Case6

Camille Peixoto Almeida n°USP: 12702259

19 de maio de 2023

1 Teste de Hipóteses: Um parâmetro

Segundo um jornalista, o time Dortmund sofre 1 gol em média quando joga em casa. Porém, a imprensa desconfia dessa informação. Desse modo, é relevante fazer o seguinte teste de hipótese:

'Verdade atual' - hipótese nula - H_0: $\mu_0 = 1$
Hipótese a ser testada - H_1: $\mu_1 \neq 1$

Em que μ_0 e μ_0 são as médias de gols tomados em casa para cada hipótese.

1.1 Expressão analítica - desvio padrão populacional conhecido

Para testar se a média de gols tomados em casa é diferente de 1, cria-se um intervalo $[X_{1_{crit}}, X_{2_{crit}}]$ em que:

$$X_{1_{crit}} = \mu_0 - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \tag{1}$$

$$X_{2_{crit}} = \mu_0 + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \tag{2}$$

Em que:

1. $X_{1_{crit}}$: valor crítico inferior
2. $X_{2_{crit}}$: valor crítico superior
3. μ_0: média de gols tomados em casa (hipótese inicial)
4. σ: desvio padrão populacional
5. n: número amostral
6. z: valor usado para representar a variável aleatória padronizada da tabela de **distribuição normal de probabilidades**

Se a média amostral dos gols tomados em casa **não** pertencer ao intervalo $[X_{1_{crit}}, X_{2_{crit}}]$ significa que se deve rejeitar a hipótese nula, ou seja, pode-se afirmar que a média de gols tomados em casa é diferente de 1 com uma chance de erro de α %. Caso contrário, se a média pertencer ao intervalo $[X_{1_{crit}}, X_{2_{crit}}]$, não se deve rejeitar H_0, pois não existem evidências estatísticas para afirmar que a média de gols tomados seja diferente de 1 com uma chance de erro de α %.

Considerando $\sigma = 0.982$ e nível de significância 5%:

$$X_{1_{crit}} = 1 - 1.96 \cdot \frac{0.982}{\sqrt{170}} = 0.8524 \qquad X_{2_{crit}} = 1 + 1.96 \cdot \frac{0.982}{\sqrt{170}} = 1.1476$$

Com:

1. $\mu_0 = 1$
2. $z_{2.5\%} = 1.96$
3. $\sigma = 0.982$
4. n = 170

Como a média amostral ($\overline{X}_{amostral} = 1.0235$) está contida no intervalo $[X_{1_{crit}}, X_{2_{crit}}]$, não se deve rejeitar H_0, porque não existem evidências estatísticas para afirmar que a média de gols é diferente de 1 com uma chance de erro de 5% de significância.

1.2 Expressão analítica - desvio padrão populacional desconhecido

Novamente, para testar se a média de gols tomados em casa é diferente de 1, cria-se um intervalo $[X_{1_{crit}}, X_{2_{crit}}]$ em que:

$$X_{1_{crit}} = \mu_0 - t_{\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n}} \tag{3}$$

$$X_{2_{crit}} = \mu_0 + t_{\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n}} \tag{4}$$

Em que:

1. $X_{1_{crit}}$: valor crítico inferior
2. $X_{2_{crit}}$: valor crítico superior
3. μ_0: média de gols tomados em casa (hipótese inicial)
4. S: desvio padrão amostral
5. n: número amostral

6. t: valor usado para representar a variável aleatória padronizada da tabela de **distribuição de probabilidades t-Student**

Mais uma vez, se a média amostral dos gols tomados em casa **não** pertencer ao intervalo $[X_{1_{crit}}, X_{2_{crit}}]$ significa que se deve rejeitar a hipótese nula, ou seja, pode-se afirmar que a média de gols tomados em casa é diferente de 1 com uma chance de erro de α %. Caso contrário, se a média pertencer ao intervalo $[X_{1_{crit}}, X_{2_{crit}}]$, não se deve rejeitar H_0, pois não existem evidências estatísticas para afirmar que a média de gols tomados seja diferente de 1 com uma chance de erro de α %.

Considerando o desvio padrão amostral (S = 1.0985) e nível de significância 5%:

$$X_{1_{crit}} = 1 - 1.9741 \cdot \frac{1.0985}{\sqrt{170}} = 0.8337 \qquad X_{2_{crit}} = 1 + 1.9741 \cdot \frac{1.0985}{\sqrt{170}} = 1.1663$$

Com:

1. $\mu_0 = 1$
2. $t_{2.5\%} = 1.9741$ (graus de liberdade = 169)
3. S = 1.0985
4. n = 170

Novamente, a média amostral ($\overline{X}_{amostral} = 1.0235$) está contida no intervalo $[X_{1_{crit}}, X_{2_{crit}}]$, por esse motivo, não se deve rejeitar H_0, porque não existem evidências estatísticas para afirmar que a média de gols é diferente de 1 com uma chance de erro de 5% de significância. Isso significa que, com uma chance de erro de 5%, a média de gols tomados em casa do time Dortmund continua 1.

1.3 Comparação entre casos: σ conhecido e σ desconhecido

Para um mesmo valor de significância (5%), calcula-se o intervalo de valores críticos no caso do desvio populacional ser conhecido e igual a 0.982 com valores de probabilidades (z) da tabela de distribuição normal e o intervalo de valores críticos no caso de desvio padrão populacionalser desconhecido em que se usa o desvio padrão amostral e valores de probabilidades (t) da tabela de distribuição t-Student. Os intervalos estão retratados na tabela abaixo:

σ **conhecido**		σ **desconhecido**	
X1crí	**X2crí**	**X1crí**	**X2crí**
0.8524	1.1476	0.8337	1.1663
Amplitude		**Amplitude**	
0,2952		0,3326	

Vê-se que, para um mesmo nível de significância, o intervalo dado pelos valores críticos no caso de σ desconhecido teve uma amplitude maior que para o caso de σ conhecido. Isso significa que se consegue diferenciar para um intervalo mais restrito de valores: aqueles que pertencem à distribuição amostral centrada na média 1 (não rejeição da hipótese inicial H_0) e aqueles que não pertecem (rejeição da hipótese inicial H_0, e aceitação da hipótese alternativa H_1).

Isso está relacionado, sob uma visão numérica, ao motivo de se desconhecer a variância e, consequentemente, o desvio padrão populacional. Isso implica numa maior incerteza (maior amplitude do intervalo definido pelos valores críticos calculados), uma vez que é necessário estimar o desvio padrão populacional pelo amostral (depende da amostra e não representa por completo a população).

Portanto, quando σ é desconhecido existe mais incerteza no teste, existe uma maior tendência de afirmar que a média do número de gols tomados em casa pelo time Dortmund vale 1 (a amplitude do intervalo é maior) quando, talvez, seja diferente de 1.

Ao final das análises, conclui-se, ao nível de 5% de significância, que o jornalista está correto, a média de gols tomados pelo time Dortmund em casa vale 1, ou seja, a hipótese do jornalista não deve ser rejeitada.

2 Teste de hipótese para a variância

Para saber como variam os gols sofridos pelo time Dortmund jogando em casa é importante analisar a variância. Supondo como hipótese inicial que a variância populacional dos gols sofridos em casa seja 2.25 gol^2. Depois do time Dortmund perder de goleada, é relevante testar se a variância é maior que o valor da hipótese inicial. Nesta parte do relatório, faz-se o teste de hipótese com nível de significância de 10%:

'Verdade atual' - hipótese nula - H_0: $\sigma^2 = \sigma_0^2$
Hipótese a ser testada - H_1: $\sigma^2 > \sigma_0^2$

Em que σ^2 e σ_0^2 são as variâncias de gols tomados em casa para cada hipótese.

2.1 Expressão analítica

$$S_{crit}^2 = b \cdot \frac{\sigma^2}{(n-1)} \tag{5}$$

Sendo:

1. S_{crit}^2: o quadrado do valor crítico de barreira para o teste
2. b: valor usado para representar a variável aleatória (variância) padronizada da tabela de **distribuição de probabilidades qui-quadrado**.
3. σ^2: variância de gols sofridos pelo Dortmund jogando em casa.

4. n: número amostral

Se a variância amostral dos gols tomados em casa for superior ao S^2_{crit}, significa que se deve rejeitar a hipótese nula, ou seja, pode-se afirmar que a variânca de gols tomados em casa é maior que o valor de referência 2.25 gol^2 com uma chance de erro de $\alpha\%$. Caso contrário, se a variância for inferior ao valor S^2_{crit}, não se deve rejeitar H_0, pois não existem evidências estatísticas para afirmar que a variância seja maior 2.25 gol^2 com uma chance de erro de $\alpha\%$.

Considerando $\sigma^2 = 2.25$ e nível de significância de 10%:

$$S^2_{crit} = 192.948 \cdot \frac{2.25}{(170-1)} = 2.5688$$

Com:

1. b = 192.948 ($\alpha = 10\%$)
2. n = 170

Como a variância amostral ($S^2 = 1.2065$) **não** ultrapassa S^2_{crit}, não se deve rejeitar H_0, porque não existem evidências estatísticas para afirmar que a variância de gols tomados pelo time de Dortmund em casa é maior que 2.25 com uma chance de erro de 10% de significância. Isso significa que, com uma chance de erro de 10%, a variância de gols tomados em casa por Dortmund continua valendo 2.25 gol^2.

Ao final da análise, pode-se dizer que a variância dos gols sofridos **não** é maior que 2.25 $gols^2$ (não se rejeita H_0) com uma chance de erro de 10%.

2.2 Para 5% de significância

Considerando $\sigma^2 = 2.25$ e nível de significância de 5%:

$$S^2_{crit} = 200.334 \cdot \frac{2.25}{(170-1)} = 2.6672$$

Com:

1. b = 200.334 ($\alpha = 5\%$)
2. n = 170

Mais uma vez, como a variância amostral ($S^2 = 1.2065$) **não** ultrapassa S^2_{crit}, não se deve rejeitar H_0, porque não existem evidências estatísticas para afirmar que a variância de gols tomados pelo time de Dortmund em casa é maior que 2.25 com uma chance de erro de 5% de significância. Isso significa que, com uma chance de erro de 5%, a variância de gols tomados em casa por Dortmund continua valendo 2.25 gol^2.

Ao final da análise, pode-se dizer que a variância dos gols sofridos **não** é maior que 2.25 $gols^2$ (não se rejeita H_0) com uma chance de erro de 5%.

2.3 Comparação dos casos de 5% e 10%

Os valores, portanto, de S^2_{crit} para 5% e 10% são:

	5%	**10%**
S^2_{crit}	**2.6672**	**2.5688**

Percebe-se que o valor de S^2_{crit} é maior para o caso em que o nível de significância (chance de erro) é menor, 5%. Isso acontece, porque ao diminuir a chance de erro é necessário, em contrapartida, aumentar a barreira (S^2_{crit}) que diferenciará os valores de variâncias na condição de não rejeição da hipótese nula ($S^2 < S^2_{crit}$) e os valores de variâncias na condição de rejeição da hipótese nula ($S^2 > S^2_{crit}$).

3 Script - Case 6

```
    # Camille Peixoto Almeida 12702259 - CASE 0

# importar a biblioteca
library(tidyverse)

# selecionarr a base de dados
df <- readRDS("H:/Meu Drive/USP/semestres_passados/1°Quadri2023/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")

# o Dortmund sofre 1 gol em média todo jogo como mandante - verdade atual
# provar que a quantidade de gols sofrida pelo time em jogos como mandante é
# diferente de 1 gol.

# TESTE DE HIPÓTESE
# H0: u = 2.6
# H1: u diferente de 2.6

media_gols_tomados <- mean(df_HomeDortmund$FullTimeAwayGoals)

dev_pad_populacional <- 0.982
media0 <- 1

n <- 170
z <- 1.96

x1crit <- media0 - z*dev_pad_populacional/(n)^0.5
```

```
x2crit <- media0 + z*dev_pad_populacional/(n)^0.5

hist(df_HomeDortmund$FullTimeAwayGoals,
     breaks = 10,
     freq = T,
     col = "yellow",
     ylab = "Frequência",
     xlab = "Gols tomados em casa",
     main = "Histograma de médias amostrais")

# como a média amostral vale 1 e pertence ao intervalo [x1crit, x2crit]
#não rejeitamos a hipótese inicial
#################################################################
# desvio padrão desconhecido

media_gols_tomados <- mean(df_HomeDortmund$FullTimeAwayGoals)
dev_pad_gols_tomados <- sd(df_HomeDortmund$FullTimeAwayGoals)

media0 <- 1

n <- 170
t <- 1.9741

x1crit_desc <- media0 - t*dev_pad_gols_tomados/(n)^0.5
x2crit_desc <- media0 + t*dev_pad_gols_tomados/(n)^0.5

# como a média amostral vale 1 e pertence ao intervalo [x1crit, x2crit]
# não rejeitamos a hipótese inicial


#qnorm(0.05)
#qt(0.025, 169)
#qchiq(0.9, df = 169)
#####################################

# TESTE DE HIPÓTESE
# H0: var0 = 1.5^2
# H1: var1 > var0

var_populacional <- 1.5^2
var_amostral <- var(df_HomeDortmund$FullTimeAwayGoals)

b10porcento <- 192.948
b5porcento <- 200.334

Scrit_aoquadrado <- b5porcento*var_populacional/(n-1)
```

```
# como Scrit_quadrado > var_amostral não podemos rejeitar H0

#considerando 5% de significância
```