Oficina de Estatística Tutorial 6

Virgilio Mendes

19/07/2019

Tutorial 6

Neste tutorial, vamos trabalhar com alguns tipos de teste de hipóteses para comparar duas amostras:

1. Teste t (pareado);
2. Teste t (para amostras independentes);
3. Teste Qui-Quadrado.

Teste t pareado

Para trabalharmos com um teste t pareado, vamos simular uma nota de avaliação do prefeito de Belo Horizonte antes e depois da inauguração de uma obra. Suponhamos que esta pesquisa de opinião peça para que cada um dos 150 respondentes avaliarem o prefeito numa escala de 0 a 10, sendo 0 a pior nota e 10 a melhor nota.

Em nossa simulação, estas notas estarão em uma distribuição normal de média 6.5 e desvio-padrão 0.5 antes da inauguração da obra e em uma distribuição normal de média 7.5 e desvio-padrão 1.0 após esta inauguração.

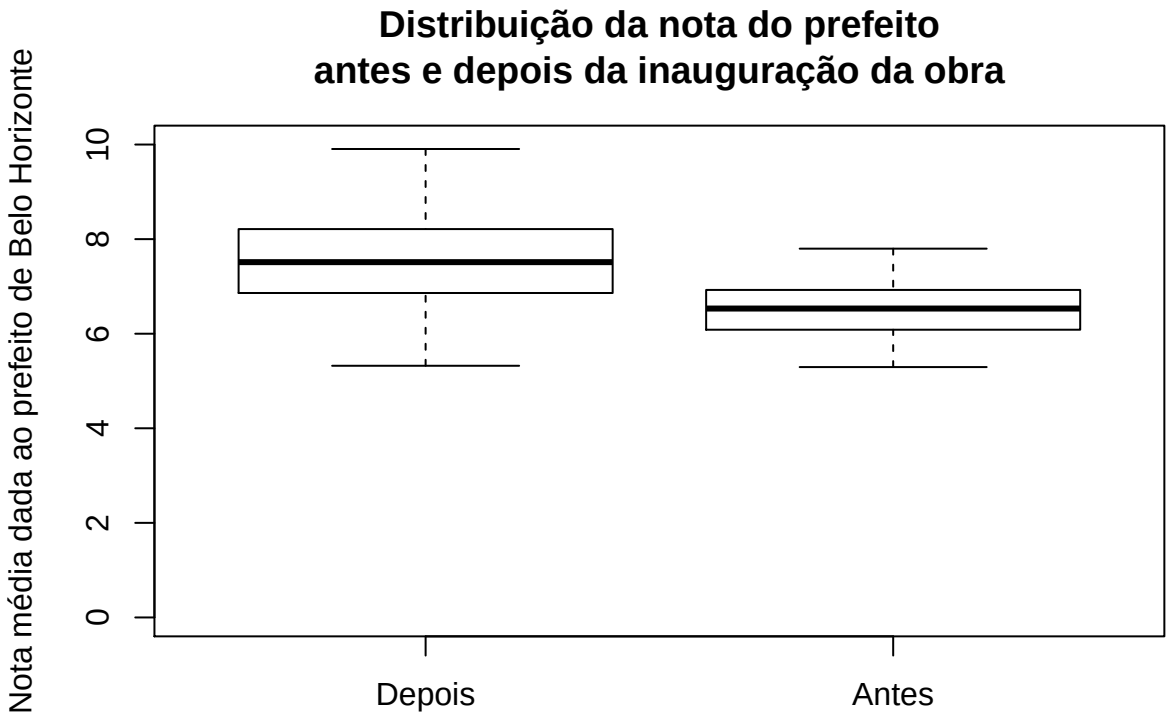
Assim como em tutoriais anteriores, vamos simular várias amostras das quais vamos sortear uma, com a qual trabalharemos. A principal diferença desta simulação para outra que fizemos é que agora nossa amostra é de fato um banco de dados e não somente um vetor. Dessa forma, precisamos criar um objeto de classe `data.frame` antes de iniciarmos o loop, para que guardemos um banco de dados em cada "prateleira" da nossa lista a cada iteração desse loop.

```
set.seed(1234)

amostras_pesquisas <- list()

n_amostra <- 150
banco_pesquisa <- data.frame(individuos = 1:150)

for (i in 1:100) {
  banco_pesquisa$antes <- rnorm(n_amostra, mean = 6.5, sd = 0.5)
  banco_pesquisa$depois <- rnorm(n_amostra, mean = 7.5, sd = 1.0)
  amostras_pesquisas[[i]] <- banco_pesquisa
}

amostra_sorteada <- sample(amostras_pesquisas, 1)[[1]]
```

Vamos então testar a H0 de que a média de avaliação do prefeito não se alterou após a inauguração da obra e uma Ha de que esta média de avaliação é maior uma vez inaugurada a obra. Para fazermos esse teste no R, também utilizamos a função `t.test`, agora especificando o argumento `paired = T`.

Antes de testarmos nossa hipótese, seria interessante visualizar a distribuição dessa nossa variável nos dois momentos analisados. Será que um box-plot nos sugeriria alguma plausibilidade para rejeitar nossa H0?

```
boxplot(amostra_sorteada$depois, amostra_sorteada$antes,
        main = "Distribuição da nota do prefeito\nantes e depois da inauguração da obra",
        names = c("Depois", "Antes"),
        ylab = "Nota média dada ao prefeito de Belo Horizonte",
        ylim = c(0,10))
```



O boxplot parece indicar a diferença entre essas médias, ainda que esses valores talvez fiquem próximos, dada a incerteza inerente à amostra.

Vamos testar então a hipótese com um nível de significância de 5%. Perceba que a ordem de inclusão dos elementos importa no teste, lembrando que nossa H0: Media_Depois = Media_Antes, e nossa Ha: Media_Depois > Media_Antes.

```
t.test(amostra_sorteada$depois, amostra_sorteada$antes,
       paired = T, conf.level = 0.95, alternative = "greater")
```

```
Paired t-test
```

data: amostra_sorteada*depoisandamostra*_s*orteada*antes t = 11.364, df = 149, p-value < 2.2e-16 alternative hypothesis: true difference in means is greater than 0 95 percent confidence interval: 0.8817737 Inf sample estimates: mean of the differences 1.032104

Nosso teste de hipótese nos mostra que podemos rejeitar a hipótese nula de que as médias não são diferentes, em favor da Ha de que a média após a inauguração da obra é maior, com um nível de significância de 0.05 (afinal, nosso p-valor é bastante baixo). Isso também pode ser verificado ao observamos a estimativa para a média dessas diferenças (1.03), além do limite inferior do intervalo de confiança calculado para essa estimativa (0.88).

Teste t para amostras independentes

A partir desse momento, vamos trabalhar com uma nova base de dados. Trata-se de um recorte da MUNIC de 2015. A MUNIC é uma pesquisa realizada pelo IBGE que apresenta um perfil dos municípios brasileiros. Nosso recorte inclui variáveis de recursos humanos, terceirização e informatização das cidades no país. Vamos então selecionar uma amostra estratificada por regiões do país de 5% do total de municípios no Brasil.

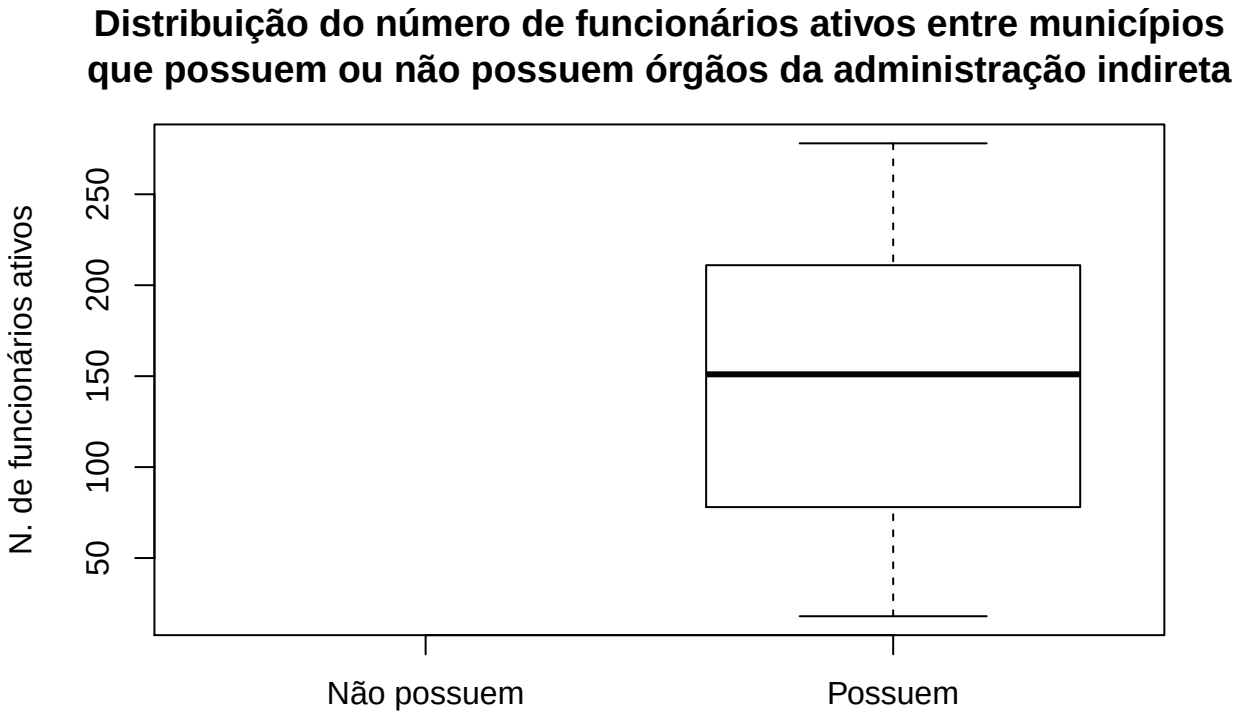
```
set.seed(1234)

library(splitstackshape)

munic_2015 <- read.csv2("https://raw.githubusercontent.com/lgelape/modus_2019/master/Bancos/munic2015_m

amostra_munic <- stratified(munic_2015, group = "regiao", size = 0.05)
```

O teste t para amostras independentes é adequado para testarmos a diferença entre alguma variável quantitativa em dois grupos. Para este tutorial, vamos testar se, em nossa amostra, há uma diferença entre a média do número de funcionários ativos na Administração Direta (variável `munic_2015$direta_ativos`) entre dois grupos de cidade: um que possui e outro que não possui órgãos da administração indireta (variável `munic_2015$adm_indireta`).

Sendo assim, nosso teste tem em H0: as médias são iguais entre esses dois grupos; e em Ha que as médias são diferentes (ou seja, é um teste bicaudal).

O que sugere um boxplot dessas duas variáveis em nossa amostra?

```
boxplot(amostra_munic[amostra_munic$direta_ativos,
                      which(amostra_munic$adm_indireta == "Não")],
        amostra_munic[amostra_munic$direta_ativos,
                      which(amostra_munic$adm_indireta == "Sim")],
        main = "Distribuição do número de funcionários ativos entre municípios\nque possuem ou não possu
        names = c("Não possuem", "Possuem"),
        ylab = "N. de funcionários ativos")
```



Vamos agora testar nossa hipótese sob um nível de significância de 0.01.

```
# Antes, precisamos testar o pressuposto de mesma variância das
# nossas duas "amostras"
var.test(direta_ativos ~ adm_indireta, data = amostra_munic)
```

```
F test to compare two variances
```

data: direta_ativos by adm_indireta F = 0.072277, num df = 226, denom df = 49, p-value < 2.2e-16 alternative hypothesis: true ratio of variances is not equal to 1 95 percent confidence interval: 0.04506778 0.10883678 sample estimates: ratio of variances 0.0722769

```
# Podemos rejeitar H0 de que as variancias sao iguais.
# Entao precisamos especificar um argumento do nosso teste

t.test(direta_ativos ~ adm_indireta, data = amostra_munic,
       conf.level = 0.99, var.equal = F)
```

```
Welch Two Sample t-test
```

data: direta_ativos by adm_indireta t = -3.9104, df = 50.57, p-value = 0.0002753 alternative hypothesis: true difference in means is not equal to 0 99 percent confidence interval: -2618.2459 -490.4223 sample estimates: mean in group NÃ£o mean in group Sim 698.5859 2252.9200

Nosso teste de hipótese nos mostra que podemos rejeitar a hipótese nula de que as médias são iguais, com um nível de significância de 0.01 (afinal, nosso p-valor é baixo, sendo igual a 0.002753). Isso também pode ser verificado ao observamos o valor reportado para o intervalo de confiança da diferença entre essas médias, que não absorve o valor de 0.

Teste Qui-Quadrado

Por fim, temos o teste qui-quadrado, no qual testamos a associação entre duas variáveis categóricas. Utilizando a MUNIC de 2015, vamos testar se existe associação entre a ocorrência de contratação de uma assessoria jurídica (`amostra_munic$assessoria_juridica`) e a existência de sistemas informatizados de execução orçamentária (`amostra_munic$folha_pagamento`).

A hipótese nula de um teste qui-quadrado é a de que as variáveis em análise são independentes na população, enquanto Ha é a de que elas não são independentes.

A função que utilizamos no R para tanto é a `chisq.test`, que funciona de maneira bastante simples.

```
chisq.test(amostra_munic$assessoria_juridica,
           amostra_munic$execucao_orcamentaria)
```

```
Pearson's Chi-squared test with Yates' continuity correction
```

data: amostra_munic$assessoria_juridica and amostra_munic$execucao_orcamentaria X-squared = 1.0329e-29, df = 1, p-value = 1

De imediato não podemos rejeitar a H0 de que essas variáveis são independentes (basta olhar para o nosso p-valor).

Além disso, podemos guardar diversas informações relacionadas a um teste qui-quadrado num objeto, para podermos acessá-los posteriormente. Para além da estatística de teste, p-valor e nível de significância, conseguimos também tabelas das frequências esperada e observada, além das tabelas de resíduos e resíduos padronizados.

```
teste_quiquadrado <- chisq.test(amostra_munic$assessoria_juridica,
                                amostra_munic$execucao_orcamentaria)

# Estatistica-teste
teste_quiquadrado$statistic
```

X-squared 1.032864e-29

```
# p-valor
teste_quiquadrado$p.value
```

[1] 1

```
# Tabela de valores observados
teste_quiquadrado$observed
```

```
   NÃ£o Sim
```

NÃ£o 10 66 Sim 22 140

```r
# Tabela de valores esperados
teste_quiquadrado$expected
```

```
       NÃ£o       Sim
```

NÃ£o 10.21849 65.78151 Sim 21.78151 140.21849

```r
# Tabela de residuos
teste_quiquadrado$residuals
```

```
          NÃ£o         Sim
```

NÃ£o -0.06834914 0.02693857 Sim 0.04681471 -0.01845116

```r
# Tabela de residuos padronizados
teste_quiquadrado$stdres
```

```
          NÃ£o         Sim
```

NÃ£o -0.08904692 0.08904692 Sim 0.08904692 -0.08904692