AULA 20 - Testes de Hipóteses para duas populações

Professor: Jorge L. Bazán Monitora: Patrícia Stülp

26/06/2023

1. Exemplo: Teste para a média de duas populações Normais (teste t de Student)

Considerando o conjunto de dados do CENSO AMERICANO (repositório UCI), podemos dizer que há diferença de idade entre os sexos? Isto é, há diferença entre as idades dos homens e das mulheres?

1.1 PASSO 1: Especificar (em termos dos parâmetros) as hipóteses H0 e Ha.

Propomos comparar as médias de idade das mulheres vs as médias de idade dos homens, considerando uma hipótese bicaudal.

- $H0 : \mu_1 = \mu_2$
- $Ha : \mu_1 \neq \mu_2$

1.1.1 Dados amostrais

Considerando as informações amostrais, temos

```
# AMOSTRA

# Mulheres
n1 = 9782
xbarra = 36.883459
s1 = 13.532427

# Homens
n2 = 20380
ybarra = 39.184004
s2 = 12.873243
```

1.2 PASSO 2: Especificar a estatística do teste e sua distribuição, sob H0.

Neste caso, temos que identificar a distribuição amostral para testar as hipóteses considerando os quatro casos na tabela a seguir.

Testes de Hipótese frequentes
2. Sobre os parâmetros
$H_0 : \mu_1 = \mu_2$

H_1	*Caso*	*Estatística de teste*	*Rejeitar H_0 se:*
$\mu_1 - \mu_2 > 0$ $\mu_1 - \mu_2 < 0$ $\mu_1 - \mu_2 \neq 0$	1. σ_1 e σ_2 são conhecidas, as amostras são independentes e cada uma das populações tem distribuição normal ou os tamanhos de amostra n_i são suficientemente grandes	$Z = \dfrac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$	$Z > z_{1-\alpha}$ $Z < z_\alpha = -z_{1-\alpha}$ $\lvert Z\rvert > z_{1-\alpha/2}$
$\mu_1 - \mu_2 > 0$ $\mu_1 - \mu_2 < 0$ $\mu_1 - \mu_2 \neq 0$	2. σ_1 e σ_2 são desconhecidas porém iguais, as amostras são independentes e as populações têm distribuição normal	$T = \dfrac{\bar{X} - \bar{Y}}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$ Con $S_p^2 = \dfrac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1 + n_2 - 2}$ e a t é com $n_1 + n_2 - 2$ g.l.	$T > t_{1-\alpha}$ $T < t_\alpha = -t_{1-\alpha}$ $\lvert T\rvert > t_{1-\alpha/2}$
$\mu_1 - \mu_2 > 0$ $\mu_1 - \mu_2 < 0$ $\mu_1 - \mu_2 \neq 0$	3. σ_1 e σ_2 são desconhecidas porém diferentes, as amostras são independentes e as populações têm distribuição normal	$T = \dfrac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$ e a t é com v g.l. con $v = \dfrac{\left(S_1^2/n_1 + S_2^2/n_2\right)^2}{\frac{(S_1^2/n_1)^2}{n_1+1} + \frac{(S_2^2/n_2)^2}{n_2+1}} - 2$	$T > t_{1-\alpha}$ $T < t_\alpha = -t_{1-\alpha}$ $\lvert T\rvert > t_{1-\alpha/2}$
$\mu_1 - \mu_2 > 0$ $\mu_1 - \mu_2 < 0$ $\mu_1 - \mu_2 \neq 0$	4. σ_1 e σ_2 são desconhecidas, as amostras são independentes e têm tamanhos suficientemente grandes	$Z = \dfrac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$	$Z > z_{1-\alpha}$ $Z < z_\alpha = -z_{1-\alpha}$ $\lvert Z\rvert > z_{1-\alpha/2}$

Figura 1: Testes de 2 médias

Para o problema, as populações de homens e mulheres são consideradas que se distribuem normalmente e, além disso, independentes porém as variâncias populacionais são desconhecidas. Assim, temos que decidir entre o caso 2 (variâncias desconhecidas porém iguais) ou caso 3 (variancias desconhecidas porém diferentes). Uma alternativa é testar previamente a igualdade ou não das variâncias.

Com isso, podemos definir as correspondentes hipóteses como:

- $H0 : \sigma_1^2 = \sigma_2^2$
- $Ha : \sigma_1^2 \neq \sigma_2^2$

e então o teste pode ser verificando usando a estatística

$$F_{cal} = \frac{S_1^2}{S_2^2} \sim F(n_1 - 1,_2 - 1)$$

sendo $F(n_1 - 1, n_2 - 1)$ a estatística F de Fisher-Snedecor com $n_1 - 1$ e $n_2 - 1$ graus de liberdade.

Se temos posse dos dados originais, usando o R utilizamos os seguintes comandos.

```
var.test(idades ~ sexo, data, alternative = "two.sided")
var.test(idadem, idadeh, alternative = "two.sided")
```

Com as informações do problema, temos

```
Fcal = s1^2 / s2^2
Fcal
```

```
## [1] 1.105034

pvalor = pf(Fcal, n1 - 1, n2 - 1)
pvalor

## [1] 1
```

então, comparando esta estatística com $\alpha = 0.05$, vemos que $p_{valor} = 1 > \alpha$ logo aceitamos H0 (não podemos rejeitar H0), e portanto, podemos considerar que as variâncias embora desconhecidas podem ser assumidas como iguais. Com isto, usaremos o Caso 2 apresentado na tabela. Assim, a correspondente estatística é dada por

$$T_{tab} = \frac{\overline{x} - \overline{y}}{\sqrt{S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

onde

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

Neste caso, a região crítica é, Rejeitar H0 se $\mid T_{tab} \mid > t_{1-\alpha/2}$.

1.3 PASSO 3: Fixar o nível de significância do teste (α)

O erro tipo 1, o nível de significância do teste α neste problema é fixado em 0.05. Como a hipótese é de duas caudas (a hipótese Ha corresponde ao caso maior que ou menor que), e para este nível de significância é obtido da seguinte forma:

```
# Erro tipo 1 = alfa
alfa = 0.05

# Graus de liberdade
gl = n1 + n2 - 2

# Valor de Ttab para o erro tipo 1
qt(alfa/2, gl)

## [1] -1.960043

qt(1 - alfa/2, gl)

## [1] 1.960043
```

Então rejeitamos H0 se $\mid T_{tab} \mid > t_{1-\alpha/2} = 1.9600426$.

1.4 PASSO 4: Calcular a região crítica do teste ou valor-p

Este passo pode ser feito de duas formas.

1.4.1 Encontrando a estatística

Devemos calcular

```
S2p = ((n1-1)*s1^2 + (n2-1)*s2^2) / (n1 - 1 + n2 - 2)
S2p

## [1] 171.371

ttab = (xbarra - ybarra) / sqrt(S2p*(1/n1+1/n2))
ttab

## [1] -14.28723
```

1.4.2 Encontrando o valor-p

Necessitamos encontrar a probabilidade de rejeitar H0 usando a estatística quando de fato a hipótese nula é verdadeira, isto é, $p_{valor} = 2 \times P(t < t_{tab}, n_1 - 1 + n_2 - 2)$. Neste caso é

```
pvalor = 2*pt(ttab, n1 - 1 + n2 - 2)
pvalor

## [1] 3.719862e-46
```

1.5 PASSO 5: Decidir entre H0 e Ha, comparando o valor-p com α (ou verificando se a estatística do teste pertence ou não à região crítica).

Este passo pode ser feito de duas formas dependendo do passo anterior.

- Usando a região crítica: como $\mid T_{tab} \mid = 14.2872277 > t_{1-\alpha/2} = 1.9600426$, rejeitamos H0 e concluímos que as idades de mulheres e homens são diferentes.
- Usando valor-p: neste caso encontramos que o $p_{valor} = 3.7198617 \times 10 - 46 < \alpha = 0.05$ e, portanto, rejeitamos H0. Concluímos que as idades de mulheres e homens são diferentes com um nível de confiança de $1 - \alpha = 0.95$ ou probabilidade de erro tipo 1 de 0.05.
- Portanto, usando a região crítica ou valor-p, concluimos que há diferenças nas idades de mulheres e homens.

2 Exemplo: Teste para a média de duas populações Normais (teste t de Student) usando conjunto de dados.

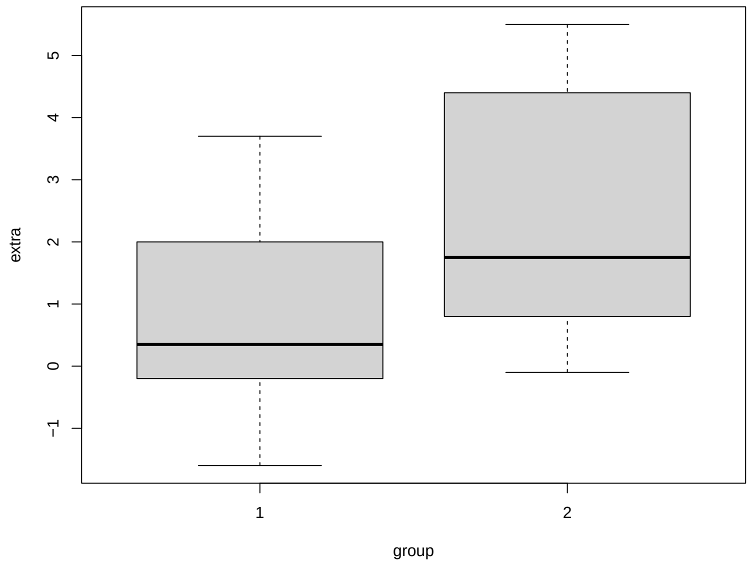
```
t.test(x, ...)

t.test(x, y = NULL, alternative = c("two.sided", "less",
       "greater"), mu = 0, paired = FALSE, var.equal = FALSE,
       conf.level = 0.95, ...)
```

```
t.test(formula, data, subset, na.action, ...)
```

2.1 Conjunto de dados: sonho de estudantes

O conjunto de dados a ser utilizado, mostra o efeito de dois medicamentos soporíficos (aumento nas horas de sono em comparação ao controle) em 10 pacientes, disponível no pacote `graphics` nomeado como `sleep`.

```
#install.packages(graphics)
require(graphics)
sleep

## extra group ID
## 1 0.7 1 1
## 2 -1.6 1 2
## 3 -0.2 1 3
## 4 -1.2 1 4
## 5 -0.1 1 5
## 6 3.4 1 6
## 7 3.7 1 7
## 8 0.8 1 8
## 9 0.0 1 9
## 10 2.0 1 10
## 11 1.9 2 1
## 12 0.8 2 2
## 13 1.1 2 3
## 14 0.1 2 4
## 15 -0.1 2 5
## 16 4.4 2 6
## 17 5.5 2 7
## 18 1.6 2 8
## 19 4.6 2 9
## 20 3.4 2 10

plot(extra ~ group, data = sleep)

with(sleep, mean(extra[group == 1]))

## [1] 0.75

with(sleep, mean(extra[group == 2]))

## [1] 2.33

with(sleep, sd(extra[group == 1]))
```




```
## [1] 1.78901
```

```
with(sleep, sd(extra[group == 2]))
```

```
## [1] 2.002249
```

Queremos saber se existe diferença entre os dois grupos. Temos duas formas equivalentes de executar o teste t de comparação de médias.

```
# Forma tradicional
with(sleep, t.test(extra[group == 1], extra[group == 2]))
```

```
##
## Welch Two Sample t-test
##
## data: extra[group == 1] and extra[group == 2]
## t = -1.8608, df = 17.776, p-value = 0.07939
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -3.3654832 0.2054832
## sample estimates:
## mean of x mean of y
## 0.75 2.33
```

```
# Formula tipo regressão
t.test(extra ~ group, data = sleep)
```

```
##
## Welch Two Sample t-test
##
## data: extra by group
```

```
## t = -1.8608, df = 17.776, p-value = 0.07939
## alternative hypothesis: true difference in means between group 1 and group 2 is no
## -3.3654832 0.2054832
## sample estimates:
## mean in group 1 mean in group 2
## 0.75 2.33
```

Em ambos casos, considerando o valor-p e comparando com um nível de significância de 0.05, aceitamos H0 e, portanto, podemos dizer que o efeito dos medicamentos no grupo teste não é diferente do grupo controle.

2.2 Dados simulados

Considere os seguintes dados simulados

```
set.seed(12345)
x = rnorm(10, 5, 1)
y = rnorm(10, 5, 2)

mean(x)

## [1] 4.867056

mean(y)

## [1] 5.571956

s1 = sd(x)
s2 = sd(y)
n1 = length(x)
n2 = length(y)

# Teste igualdade de variancias
var.test(x, y, alternative = "two.sided")

##
## F test to compare two variances
##
## data: x and y
## F = 0.23356, num df = 9, denom df = 9, p-value = 0.0413
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
## 0.05801292 0.94030993
## sample estimates:
## ratio of variances
## 0.2335597

Fcal = s1^2/s2^2
Fcal

## [1] 0.2335597
```

```
pvalor = 2*pf(Fcal, n1 - 1, n2 - 1)
pvalor

## [1] 0.04130414

# Variancias são diferentes, uma vez que pvalor < alfa. Portanto, usaremos
# o teste de igualdade de medias para variancias desconhecidas e diferentes

t.test(x, y, alternative = "two.sided", var.equal = FALSE)

##
## Welch Two Sample t-test
##
## data: x and y
## t = -1.1921, df = 12.987, p-value = 0.2546
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -1.9825002 0.5727008
## sample estimates:
## mean of x mean of y
## 4.867056 5.571956

nu = (s1^2/n1 + s2^2/n2)^2/((s1^2/n1)^2/(n1-1) +(s2^2/n2)^2/(n2-1))
nu

## [1] 12.9866

2*pt(-1.1921, nu)

## [1] 0.2545508
```

Neste caso, primeiro concluimos que as variâncias populacionais desconhecidas são diferentes e então usando o caso 3 (apresentado na tabela) para a comparação das médias concluimos que as médias não são diferentes. Note também, que o cálculo dos graus de liberdade foi obtido usando

$$\nu = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1-1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2-1}}$$

e a estatística foi obtida usando

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

Agora, considere os seguintes dados com diferentes tamanhos de amostra

```
# DADOS 1
set.seed(12345)
x = rnorm(10, 5, 1)
y = rnorm(10, 4, 1)

mean(x)

## [1] 4.867056

mean(y)

## [1] 4.285978

sd(x)

## [1] 0.8136554

sd(y)

## [1] 0.8418052

length(x)

## [1] 10

length(y)

## [1] 10

t.test(x, y, alternative = "two.sided", var.equal = FALSE)

##
## Welch Two Sample t-test
##
## data: x and y
## t = 1.5695, df = 17.979, p-value = 0.134
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -0.1968018 1.3589580
## sample estimates:
## mean of x mean of y
## 4.867056 4.285978

# DADOS 2
x = rnorm(50, 5, 1)
y = rnorm(50, 4, 1)

mean(x)

## [1] 5.20829
```

```
mean(y)

## [1] 4.27182

sd(x)

## [1] 1.234496

sd(y)

## [1] 1.046154

length(x)

## [1] 50

length(y)

## [1] 50

t.test(x, y, alternative = "two.sided", var.equal = FALSE)

##
## Welch Two Sample t-test
##
## data: x and y
## t = 4.0922, df = 95.432, p-value = 8.935e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## 0.4821888 1.3907509
## sample estimates:
## mean of x mean of y
## 5.20829 4.27182
```

Neste caso, concluímos que as médias não são diferentes quando consideramos um tamanho de amostra pequeno de 10 dados de cada população, porém as médias são consideradas diferenes quando consideramos um tamanho de amostra grande de 50 dados de cada população. Assim, percebemos que o tamanho de amostra é importante para aceitar ou rejeitar uma hipótese. Note também que, por default, temos usado testes bicaudais assumindo o caso 3 com variâncias populacionais desconhecidas, porém diferentes. Temos que verificar se as variâncias são diferentes. Ccaso não sejam diferentes, teremos que usar o caso 2 por meio do comando

```
t.test(x, y, alternative = "two.sided", var.equal = TRUE)
```