Minicurso

Luiz Santos

26/08/2019

Análise Exploratória e Teste de Hipótese no R

1 Análise Exploratória ou Descritiva

A análise estatistica descritiva é um ramo da estatística que aplica várias técnicas para descrever e sumarizar um conjunto de dados. Ela é composta por medidas de tendência central e de dispersão. No R, as funções summary() e boxplot() apresentam um bom resumo dessa análise.

1.1 Medidas de Tendência Central e Separatrizes

Medidas de tendência central ou promédios são valores que servem para representar a distribuição como um todo, além de possibilitarem o confronto entre distribuições. Das principais medidas de tendência central destacamos aqui a média aritmética e a mediana.

Separatrizes: são os valores da distribuição nomeados por suas posições na série ordenada. Exemplos:

1.1.1 Média Aritmética

A média aritmética $(\bar{X})$ é o quociente entre a soma dos valores do conjunto e o número total de valores.

$$(\bar{X}) = \frac{soma\ dos\ valores\ do\ conjunto}{número\ total\ de\ valores}$$

Exemplo 1:

Suponhamos que os números de questões respondidas corretamente (nqrc) por 20 alunos de ciências atuariais em uma prova de Introdução às Ciências Atuariais foram os seguintes:

$$nqrc = (7, 6, 7, 6, 7, 4, 5, 7, 5, 8, 6, 5, 5, 7, 8, 4, 7, 7, 7, 6).$$

A média da turma será:

$$(\bar{X}) = \frac{7+6+7+6+7+4+5+7+5+8+6+5+5+7+8+4+7+7+7+6}{20} = 6,2$$

Alternativamente, no R:

```
nqrc=c(7,6,7,6,7,4,5,7,5,8,6,5,5,7,8,4,7,7,7,6)
mean(nqrc)
```

```
## [1] 6.2
```

1.1.2 Mediana

A mediana é a medida de tendência central que divide a distribuição em duas partes iguais, ou seja, é o valor que fica no meio da série ordenada.

Em geral, usa-se o seguinte procedimento para determinar o elemento mediano:

i. Se o número de observações N é ímpar, então $Emd = \frac{N+1}{2}$, sendo Emd o elemento mediano;

ii. Se o número de observações N é par, então $Emd = \frac{N}{2}$, e, neste caso, a mediana é igual à média aritmética dos dois valores centrais.

Exemplo 2:

Suponha uma empresa que observa os seguintes fluxos de caixa mensais ao longo de de 5 meses: R$ 1.000,00, R$ 3.000,00, R$ 2.500,00, R$ 2.700,00 e R$ 3.100,00. Aqui, Emd= (N+1)/2 = 6/2 = 3. Logo, ordenando os valores obtemos que o fluxo de caixa mediano da empresa é R$ 2.700,00.

Alternativamente, no R:

```
fluxo=c(1000,2500,2700,3000,3100)
median(fluxo)
```

```
## [1] 2700
```

1.1.3 Quartis

São medidas que dividem os dados ordenados em quatro partes iguais: primeiro quartil (Q1), segundo quartil (Q2) e terceiro quartil (Q3).

Pode-se dizer que 25% dos valores estão abaixo de Q1 e 75% dos valores estão acima de Q1. A diferença entre Q3 e Q1 é chamada de amplitude interquartílica. O segundo quartil é exatamente igual à mediana.

$$Q_1 = x_{\frac{n}{4}}$$

$$Q_2 = x_{\frac{2n}{4}}$$

$$Q_3 = x_{\frac{3n}{4}}$$

Exemplo 3:

Considere as seguintes observações, em que a variável y representa a idade de beneficiários num plano de saúde e n representa o número de elementos observados:

$$y = (7, 16, 18, 18, 19, 20, 20, 22, 31, 34, 38, 58)$$

$$n = 12$$

$$Q_1 = x_{\frac{n}{4}} = x_{\frac{12}{4}} = x_3 = 18$$

$$Q_2 = x_{\frac{3n}{4}} = x_{\frac{2\times12}{4}} = x_6 = 20$$

$$Q_3 = x_{\frac{3n}{4}} = x_{\frac{3\times12}{4}} = x_9 = 31$$

Alternativamente, no R:

```
y=c(7,16,18,18,19,20,20,22,31,34,38,58)
quantile(y,type=4)
```

```
##   0%  25%  50%  75% 100%
##    7   18   20   31   58
```

```
#Ou por
summary(y)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    7.00   18.00   20.00   25.08   31.75   58.00
```

1.1.4 Percentis ou centis

São medidas que dividem os dados ordenados em 100 partes iguais. Em uma amostra, são possíveis de serem calculados 99 percentis.

O k-ésimo percentil, denotado por P_k, é o valor na posição, de forma que k das medidas são menores que a posição P_k, ou seja, (100-k)% das observações são maiores que P_k. O k-ésimo percentil é determinado por:

P_k = é o valor do kn/100 −ésimo termo no conjunto de dados ordenados, onde k é o percentil e n é o tamanho da amostra.

Exemplo 4:

Considere os dados do Exemplo 3. O percentil 62 (P62) é dado por

$$P_62 = \frac{kn}{100} = \frac{62 \times 12}{100} = 7,44$$

Nesse caso, o percentil 62 está entre os números nas posições 7 e 8. Pode-se obter esse percentil interpolando-se as observações nestas posições:

$$P_62 = y_7 + 0,44(y_8 - y_7) = 20 + 0.44(22 - 21) = 20,88$$

No R, a função quantile() pode ser utilizada para obtenção do percentil P_62:

```
quantile(y,.62,type=4)
```

```
##   62%
## 20.88
```

1.2 Medidas de dispersão

Podemos definir variabilidade de um conjunto de dados como sendo a maior ou menor diversificação dos valores em torno de uma medida de tendência central. Considerando, por exemplo, as notas de dois alunos A e B, em cinco disciplinas diferentes:

1.2.1 Desvio padrão

É a raiz quadrada da média dos quadrados dos desvios, tomados em relação à média aritmética.

Noutros termos, o desvio padrão fornece uma medida de dispersão das observações ao redor da média. Um desvio padrão pequeno indica que os dados possuem uma amplitude pequena ao redor da média. Já um desvio padrão grande, indica que os dados possuem uma amplitude grande ao redor da média.

O desvio padrão é fornecido sempre na mesma escala da variável resposta e é obtido pela raiz quadrada da variância.

$$S = \sqrt{\frac{\sum_{i=1}^{n}(X - \bar{x})^2)}{n-1}}$$

Exemplo 5:

Considerando os dados do Exemplo 1,

```
nqrc=c(7,6,7,6,7,4,5,7,5,8,6,5,5,7,8,4,7,7,7,6)
sd(nqrc)
```

```
## [1] 1.196486
```

1.2.2 Variância

É o quadrado do desvio padrão.

$$S = \frac{\sum_{i=1}^{n}(X - \bar{x})^2)}{n-1}$$

Exemplo 6:

Considerando os dados do Exemplo 1:

```
nqrc=c(7,6,7,6,7,4,5,7,5,8,6,5,5,7,8,4,7,7,7,6)
var(nqrc)
```

```
## [1] 1.431579
```

1.2.3 Coeficiente de variação

O coeficiente de variação expressa o desvio padrão como percentual da média.

O CV fornece uma idéia de precisão experimental: quanto menor o CV, menor a variabilidade e melhor a precisão experimental. Por outro lado, quanto maior o CV, maior será a variabilidade experimental e pior será a precisão experimental.

O CV de variação é extremamente afetado pela escala da variável resposta. Por esse motivo ele é, em geral, apenas um bom indicador para comparar variáveis semelhantes.

$$CV = \frac{s}{\bar{x}}$$

Exemplo 7:

Considerando os dados do Exemplo 1:

```
nqrc=c(7,6,7,6,7,4,5,7,5,8,6,5,5,7,8,4,7,7,7,6)
cv=sd(nqrc)/mean(nqrc)
cv
```

```
## [1] 0.1929816
```

1.3 Abordagem Gráfica

As variáveis, tanto qualitativas quanto quantitativas, podem ser resumidas em tabelas e gráficos. Para cada variável, existem maneiras mais adequadas de representação dos dados. Veremos algumas nesse curso.

Para variáveis qualitativas, em geral são utilizadas tabelas de frequências para representar as frequências de cada categoria. Para as mesmas variáveis, podem ser utilizados gráficos como o gráfico de barras e de setores (pizza).

Para variáveis quantitativas, podem-se ser utilizadas tabelas de frequências para representar a ocorrência de valores em classes prée-estabelecidas. Também podem ser utilizados gráficos como o histograma ou ramo e folhas.

1.3.1 Gráfico de barra

Variáveis qualitativas podem ser representadas por gráficos, tais como o de barras e o de setores (pizza).

Para obter um gráfico de barras no R utilize o seguinte procedimento:

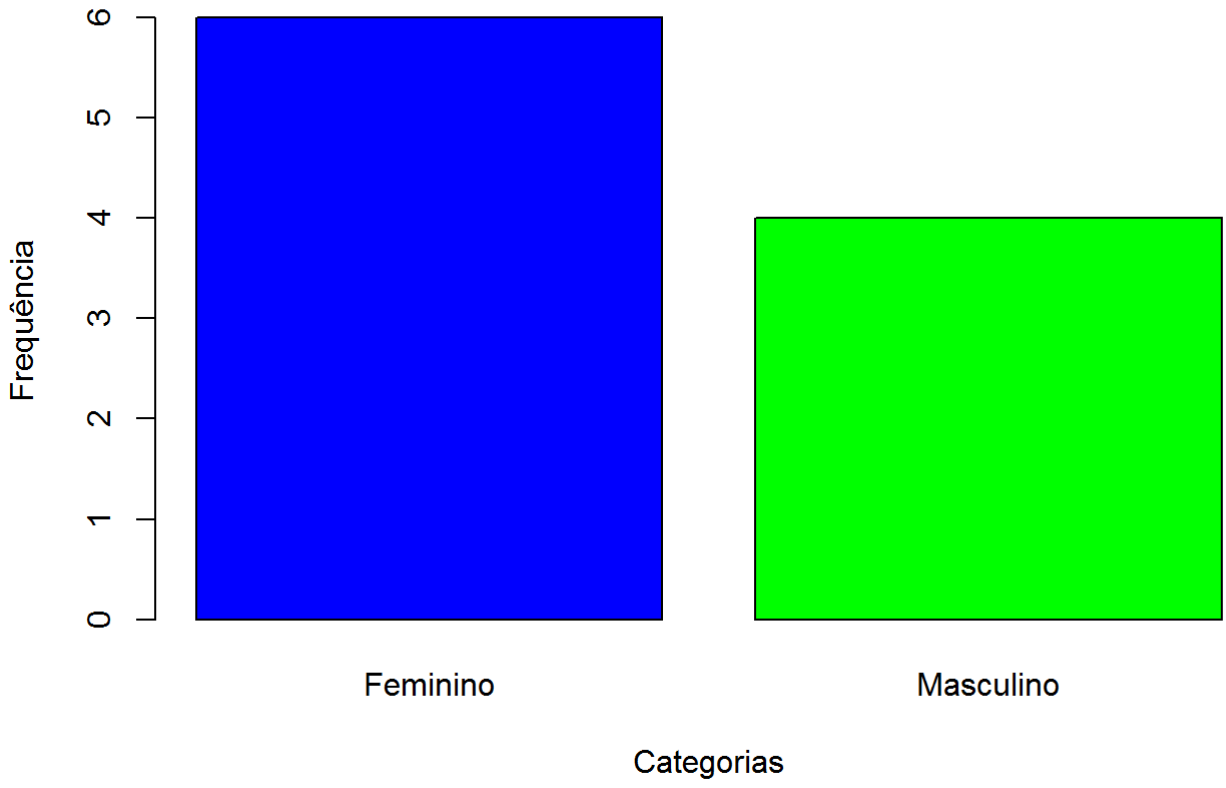
Exemplo 8:

Suponha que você entrevistou um grupo de clientes que apresentou o seguitne vetor de sexo:

$$sexo = (F, M, F, F, M, F, F, F, M, M)$$

Se for necessário que você sintetize essa informação num gráfico em barras, é possível fazê-lo no R por meio do comando:

```
x=c(6,4)
barplot(x,ylab="Frequência",xlab="Categorias", names=c("Feminino","Masculino"),col=c("blue",
"green"))
```



1.3.2 Gráfico de Setores (pizza)

Um gráfico de setores é um círculo dividido em partes que representam as frequências relativas ou percentagens de cada classe ou categoria.Para obtê-lo, use:

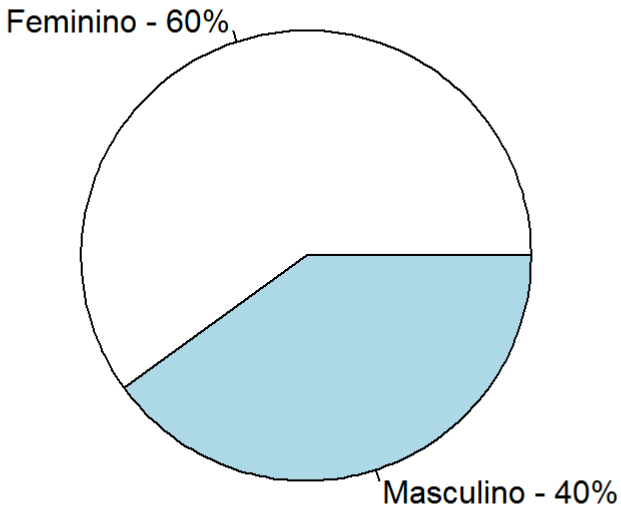
Exemplo 9:

A partir do Exemplo 8, sintetiza-se o gráfico em pizza por meio do código:

```
x=c(6,4)
names(x)<-c("Feminino - 60%","Masculino - 40%")
table(x)
```

```
## x
## 4 6
## 1 1
```

```
xp<-prop.table(x)*100
pie(xp,labels=names(x))
```



1.3.3 Histograma

Histograma é um gráfico que representa a distribuição de frequência absoluta, relativa ou percentual. Observe que as barras estão juntas. Isso ocorre porque um histograma é utilizado para representar uma variáavel quantitativa contínua.

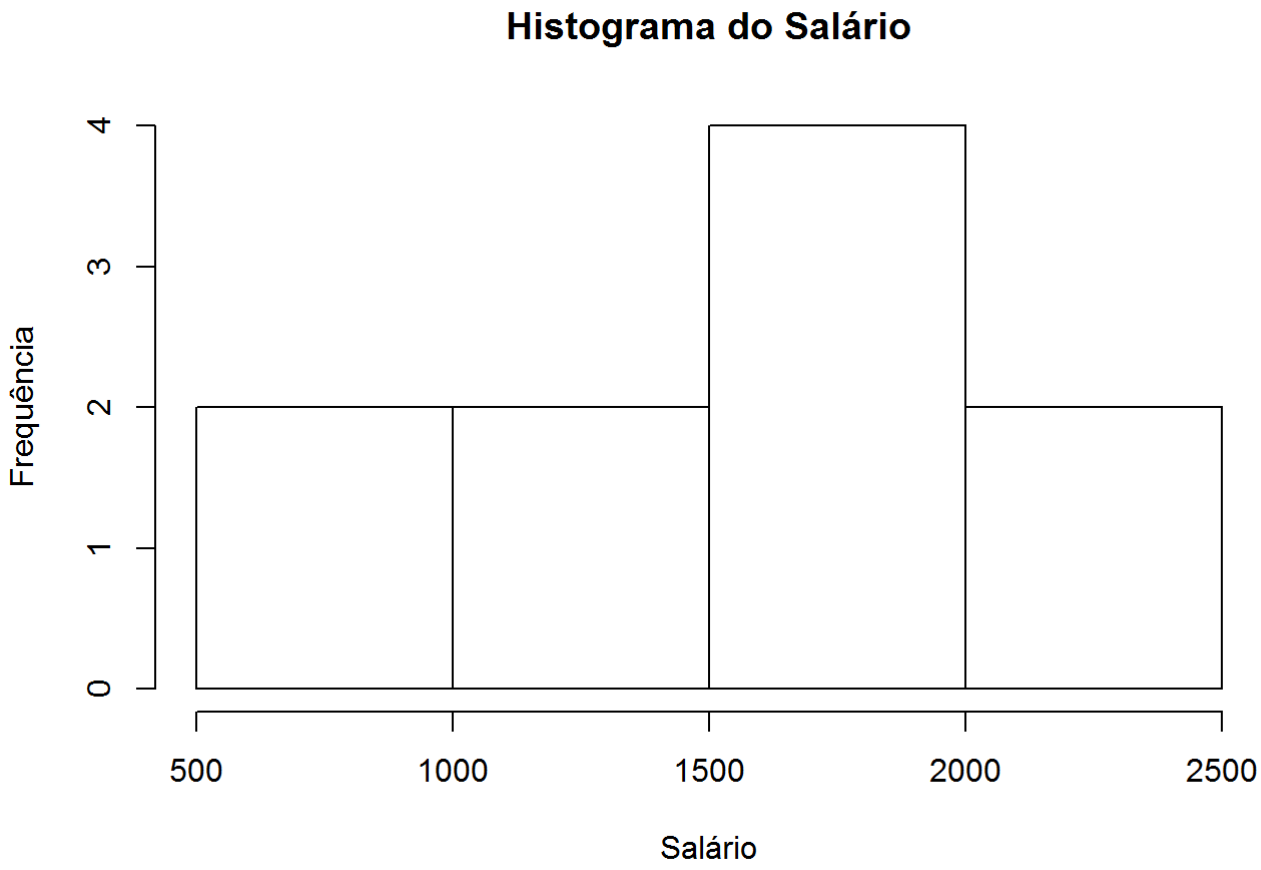
Exemplo 10:

Uma empresa, que emprega 10 pessoas, apresenta o seguinte vetor de salários (em reais):

$$Salário = (1000, 2000, 1100, 1550, 2300, 1980, 1875, 2500, 900, 1340)$$

O histograma referente a esses dados é criado a partir de:

```
salario=c(1000,2000,1100,1550,2300,1980,1875,2500,990,1340)
hist(salario,ylab="Frequência",xlab="Salário",main="Histograma do Salário")
```



1.3.3 Box plot

O box-plot é um gráfico que verifica a distribuição dos dados: mostra a posição central (média e mediana), a dispersão (amplitude), a simetria dos dados de uma amostra e a presença de outliers.

Em um box plot são apresentados 5 estatísticas: o mínimo, o primeiro quartil (Q1), a mediana, o terceiro quartil (Q3) e o máximo. Esse valores também são chamados de resumo dos cinco números.

Para construir um box plot desenha-se um retângulo alinhado verticalmente (ou horizontalmente) com duas semirretas, uma em cada um dos lados opostos do retângulo. A altura do retângulo é definida pelos quartis Q1 e Q3. Uma linha secciona o retângulo no valor da mediana (ou Q2). As semirretas ligam respectivamente os quartis Q1 e Q3 ao valor mínimo e ao máximo do conjunto de dados. Confira no exemplo o Box Plot “desenhado” com as estatísticas do resumo de cinco pontos.

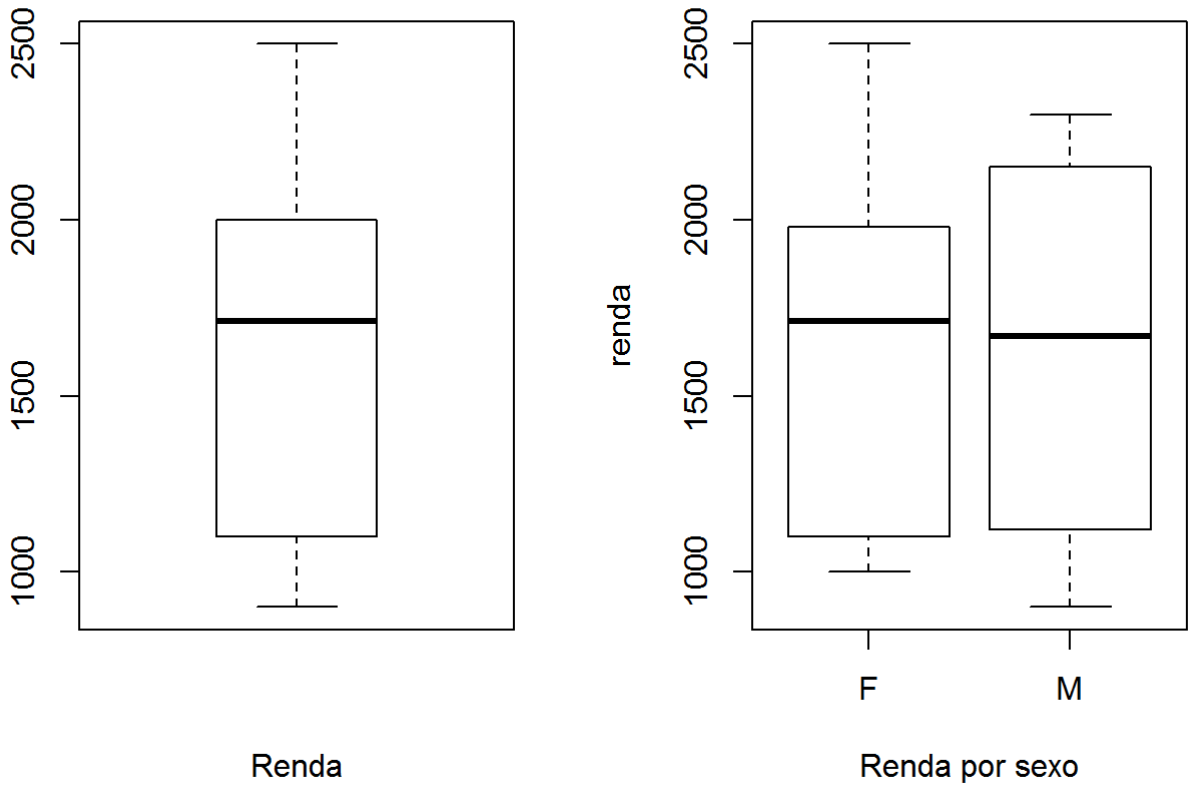
O retângulo contém 50% dos valores do conjunto de dados. A posição da linha mediana no retângulo informa sobre a assimetria da distribuição. Uma distribuição simétrica teria a mediana no centro do retângulo. Se a mediana é próxima de Q1, então, os dados são positivamente assimétricos. Se a mediana é próxima de Q3 os dados são negativamente assimétricos.

Os outliers em um box plot aparecem como pontos ou asteriscos fora das “linhas” desenhadas.

Exemplo 11:

A partir dos dados do Exemplo 10, fabricam-se dois box plots para a renda, uma para a totalidade dos dados, outra desmembrando-se por sexo:

```
renda=c(1000,2000,1100,1550,2300,1980,1875,2500,900,1340)
sexo=c("F","M","F","F","M","F","F","F","M","M")
par(mfrow=c(1,2))
boxplot(renda,xlab="Renda")
boxplot(renda~sexo,xlab="Renda por sexo")
```



2 Teste de Hipótese no R

2.1 Teste paramétrico de comparação de variância entre duas populações

Considere-se uma amostra de tamanho n_1 de uma população normal, com variância desconhecida σ_1^2, e outra de tamanho n_2, também de uma população normal e com variância desconhecida σ_2^2, sendo essas amostras de populações independentes. Para testar H_0: $\sigma_1^2 = \sigma_2^2$ contra a alternativa H_1: $\sigma_1^2 \neq \sigma_1^2$, tem-se que a estatística do teste, supondo H_0 verdadeira, é:

$$F = \frac{S_1^2}{S_2^2}$$

cuja distribuição se mostra como sendo uma $F(n_1 - 1; n_2 - 1)$, em que s_1^2 e s_2^2 são as respectivas variâncias amostrais.

A região crítica nesse caso é da forma:

$$RC = \{y |\ y < x_1 \ ou\ y > x_2\}$$

em que x_1 e x_2 são obtidos de maneira que

$$P(F < x_1) = P(F > x_2) = \alpha/2,$$

sendo α o nível de significância do teste.

Exemplo 11:

Certo pesquisador estava interessado em investigar se dois grupos (grupo A e B) tinham o mesmo grau de homogeneidade quanto à tendência para realizar investimentos. Para isso, o pesquisador pegou uma amostra de cinco pessoas do grupo A e sete do grupo B,submetendo-as a um teste psicológico que fornece informações numéricas relacionadas com a tendência para investir, obtendo os seguintes resultados:

GrupoA: 1 2 1 2 3

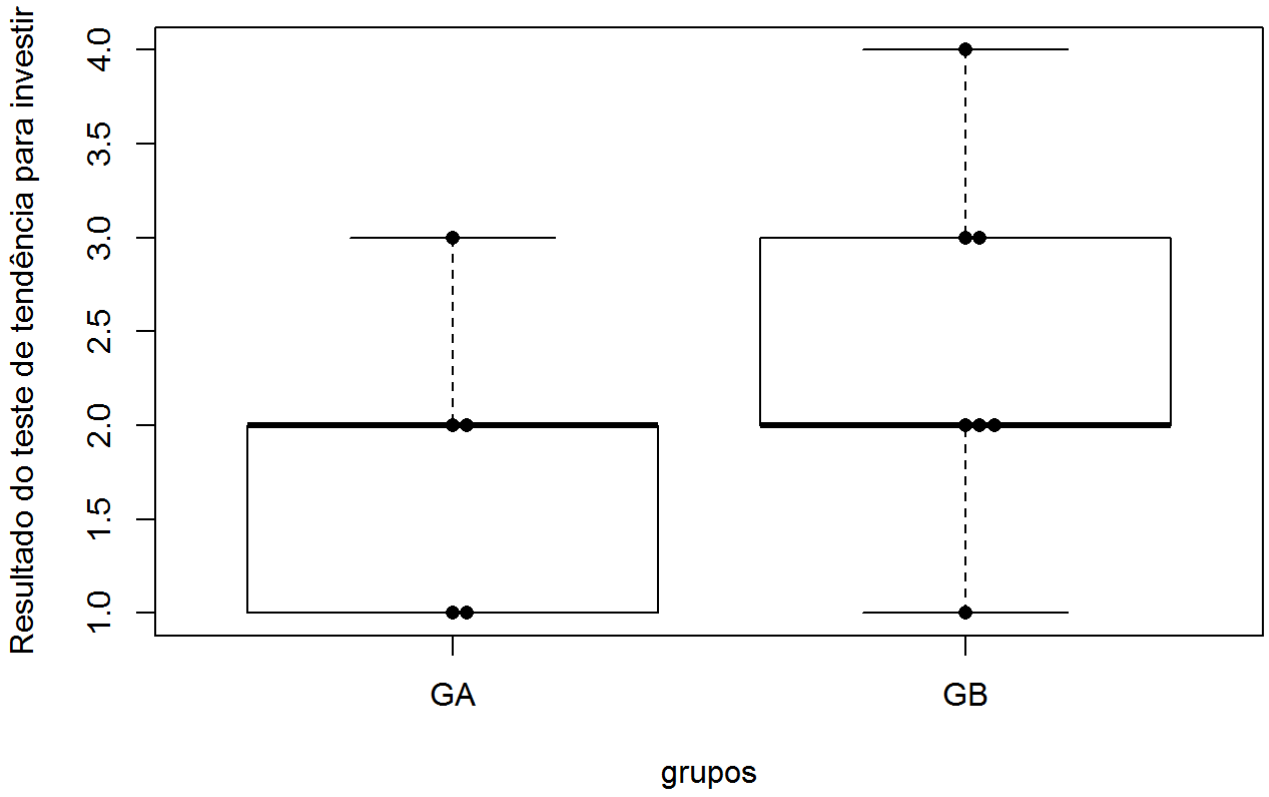
GrupoB: 4 1 3 3 2 2 2

Considerando que as informações numéricas relacionadas com a tendência para investir são normalmente distribuídas, pode-se acreditar, ao nível de 10%, que os grupos 1 e 2 têm o mesmo grau de homogeneidade quanto à tendência para investir?

```
#Informando os dados coletados
GA<-c(1,2,1,2,3)
GB<-c(4,1,3,3,2,2,2)
grupos<-c(rep(c("GA","GB"),c(5,7)))
```

Inicialmente, realiza-se uma breve descritiva. Com essa análise, tem-se a possibilidade de identificar padrões que podem ser interessantes e que venham a contribuir para uma melhor compreensão do fenômeno em estudo.

```
boxplot(c(GA,GB)~grupos,ylab="Resultado do teste de tendência para investir",xlab="grupos")
stripchart(c(GA,GB)~grupos,method="stack",pch=16,offset=.5,vertical=T,add=T)
```



Pela visualização do gráfico, parece haver diferença entre as variâncias. Mas esse indício precisar ser testado. Manualmente, a estatística F pode ser calculada pelo razão das variâncias do grupos, ou seja,

```
a=var(GA)
a
```

```
## [1] 0.7
```

```
b=var(GB)
b
```

```
## [1] 0.952381
```

```
a/b
```

```
## [1] 0.735
```

De forma automática, o procedimento pode ser feito por meio do código:

```
var.test(GA,GB)
```

```
##
##  F test to compare two variances
##
## data:  GA and GB
## F = 0.735, num df = 4, denom df = 6, p-value = 0.7989
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.1180313 6.7600236
## sample estimates:
## ratio of variances
##              0.735
```

Como o valor p = 0,7989 é maior que = 0,1, não se rejeita a hipótese nula de variâncias iguais.

2.2 Teste paramétrico de comparação de média entre duas populações

2.2.1 Duas populações independentes

Ver-se-á aqui somente os casos em que as variâncias das populações são desconhecidas, pois, na realidade, essas são as situações mais frequentemente encontradas nos problemas práticos.

i. As variâncias são desconhecidas, porém supostamente iguais:

Sejam $X_1, X_2, \ldots, X_n$ uma amostra aleatória de uma normal com média μ_x e variância desconhecida σ^2 e $Y_1, Y_2, \ldots, Y_m$ uma amostra aleatória de uma normal com média μ_y e variância também igual a σ^2, sendo essas duas amostras de populações independentes. Ao se fazer:

$$S_p^2 = \frac{(n-1)S_x^2 + (m-1)S_y^2}{n+m+2}$$

pode-se mostrar que a estatística:

$$t = \frac{(\bar{X} - \bar{Y}) - (\mu_x - \mu_y)}{S_p \times \sqrt{\frac{1}{n} + \frac{1}{m}}}$$

tem distribuição t de Student, com $n + m - 2$ graus de liberdade. De acordo com esse resultado, tem-se que a estatística a ser usada para testar $H_0 : \mu_x - \mu_y = 0$, supondo essa hipótese verdadeira, é:

$$t = \frac{(\bar{X} - \bar{Y})}{S_p \times \sqrt{\frac{1}{n} + \frac{1}{m}}}$$

Exemplo 12:

Caso se considere o exemplo 11, pode-se acreditar, ao nível de 5%, que o grupo 2 tem maior tendência para investir que o grupo 1?

a. Pretende-se agora testar se, em média, o investimento do grupo 2 é maior que o investimento do grupo 1, ou seja:

$H_0 : \mu_y = \mu_x$ ou $\mu_y - \mu_x = 0$

$H_1 : \mu_y > \mu_x$ ou $\mu_y - \mu_x > 0$

sendo μ_x e μ_y as tendências médias para investir dos grupos 1 e 2, respectivamente, conforme é definido no exemplo 1.

b. Pelas condições do experimento, tem-se que a estatística do teste, supondo H_0 verdadeira, é:

$$t = \frac{(\bar{X} - \bar{Y})}{S_p \times \sqrt{\frac{1}{5} + \frac{1}{7}}}$$

sendo

$$S_p^2 = \frac{4s_x^2 + 6s_y^2}{10}$$

c. Ao se fixar $\alpha = 5$ e sendo esse um teste unilateral à direita, da tabela da distribuição t de Student, obtém-se que o valor y, tal que P(t>y|v=10)=0,05 é 1,812 , ou seja, aqui a região crítica é dada por:

$$RC = y|\ y > 1,812$$

De acordo com os cálculos feitos no exemplo 11, ter-se-á:

```
sp2=(4*a+6*b)/10
sp2
```

```
## [1] 0.8514286
```

```
c=mean(GA)
c
```

```
## [1] 1.8
```

```
d=mean(GB)
d
```

```
## [1] 2.428571
```

```
t=(c-d)/(sqrt(sp2*((1/5)+(1/7))))
t
```

```
## [1] -1.163386
```

Automaticamente, o valor t pode ser calculado por meio do comando

```
t.test(GA,GB,var.eq=T,alternative="greater")
```

```
##
##  Two Sample t-test
##
## data:  GA and GB
## t = -1.1634, df = 10, p-value = 0.8642
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -1.607835       Inf
## sample estimates:
## mean of x mean of y
##  1.800000  2.428571
```

Como valor p = 0,8642 é maior que = 0,05, não existem evidências que levem a rejeitar a hipótese de que, em média, os grupos se difiram quanto à inclinação para o investimento.

ii. As variâncias são desconhecidas e supostamente desiguais:

Sejam $X_1, X_2, \ldots, X_n$ e $Y_1, Y_2, \ldots, Y_m$ amostras aleatórias de distribuições normais independentes, com médias μ_x e μ_y e variâncias σ_x^2 e σ_y^2, sendo $\sigma_x^2 \neq \sigma_y^2$. Para testar $H_0 : \mu_x = \mu_y$, tem-se, nesse caso, que a estatística do teste, supondo H0 verdadeira, é:

$$t = \frac{(\bar{X} - \bar{Y})}{\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}}$$

Cuja distribuição pode-se provar que é aproximadamente uma t de Student, com graus de liberdade dados por:

$$\nu = \frac{(\frac{S_x^2}{n} + \frac{S_y^2}{m})^2}{\frac{(\frac{S_x^2}{n})^2}{n-1} + \frac{(\frac{S_y^2}{m})^2}{m-1}}$$

Exemplo 13:

Dois grupos constituídos por 15 alunos, cada um, obtiveram em um teste de inteligência as seguintes estatísticas: o grupo 1 obteve média 23,4 e variância 8672 e o grupo 2, 11 e 2211, respectivamente. Ao nível de 1%, esses dois grupos têm, em média, a mesma inteligência?

Ao se considerarem μ_1 e μ_2 como as médias das populações de onde foram retirados os grupos 1 e 2, tem-se as seguintes hipóteses:

$H_0 : \mu_1 = \mu_2$ ou $\mu_1 - \mu_2 = 0$

$H_1 : \mu_1 \neq \mu_2$ ou $\mu_1 - \mu_2 \neq 0$

Supondo H_0 verdadeira, e de acordo com as considerações feitas por esse experimento, tem-se que o valor da estatística do teste, para os dados desta amostra, é:

$$t = \frac{(23,4 - 11)}{\sqrt{\frac{8672}{15} + \frac{2211}{15}}} = 0,46$$

sendo os graus de liberdade (g.l.) obtidos por:

$$\nu = \frac{(\frac{8672}{15} + \frac{2211}{15})^2}{\frac{(\frac{8672}{15})^2}{14} + \frac{(\frac{2211}{15})^2}{14}} = 20,7 = 21$$

Vê-se, portanto, que $t_c \notin RC$, logo não se rejeita H0, ou seja, há evidências de que, em média, os grupos 1 e 2 têm a mesma inteligência.

Automaticamente, o valor t pode ser calculado por meio do comando

```
t.test(GA,GB,var.eq=F,alternative="greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  GA and GB
## t = -1.1963, df = 9.5435, p-value = 0.8698
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -1.585498       Inf
## sample estimates:
## mean of x mean of y
##  1.800000  2.428571
```

2.2.2 Duas populações dependentes (pareadas)

Em alguns casos, fatores externos podem influenciar de forma significativa a comparação de duas médias populacionais. Como exemplo, suponha a verificação da eficácia de dois métodos diferentes de ensino, em que fatores capazes de afetar o rendimento acadêmico, como motivação, inteligência, idade, classe social etc., poderão influenciar no resultado do teste. Ou seja, em situações como essas, é recomendável coletar as observações em pares, de modo que os dois elementos de cada par sejam homogêneos em todos os sentidos, exceto no que diz respeito ao fator que se queira comparar. Em um grande número desses experimentos, no entanto, o mesmo indivíduo é usado nas duas amostras, de forma que estes são conhecidos por experimentos “antes e depois do tratamento”. A seguir, formaliza-se o procedimento de comparação nesses casos.

Considerem-se $X_1, X_2, \ldots, X_n$ e $Y_1, Y_2, \ldots, Y_n$ amostras aleatórias de distribuições normais, com médias μ_x e μ_y, e variâncias desconhecidas. Se as observações forem utilizadas de forma pareada, isto é, $(X_1, Y_1), (X_1, Y_2), \ldots, (X_n, Y_n)$, e fizer-se $D_i = X_i - Y_i$, $1 \leq i \leq n$, poder-se-á evidenciar que $D_1, D_2, \ldots, D_n$ é uma amostra aleatória de uma distribuição normal, cujo valor esperado é $\mu_D = \mu_x - \mu_y$. Um estimador não viciado da variância dessa amostra é dado por:

$$S_D^2 = \frac{1}{n-1}\sum_{i=1}^{n}(D_i - \bar{D})^2$$

sendo:

$$\bar{D} = \frac{1}{n}\sum_{i=1}^{n} D_i$$

Tem-se, portanto, que a estatística para testar $H_0 : \mu_x = \mu_y$, ao se supor essa hipótese verdadeira, é:

$$t = \frac{\bar{D}}{\frac{S_D}{\sqrt{n}}}.$$

Cuja distribuição pode-se facilmente provar que é uma t de Student, com n-1 graus de liberdade.

Exemplo 14:

Foram realizados testes, com um grupo de oito pessoas, sobre o efeito da frustração no comportamento agressivo. Para isso, primeiramente foi aplicado um teste de agressividade, sem que as pessoas tivessem passado por alguma frustração. Porém, numa outra etapa do experimento, os indivíduos que se submeteram ao teste passaram primeiro por um momento de frustração e em seguida foi aplicado o mesmo teste de agressividade. Obtiveram-se assim os escores da tabela a seguir em que, quanto maior o valor, maior a reação agressiva. Supõe-se aqui que esses escores têm distribuição normal.

```
require(tibble)
```

```
## Loading required package: tibble
```

```
## Warning: package 'tibble' was built under R version 3.6.1
```

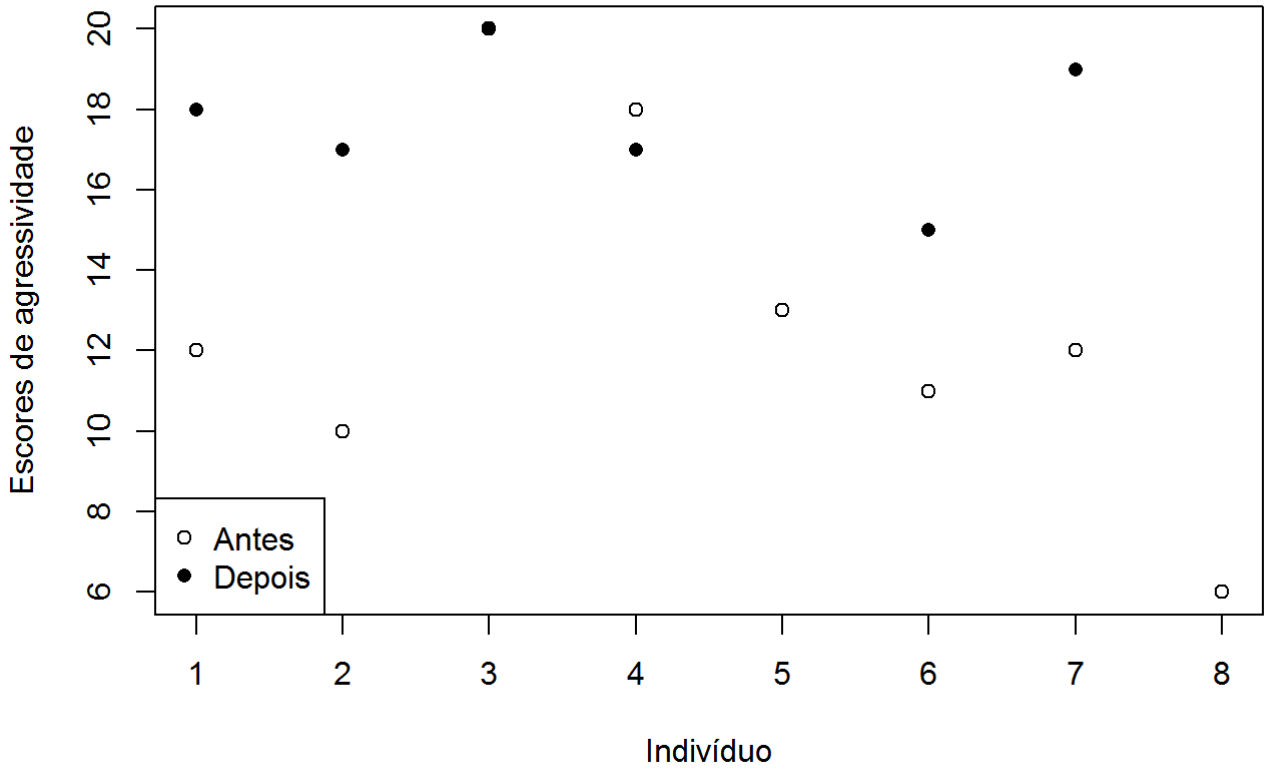
```
frustracao <- tibble(
  pessoas = 1:8,
  Antes = c(12,10,20,18,13,11,12,6),
  Depois = c(18,17,20,17,22,15,19,4)
)
frustracao
```

```
## # A tibble: 8 x 3
##   pessoas Antes Depois
##     <int> <dbl>  <dbl>
## 1       1    12     18
## 2       2    10     17
## 3       3    20     20
## 4       4    18     17
## 5       5    13     22
## 6       6    11     15
## 7       7    12     19
## 8       8     6      4
```

Conforme a análise gráfica a seguir, observa-se que, em um total de oito indivíduos, seis apresentaram escores bem maiores de agressividade após a exposição a uma atividade que causasse frustração. Nas duas exceções, em que os escores de agressividade foram maiores antes da exposição, constata-se que estes

estão próximos dos valores medidos depois da exposição. Com isso, graficamente, pode-se observar indícios de que a exposição à frustração provoca um aumento na agressividade.

```
Antes = c(12,10,20,18,13,11,12,6)
Depois = c(18,17,20,17,22,15,19,4)
plot(Antes,ylab="Escores de agressividade",xlab="Indivíduo")
points(Depois,pch=16)
legend("bottomleft",pch=c(1,16),legend=c("Antes","Depois"))
```



Caso se considere X=agressividade da pessoa antes de ser frustrada e Y=agressividade após a frustração, então:

a. Pretende-se aqui testar, se não altera a agressividade da pessoa, o fato desta ter sido frustrada ou não, contra a alternativa de que depois de frustrada a pessoa fica mais agressiva, ou seja:

$H_0 : \mu_y = \mu_x$ ou $\mu_y - \mu_x = 0$ $H_1 : \mu_y > \mu_x$ ou $\mu_y - \mu_x > 0$

b. Pelas suposições do experimento, tem-se que a estatística do teste, ao se supor H0 verdadeira, é:

$$t = \frac{\bar{D}}{\frac{S_D}{\sqrt{8}}}.$$

e

$$S_D^2 = \frac{1}{7}\sum_{i=1}^{8}(D_i - \bar{D})^2 = \frac{1}{7}\sum_{i=1}^{8}(D_i)^2 - \frac{7}{8}\bar{D}^2$$

Assim,

$$\bar{D} = \frac{30}{8} = 3,75 S_D^2 = \frac{236}{7} - \frac{8}{7}(3,75)^2 = 17,64$$

Então

$$t_c = \frac{3,75}{\sqrt{\frac{17,64}{8}}} = 2,52$$

Logo, existem evidências de que depois de frustradas as pessoas ficam mais agressivas. Dado que “Antes” e “Depois” foram as variáveis criadas para armazenar os dados na análise gráfica desse último exemplo, o comando do R a ser utilizado nesse caso é:

```
t.test(Depois,Antes,paired=T)
```

```
##
##  Paired t-test
##
## data:  Depois and Antes
## t = 2.5252, df = 7, p-value = 0.03951
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.2384278 7.2615722
## sample estimates:
## mean of the differences
##                    3.75
```

2.3 Comparação entre duas proporções de populações independentes

Sejam π_1 e π_2 duas populações independentes, com π_1 e π_2 as respectivas proporções de certas características de interesse dessas populações. Para se testar $H_0 : p_1 = p_2 = p$, retira-se uma amostra de tamanho n_1 de π_1 e outra de tamanho n_2 de π_2 , e considerando $\hat{p}_i$ a proporção de ocorrência da característica em estudo na amostra de tamanho $n_i, i = 1, 2$. Dessa forma, obter-se-á um estimador de p, dado por:

$$\hat{p} = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2}$$

Considere que n_1 e n_2 sejam grandes e $\hat{p}_i$ nem muito próximas de zero nem também muito próximas de um. Para efeitos práticos, isso corresponde à condição de que o mínimo entre $\hat{p}_i$ e 1-$\hat{p}_i$ multiplicado pelo mínimo entre n_1 e n_2 seja maior que 5. Dessa forma, tem-se que a distribuição de $\hat{p}_1 - \hat{p}_2$ é normal com média $p_1 - p_2$ e variância $p_1(1 - p_1)/n_1 + p_2(1 - p_2)/n_2$, implicando que a estatística:

$$= \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$$

tem distribuição N(0,1).

Dado que p_1 e p_2 são desconhecidas, ao se substituir p por seu estimador e considerando H_0 verdadeira, verifica-se que a estatística do teste será dada por:

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}}$$

cuja distribuição é aproximadamente N(0,1).

Exemplo 15:

Uma pesquisa realizada por uma aluna concluinte do curso de Psicologia de uma universidade de Natal teve como objetivo investigar como os pacientes alcoólatras, internados em dois hospitais psiquiátricos de Natal percebem sua situação. Foram então selecionadas duas amostras, sendo uma de 15 alcoólatras do sexo masculino, de classe média, e outra de 17 alcoólatras do sexo masculino, de classe de baixa renda. Entre outras variáveis investigadas, houve interesse em comparar as razões alegadas pelos pacientes para o início da ingestão de bebidas alcoólicas, de modo que os resultados obtidos foram os seguintes:

```
require(tibble)
motivacao <- tibble(
  media = c(7,8),
  baixarenda = c(6,11)
)
motivacao
```

```
## # A tibble: 2 x 2
##   media baixarenda
##   <dbl>      <dbl>
## 1     7          6
## 2     8         11
```

É possível concluir, ao nível de 5%, que os problemas familiares (afetivos) são indicados em igual proporção pelos alcoólatras de classe média e de baixa renda, da cidade de Natal?

Considere p_1 e p_2 as proporções de alcoólatras da cidade de Natal, da classe média e de baixa renda, respectivamente, que alegam os problemas familiares (afetivos) como a razão para o início da ingestão de bebidas alcoólicas. Nesse caso, tem-se que as hipóteses que se deseja testar são:

$H_0 : p_1 = p_2 = p$

$H_1 : p_1 \neq p_2$

Pelos dados obtidos, tem-se que

$$\hat{p}_1 = \frac{7}{15} = 0,47$$

$$\hat{p}_2 = \frac{6}{17} = 0,35$$

$$\hat{p} = \frac{15\hat{p}_1 + 17\hat{p}_2}{15 + 17} = 0,41$$

Assim, o mínimo entre $\hat{p}$ e 1-$\hat{p}$, multiplicado pelo mínimo entre n_1 e n_2, será igual a 6,15, que é maior que 5. Logo, ao se supor H_0 verdadeira, tem-se que a estatística do teste é:

$$z = \frac{0,47 - 0,35}{\sqrt{(0,41)(0,59)(\frac{1}{15} + \frac{1}{17})}} = 0,69$$

Assim, há evidências de que são iguais as proporções de alcoólatras de classe média e de baixa renda, da cidade de Natal, que alegam os problemas familiares (afetivos) como as razões para o início da ingestão de bebidas alcoólicas.

Para realizar o teste desse exemplo utilizando o R, os comandos seriam:

```
tabela=matrix(c(7,6,8,11),2,2,byrow=T)
prop.test(t(tabela))
```

```
##
##  2-sample test for equality of proportions with continuity
##  correction
##
## data:  t(tabela)
## X-squared = 0.085862, df = 1, p-value = 0.7695
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.2886443  0.5160953
## sample estimates:
##    prop 1    prop 2
## 0.4666667 0.3529412
```

Referências

ANJOS, Adilson dos. Estatística básica com o uso do software R. 2015. Disponível em: https://docs.ufpr.br/~aanjos/TRI/R/rbasico.pdf (https://docs.ufpr.br/~aanjos/TRI/R/rbasico.pdf). Acesso em: 26 ago. 2019.

AZEVEDO, Paulo Roberto Medeiros de. **Introdução à estatística**. 3. ed. Natal: Edufrn, 2016. 235 p. Disponível em: https://repositorio.ufrn.br/ (https://repositorio.ufrn.br/). Acesso em: 26 ago. 2019.

AZEVEDO, Paulo Roberto Medeiros de; MORALES, Fidel Ernesto Castro; PINHO, André Luís Santos de. Métodos Básicos de Estatística. Natal: Edufrn, 2018. 123 p. Disponível em: http://repositorio.ufrn.br/ (http://repositorio.ufrn.br/). Acesso em: 26 ago. 2019.