Case 5 - Teste de Hipóteses I

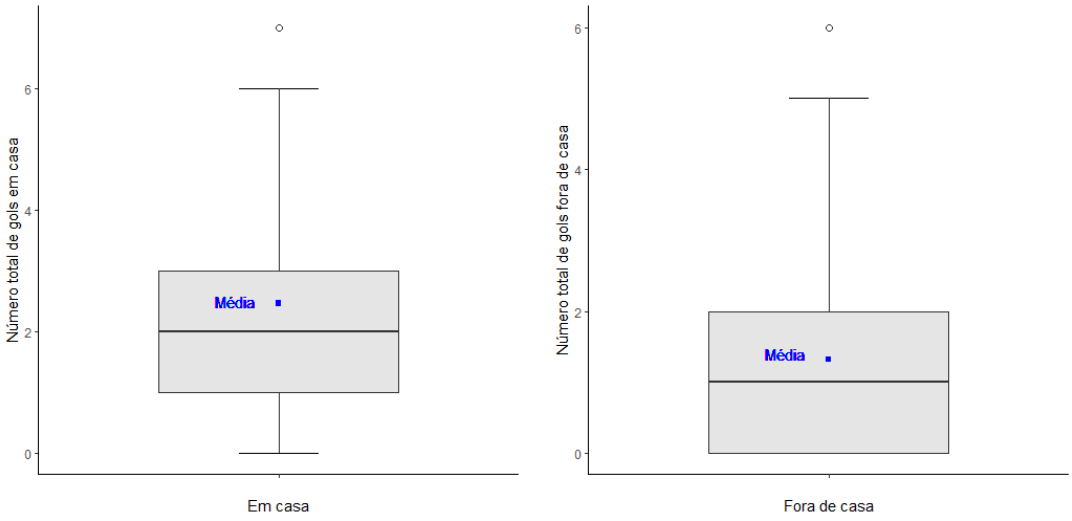
Camille Peixoto Almeida

14 de maio de 2023

Para este relatório, analisa-se o desempenho de gols da equipe Dortmund a partir da planilha "bundesliga.rds"do banco de dados kaggle.

1 Análise boxplot

Para analisar o desempenho de gols do time Dortmund é interessante questionar se existe diferença no cenário de jogo dentro e fora de casa. Para elucidar isso, é válido usar a representação boxplot abaixo.



Dos dois boxplot acima, pode-se ver que a média do número de gols feitos dentro de casa é maior que a média de gols fora de casa. Assim, vê-se um indício de que o time Dortmund tende a fazer mais gols em casa.

Além disso, nos dois casos, a mediana (2° quartil) fica, aproximadamente, bem centralizada no conjunto de dados, entre o primeiro e o terceiro quartis, tanto para dentro como para fora de casa. Isso confere a característica de

simetria à distribuição, ou seja, as distribuições para os dois casos possuem essa simetria.

Visualmente, a dispersão dos dados pode ser representada pelo intervalo interquartílico que é a diferença entre o terceiro quartil e o primeiro quartil. Para ambos, esse intervalo foi de, aproximadamente, 2 gols. Isso indica que a variância (ou também desvio padrão) podem ser semelhantes.

Analisando também a amplitude (os valores máximos e mínimos), vê-se que em casa a amplitude foi de 7 (de 0 a 7 gols, aproximadamente) e fora de casa foi de 6 (0 a 6 gols, aproximadamente). A amplitude pode ser de mais fácil entendimento, porém ela é influênciada pelos outiers, o que a torna uma medida estatística menos robusta que o intervalo interquartílico.

2 Medidas de tendência central

Medidas de Disperção Centrais	**Em casa**	**Fora de Casa**
Mínimo	0	0
1°Quartil	1	1
Mediana	2	2
Média	2,465	1,924
3°Quartil	3	3
Máximo	7	6

Com as medidas da tabela acima, de fato, quando o time Dortmund joga em casa, ele faz mais gols, pois sua média em casa foi maior que fora de casa.

Fora isso, o intervalo interquartílico para os dois casos foi de 2, com terceiro quartil igual a 3 e primeiro quartil igual a 1. Assim, a mediana (o valor central que divide 50% dos dados) é o valor médio entre os quartis e isso indica uma simetria da distribuição dos dados.

3 Teste de hipótese

Supondo que um antigo jornalista afirmou que a média de gols **em casa** do time Dortmund seja igual a 2.6 gols, gostaria de se testar se essa média foi alterada com as recentes derrotas:

H_0 (referência - "verdade atual"): $\mu_0 = 2.6$
H_1 (hipótese testada - "descrédito recente"): $\mu_1 < \mu_0$
com nível de significância de 5%

3.1 Teste de hipótese com desvio populacional conhecido

$$H_0 : \mu_0 = 2.6$$
$$H_1 : \mu_1 < \mu_0$$

Com a expressão analítica para uma **distribuição normal de probabilidades**:

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

Para desvio padrão populacional (σ) igual a 0.5, nível de significância de 5% (z = 1.645) e n (número amostral) = 170, temos:

$$X_{crit} = 2,6 - 1,645 \cdot \frac{0,5}{\sqrt{170}} = 2.537$$

Como a média amostral vale $\overline{X} = 2.465$, então:

$$\overline{X} = 2.465 < X_{crit} = 2.537$$

Isso significa que se deve rejeitar H_0 (a hipótese inicial, hipótese do jornalista), ou seja, têm-se evidências estatísticas para afirmar que, de fato, a média de gols é **menor** que 2.6 com a chance de erro de, no máximo, 5% de significância, porque a média de gols da amostra foi ainda menor que a barreira definida pelo valor crítico calculado.

3.2 Teste de hipótese com desvio populacional desconhecido

$$H_0 : \mu_0 = 2.6$$
$$H_1 : \mu_1 < \mu_0$$

Com a expressão analítica para uma **distribuição de probabilidades t-Student**:

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

Em que S é o desvio amostral igual a 1.562.

Para desvio padrão populacional (σ) desconhecido, nível de significância de 5% (t = 1.653) e n (número amostral) = 170, temos:

$$X_{crit} = 2,6 - 1,653 \cdot \frac{1.562}{\sqrt{170}} = 2.402$$

Desse modo,

$$\overline{X} = 2.465 > X_{crit} = 2.402$$

Isso significa que **não** se pode rejeitar a hipótese inicial H_0 (não se pode "confrontrar"com a hipótese alternativa H_1), ou seja, não se têm evidências estatísticas para afirmar que a média de gols foi reduzida com a influência das derrotas recentes do time Dortmund, pois a média amostral calculada não foi menor que a barreira do valor crítico.

3.3 Comparação entre casos: σ conhecido e σ desconhecido

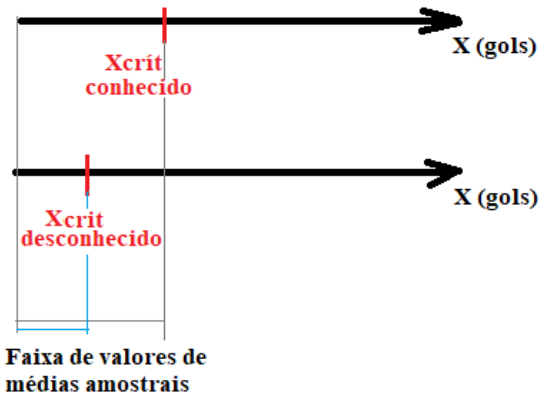
Para um mesmo valor de significância (5%), calcula-se os valores críticos no caso do desvio populacional ser conhecido e igual a 0.5 com valores de probabilidades (z) da tabela de distribuição normal e valores críticos no caso de desvio padrão populacional desconhecido em que se usa o desvio padrão amostral e valores de probabilidades (t) da tabela de distribuição t-Student:

5 %	**Sigma conhecido**	**Sigma desconhecido**
X crítico	2.537	2.402

Vê-se que o valor crítico para σ conhecido foi maior que o valor crítico para σ desconhecido a um mesmo nível de significância. Isso significa que, para um mesmo erro de 5%, se consegue afirmar, para valores mais altos de médias amostrais, se as derrotas recentes impactaram ou não na diminuição da média de gols. Em resumo, é possível diferenciar mais valores de médias amostrais que implicam na rejeição da hipótese nula para um mesmo nível de significância.

Isso está relacionado, sob uma visão numérica, pelo motivo de se desconhecer a variância e, consequentemente, o desvio padrão populacional. Isso implica numa maior incerteza, uma vez que é necessário estimar o desvio padrão populacional pelo amostral (depende da amostra e amostral não representa por completo a população). Portanto, quando σ é desconhecido existe mais incerteza no teste, existe uma maior tendência de afirmar que as derrotas recentes não impactaram na média quando, realmente, influenciaram na diminuição da média.

Figura 2



Ao final da análise, foi dito que, quando σ populacional é conhecido, existem evidências estatísticas para afirmar que a média de gols foi reduzida com a influência das derrotas recentes do time Dortmund com um nível de significância de 5% (ou seja, rejeita-se a afirmação do jornalista de que a média seja 2.6). Porém, quando σ é desconhecido, não há evidências estatísticas para rejeitar H_0 com o mesmo nível de significância.

3.4 Teste de hipótese com desvio populacional conhecido e desconecido com 1% de significância

Para 1% de significância (z = 2,325 e t = 2,326), calcula-se os valores críticos da média de gols do time Dortmund:

1 %	**Sigma conhecido**	**Sigma desconhecido**
X crítico	2.512	2.321

Vê-se que para σ conhecido:

$$\overline{X} = 2.465 < X_{crit} = 2.512$$

Portanto, pode-se concluir que se deve rejeitar H_0 (a hipótese inicial, hipótese do jornalista), ou seja, têm-se eviências estatística para afirmar que, de fato, a média de gols é menor que 2.6 com chance de erro máximo 1%, porque a média de gols da amostra foi ainda menor que a barreira definida pelo valor crítico calculado.

E para σ desconhecido:

$$\overline{X} = 2.465 > X_{crit} = 2.321$$

Para este caso, não se pode rejeitar a hipótese H_0, ou seja, não existem eviências estatísticas para afirmar que a média de gols foi diminuída pelas derrotas recentes de Dortmund, pois a média amostral calculada não foi menor que a barreira do valor crítico.

3.5 Comparação entre casos: σ conhecido ou desconecido com nível de significância 1%

Novamente, o valor crítico calculado para σ conhecido foi maior que para o caso de σ desconhecido para 1% de significância. Isso significa que é possível afirmar se as derrotas recentes de Dortmund impactaram ou não na diminuição da média de gols para valores mais altos de médias amostrais, ou seja, é possível diferenciar uma faixa maior de valores de médias amostrais que implicam na rejeição de H_0.

Mais uma vez isso está relacionado com o desconhecimento do desvio padrão populacional em que a estimativa é feita pelo desvio amostral que está ligado a uma amostra que não é de fato a população. Logo, para σ desconhecido existe

maior incerteza no teste, existe uma maior tendência de afirmar que as derrotas recentes não diminuíram a média.

Ao final das análise, pode-se concluir que quando σ populacional é conhecido, existem evidências estatísticas para rejeitar a afirmação do jornalista. Porém, quando σ é desconhecido, não há evidências estatísticas para rejeitar H_0 com o mesmo nível de significância.

Essa conclusão foi a mesma que para 5% de significância, ou seja, diminuir a chance de erro para 1% não implicou na não rejeição para quando σ é conhecido. É interessante que o $\overline{X}_{crit}$ conhecido para 5% de significância foi maior que para 1% de significância. Isso é esperado, porque ao diminuir sua chance de erro é normal que para rejeitar H_0 seja necessário uma valor crítico de barreira um pouco menor. Essa análise também serve para quando o σ é desconhecido e é exemplificada já que $\overline{X}_{crit}$ desconhecido para 5% de significância é maior que para 1%.

4 Script-Case5

```
    # 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")


#boxplot(df$FullTimeHomeGoals, df$FullTimeAwayGoals, names= c("Em casa",
"Fora de casa"), main = "Número de gols")


# desempenho do Dortmund em casa
ggplot(df_HomeDortmund, aes(y = df_HomeDortmund$FullTimeHomeGoals, x = "")) +

  geom_errorbar(stat = "boxplot", width = 0.2)+ # barras

  geom_boxplot(width = 0.6,
               fill = "grey90",
               outlier.size = 2,
               outlier.shape = 1) + # tamanho da caixa

  geom_point(stat="summary", fun= "mean", col = "blue", shape = 15)+
```

```
  labs(title = "",y = "Número total de gols em casa", x = "Em casa")+
  geom_text(x = 0.89, y=2.5, label = "Média", col = "blue")+
  theme_classic()

# desempenho do Dortmund fora de casa
ggplot(df_AwayDortmund, aes(y = df_AwayDortmund$FullTimeHomeGoals, x = "")) +

  geom_errorbar(stat = "boxplot", width = 0.2)+ # barras

  geom_boxplot(width = 0.6,
               fill = "grey90",
               outlier.size = 2,
               outlier.shape = 1) + # tamanho da caixa

  geom_point(stat="summary", fun= "mean", col = "blue", shape = 15)+
  labs(title = "",y = "Número total de gols fora de casa", x = "Fora de casa")+
  geom_text(x = 0.89, y = 1.4, label = "Média", col = "blue")+
  theme_classic()


medidas_estat_Home <- summary(df_HomeDortmund$FullTimeHomeGoals)
medidas_estat_Away <- summary(df_AwayDortmund$FullTimeAwayGoals)


#teste de hipótese em casa

# H0: media0_Home = 2.6
# H1: media0_Home < 2.6
# nível de significância: 5% -> z = 1,645
summary(df_HomeDortmund$FullTimeHomeGoals)

desv_Home <- 0.5
media0_Home <- 2.6
z95 <- 1.645
z99 <- 2.325
z<- z99
n_Home <- 170
Xcrit_Home <- media0_Home - z*desv_Home/(n_Home^0.5)
Xcal<- 2.465

# Como X amostral < Xcrit -> rejeito H0


# desvio padrão desconhecido
# se eu não conheco a variância eu tenho mais incerteza e o intervalo aumenta
#a amplitude do intervalo aumenta
```

```
desv_Home_desc <- sd(df_HomeDortmund$FullTimeHomeGoals)
t <- 2.326
Xcrit_Home_desc <- media0_Home - t*desv_Home_desc/(n_Home^0.5)

# Como Xcrit_Home_desc > Xcalc : rejeito H0
```

5 Referências

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