Estudo de caso: Grupo D 3

Gilmar Pereira, Maressa Tavares e Victor Ruela

30 de Setembro, 2019

1 Resumo

0) separar entre homens e mulheres em cada ano. Realizar um teste de hipoteses para cada sexo, comparando os dois anos
1) H0: nao ha diferenca entre as medias H1: as medias sao diferentes
2) Validar as premissas Normalidade Igualdade de variancia (se diferente, o teste T nao podera ser usado) Independencia
3) Tamanho de efeito: 5 (de acordo com tamanho das faixas, primeiro resultado do google)
   Numero de amostras conhecido: calcular a potencia e estimar nivel de significancia adequado (alpha = 0.05)
4) Como melhorar: verificar aumento da potencia, levando a aumento da quantidade de amostras
5) Discutir sobre diferencas na potencia entre h omens e mulheres (menos mulheres na engenharia, poucas amostras)

2 Summary

O presente trabalho realizou o delineamento e executou os testes estatísticos para avaliar as diferenças no IMC médio entre duas populações de estudantes de pós-graduação em Engenharia elétrica, nos semetres de 2016-2 e 2017-2. As sub-populações masculina e feminina foram analisadas separadamente. Após os testes verificou-se que …

3 Planejamento do experimento

3.1 Objetivo do experimento

O objetivo é estudar as diferenças entre o IMC médio entre duas turmas de estudantes de pós-graduação em Engenharia elétrica na UFMG, para dados coletados nos semestres de 2016-2 e 2017-2. A análise será dividada entre as duas sub-populações (homens e mulheres), uma vez que é experado diferenças no ICM e também no tamanho amostral. Portanto, para verificar estas diferenças, as seguintes hipóteses serão testadas:

$$\begin{cases} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 \neq \mu_2 \end{cases}$$

Onde μ_1 e μ_2 são as turmas de 2016-7 e 2017-2, respectivamente. A mesma hipótese será testada separadamente para cada sub-população.

Os testes serão realizados para um tamanho de efeito $\delta^* = 5$, visto que essa é a diferença de valor entre os níveis de classificação IMC encontrados na literatura [1].

3.2 Análise Exploratória dos Dados

Antes de iniciar a análise exploratória dos dados, é precisa realizar um breve pré-processamento dos dados, visto que os dados de cada ano possuem formatos ligeiramente diferentes. Por exemplo, para os dados de 2016-2, é necessário remover as linhas referentes aos alunos de graduação. Ambos os arquivos foram padronizados para possuir o mesmo nome de coluna, e uma nova coluna com o cálculo do IMC foi adicionada em cada. Novos datasets também foram criados contendo somente os dados de cada sexo e todos os dados, para auxiliar nas análises.

```
imc <- function(height, weight){
  return(weight/(height^2))
}

data.2016.2 <- read.csv('imc_20162.csv', header = T, sep = ",")
data.2017.2 <- read.csv('CS01_20172.csv', header = T, sep = ";")
data.2017.2 <- data.2017.2 %>% rename(Gender = Sex)

# calculate the IMC
data.2016.2 <- data.2016.2 %>% mutate(imc = imc(Height.m, Weight.kg))
data.2017.2 <- data.2017.2 %>% mutate(imc = imc(height.m, Weight.kg))

# remove undergraduate students for 2016.2 data
data.2016.2 <- data.2016.2 %>% filter(Course == 'PPGEE')

# remove unecessary columns and rename
data.2016.2 <- data.2016.2 %>% select(Gender, imc)
data.2017.2 <- data.2017.2 %>% select(Gender, imc)

# split the 2016.2 data by gender and remove undergraduate students
data.2016.2.Females <- data.2016.2 %>% filter(Gender == 'F') %>% select(imc)
data.2016.2.Males <- data.2016.2 %>% filter(Gender == 'M') %>% select(imc)

# split the 2017.2 data by gender
data.2017.2.Females <- data.2017.2 %>% filter(Gender == 'F') %>% select(imc)
data.2017.2.Males <- data.2017.2 %>% filter(Gender == 'M') %>% select(imc)

# create a single dataset
data.all <- data.2016.2 %>% mutate(Class = '2016-2')
data.all <- data.all %>% bind_rows(data.2017.2 %>% mutate(Class = '2017-2'))
```

Para ter uma ideia inicial dos dados, suas estatísticas básicas são calculadas:

```
summary(data.2016.2)
```

```
##  Gender      imc
##  F: 7   Min.   :17.58
##  M:21   1st Qu.:21.27
##         Median :23.52
##         Mean   :23.97
##         3rd Qu.:25.29
##         Max.   :37.55
```

```
summary(data.2017.2)
```

```
##  Gender      imc
##  F: 4   Min.   :17.36
##  M:21   1st Qu.:20.83
##         Median :23.26
##         Mean   :23.35
##         3rd Qu.:25.18
##         Max.   :30.42
```

É possível fazer as seguintes observações que a quantidade de amostras do sexo feminimo é bem menor do que a do masculino, o que pode afetar a potência dos testes a serem executados.

Para melhor visualização, um gráfico boxplot é gerado na figura abaixo. Nele é possível ver que há alguns outliers nos valores de IMC em alguns anos, que verão ser removidos posteriormente.

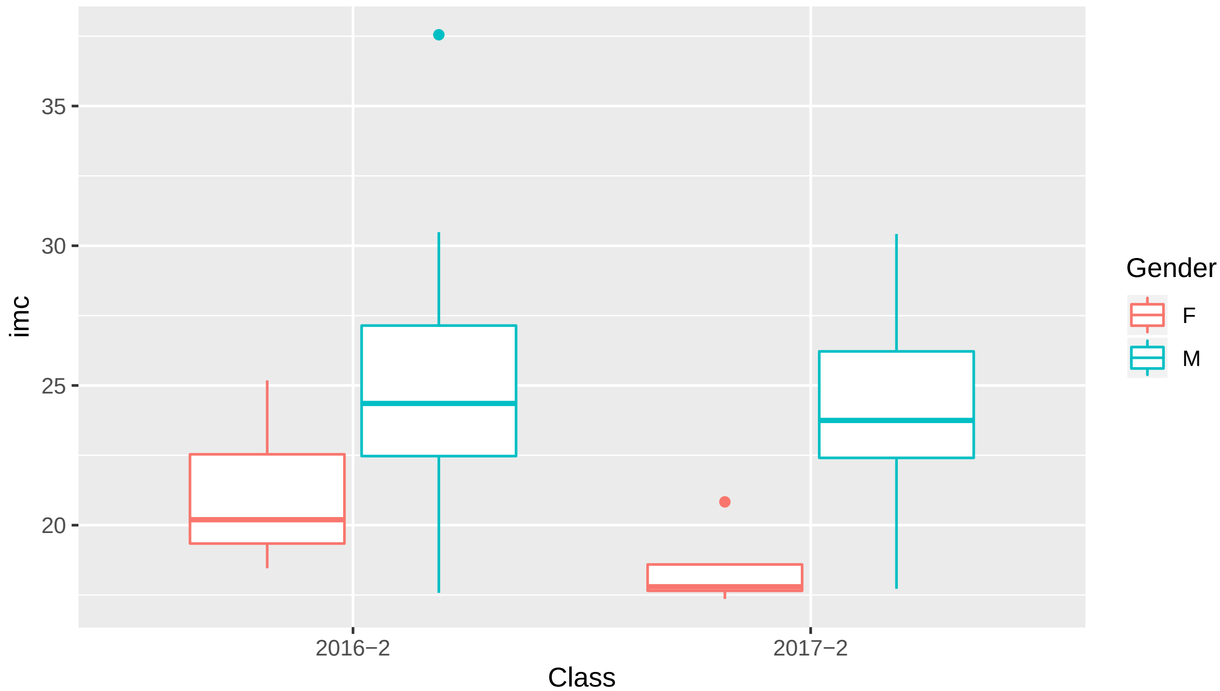

Figure 1: Boxplot dos dados

4 Resultados

4.1 Validação das premissas

Para realizar as inferências estatísticas sobre o IMC das duas populações é necessário validar as premissas antes de executar o teste. Neste caso, como as variâncias das duas populações é desconhecida, utiliza-se a distribuição t para o teste de hipóteses e para os intervalos de confiança [2]. A seguir são apresentados os testes realizados para validar as premissas exigidas pelo teste t. Para facilitar as análises optou-se por separar o grupo na população feminina e masculina.

4.1.1 Subpopulação Feminina

1 - Normalidade:

Para avaliar a normalidade dos dados das duas subpopulações, utilizou-se o gráfico quantil-quantil e o teste de Shapiro Wilk com $\alpha = 0.05$, como apresnetado a seguir.

```
## [1] 6 1
```

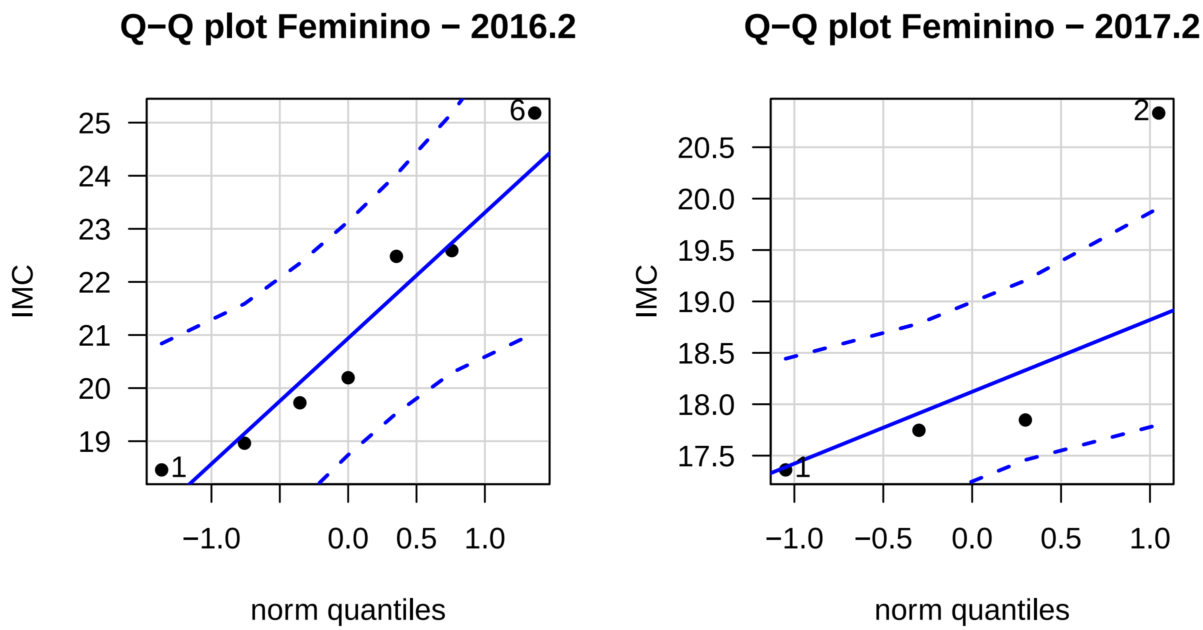

Figure 2: Gráfico quantil-quantil das populações - Feminino

```
## [1] 2 1
```

```r
shapiro.test(data.2016.2.Females$imc)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data.2016.2.Females$imc
## W = 0.91974, p-value = 0.4674
```

```r
shapiro.test(data.2017.2.Females$imc)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data.2017.2.Females$imc
## W = 0.7475, p-value = 0.03659
```

Pela análise do teste de Shapiro-Wilk e do gráfico concluímos que não há evidências para rejeitar a normalidade dos dados de 2016.2, pois o p-valor do teste de Shapiro Wilk foi maior que 0.05. Por outro lado, os dados de 2017.2 apresentam um ponto fora dos limites e o p-valor foi inferior a 0.05, por isso, há evidências para rejeitar a normalidade dos dados.

Diante disso, é necessário fazer uma transformação nos dados de 2017.2 para tentar normalizar os dados, a fim de possibilitar a realização dos testes estatísticos.

```r
qqnorm(log(data.2017.2.Females$imc), pch = 16, las = 1)
qqline(log(data.2017.2.Females$imc))
```

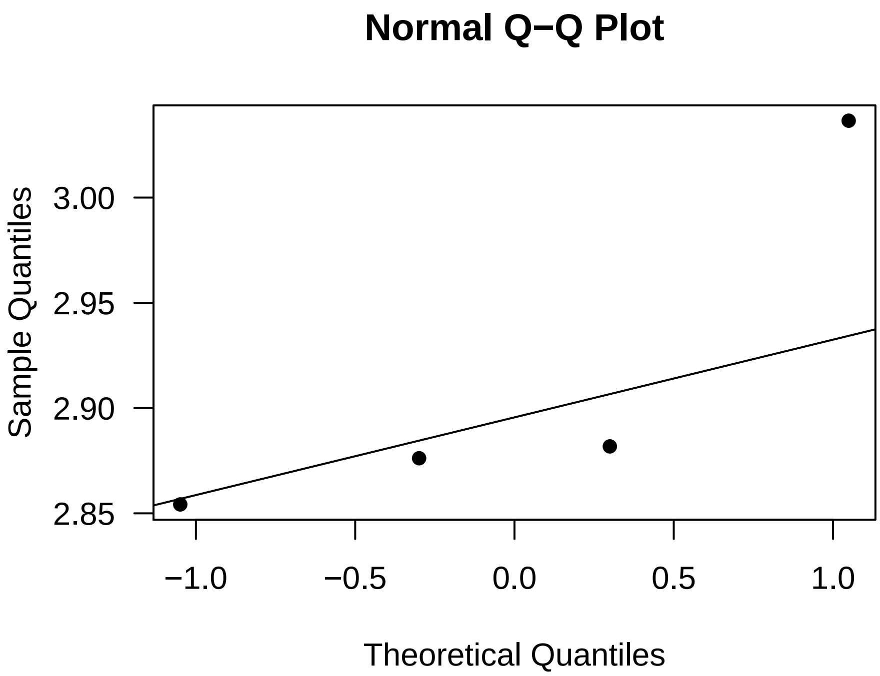

Figure 3: Gráfico quantil-quantil para tranformação logarítimica dos dados

```r
shapiro.test(log(data.2016.2.Females$imc))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  log(data.2016.2.Females$imc)
## W = 0.93097, p-value = 0.5592
```

Embora o gráfico dos quantis não deixe claro a normalidade o teste de Shapiro mostra que, após o ajuste logarítmico dos dados, não há evidências para rejeitar a normalidade.

2- Igualdade de Variâncias:

A segunda premissa a ser avaliada é a igualdade de variâncias das duas populações, homocedasticidade. Para tanto, utilizou-se o teste F com a função var.test e $\alpha = 0.05$, que tem como hipótese nula a igualdade da variância das duas amostras.

```r
var.test(data.2016.2.Females$imc,data.2017.2.Females$imc)
```

```
##
##  F test to compare two variances
##
## data:  data.2016.2.Females$imc and data.2017.2.Females$imc
## F = 2.2688, num df = 6, denom df = 3, p-value = 0.5353
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##   0.1539783 14.9715344
## sample estimates:
## ratio of variances
```

```
##           2.268827
```

Pelo teste F conclui-se que não há evidências para rejeitar a hipótese nula de igualdade das variâncias das duas populações, portanto, as amostras são consideradas homocedásticas.

3 - Independência

Considerando que as populações referem-se a dois grupos distintos da pós-graduação, sabe-se a priori, que as amostras são independentes. Para concluir com precisão em relação à independência realizou-se o teste Qui-quadrado com $\alpha = 0.05$.

```
data.chi.Female <- c(data.2016.2.Females$imc, data.2017.2.Females$imc)
chisq.test(data.chi.Female)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  data.chi.Female
## X-squared = 3.0047, df = 10, p-value = 0.9813
```

Como era esperado, o teste Qui-quadrado reafirma a independência entre as duas amostras, como o p-valor = 0.9813.

Diante dessas análises, percebemos que provavlemente o tamanho pequeno das amostras de IMC feminino pode ter influenciado nas conclusões em relação às premissas. Portanto, esse fator pode influenciar também nos testes de hipóteses que serão apresentados nas próximas seções.

4.1.2 Subpopulação Masculina

Assim como na validação das premissas da subpopulação feminina, foram realizados os testes para a subpopulação masculina, como descrito a seguir.

1 - Normalidade:

Para validar a normalidade dos dados utilizou-se o gráfico quantil-quantil e o teste de Shapiro Wilk com $\alpha = 0.05$.

```
## [1]  4 12
```

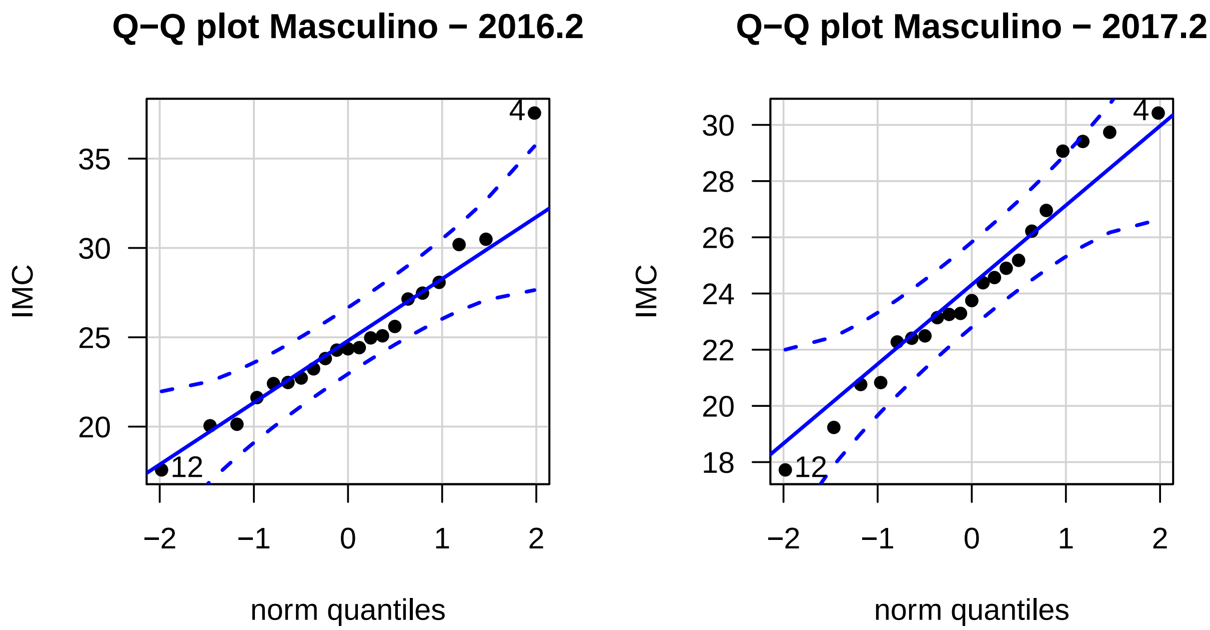

Figure 4: Gráfico quantil-quantil das populações - Feminino

```
## [1] 12  4
```

```r
shapiro.test(data.2016.2.Males$imc)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data.2016.2.Males$imc
## W = 0.92833, p-value = 0.1275
```

```r
shapiro.test(data.2017.2.Males$imc)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data.2017.2.Males$imc
## W = 0.96494, p-value = 0.6206
```

Pela análise do teste de Shapiro-Wilk e dos gráficos concluímos que não há evidências para rejeitar a hipótese nula para as duas amostras, pois em ambas o p-valor do teste foi superior a 0.05. Portanto, para a subpopulação masculina dos dois semestres os dados estão normalmente distribuídos.

2- Igualdade de Variâncias:

Para validar a igualdade de variâncias das duas amostras utilizou-se o teste F com $\alpha = 0.05$. A hipótese nula considera a igualdade da variância das duas amostras.

```r
var.test(data.2016.2.Males$imc,data.2017.2.Males$imc)
```

```
##
##  F test to compare two variances
```

```
##
## data:  data.2016.2.Males$imc and data.2017.2.Males$imc
## F = 1.5839, num df = 20, denom df = 20, p-value = 0.3119
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.6426853 3.9034665
## sample estimates:
## ratio of variances
##           1.583888
```

Pelo teste F conclui-se que as populações são homocedásticas, pois o p-valor foi superior a 0.05, logo, não há evidências para rejeitar a hipótese nula.

3 - Independência

Assim como na subpopulação feminina, sabe-se a priori, que as amostras são independentes. Para concluir com precisão em relação à independência realizou-se o teste Qui-quadrado com $\alpha = 0.05$.

```r
data.chi.male <- c(data.2016.2.Males$imc,data.2017.2.Males$imc)
chisq.test(data.chi.male)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  data.chi.male
## X-squared = 24.96, df = 41, p-value = 0.9772
```

Como era esperado, o teste Qui-quadrado reafirma a independência entre as duas amostras, como o p-valor = 0.9772.

Ao contrário do que aconteceu com as amostras da subpopulação feminina, nesse caso o tamanho das amostras foram maiores, e por isso, não houve nenhum problema em relação às premissas do teste.

5 Resultados

Nesta seção são apresentados os resultados do teste para as hipóteses estabelecidas anteriormente com $\alpha = 0.05$. Os resultados estão subdivididos entre a população feminina e masculina devido às diferenças de IMC que existem entre eles.

5.1 Teste de Hipóteses Subpopulação Feminina

```r
#teste para custo médio:
(mean.t.teste <- t.test(data.2016.2.Females,
                  data.2017.2.Females,
                   var.equal = TRUE))
```

```
##
##  Two Sample t-test
##
## data:  data.2016.2.Females and data.2017.2.Females
## t = 1.9308, df = 9, p-value = 0.08556
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.4527037  5.7283762
## sample estimates:
## mean of x mean of y
##  21.08443  18.44660
```

5.2 Teste de Hipóteses Subpopulação Masculina

```
#teste para custo médio:
(mean.t.teste <- t.test(data.2016.2.Males,
                  data.2017.2.Males,
                   var.equal = TRUE))
```

```
##
##  Two Sample t-test
##
## data:  data.2016.2.Males and data.2017.2.Males
## t = 0.53979, df = 40, p-value = 0.5923
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.784943  3.085836
## sample estimates:
## mean of x mean of y
##  24.93595  24.28551
```

```
pwr.t2n.test(n1 = 7, n2= 4, d = 0.3, sig.level = 0.05, power = NULL,
    alternative = "two.sided")
```

```
##
##      t test power calculation
##
##              n1 = 7
##              n2 = 4
##               d = 0.3
##       sig.level = 0.05
##           power = 0.07141217
##     alternative = two.sided
```

```
pwr.t2n.test(n1 = 21, n2= 21, d = 0.8, sig.level = 0.05, power = NULL,
    alternative = "two.sided")
```

```
##
##      t test power calculation
##
##              n1 = 21
##              n2 = 21
##               d = 0.8
##       sig.level = 0.05
##           power = 0.7155555
##     alternative = two.sided
```

6 Discussão e Conclusões

Neste trabalho foi feito um estudo estatístico …

7 Divisão das Atividades

- Victor - Verificador e Monitor
- Maressa - Relatora
- Gilmar - Coordenador

Referências

[1] *Tabela de imc.* http://www.calculoimc.com.br/tabela-de-imc/.

[2] D. C. Montgomery and G. C. Runger, *Applied statistics and probability for engineers, (with cd).* John Wiley & Sons, 2007.