Análise estatística de várias populações: 2ª Aula Prática - Parte 1

Profa. Dra. Tatiane F. N. Melo

14/09/2024

Inferência estatística para duas populações:

1º Caso) Comparação das variâncias dos grupos

Exemplo 1: Suponha que estamos interessados em verificar se há uma diferença significativa na variabilidade do número de vacinas aplicadas diariamente em dois municípios do estado de Goiás, a saber: Goiânia e Aparecida de Goiânia, em 2022, $n = 331$ (Ministério da Saúde - Vacinômetro COVID-19).

Inicialmente vamos carregar os dados no R, utilizando os códigos a seguir.

```
# Pacote necessario para leitura dos dados
library(readxl)

# Dados referentes ao numero de vacinas contra COVID19
# aplicadas, diariamente, em 2022 nos municípios de
# Goiânia e Aparecida de Goiânia
dados.Ex1 <- read_excel("Exemplo1Aula2Parte1.xls")

print(dados.Ex1)
```

```
## # A tibble: 331 x 3
##    `Data da Vacina`    Total de Doses Aplicadas - Goiân~1 Total de Doses Aplic~2
##    <dttm>                                           <dbl>                   <dbl>
##  1 2022-01-01 00:00:00                                 9                       3
##  2 2022-01-02 00:00:00                                 1                       6
##  3 2022-01-03 00:00:00                              9739                    2479
##  4 2022-01-04 00:00:00                             10619                    2545
##  5 2022-01-05 00:00:00                             10267                    2130
##  6 2022-01-06 00:00:00                              8672                    2126
##  7 2022-01-07 00:00:00                              7849                    2131
##  8 2022-01-08 00:00:00                              1577                     569
##  9 2022-01-10 00:00:00                             10382                    2626
## 10 2022-01-11 00:00:00                              8603                    2187
## # i 321 more rows
## # i abbreviated names: 1: `Total de Doses Aplicadas - Goiânia`,
## #   2: `Total de Doses Aplicadas - Aparecida de Goiânia`
```

```r
Vac.Goiania           <- dados.Ex1$`Total de Doses Aplicadas - Goiânia`

Vac.Aparecida.Goiania <- dados.Ex1$`Total de Doses Aplicadas - Aparecida de Goiânia`
```

Inicialmente, vamos realizar um teste bilateral, ou seja,

$$H_0 : \sigma^2_G = \sigma^2_{AG} \quad \text{contra} \quad H_1 : \sigma^2_G \neq \sigma^2_{AG}.$$

No R, podemos realizar o teste de duas formas: (i) criando uma função usando as fórmulas vista na aula teórica ou (ii) usar uma função já existente no R. Como visto anteriormente, sempre é mais recomendável usar funções próprias do R. Somente em casos onde não existe no R, sua implementação, devemos criar novas funções.

Se no R não existisse a função pronta, poderíamos usar:

```r
##### Realizar o teste F (aplicando as fórmulas)
TH.VAR <- function(x,y,alfa)
{
  n.1 = length(x)
  n.2 = length(y)
  var.1 = var(x)
  var.2 = var(y)

  if(var.1 > var.2)
  {
    f_obs = var.1/var.2
    quantil_f_obs = pf(f_obs,n.1-1,n.2-1)
  }
  else
  {
    f_obs = var.2/var.1
    quantil_f_obs = pf(f_obs,n.2-1,n.1-1)

  }

  Valor.p = 2*min(quantil_f_obs, 1-quantil_f_obs)
  print(c("f.obs:", round(f_obs,3)))
  print(c("Valor.p:", round(Valor.p,100)))

  if(Valor.p <= alfa) print("Rejeitamos H.0")
  else print("Nao rejeitamos H.0")
}

TH.VAR(Vac.Goiania,Vac.Aparecida.Goiania,0.01)
```

```
## [1] "f.obs:" "12.176"
## [1] "Valor.p:" "0"
## [1] "Rejeitamos H.0"
```

A segunda maneira de realizar o teste de igualdade de variâncias é usando uma função já existente no R. Ou seja,

```r
########### Realizar o teste F (funcao já existente no R)
# Teste bilateral
var.test(Vac.Goiania,Vac.Aparecida.Goiania)
```

```
##
##  F test to compare two variances
##
## data:  Vac.Goiania and Vac.Aparecida.Goiania
## F = 12.176, num df = 330, denom df = 330, p-value < 2.2e-16
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##   9.808772 15.113577
## sample estimates:
## ratio of variances
##           12.17562
```

Note que rejeitamos H_0, então temos uma indicação de que ambos os municípios, Goiânia e Aparecida de Goiânia, não têm uma consistência semelhante nas suas quantidades de doses aplicadas, ao longo do tempo. Pode ser que um dos municípios tenha uma estratégia de vacinação mais consistente, enquanto o outro pode ter flutuações maiores nas quantidades de doses aplicadas, ao longo do tempo.

Suponha que temos o interesse em saber qual munícipio apresenta uma variabilidade maior nas doses aplicadas. Neste caso, precisamos realizar um teste unilateral, por exemplo, à direita. Ou seja,

$$H_0 : \sigma_G^2 \leq \sigma_{AG}^2 \quad \text{contra} \quad H_1 : \sigma_G^2 > \sigma_{AG}^2.$$

```r
# Teste unilateral à direita
var.test(Vac.Goiania,Vac.Aparecida.Goiania, alternative="greater")
```

```
##
##  F test to compare two variances
##
## data:  Vac.Goiania and Vac.Aparecida.Goiania
## F = 12.176, num df = 330, denom df = 330, p-value < 2.2e-16
## alternative hypothesis: true ratio of variances is greater than 1
## 95 percent confidence interval:
##  10.15617      Inf
## sample estimates:
## ratio of variances
##           12.17562
```

Aqui, também rejeitamos a hipótese nula. Logo, ao nível de 1%, o município de Goiânia tem flutuações maiores nas quantidades de doses aplicadas, ao longo do tempo, do que Aparecida de Goiânia.

Exemplo 2: Agora vamos supor que estamos interessados em verificar se há uma diferença significativa na variabilidade das idades das mães, de nascidos vivos, entre aquelas que fizeram parto vaginal com aquelas que foram submetidas ao parto cesário. Os dados foram obtidos da Base de dados SINASC, do Estado de São Paulo, em 2023.

Neste caso, foram 28.812 mães que fizeram partos vaginais e 36.718 foram submetidas ao parto cesário. Logo, $n_1 = 28.812$ e $n_2 = 36.718$. Realizaremos o teste bilateral

$$H_0 : \sigma_V^2 = \sigma_C^2 \quad \text{contra} \quad H_1 : \sigma_V^2 \neq \sigma_C^2,$$

onde σ_V^2 é a variabilidade das idades das mães que fizeram parto vaginal e σ_C^2 é a variabilidade das idades das mães que foram submetidas ao parto cesário. Implementando, este teste no R:

```
# Dados referentes às idades das mães que fizeram partos
# vaginal ou cesário no Estado de São Paulo, em 2023.
dados.Ex2 <- read_excel("Exemplo2Aula2Parte1.xls")

print(dados.Ex2)
```

```
## # A tibble: 65,535 x 2
##    IDADE PARTO
##    <chr> <chr>
##  1 25    1
##  2 25    1
##  3 31    1
##  4 34    2
##  5 32    1
##  6 33    2
##  7 17    2
##  8 30    2
##  9 16    1
## 10 23    1
## # i 65,525 more rows
```

```
idade_vaginal  <- as.numeric(subset(dados.Ex2, PARTO == 1)$IDADE)

idade_cesario  <- as.numeric(subset(dados.Ex2, PARTO == 2)$IDADE)

# Teste bilateral - Igualdade de variâncias
var.test(idade_vaginal,idade_cesario)
```

```
##
##  F test to compare two variances
##
## data:  idade_vaginal and idade_cesario
## F = 1.0223, num df = 28811, denom df = 36717, p-value = 0.04716
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  1.000277 1.044887
## sample estimates:
## ratio of variances
##           1.022324
```

Neste caso, não rejeitamos H_0, ao nível de 1%, então temos uma indicação de que em ambos os partos (vaginal e cesário), as variabilidades nas idades das mães são iguais.

Exemplo 3: Consideremos os dados referentes ao número de vacinas aplicadas contra COVID-19 (Ministério da Saúde - Vacinômetro COVID-19) em 20 dias do mês novembro de 2022, nos municípios de Formosa e Catalão - estados de Goiás. Neste caso, temos que o tamanho amostral é igual a $n = 20$.

```
# Dados referentes ao numero de vacinas contra COVID19
# aplicadas nos 20 dias de novembro de 2022 nos municípios de
# Formosa e Catalão
dados.Ex3 <- read_excel("Exemplo3Aula2Parte1.xls")

print(dados.Ex3)
```

```
## # A tibble: 20 x 4
##      Obs `Data da Vacina`    Total de Doses Aplicadas -~1 Total de Doses Aplic~2
##    <dbl> <dttm>                                    <dbl>                   <dbl>
##  1     1 2022-11-01 00:00:00                          17                      29
##  2     2 2022-11-03 00:00:00                          18                      18
##  3     3 2022-11-04 00:00:00                          27                      15
##  4     4 2022-11-07 00:00:00                          17                      25
##  5     5 2022-11-08 00:00:00                          29                      45
##  6     6 2022-11-09 00:00:00                          67                      54
##  7     7 2022-11-10 00:00:00                          69                      52
##  8     8 2022-11-11 00:00:00                         100                      61
##  9     9 2022-11-12 00:00:00                         214                       0
## 10    10 2022-11-16 00:00:00                           0                      96
## 11    11 2022-11-17 00:00:00                         364                      69
## 12    12 2022-11-18 00:00:00                         251                      88
## 13    13 2022-11-21 00:00:00                         214                      94
## 14    14 2022-11-22 00:00:00                         246                     104
## 15    15 2022-11-23 00:00:00                         226                      77
## 16    16 2022-11-24 00:00:00                         132                      39
## 17    17 2022-11-25 00:00:00                         149                     102
## 18    18 2022-11-28 00:00:00                          61                      41
## 19    19 2022-11-29 00:00:00                         144                      84
## 20    20 2022-11-30 00:00:00                         155                       0
## # i abbreviated names: 1: `Total de Doses Aplicadas - Catalao`,
## #   2: `Total de Doses Aplicadas - Formosa`
```

```
Vac.Formosa <- dados.Ex3$`Total de Doses Aplicadas - Formosa`

Vac.Catalao <- dados.Ex3$`Total de Doses Aplicadas - Catalao`
```

Como a amostra tem tamanho 20, a premissa de normalidade é necessária. Para isso, realizamos o teste de normalidade Shapiro-Wilk, que indicou, ao nível de 1%, que os dados são provenientes de uma população normal.

```
# Teste para normalidade dos dados (tamanho da amostra < 30)
# H.0: Os dados vieram de uma população normal
shapiro.test(Vac.Formosa)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Vac.Formosa
## W = 0.9477, p-value = 0.3335
```

```r
shapiro.test(Vac.Catalao)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Vac.Catalao
## W = 0.92031, p-value = 0.1005
```

O código em R, para o teste bilateral de igualdade entre as variâncias teste unilateral à esquerda, ou seja,

$$H_0 : \sigma^2_{Form} = \sigma^2_{Cat} \quad \text{contra} \quad H_1 : \sigma^2_{Form} \neq \sigma^2_{Cat},$$

é dado por:

```r
# Teste bilateral - Igualdade de variâncias
var.test(Vac.Formosa,Vac.Catalao)
```

```
##
##  F test to compare two variances
##
## data:  Vac.Formosa and Vac.Catalao
## F = 0.11026, num df = 19, denom df = 19, p-value = 1.264e-05
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.04364323 0.27857276
## sample estimates:
## ratio of variances
##          0.1102625
```

Como rejeitamos H_0, então temos uma indicação de que ambos os municípios, Formosa e Catalão, não têm uma consistência semelhante nas suas quantidades de doses aplicadas, ao longo do mês de novembro de 2022.

Quando realizamos o teste unilateral à esquerda, ou seja,

$$H_0 : \sigma^2_{Form} \geq \sigma^2_{Cat} \quad \text{contra} \quad H_1 : \sigma^2_{Form} < \sigma^2_{Cat},$$

rejeitamos a hipótese nula. Veja:

```r
# Teste unilateral à esquerda
var.test(Vac.Formosa,Vac.Catalao, alternative="less")
```

```
##
##  F test to compare two variances
##
## data:  Vac.Formosa and Vac.Catalao
## F = 0.11026, num df = 19, denom df = 19, p-value = 6.319e-06
## alternative hypothesis: true ratio of variances is less than 1
## 95 percent confidence interval:
##  0.0000000 0.2390768
## sample estimates:
## ratio of variances
##          0.1102625
```

Então, ao nível de 1%, o município de Formosa tem flutuações menores nas quantidades de doses aplicadas, ao longo do mês de novembro de 2022, do que Catalão.

2º Caso) Teste t para comparar médias quando as variâncias são iguais (homocedásticas)

Voltando ao Exemplo 2: Nesse exemplo, não rejeitamos a hipótese de igualdade entre as variâncias, ao nível de 1%, então podemos usá-lo neste caso. Primeiramente, vamos realizar o teste bilateral: %

$$H_0 : \mu_V = \mu_C \quad \text{contra} \quad H_1 : \mu_V \neq \mu_C,$$

onde μ_V é a idade média das mães que fizeram parto vaginal e μ_C é a idade média das mães que foram submetidas ao parto cesário.

```
t.test(idade_vaginal,idade_cesario, var.equal = TRUE, paired = FALSE)
```

```
##
##  Two Sample t-test
##
## data:  idade_vaginal and idade_cesario
## t = -67.385, df = 65528, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.498936 -3.301144
## sample estimates:
## mean of x mean of y
##  27.38595  30.78599
```

Neste caso, rejeitamos $H_0 : \mu_V = \mu_G$. Então, temos uma indicação de que em ambos os partos (vaginal e cesário), a idade média das mães não é a mesma, ao nível de 1%. A média das idades é significativamente diferente nos partos vaginal e cesário. Isto indica que, em média, em um dos partos, a idade da mãe é maior que o outro.

Logo, podemos realizar um teste unilateral, para verificar em qual tipo de parto a idade média da mãe é maior. Por exemplo, um teste unilateral à esquerda,

$$H_0 : \mu_V \geq \mu_C \quad \text{contra} \quad H_1 : \mu_V < \mu_C.$$

```
t.test(idade_vaginal,idade_cesario, var.equal = TRUE, alternative = "less", paired = FALSE)
```

```
##
##  Two Sample t-test
##
## data:  idade_vaginal and idade_cesario
## t = -67.385, df = 65528, p-value < 2.2e-16
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -3.317044
## sample estimates:
## mean of x mean of y
##  27.38595  30.78599
```

Neste caso, rejeitamos $H_0 : \mu_V \geq \mu_C$. Então, há indicação de que, em média, as mães que foram submetidas ao parto cesário têm idades maiores do que as que fizeram parto vaginal, ao nível de 1%.

3º Caso) Teste t para comparar médias quando as variâncias são diferentes (heterocedásticas)

Voltando ao Exemplo 1: Tivemos, anteriormente, a indicação de variâncias diferentes. Agora, queremos testar as hipóteses: %

$$H_0 : \mu_G = \mu_{AG} \quad \text{contra} \quad H_1 : \mu_G \neq \mu_{AG},$$

onde μ_G é o número médio de vacinas aplicadas na cidade de Goiânia e μ_{AG} é o número médio de vacinas aplicadas em Aparecida de Goiânia.

```
t.test(Vac.Goiania,Vac.Aparecida.Goiania, var.equal = FALSE, paired = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  Vac.Goiania and Vac.Aparecida.Goiania
## t = 13.795, df = 383.84, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1850.432 2465.574
## sample estimates:
## mean of x mean of y
## 3122.0574  964.0544
```

Rejeitamos $H_0 : \mu_G = \mu_{AG}$. Logo, ao nível de 1%, o número médio de vacinas aplicadas em Goiânia e Aparecida de Goiânia são diferentes. Vamos realizar um teste unilateral, para verificar em qual dos dois municípios um número médio de doses aplicadas maior. Por exemplo, um teste unilateral à direita, ou seja,

$$H_0 : \mu_G \leq \mu_{AG} \quad \text{contra} \quad H_1 : \mu_G > \mu_{AG}.$$

```
t.test(Vac.Goiania,Vac.Aparecida.Goiania, alternative = "greater",var.equal = FALSE, paired = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  Vac.Goiania and Vac.Aparecida.Goiania
## t = 13.795, df = 383.84, p-value < 2.2e-16
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  1900.073      Inf
## sample estimates:
## mean of x mean of y
## 3122.0574  964.0544
```

Neste caso, rejeitamos H_0. Então, há indicação de que a quantidade média de doses, da vacina contra COVID-19, aplicadas em Goiânia foi maior do que as aplicadas em Aparecidade de Goiânia, ao nível de 1%.

Conclusão:

- O município de Goiânia alcançou um número médio maior de vacinas aplicadas, mas enfrentou desafios com a estabilidade da vacinação. Indicando haver períodos em que a vacinação foi muito alta e outros em que foi mais baixa.

- Já Aparecida de Goiânia teve um número médio de doses aplicadas menor, porém foi mais estável, na aplicação, ao longo do tempo.