TPC 8: Teste de Hipóteses - Turmas 1 e 2

André Filipe Gomes Silvestre Nº104532 CDB1

Exercício

Num inquérito sobre óculos de sol foram colocadas várias questões aos inquiridos. Para além de características sociodemográficas (sexo, idade e nível de educação), perguntou-se o tipo de óculos de sol que possuíam, quando tinham sido adquiridos, onde tinham sido adquiridos, quanto tinham custado e se eram da marca *SoleMio(SM/RB)*.

Para além destas questões, ainda foram colocadas outras que originaram a construção de um conjunto de indicadores, cada um numa escala contínua de 0 a 10 – fatores que influenciam a compra de óculos de sol.

Para este TPC, irão apenas analisar duas questões: **1.** O indicador *"Importância do Preço na compra de óculos de sol"* – variável ***Price***; e, **2.** a questão ***"are_RB"***, que indica se os óculos são ou não da marca *SoleMio*

Os "Fatores que influenciam a compra de óculos de sol" são variáveis que assumem valores reais no intervalo 0-10, onde 0 corresponde a "nada importante" e 10 corresponde a "extremamente importante".

1. Será que a importância concedida ao preço está, em termos médios, acima do ponto intermédio da escala (i.e. 5)?
2. Será que homens e mulheres diferem, em termos médios, na importância concedida ao preço? (Suponha que as variâncias do preço nos dois grupos, embora sejam desconhecidas, podem ser consideradas iguais)

Responda a estas questões através da aplicação e interpretação de um teste de hipóteses adequado.

Defina as populações em análise e os parâmetros de interesse.

Siga os passos indicados nos slides (Etapas de um teste de hipóteses).

Considere uma **significância de referência, α, de 5%**.

Resolução

1.

```
# Leitura do ficheiro Estudo_Oculos_Sol.rds
bd_oculos_sol <-readRDS("Estudo_Oculos_Sol.rds")
```

No caso em estudo, a população em análise é

> X **-** Importância concedida, pelos inquiridos, ao Preço enquanto fator determinante na compra de óculos de sol (variável *Price*)

E o parâmetro de interesse é

> μ (*importância concedida ao preço, em termos médios*)

(**TC** $\rightarrow$ A afirmação é *"a importância concedida ao preço está, em termos médios, acima do ponto intermédio da escala"*, ou seja $\mu > 5$, pelo que esta será a **hipótese alternativa** (não tem $=$))

Dado que a Questão Problema é *Será que a importância concedida ao preço está, em termos médios, acima do ponto intermédio da escala (i.e. 5)*, podemos definir as hipóteses como:

- $H_0 : \mu \leq 5$ (hipótese nula)

  Ou seja, a importância média concedida ao preço é, no máximo, igual ao ponto intermédio da escala.

- $H_1 : \mu > 5$ (hipótese alternativa)

  Ou seja, a importância concedida ao preço está, em termos médios, acima do ponto intermédio da escala.

Logo este Teste de Hipóteses é **Unilateral Direito**

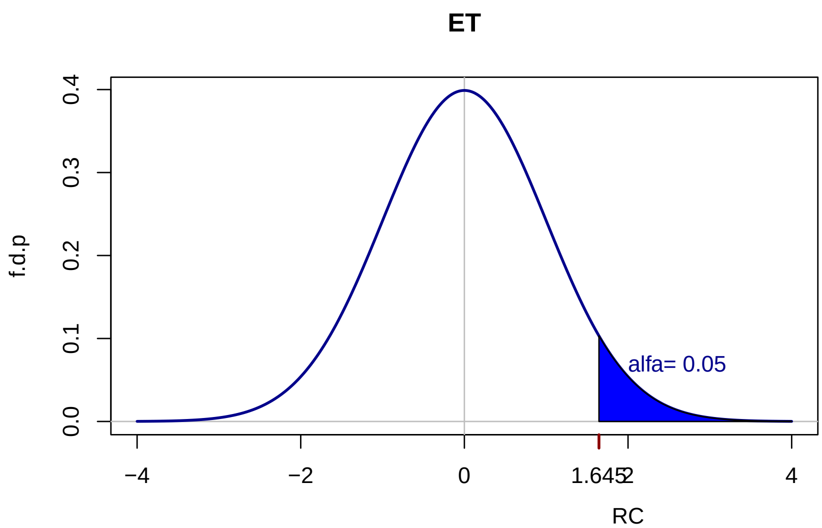
Pelo que, sendo $n = 640$ e não conhecendo σ^2, a VF a usar é

$$VF = \frac{\overline{X} - \mu_0}{\frac{S'}{\sqrt{n}}} \sim N(0, 1)$$

(**TC -** Podemos tomar decisões quer identificando a Região Crítica e a Região Não Crítica, com base no α dado (0.05), quer calculando (ou lendo, se usarmos o *t.test*) o *p-value* após termos obtido o valor concreto da ET para este caso, comparando-o com o α.)

Considerando $\alpha = 0.05$ (*significância de 5%*) as regiões serão

```r
eixo_x<-c(-4,4)
eixo_y<-c(0,dnorm(0))

# Preparar o Espaço
plot(1,
     xlim = eixo_x, ylim = eixo_y,
     type = "n",
     main = "ET",
     ylab = "f.d.p", xlab = "")

# Add x and y-axis lines
abline(h = 0, col="grey")
abline(v = 0, col="grey")

# Desenhar a Função
curve(dnorm(x),
      from = eixo_x[1], to = eixo_x[2],
      n = 1000,
      col = "darkblue",
      lwd = 2,
      add=TRUE)

# Sombrear a área da RC
zcrit<-qnorm(0.05,lower.tail = FALSE)

x1 <- seq(zcrit,eixo_x[2],0.01) # sequência de pontos, separados por 0.01
                                 # a começar em zcrit e até ao extremo direito
y1 <- dnorm(x1)                  # imagem desses x

coord_x <- c(zcrit,x1,eixo_x[2])
coord_y <- c(0,y1,0)
polygon(coord_x,coord_y,col='blue')
text(x=2,
     y=.06,
     labels=paste("alfa=",0.05),
     adj=c(0,0),
     col="darkblue")

axis(1,at=zcrit,labels = round(zcrit,3),col.ticks = "darkred",lwd.ticks = 2)
mtext("RC",side=1,line=2.5,at=2)
```



Temos então

$$RC = [1.64, +\infty[\qquad e \qquad RNC =]-\infty, 1.64[$$

Cálculo da Estatística de Teste (**ET**)

```
#Definição da RC e da RNC
significancia = 0.05

RC <- qnorm(1-significancia)

# Cálculo do valor do teste
H0 <- 5

# Calcular a média, variância amostral e dimensão da amostra
# (tudo o que precisamos para ter o teste t)
media <- mean(bd_oculos_sol$Price)       # Média da Variável
n <- length(bd_oculos_sol$Price)         # Dimensão da Amostra
dp <- sd(bd_oculos_sol$Price)            # Desvio-Padrão da Variável
```

Média Amostral	DP Amostral	n
6.62	1.08	640

```
# Alternativa 1: Calcular ET e usar RC e RNC
erro <- dp/sqrt(n)                # Erro Padrão do Estimador
ET <- (media - H0)/erro           # Estatística de Teste
ET
```

```
## [1] 38.12991
```

```r
# Alternativa : Calcular p-value
p_value <- pnorm(ET,lower.tail = FALSE)  # porque Teste à Direita
p_value
```

```
## [1] 0
```

```r
# = 0 -> O que não é de estranhar visto a ET ter um valor extremamente grande.
```

```r
# Alternativa 3: Fazer o t.test
teste1 <- t.test(bd_oculos_sol$Price,alternative = "greater", mu = 5,
                 conf.level = 1-significancia)
teste1
```

```
##
##  One Sample t-test
##
## data:  bd_oculos_sol$Price
## t = 38.13, df = 639, p-value < 2.2e-16
## alternative hypothesis: true mean is greater than 5
## 95 percent confidence interval:
##  6.553947      Inf
## sample estimates:
## mean of x
##   6.62411
```

Como o teste é **Unilateral Direito**, e $ET \sim N(0,1)$, o ponto fronteira entre RC e RNC é o quantil de probabilidade $1-\alpha = 1-0.05 = 0.95$ de uma normal standard, $z_{crit} = 1.64$

Logo, temos que

$$RC = [1.64, +\infty[\qquad e \qquad RNC =]-\infty, 1.64[$$

$\therefore z = 38.13 \in RC$, então rejeitamos H_0, para o nível de significância $\alpha = 0.05$

OU

$\therefore$ Como $pvalue \simeq 0 \leq \alpha = 0.05$, logo rejeitamos H_0

> No contexto do problema, rejeitar H_0 significa que, com uma significância de 5%, a verdadeira importância concedida, pelos inquiridos, ao preço na compra de óculos de sol está, em termos médios, acima do ponto intermédio da escala (> 5)

2.

No 2º caso em estudo, a população em análise mantém-se

X **-** Importância concedida, pelos inquiridos, ao Preço na compra de óculos de sol (variável *Price*)

Porém, agora dividimos consoante o sexo

X_H **-** Importância concedida, pelos Homens, ao Preço na compra de óculos de sol

X_M **-** Importância concedida, pelas Mulheres, ao Preço na compra de óculos de sol

Em que

$$X_H \sim N(\mu_H, \sigma_H) \quad e \quad X_M \sim N(\mu_M, \sigma_M)$$

E o parâmetro de interesse é

- $\mu_H \ - \ \mu_M$ ($H-$Homem e $M-$Mulher)

  Em que consideramos $\sigma_H^2 \ = \ \sigma_M^2$

Dado que a Questão Problema é *Será que homens e mulheres diferem, em termos médios, na importância concedida ao preço?*, podemos definir as hipóteses como:

- $H_0 : \mu_H \ = \mu_M \quad \Leftrightarrow \quad \mu_H - \mu_M = 0$ (hipótese nula)

  Ou seja, a importância média concedida ao preço é igual para homens e mulheres.

- $H_1 : \mu_H \neq \mu_M \quad \Leftrightarrow \quad \mu_H - \mu_M \neq 0$ (hipótese alternativa)

  Ou seja, a importância média concedida ao preço não é a mesma para homens e mulheres.

Logo este Teste de Hipóteses é **Bilateral**

Pelo que, sendo $n_1 + n_2 = 640 > 30$ e não conhecendo σ^2, mas assume-se $\sigma_H^2 \ = \ \sigma_M^2$, a VF a usar é

$$VF = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)_0}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\sqrt{\frac{(n_1-1)s_1'^2 + (n_2-1)s_2'^2}{n_1+n_2-2}}} \sim t_{(n1+n2-2)}$$

Iremos considerar que a distribuição da ET é $t_{(n1+n2-2)}$ e $\alpha = 0.05$ (*significância de 5%*)

```
# Cálculos Amostrais
med1 <- mean(bd_oculos_sol$Price[bd_oculos_sol$sex == "Male"])
n1 <- length(bd_oculos_sol$Price[bd_oculos_sol$sex == "Male"])
dp1 <- sd(bd_oculos_sol$Price[bd_oculos_sol$sex == "Male"])

med2 <- mean(bd_oculos_sol$Price[bd_oculos_sol$sex == "Female"])
n2 <- length(bd_oculos_sol$Price[bd_oculos_sol$sex == "Female"])
dp2 <- sd(bd_oculos_sol$Price[bd_oculos_sol$sex == "Female"])
```

Média Amostral H	DP Amostral H	n H
6.71	1.03	308

Média Amostral M	DP Amostral M	n M
6.54	1.11	332

```r
# -------------------------- Através do t.test --------------------------
teste2 <- t.test(bd_oculos_sol$Price ~ bd_oculos_sol$sex,
                 alternative = "two.sided",
                 mu = 0,
                 conf.level = 1- significancia,
                 var.equal = TRUE)
teste2
```

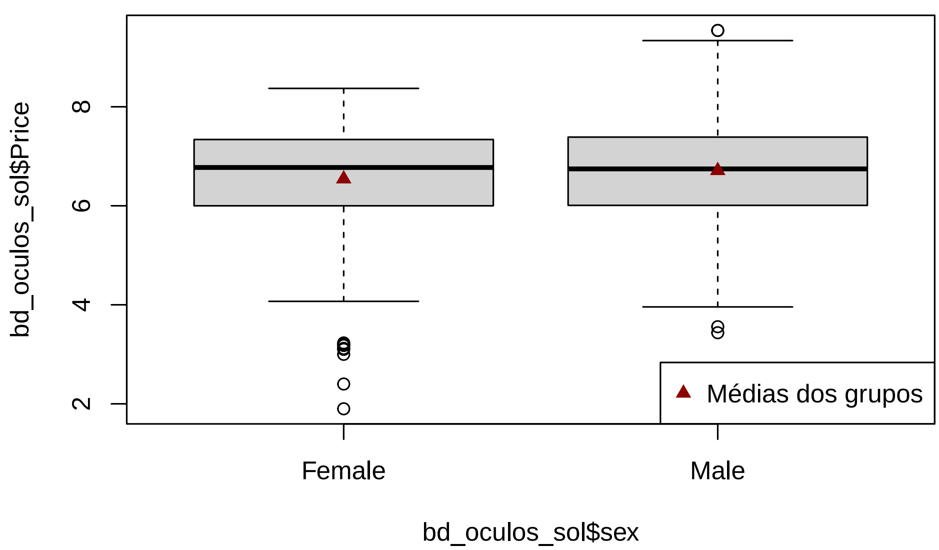
```
##
##  Two Sample t-test
##
## data:  bd_oculos_sol$Price by bd_oculos_sol$sex
## t = -1.9826, df = 638, p-value = 0.04784
## alternative hypothesis: true difference in means between group Female and group Male is not equal to
## 95 percent confidence interval:
##  -0.335643601 -0.001606883
## sample estimates:
## mean in group Female   mean in group Male
##             6.542959             6.711584
```

∴ Como $pvalue = 0.048 \ \leq \ \alpha = 0.05$, logo rejeita-se H_0 (igualdade de médias) e aceita-se a alternativa

No contexto do problema, rejeitar H_0 significa que, com uma significância de 5%, existe evidência estatística que permite concluir que Homens e Mulheres dão, em termos médios, importância **diferente** ao Preço enquanto fator determinante na compra de óculos de Sol.

EXTRA

```
medias<-teste2$estimate

plot(bd_oculos_sol$Price~bd_oculos_sol$sex)

points (medias,
        pch= 17,
        col="darkred",
        add=TRUE)

legend("bottomright",
       legend = "Médias dos grupos",
       col = "darkred",
       pch = 17)
```



É importante neste caso fazer a análise descritiva dos dados para perceber qual a tendência das diferenças.

Como se pode ver no *output* obtido, na amostra observada, as mulheres dão menor importância média ao fator *Preço* do que os homens ($\overline{x}_M = 6.54$, $\overline{x}_H = 6.71$), contudo em ambos os casos acima do ponto intermédio da escala.

Uma análise rápida do diagrama de extremos e quartis permite verificar que a divergência nas médias se deve provavelmente aos valores extremos:

- A metade central das observações é basicamente idêntica nos dois grupos;
- Os 25% superiores espalham-se num intervalo mais amplo no grupo dos homens;
- os 25% inferiores revelam mais *outliers* no grupo das mulheres.

Validação do pressuposto $\sigma_1 = \sigma_2$ através de um Teste com Hipóteses

- $H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1$
- $H_1 : \frac{\sigma_1^2}{\sigma_2^2} \neq 1$

```
var.test(bd_oculos_sol$Price  ~bd_oculos_sol$sex)
```

```
##
##  F test to compare two variances
##
## data:  bd_oculos_sol$Price by bd_oculos_sol$sex
## F = 1.1678, num df = 331, denom df = 307, p-value = 0.1677
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.9367021 1.4544917
## sample estimates:
## ratio of variances
##           1.167768
```

Com um $p-value$ de 0.168 e com $\alpha = 0.05$ não se rejeita a H_0 de igualdade de variâncias, o que valida a opção tomada no teste à comparação das médias.