# Case 7 - Teste de Hipóteses: III
# PRO3200 - Estatística

Camille Peixoto Almeida n°USP: 12702259

27 de maio de 2023

## 1 Teste de Hipótese: Dois Parâmetros – Média

Um antigo jornalista afirma que a média de gols sofridos em casa é igual a média de gols sofridos fora de casa. Já seus colegas de redação afirmam que é difícil ter o mesmo número de gols tomados dentro e fora de casa. Sendo assim, é interessante fazer o teste de hipótese considerando como verdade atual a igualdade das médias de gols sofridos dentro e fora de casa e como hipótese alternativa que as médias são diferentes, assim:

$$\begin{array}{lcl} H_0 : \mu_1 = \mu_2 & \text{ou} & \mu_1 - \mu_2 = 0 \\ H_1 : \mu_1 \neq \mu_2 & \text{ou} & \mu_1 - \mu_2 \neq 0 \end{array}$$

Sendo:

1. $\mu_1$: média de gols tomados fora de casa
2. $\mu_1$: média de gols tomados em casa

## 2 Expressão analítica para o valor crítico

### 2.1 Desvios padrão populacionais conhecidos

Considerando **desvios padrão populacionais conhecidos** para os gols sofridos dentro e fora de casa:

$$C = z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \tag{1}$$

Sendo:

1. C: valor crítico relacionado a um nível de significância
2. z: valor usado para representar a variável aleatória padronizada da tabela de **distribuição normal de probabilidades**

3. $\sigma_1$ e $\sigma_2$: desvios padrão populacionais dos número de gols sofridos em casa e fora de casa

4. $n_1$ e $n_2$: números amostrais de cada população (gols sofridos dentro e fora de casa)

**Critério de decisão**: Se C pertence ao intervalo $-C < \mu_1 - \mu_2 < C$ não existem evidências estatísticas para afirmar que se deve rejeitar a hipótese nula ao nível de significância de $\alpha\%$, caso contrário se rejeita $H_0$.

Com $\sigma_1 = 1.183$ e $\sigma_2 = 0.982$ ao nível de significância de 5% ($z_{\frac{\alpha}{2}} = 1.96$)e $n_1$=$n_2$=170, temos:

$$C = 1.96 \cdot \sqrt{\tfrac{(1.183)^2}{170} + \tfrac{(0.982)^2}{170}} = 0.2203$$

$$\mu_1 - \mu_2 = 1.3235 - 1.0235 = 0,3$$

Desse modo, é possível perceber que $\mu_1 - \mu_2 = 0,3$ não pertence ao intervalo [-0.2203, 0.2203]. Logo, existem evidências estatísticas para afirmar que a média de gols sofridos em casa é diferente da média de gols sofridos fora de casa ao nível de significância de 5%, ou seja, deve-se rejeitar a hipótese nula ($H_0$).

## 2.2 Desvios padrão populacionais desconhecidos e iguais

As expressões analíticas para definir o valor crítico relacionado ao teste de hipótese homocedástica (abaixo), considerando desvios padrão populacionais desconhecidos, são (eq.2 e eq.3):

$$H_0 : \mu_1 - \mu_2 = 0$$
$$H_1 : \mu_1 - \mu_2 \neq 0$$

$$C = t_{\frac{\alpha}{2}} \cdot S_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \tag{2}$$

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} \tag{3}$$

Sendo:

1. C: valor crítico relacionado a um nível de significância

2. t: valor usado para representar a variável aleatória padronizada da tabela de **distribuição de probabilidades t-Student** para $n_1$ + $n_2$ - 2 graus de liberdade

3. $S_p$: estimador para $\sigma_1 = \sigma_2 = \sigma$ que agrega as informações das duas amostras

4. $n_1$ e $n_2$: números amostrais de cada população (gols sofridos dentro e fora de casa)

Novamente, o critério de decisão consiste em que o valor crítico C pertença, ou não, ao intervalo $-C < \mu_1 - \mu_2 < C$. Se $\mu_1 - \mu_2$ pertence a [-C, C] então não existem evidências estatísticas para afirmar que as médias $\mu_1$ seja diferente de $\mu_2$ a um nível de significância de $\alpha\%$. Caso contrário, se $\mu_1 - \mu_2$ não pertencer a [-C, C] deve-se rejeitar a hipótese nula com uma chance de erro de $\alpha\%$.

Calculando as variâncias amostrais via o banco de dados "budesliga.rds", temos: $S_1^2 = 1.3444$ e $S_2^2 = 1.2065$. Além disso, sabendo que o número de jogos fora e dentro de casa do ime Dortmund na amostra foi de 170, calcula-se $S_p$:

$$S_p^2 = \frac{(170-1)\cdot(1.3444)^2 + (170-1)\cdot(1.2065)^2}{(170+170-2)} = 1.2755 \rightarrow S_p = 1.1294$$

Com o valor de $S_p$, calcula-se o valor crítico C com um nível de significância de 5% ($t_{\frac{\alpha}{2}} = 1.967$ - 338 graus de liberdade):

$$C = 1.9674 \cdot 1.1294 \cdot \sqrt{\frac{1}{170} + \frac{1}{170}} = 0.2410$$

Portanto, como $\mu_1 - \mu_2 = 0.3$, existem evidências estatísticas para rejeitar $H_0$ ao nível de 5% de significância, uma vez que 0.3 não pertence ao intervalo [-0.241, 0.241], ou seja, as médias $\mu_1$ e $\mu_2$ podem ser ditas diferentes ao nível de significância de 5%.

## 2.3 Desvios padrão populacionais desconhecidos e diferentes

A equação do valor crítico é:

$$C = t_{\frac{\alpha}{2}} \cdot \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \tag{4}$$

Para este caso, desvios desconhecidos e desiguais, o número de graus de liberdade para obter o valor t da tabela de distribuição de probabilidades t-Studentes será dado por:

$$GL = \frac{(w_1 + w_2)^2}{(\frac{w_1^2}{n_1+1}) + (\frac{w_2^2}{n_2+1})} - 2 \tag{5}$$

Em que:

$$w_1 = \frac{S_1^2}{n_1} \qquad \text{e} \qquad w_2 = \frac{S_2^2}{n_2}$$

Sendo:

1. GL: número de graus de liberdade

2. $n_1$ e $n_2$: são os tamanhos das amostras de número de gols sofridos dentro e fora de casa.

Desse modo, calcula-se $w_1$, $w_2$ e, consequentemente, GL:

$$w_1 = \frac{1.344}{170} = 0.0071 \qquad \text{e} \qquad w_2 = \frac{1.2065}{170} = 0.0079$$

$$GL = \frac{(0.0071+0.0079)^2}{(\frac{0.0071^2}{170+1})+(\frac{0.0079^2}{170+1})} = 339.0039$$

Sendo assim, o valor t da distribuição t-Student de 5% de significância para 339 graus de liberdade valerá, aproximadamente, 1.9670. Portanto,

$$C = 1.967 \cdot \sqrt{\frac{1.3444}{170} + \frac{1.2065}{170}} = 0.2410$$

Dessa maneira, percebe-se que $\mu_1 - \mu_2 = 0.3$ não pertence ao intervalo definido pelos extremos de valor crítico [-0.241, 0.241]. Portanto, deve-se rejeitar a hipótese inicial de que a média de gols sofridos em casa é igual a média de gols sofridos fora de casa com uma chance de erro de 5%.

# 3 Comparação entre expressões analíticas e os resultados entre os três casos

Para cada caso foi utilizada uma expressão analítica:

1°Caso: Desvios padrão conhecidos

$$C = z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \tag{6}$$

2°Caso: Desvios padrão desconhecidos e iguais

$$C = t_{\frac{\alpha}{2}} \cdot S_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \tag{7}$$

3°Caso: Desvios padrão desconhecidos e diferentes

$$C = t_{\frac{\alpha}{2}} \cdot \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \tag{8}$$

Para o 1° caso, os valores de desvio padrão e variância são conhecidos, por esse motivo, o valor z da tabela de distribuição normal de probabilidades é usado, assim como os próprios valores de variâncias populacionais.

Para o 2° e 3° casos, é usado o valor t da distribuição de probabilidades t-Student, uma vez que não são conhecidos os valores de desvios padrão populacionais.

A diferença do 2° caso para o 3° caso é que as variâncias populacionais são iguais para o 2° caso e diferentes para o 3° caso.

Na prática para o 2°caso, como as variâncias são iguais, pode-se estimá-la por meio da variável usada $S_p$ com número de graus de liberdade igual a 338 (170 + 170 -2). Porém, para o terceiro caso, como as variâncias são diferentes, o número de graus de liberdade é calculado pela equação 5. Nesse 3° caso o número de graus de liberdade resultou em 339. Próximo ao número de graus de liberdade do 2° caso quando se assumiu que as variâncias são iguais.

Da tabela abaixo é possível perceber que a amplitude do intervalo para o caso de desvios padrão conhecidos é a menor em comparação com os dois outros casos para desvio padrão desconhecido. Isso acontece, porque, como já foi dito, não se conhece os valores de desvio padrão. Desse modo é preciso estimá-los por meio do desvio padrão amostral.

O desvio amostral, como o próprio nome já diz, foi obtido da amostra que não necessariamente representa verdadeiramente a população o que implica em maior incerteza dos resultados. Por esse motivo, é preciso aumentar o intervalo de valores críticos.

Não houve diferença de amplitude para o segundo e terceiro casos até a quarta casa decimal porque, primeiramente, os número de graus de liberdade são muito próximos (338 e 339). Além disso, o valor de $S_p = 1.1294$ é próximo das variâncias amostrais $S_1 = 1.3444$ e $S_2 = 1.2065$, uma vez que é possível analisar as diferenças relativas de $S_1$ e $S_2$ a $S_p$ abaixo:

$$D_1 = \frac{|S_1 - S_p|}{S_1} = \frac{|1.3444 - 1.1294|}{1.3444} = 0,16$$
$$D_2 = \frac{|S_1 - S_p|}{S_1} = \frac{|1.2065 - 1.1294|}{1.2065} = 0,0725$$

Sendo:

1. $D_1$: diferença relativa de $S_1$ para $S_p$
2. $D_2$: diferença relativa de $S_2$ para $S_p$

| | **Intervalos dados pelos valores críticos** | **Amplitude do intervalo** |
|---|---|---|
| **1º Caso: Desvios padrão conhecidos** | [- 0.2203 , 0.2203] | 0.4406 |
| **2º Caso: Desvios padrão desconhecidos e iguais** | [- 0.2410 , 0.2410] | 0.4819 |
| **3º Caso: Desvios padrão desconhecidos e diferentes** | [-0.2410 , 0.2410] | 0.4819 |

## 3.1 Aumento ou diminuição do nível de significância

O nível de significância significa para teste de hipótese a chance de erro em rejeição ou não rejeição da hipótese. Se aumentar o nível de significância, aumenta-se a chance de erro de decisão do teste e, consequentemente, isso significa que se diminui o tamanho do intervalo dado pelo valor crítico calculado. Existe, portanto, um menor intervalo de onde a variável amostral pode estar para não rejeitar o teste, mas por outro lado, a chance de erro do teste é maior.

Para o caso de diminuir a chance de erro, o intervalo definido pelos valores críticos aumenta. Dá-se um intervalo maior de onde a variável amostral pode pertencer para não rejeitar o teste, mas por outro lado, a chance de erro é menor.

# 4 Teste de hipótese de duas variâncias

Jornalistas de uma redação esportiva gostariam de saber se a dispersão de gols sofridos em casa é igual à dispersão de gols sofridos fora de casa. Para isso faz-se o teste de hipótese:

$$H_0 : \sigma_1^2 = \sigma_2^2 = \sigma^2$$
$$H_1 : \sigma_1^2 \neq \sigma_2^2$$

Em que:

1. $\sigma_1^2$: variância de gols sofridos fora de casa
2. $\sigma_2^2$: variância de gols sofridos em casa

A hipótese nula significa que a dispersão de gols sofridos em casa é igual a dispersão de gols sofridos fora de casa. Já a hipótese alternativa diz que essas dispersões são diferentes.

## 4.1 Expressão analítica

Para populações que seguem uma distribuição normal, o teste de hipótese para a variância segue a distribuição de **Fisher-Snedecor (distribuição F)** que é definido por:

$$F_{amostral} = F_{n_1-1,n_2-1} = \frac{S_1^2}{S_2^2} \tag{9}$$

e para $\alpha\%$ de significância:

$$F_{tabelado} = F_{crit} = F_{n_1-1,n_2-1,\alpha\%} \tag{10}$$

Como critério de decisão do teste de hipótese:

1. Se $F_{amostral} \geq F_{tabelado} = F_{crit}$, deve-se rejeitar a hipótese nula ($H_0$) com uma chance de erro de $\alpha\%$.
2. Caso contrário, se $F_{amostral} \leq F_{tabelado} = F_{crit}$, não existem evidências estatísticas para afirmar que as variâncias são diferentes entre si, com um nível de significância de $\alpha\%$.

Considerando $\alpha = 5\%$ e os **dois** graus de liberdade para a tabela dde distribuição Fischer: $GL_1 = \mathbf{169}$ **e** $GL_2 = \mathbf{169}$:

$$F_{tabelado} = F_{crit} = F_{169,169,5\%} = 1.3533$$
e
$$F_{amostral} = \frac{1.3444}{1.2065} = 1.1143$$

Desse modo, $F_{amostral} = 1.1143 \leq F_{tabelado} = F_{crit} = 1.3533$ e, assim, não existem evidências estatísticas para afirmar que as variâncias são diferentes entre si a um nível de significância de 5%. Em resumo, não se deve rejeitar a hipótese nula com uma chance de erro de 5%.

### 4.2 Graus de Liberdade

Grau de liberdade indica a quantidade de informações disponíveis pelo tamanho da amostra para que seja possível estimar variáveis desconhecidas. Os graus de liberdade são definidos pelo tamanho da amostra, quanto maior for o número de elementos da amostra maior é a precisão da estatística usada.

Por exemplo, para usar os valores z da tabela de probabilidades da distribuição normal devem ser conhecidos os parâmetros: desvio padrão e variância populacionais. Quando, por exemplo, o desvio padrão e a variância populacionais são desconhecidos, a distribuição de probabilidades muda para a distribuição t-Student, uma vez que há dependência do tamanho da amostra, ou seja, se a variância é desconhecida e pretende-se estimá-la tem-se maior precisão na estimativa se o tamanho da amostra aumentar (se o número de graus de liberdade aumentar).

Logo, se o número do tamanho da amostra aumenta para infinito significa que a variância e desvio amostrais tendem aos valores de variância e desvio padrão populacionais, ou seja, com o aumento do número da amostra, a distribuição t-Student tende à distribuição normal de probabilidades.

## 5 Conclusão

## 6 Script - Case 7

```
    # Camille Peixoto Almeida 12702259 - CASE 7

# importar a biblioteca
library(tidyverse)

# selecionar a base de dados
df <- readRDS("H:/Meu Drive/USP/semestres_passados/1ºQuadri2023/reof_estat/
Estudo de Caso 5 - Teste de Hipóteses I-20230510/Case5/bundesliga.rds")

df_HomeDortmund <- subset(df,df$HomeTeam == "Dortmund")
df_AwayDortmund <- subset(df, df$AwayTeam == "Dortmund")
```

```
# PARTE 1: TESTE DE HIPÓTESES 2 PARÂMETROS - MÉDIA
# PARA DESVIOS PADRÃO POPULACIONAIS CONHECIDOS
# H0: média sofridos em casa = média sofridos fora de casa
# H1: média sofridos em casa DIFERENTE média sofridos fora de casa

media_tomados_emCasa <- mean(df_HomeDortmund$FullTimeAwayGoals)
media_tomados_foraDeCasa <- mean(df_AwayDortmund$FullTimeHomeGoals)

diferenca_media <- media_tomados_foraDeCasa - media_tomados_emCasa

desv_pad_pop_tomados_emCasa <- 0.982
desv_pad_pop_tomados_foraDeCasa <- 1.183
n1<-170
n2<-170

# Critério: não se rejeita H0 se: Crítico < media1-media2 < Crítico
# 5% de significância
z5 <- 1.96

Critico_conhe <- z5*((desv_pad_pop_tomados_emCasa^2)/n1 +
(desv_pad_pop_tomados_foraDeCasa)/n2)^0.5

# neste caso -0,3 não pertence ao intervalo [0,22;0,22] logo se rejeita H0
# existem evidências estatísticas para afirmar que as médias não são iguais de
# gols sofridos dentro e fora de casa com um nível de significância de 5%


# PARA DESVIOS PADRÃO POPULACIONAIS DESCONHECIDOS MAS IGUAIS
# H0: média sofridos em casa = média sofridos fora de casa
# H1: média sofridos em casa DIFERENTE média sofridos fora de casa

var_tomados_emCasa_amostral <- var(df_HomeDortmund$FullTimeAwayGoals)
var_tomados_foraDeCasa_amostral <- var(df_AwayDortmund$FullTimeHomeGoals)

Sp_quadrado <- ((n1-1)*var_tomados_emCasa_amostral +
(n2-1)*var_tomados_foraDeCasa_amostral)/(n1+n2-2)
Sp <- Sp_quadrado^0.5
# 336 graus de liberdade
t5 <- 1.9670
Critico_desc_igual <- t5*(Sp_quadrado)^0.5*(1/n1+1/n2)^0.5

# neste caso -0,3 não pertence ao intervalo [0,24;0,24] logo se rejeita H0
# existem evidências estatísticas para afirmar que as médias não são iguais de
# gols sofridos dentro e fora de casa com um nível de significância de 5%
```

```
# PARA DESVIOS PADRÃO POPULACIONAIS DESCONHECIDOS E NÃO IGUAIS
# H0: média sofridos em casa = média sofridos fora de casa
# H1: média sofridos em casa DIFERENTE média sofridos fora de casa

w1 <- var_tomados_emCasa_amostral/n1
w2 <- var_tomados_foraDeCasa_amostral/n2

GL <- (w1+w2)^2/(((w1^2)/(n1+1))+((w2^2)/(n2+1))) - 2
# Para um grau de liberdade de 339,0 tem-se t5 =
t5GL339 <- -qt(0.025, 339.0039)


Critico_desc_nao_igual <- t5GL339*(var_tomados_emCasa_amostral/n1 +
var_tomados_foraDeCasa_amostral/n2)^0.5

#tamanhos dos intervalos:
# desvios padrão conhecidos:
tamanho_intervalo_conhec <- 2*Critico_conhe

# desvios padrão desconhecidos e iguais:
tamanho_intervalo_desc_igual <- 2*Critico_desc_igual

# desvios padrão desconhecidos e diferentes:
tamanho_intervalo_desc_nao_igual <- 2*Critico_desc_nao_igual

# PARTE 2:
# H0: variância sofridos em casa = variância sofridos fora de casa
# H1: variância sofridos em casa DIFERENTE variância sofridos fora de casa

# Para 5% de significância
F5GL169169 <- qf(0.025,169,169)^(-1)
Famostral <- var_tomados_foraDeCasa_amostral/var_tomados_emCasa_amostral
```

# 7 Referências


1. RAMOS, Alberto. Apostila de Estatística-PRO3200. Escola Politécnica da Universidade de São Paulo, Departamento de Engenharia de Produção, São Paulo.2021

2. HO, Linda Lee; RIBEIRO, Celma de Oliveira. Intervalo de confiança.PRO3200 - Estatística, Departamento de Engenharia e Produção, Universidade de São Paulo.2022.