Inferência Estatística II

Testes de hipótestes específicos

Prof. Paulo Cerqueira Jr - cerqueirajr@ufpa.br
Faculdade de Estatística - FAEST
Instituto de Ciências Exatas e Naturais - ICEN

https://github.com/paulocerqueirajr

Introdução

Introdução

- Existem uma série de testes de hipóteses, cada um com sua finalidade.
- Para uma amostra, temos alguns testes:
  - Teste de normalidade
  - Teste-t comparação de grupos (indedependentes e pareados);
  - Testes não-paramétricos Wilcoxon;
  - Teste para média populacional com variância desconhecida (Teste t de Student).

Testes para duas amostra

Testes para duas amostra

- Existem uma série de testes de hipóteses, cada um com sua finalidade.
- Para duas amostra, temos alguns testes:
  - Teste para a comparação de médias com variância conhecida (Teste Z);
  - Teste para a comparação de médias com variância desconhecida (Teste t de **Student** comparação de grupos);
  - Teste para a comparação de variâncias;
  - Teste para comparação de amostras dependentes (Teste t de **Student** comparação de grupos pareado)

Teste para Comparação das médias

- Queremos comparar:

$$H_0 : \mu_1 = \mu_2 \quad \begin{array}{ll} \times & (i) H_1 : \mu_1 \neq \mu_2. \text{ (bilateral)} \\ \times & (ii) H_1 : \mu_1 > \mu_2. \text{ (Unilateral à direita)} \\ \times & (iii) H_1 : \mu_1 < \mu_2. \text{ (Unilateral à esquerda)} \end{array}$$

CASO 1: Mesma variância, conhecida. $(\sigma_1^2 = \sigma_2^2 = \sigma^2)$.

- Estatística para o teste:

$$Z = \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sigma \sqrt{\frac{1}{n} + \frac{1}{m}}} \sim N(0, 1).$$

Exemplo

Exemplo: Vamos supor que os níveis de QI entre meninos e meninas da décima série sejam normalmente distribuídos, cada um com desvio padrão populacional de 25. Uma professora quer saber se a média do QI entre meninos e meninas da turma é diferente, então ela seleciona duas amostras aleatórias — uma de meninos e outra de meninas — cada uma com tamanho 40, e registra os níveis de QI. Vamos realizar o teste Z para duas amostras para determinar se a média dos níveis de QI é diferente entre meninos e meninas, com nível de significância de 5%.

Os dados:

```r
# Definindo os dados:
qi_meninos = c( 79,118,99,117,98,102,112,111,102,73,114,97,114,122,
                  90,115,84,105,84,126,83,96,111,151,147,103,104,118,
                  132,108,95,118,121,88,94,92,94,109,105,123)
qi_meninas = c( 99,128,89,107,99,104,119,112,105,93,84,91,113,129,
                   100,105,94,115,114,106,113,116,116,131,117,123,134,
                   128,112,101,105,89,101,118,124,72,104,119,145,133)
dados <- data.frame(grupo = rep(c("Meninas", "Meninos")),
                qi = c(qi_meninos, qi_meninas))
head(dados)
```

```
    grupo  qi
1 Meninas  79
2 Meninos 118
3 Meninas  99
4 Meninos 117
5 Meninas  98
6 Meninos 102
```

Exemplo

Passo 1: $H_0 : \mu_A = \mu_B$ contra $H_1 : \mu_A \neq \mu_B$

Passo 2:

$$Z = \frac{\overline{X}_A - \overline{X}_B - (\mu_A - \mu_B)}{\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim N(0,1)$$

Passo 3: Sob H_0, temos $Z = \frac{\overline{X}_A - \overline{X}_B}{\sqrt{50}\sqrt{\frac{1}{n}+\frac{1}{m}}} \sim N(0,1)$,

$$RC = \{Z \in \mathbb{R} : |Z| < z_c\}$$

Logo,

$$0,025 = P(Z < z_1 | Z \sim N(0,1)) \quad \text{e} \quad 0,025 = P(Z > z_2 | Z \sim N(0,1)).$$

Assim, temos $z_1 = -1,96$ e $z_2 = 1,96$. E a regra de decisão é dada por

$$\text{Rejeitar } \; H_0 \; \text{ se } \; |Z| < 1,96.$$

Exemplo

Passo 4: O valor observado da estatística é

$$z_0 = \frac{106.350 - 110.175}{\sqrt{50}\sqrt{\frac{1}{40} + \frac{1}{40}}} = -1.710592$$

Passo 5: Como o valor observado de Z pertence a RC, rejeitamos H_0, e concluímos que há evidências de que a média dos níveis de QI é a mesma entre meninos e meninas.

> **Importante**
>
> - Poderíamos construir um I.C. para a diferença $\theta = \mu_A - \mu_B$!!!
> - Em que:
>
> $$\mathrm{IC}(\theta, \gamma) = \left(\overline{X}_A - \overline{X}_B \pm z_{\alpha/2}\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}\right)$$

Exemplo

```r
library(BSDA)
z.test(x=qi_meninos, y=qi_meninas, mu=0, sigma.x=25, sigma.y=25, alternative = "two.sided")
```

```
    Two-sample z-Test

data:  qi_meninos and qi_meninas
z = -0.68424, p-value = 0.4938
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -14.781532   7.131532
sample estimates:
mean of x mean of y
  106.350   110.175
```

Teste para Comparação das médias

CASO 2: Mesma variância, desconhecida.

- Neste caso, a estatística para o teste:

$$T = \frac{\overline{X}_A - \overline{X}_B - (\mu_A - \mu_B)}{S_p\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim t_{n+m-2},$$

onde

$$S_p^2 = \frac{(n-1)S_A^2 + (m-1)S_B^2}{n+m-2}.$$

- Este teste é conhecido como `teste t para duas amostras!`

Teste para Comparação das médias

CASO 3: Variâncias desiguais e desconhecidas.

- Pode-se provar que a estatística

$$T = \frac{\overline{X}_A - \overline{X}_B - (\mu_A - \mu_B)}{\sqrt{\frac{S_A^2}{n} + \frac{S_B^2}{m}}},$$

sob H_0, tem uma distribuição aproximadamente t-Student com graus de liberdade, dados aproximadamente por

$$v = \frac{(x+y)^2}{\frac{x^2}{n-1} + \frac{y^2}{m-1}},$$

com $x = \frac{S_A^2}{n}$ e $y = \frac{S_B^2}{m}$.

- Este teste é conhecido como `Problema de Behrens-Fisher!`.

Exemplo

- Os dados:

```r
# Peso de homens e mulheres
peso.m <- c(38.9, 61.2, 73.3, 21.8, 63.4,
            64.6, 48.4, 48.8, 48.5)
peso.h <- c(67.8, 60, 63.4, 76, 89.4,
            73.3, 67.3, 61.3, 62.4)
# Creando um data frame
dados <- data.frame(
                grupo = rep(c("Mulheres",
                              "Homens"), ea
                peso = c(peso.m,  peso.h)
                )
head(dados)
```

```
     grupo peso
1 Mulheres 38.9
2 Mulheres 61.2
3 Mulheres 73.3
4 Mulheres 21.8
5 Mulheres 63.4
6 Mulheres 64.6
```

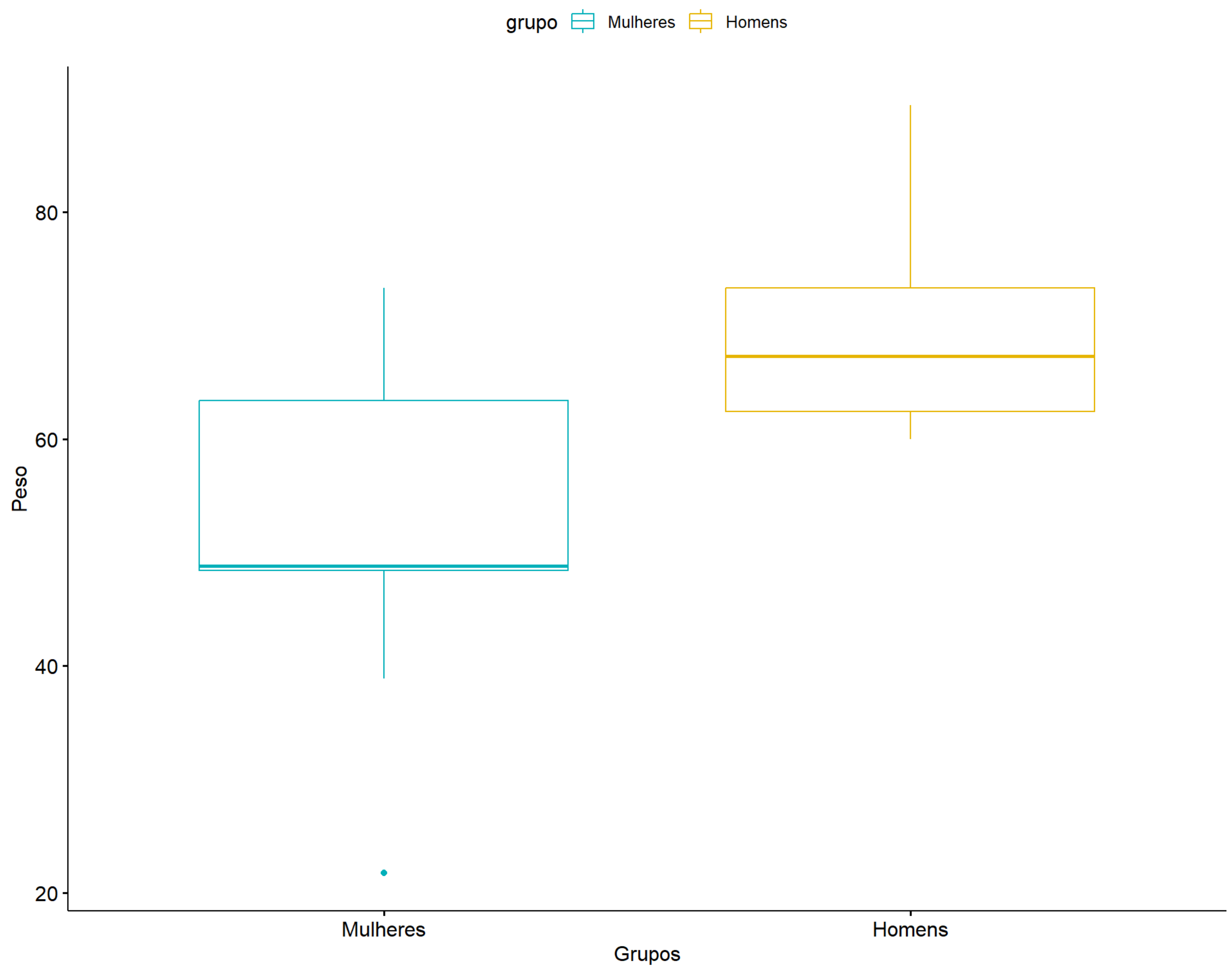
- Medidas de resumo:

```r
library(dplyr)
group_by(dados, grupo) %>%
  summarise(
    freq = n(),
    media = mean(peso, na.rm = TRUE),
    vari = var(peso, na.rm = TRUE)
  )
```

```
# A tibble: 2 × 4
  grupo     freq media  vari
  <chr>    <int> <dbl> <dbl>
1 Homens       9  69.0  87.9
2 Mulheres     9  52.1 243. 
```

Exemplo

```r
library("ggpubr")
ggboxplot(dados, x = "grupo", y = "peso",
          color = "grupo", palette = c("#00AFBB", "#E7B800"),
        ylab = "Peso", xlab = "Grupos")
```



Exemplo

> A construção formal do teste.

Passo 1: $H_0 : \mu_A = \mu_B$ contra $H_1 : \mu_A \neq \mu_B$

Passo 2:

$$T = \frac{\overline{X}_A - \overline{X}_B - (\mu_A - \mu_B)}{\sqrt{\frac{S_A^2}{n} + \frac{S_B^2}{m}}}$$

Exemplo

Passo 3: Sob H_0, temos $T = \frac{\overline{X}_A - \overline{X}_B}{\sqrt{\frac{S_A^2}{n} + \frac{S_B^2}{m}}} \sim t_v$

$$RC = \{T \in \mathbb{R} : T < t_1 \text{ ou } T > t_2\},$$

com

$$0,025 = P(T < t_1 | T \sim t_v) \quad \text{e} \quad 0,025 = P(T > t_2 | T \sim t_v).$$

Agora,

$$v = \frac{((87.9/9) + (243/9))^2}{(87.9/9)^2/8 + (243/9)^2/8} = 13.1179798 \simeq 14$$

Portanto,

$$t_1 = -2.1447867 \quad \text{e} \quad t_2 = 2.1447867.$$

E a regra de decisão é dada por: Rejeitar H_0 se $T < -2.1447867$ ou $T > 2.1447867$.

Exemplo

Passo 4: O valor observado da estatística é

$$t_0 = \frac{69.0 - 52.1}{\sqrt{\frac{87.9}{9} + \frac{243}{9}}} = 2.7871451.$$

Passo 5: Como o valor observado de T pertence a RC, rejeitamos H_0, e concluímos que há evidências de que os dois tipos de vigas têm resistências médias diferentes.

Exemplo

```r
res <- t.test(peso ~ grupo, data = dados, var.equal = FALSE)
res
```

```
	Welch Two Sample t-test

data:  peso by grupo
t = 2.7842, df = 13.114, p-value = 0.01538
alternative hypothesis: true difference in means between group Homens and group Mulheres is not equal to 0
95 percent confidence interval:
  3.795858 29.981920
sample estimates:
  mean in group Homens mean in group Mulheres 
              68.98889               52.10000 
```

Teste para Comparação das Variâncias

Teste para Comparação das Variâncias

$$H_0 : \sigma_1^2 = \sigma_2^2$$

A estatística do teste será

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F_{n-1,m-1}$$

Exemplo: Queremos verificar se duas máquinas produzem peças com a mesma homogeneidade quanto à resistência à tensão. Para isso, sorteamos duas amostras de seis peças de cada máquina, e obtivemos as seguintes resistências:

- Máquina A: 145, 127, 136, 142, 141, 137
- Máquina B: 143, 128, 132, 138, 142, 132

Exemplo

Passo 1: $H_0 : \sigma_A^2 = \sigma_B^2$ contra $H_1 : \sigma_A^2 \neq \sigma_B^2$

Passo 2:

$$F = \frac{S_A^2/\sigma_A^2}{S_B^2/\sigma_B^2} \sim F_{5,5}$$

Passo 3: Sob H_0, temos que $F = \frac{S_A^2}{S_B^2} \sim F_{5,5}$

Fixando $\alpha = 0,05$, a RC é dada por

$$RC = \{F < F_1 \text{ ou } F > F_2\},$$

com F_1 e F_2 tais que

$$0,025 = P(F < F_1 | F \sim F_{5,5}) \quad \text{e} \quad 0,025 = P(F > F_2 | F \sim F_{5,5})$$

Exemplo

Temos então, $F_2 = 7,15$ e $F_1 = 1/7,15 = 0,14$. Assim, a regra de decisão é:

$$\text{Rejeitar } \; H_0 \; \text{ se } \; F < 0,14 \; \text{ ou } \; F > 7,15.$$

Passo 4: Com os dados apresentados, temos $S_A^2 = 40$ e $S_B^2 = 37$. Portanto, o valor observado da estatística é $F_o = 40/37 = 1,08$.

Passo 5: Como o valor observado da estatística não pertence a RC, aceitamos H_0 e concluímos que as máquinas produzem com a mesma variabilidade.

Exemplo

```r
A <- c(145, 127, 136, 142, 141, 137)
B <-  c(143, 128, 132, 138, 142, 132)
dados <- data.frame(
                grupo = rep(c("A",
                              "B")),
                res = c(A,  B)
                )
head(dados)
```

```
  grupo res
1     A 145
2     B 127
3     A 136
4     B 142
5     A 141
6     B 137
```

```r
res.ftest <- var.test(res ~ grupo, data = dados)
res.ftest
```

```

    F test to compare two variances

data:  res by grupo
F = 0.6676, num df = 5, denom df = 5, p-value = 0.6683
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.09341832 4.77094592
sample estimates:
ratio of variances
          0.667603
```

Duas Populações Normais dependentes

Duas Populações Normais dependentes

Aqui temos duas amostras $X_1, X_2, \cdots, X_n$ e $Y_1, Y_2, \cdots, Y_n$, só que agora as observações são pareadas, isto é, temos uma amostra de pares

$$(X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n)$$

Se definirmos a v.a. $D = X - Y$, teremos uma amostra $D_1, D_2, \cdots, D_n$, resultante da diferença dos valores entre cada par. Reduzimos o problema de duas populações a um problema de uma única população, já visto anteriormente. Assim,

$$\overline{D} = \frac{1}{n}\sum_{i=1}^{n} D_i = \frac{1}{n}\sum_{i=1}^{n}(X_i - Y_i) = \frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{n}\sum_{i=1}^{n} Y_i = \overline{X} - \overline{Y}$$

terá distribuição $N(\mu_D, \sigma_D^2/n)$.

Duas Populações Normais dependentes :

Considerando

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

temos que

$$T = \frac{\sqrt{n}(\overline{D} - \mu_D)}{S_D} \sim t_{n-1}$$

Como $\mu_D = E(D) = E(X - Y) = E(X) - E(Y) = \mu_1 - \mu_2$, testar $H_0 : \mu_D = 0$ é equivalente a testar $H_0 : \mu_1 = \mu_2$.

Exemplo

Exemplo: Cinco operadores de certo tipo de máquina são treinados em máquinas de duas marcas diferentes, A e B. Mediu-se o tempo em que cada um deles gasta na realização de uma mesma tarefa, e os resultados estão na tabela a seguir.

Operador	**Marca A**	**Marca B**
A	80	75
B	72	70
C	65	60
D	78	72
E	85	78

Ao nível de significância de 10%, poderíamos afirmar que a tarefa realizada na Máquina A demora mais que na Máquina B?

Exemplo

Passo 1:

$$H_0 : \mu_A = \mu_B \times H_1 : \mu_A > \mu_B$$

Essas hipóteses são equivalentes a

$$H_0 : \mu_D = 0 \times H_1 : \mu_D > 0$$

Passo 2:

$$T = \frac{\sqrt{n}(\overline{D} - \mu_D)}{S_D} \sim t_4$$

Passo 3: Como é o mesmo operador que realiza a tarefa nas duas máquinas, dizemos que as variáveis são emparelhadas. Sob H_0, temos $T = \frac{\sqrt{n}\overline{D}}{S_D} \sim t_4$. Assim, com $\alpha = 0,10$, temos

$$P(T > t_c | T \sim t_4) = 0,10.$$

Portanto, $t_c = 1,533$, logo, a regra de decisão é: Rejeitar H_0 se $T > 1,533$.

Exemplo

Passo 4: Da Tabela de dados acima, obtemos os valores de D:

$$d_i: \quad 5, 2, 5, 6, 7$$

e, portanto,

$$\overline{d} = 5, \quad \text{e} \quad s_D^2 = 3,5$$

Logo, o valor observado da estatística $\hat{e}$

$$t_o = (\sqrt{5} \times 5)/\sqrt{3,5} = 5,98$$

Passo 5: Como o valor observado pertence a RC, rejeitamos H_0, ou seja, demora-se mais para realizar a tarefa na máquina A.

Podemos construir um I.C. para μ_D, adotando $\gamma = 0,90$:

$$IC(\mu_D; 90\%) = \bar{D} \pm t_{\alpha/2} \times \sqrt{s_D^2}/\sqrt{n}$$

$$IC(\mu_D; 90\%) = 5 \pm 1,78 = [3,22\ ;\ 6,78]$$

Exemplo

```r
ma <- c(80, 72, 65, 78, 85)
mb <- c(75, 70, 60, 72, 78)

dados <- data.frame(
                grupo = rep(c("antes", "depois"), each=5),
                tempo = c(ma, mb)
                )
head(dados)
```

```
   grupo tempo
1  antes    80
2  antes    72
3  antes    65
4  antes    78
5  antes    85
6 depois    75
```

```r
group_by(dados, grupo) %>%
  summarise(
    freq = n(),
    media = mean(tempo, na.rm = TRUE),
    desvio = sd(tempo, na.rm = TRUE)
  )
```

```
# A tibble: 2 × 4
  grupo   freq media desvio
  <chr>  <int> <dbl>  <dbl>
1 antes      5    76   7.71
2 depois     5    71   6.86
```