.4362447 0.5672211
## sample estimates:
##       cor
## 0.5046308
```

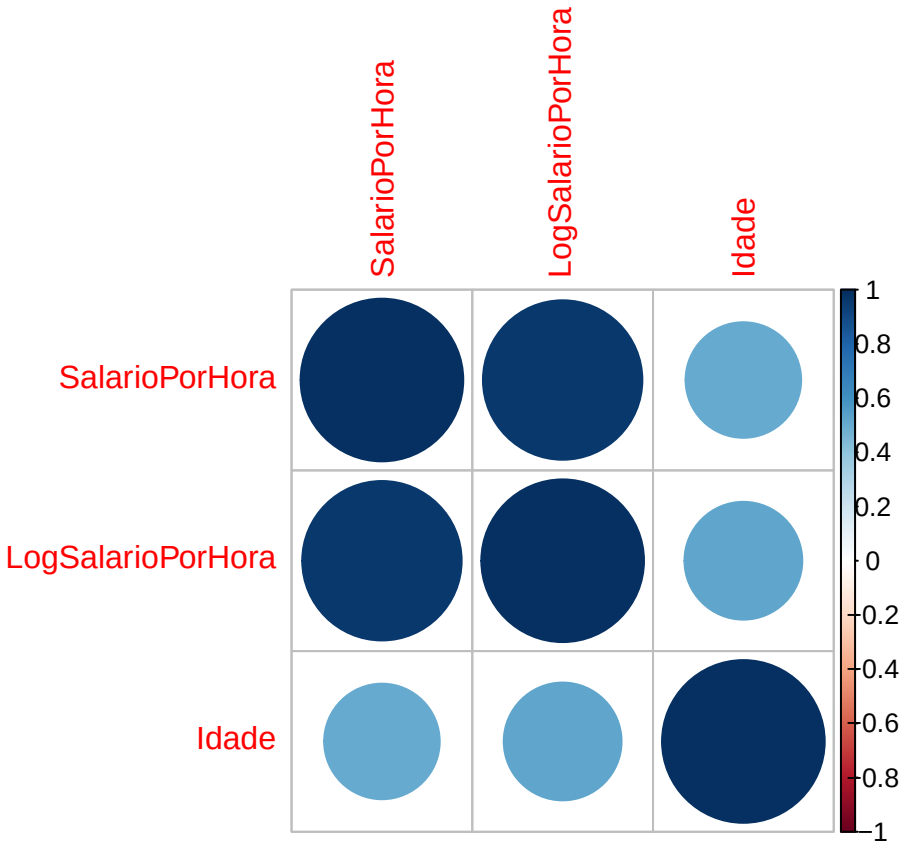
### 2.2. Exemplo2: LogSalarioPorHora X Idade

```
# cor=0.50 / p-value = 2.2e-16
# p-value < 0.05, entao significante
cor.test(base$LogSalarioPorHora, base$Idade)
```

```
##
##  Pearson's product-moment correlation
##
## data:  base$LogSalarioPorHora and base$Idade
## t = 13.742, df = 498, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.4576973 0.5851294
## sample estimates:
##      cor
## 0.524343
```

### 2.3. Plotar correlações

```
m <- cor(base[, c("SalarioPorHora","LogSalarioPorHora", "Idade")])
corrplot(m)
```




**3. Exercício**

**E1: Verifique se há diferença de médias de Idade entre os grupos Mulheres e Homens**

**E2: Verifique se há diferença de médias de Idade entre os grupos 1, 2, 3 e 4 de níveis de educação**