Estudo de Caso 2: Planejamento e Análise de Experimentos

Matheus Marzochi, Mayra Mota, Rafael Ramos e Victor Magalhães

30 de Setembro de 2019

Resumo

O experimento disponibilizado para este trabalho consiste em avaliar se o estilo de vida entre duas populações de estudantes se alterou ao longo do tempo. Nessa avaliação duas amostras foram utilizadas, uma contendo informações (peso e altura dos alunos) referentes ao ano de 2016 e a outra com informações referentes a 2017.

Neste trabalho o IMC (Índice de Massa Corporal) é usado como parâmetro de avaliação do estilo de vida. Vale ressaltar que esse indicador possui limitações como pode ser verificado em ("How Often Is B.M.I. Misleading?" 2015)("The Health Risk of Obesity—Better Metrics Imperative" 2013). Além disso, as duas populações foram subdivididas por gênero. A motivação desta divisão é a consideração de que, em média, o IMC masculino pode ser diferente do feminino.

Papéis Desempenhados

A divisão de tarefas no grupo segue a descrição da *Declaração de Políticas de Equipe*. Estando aqui organizada da seguinte forma:

- Matheus: Verificador
- Mayra: Monitora
- Rafael: Coordenador
- Victor: Revisor

Planejamento do Experimento

A hipótese nula (H_0) usada neste experimento é de que a diferença das médias de IMC das duas populações é nula. A hipótese alternativa (H_1), por sua vez, afirma que existe sim uma alteração entre as médias dos dois semestres.

$$\begin{cases} H_0 : \mu_{2016/2} - \mu_{2017/2} = 0 \\ H_1 : \mu_{2016/2} - \mu_{2017/2} \neq 0 \end{cases}$$

Como a comparação é relativa a duas amostras, em função da influência do gênero no valor do IMC, as análises são realizadas de forma independente para cada sexo. Como consequência dessa independência, neste trabalho são realizados testes de comparação simples entre as populações.

Coleta dos Dados

O procedimento de coleta de dados foi baseado na rotina presente abaixo.

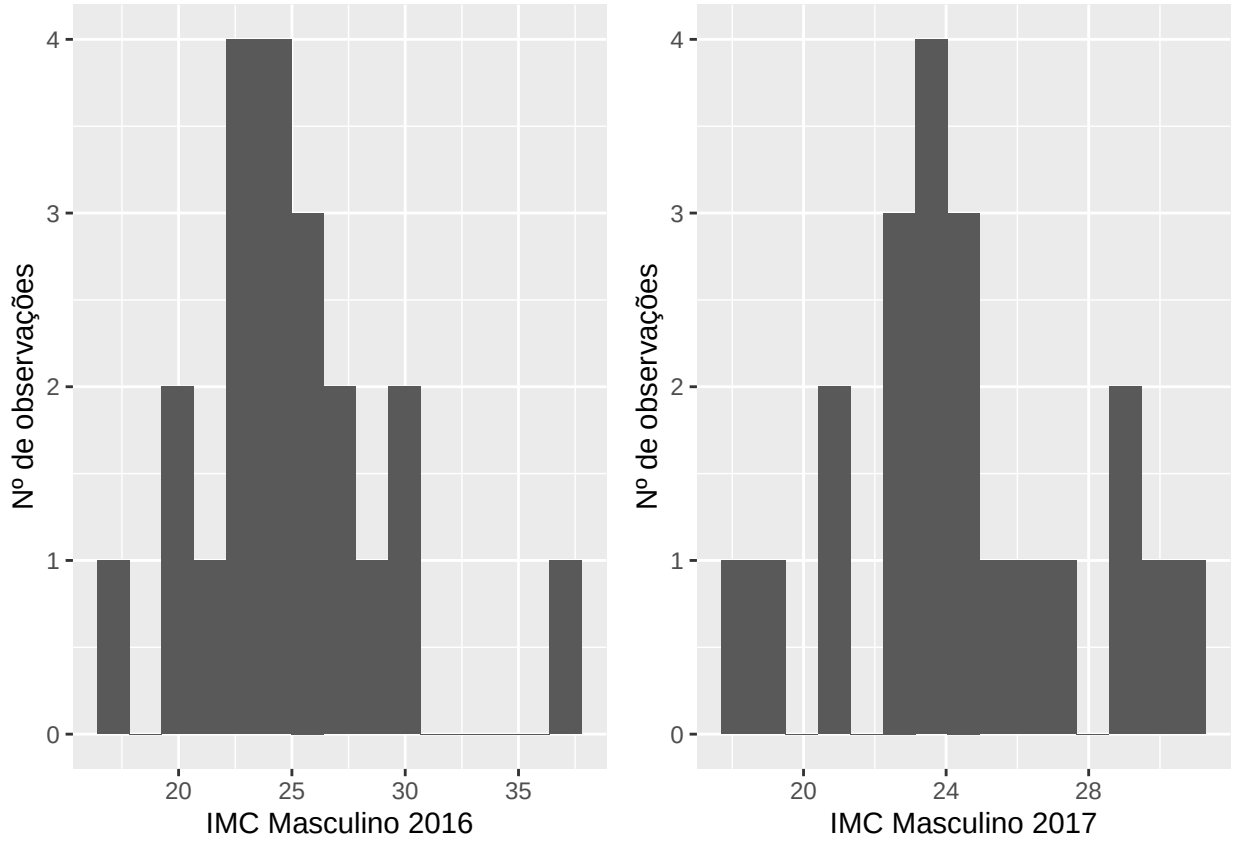
```
data_2017 <- read.csv(file='CS01_20172.csv', sep=';')#Dados ano 2016
data_2016<-read.csv("imc_20162.csv");#Dados ano 2017

PPGEE_dados=data_2016[data_2016[2]=='PPGEE',];

#Dados população feminina e masculina do ano de 2016.
Dados_Masculino_2016=PPGEE_dados[PPGEE_dados[3]=='M',];
Dados_Feminino_2016=PPGEE_dados[PPGEE_dados[3]=='F',];

Height_Masc_2016=Dados_Masculino_2016[,4];
Heigh_Feminino_2016=Dados_Feminino_2016[,4];
Weight_Masculino_2016=Dados_Masculino_2016[,5];
Weight_Feminino_2016=Dados_Feminino_2016[,5];

#Calculo do IMC masculino 2016
IMC_masculino_2016=(Weight_Masculino_2016/((Height_Masc_2016)*(Height_Masc_2016)));
#Calculo do IMC feminino 2016
IMC_Feminino_2016=(Weight_Feminino_2016/((Heigh_Feminino_2016)*(Heigh_Feminino_2016)));


#Dados população feminina e masculina do ano de 2017.
Dados_Masculino_2017=data_2017[data_2017[3]=='M',];
Dados_Feminino_2017=data_2017[data_2017[3]=='F',];

Height_Masculino_2017=Dados_Masculino_2017[,2];
Height_Feminino_2017=Dados_Feminino_2017[,2];
Weight_Masc_2017=Dados_Masculino_2017[,1];
Weight_Feminino_2017=Dados_Feminino_2017[,1];

#Calculo do IMC masculino 2017
IMC_masculino_2017=(Weight_Masc_2017/((Height_Masculino_2017)*(Height_Masculino_2017)));
#Calculo do IMC feminino 2017
IMC_Feminino_2017=(Weight_Feminino_2017/((Height_Feminino_2017)*(Height_Feminino_2017)));
```

Análise Exploratória dos Dados

Antes da análise estatística, os dados são avaliados de forma qualitativa na análise exploratória com o objetivo de se extrair informações úteis. Entre as ferramentas usadas para se explorar os dados existentes está o Histograma.

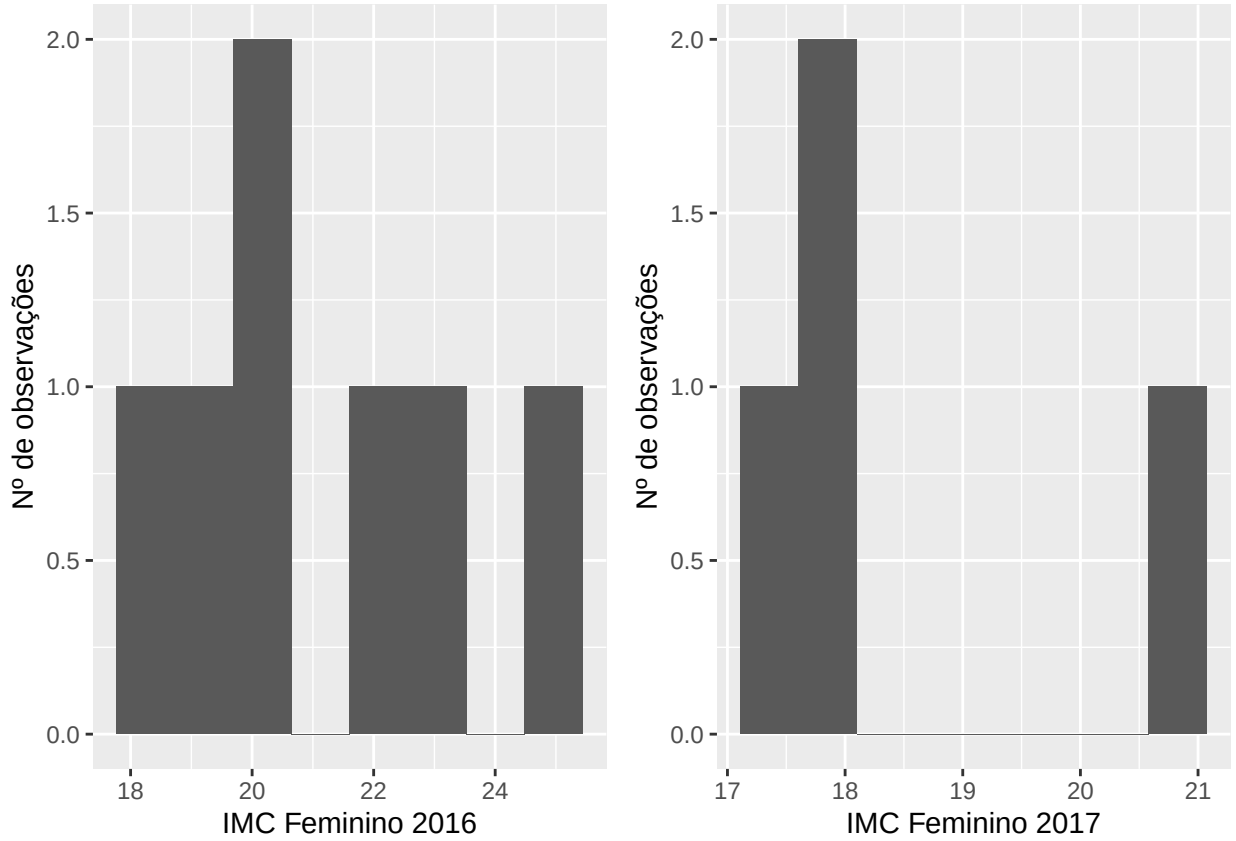
Segue abaixo o trecho de código usado para gerar o histograma da população masculina dos anos de 2016 e 2017.

```
p1 <- ggplot(as.data.frame(IMC_masculino_2016), aes(x = IMC_masculino_2016))
p1 <- p1 +geom_histogram(bins = 15)+xlab("IMC Masculino 2016")+ylab("Nº de observações")
p2 <- ggplot(as.data.frame(IMC_masculino_2017), aes(x = IMC_masculino_2017))
p2 <- p2 + geom_histogram(bins = 15)+xlab("IMC Masculino 2017")+ylab("Nº de observações")
ggarrange(p1, p2, nrow = 1, ncol = 2)
```



Segue abaixo o trecho de código usado para gerar o histograma da população feminina dos anos de 2016 e 2017.

```
p1 <- ggplot(as.data.frame(IMC_Feminino_2016), aes(x = IMC_Feminino_2016))
p1 <- p1 +geom_histogram(bins = 8)+xlab("IMC Feminino 2016")+ylab("Nº de observações")
p2 <- ggplot(as.data.frame(IMC_Feminino_2017), aes(x = IMC_Feminino_2017))
p2 <- p2 + geom_histogram(bins = 8)+xlab("IMC Feminino 2017")+ylab("Nº de observações")
ggarrange(p1, p2, nrow = 1, ncol = 2)
```



Através da análise exploratória dos dados é possível observar que o número de amostras fornecidos é pequeno, especialmente no caso feminino. Para a população masculina apesar do número de amostras não ser elevado, a distribuição das amostras leva à indícios que a distribuição pode ser normal. No caso feminino há indícios de que seja necessário o uso de estatística não paramétrica, uma vez que pode haver não normalidade. Essas hipótes são analisadas mais profundamente na análise estatística.

A seguir temos outra análise dos dados. Leva-se em consideração também gênero e ano amostral. Esta análise é feita a partir de uma série de gráficos do tipo *boxplot*.

```
boxplot(IMC_Feminino_2016,IMC_masculino_2016,
names = c("Feminino","Masculino"))
```

A primeira informação que temos, observando o primeiro gráfico, é a grande diferença entre a média do IMC entre os alunos do sexo feminino e masculino, reforçando a ideia de trabalhar com sub-populações para diminuir resíduos e aumentar a potência dos testes.

Outro ponto é em relação à distribuição das observações. Note-se uma certa simetria para a sub-população masculina e uma pequena assimetria para a feminina. Isso parece não comprometer a premissa de normalidade da população.

Além disso, observa-se um *outlier* na sub-população masculina, que pode afetar a verificação da Premissa da Normalidade e de inferências sobre o parâmetro avaliado. Já para o caso feminino, isso não ocorre.

```
boxplot(IMC_Feminino_2017,IMC_masculino_2017,
names = c("Feminino 2017-2","Masculino 2017-2"))
```

No caso dos dados do semestre 2017-2, *outliers* não são observados. Porém, um ponto importante a ser ressaltado é a elevada assimetria na sub-população feminina, o que pode ser causada pelo baixo número de

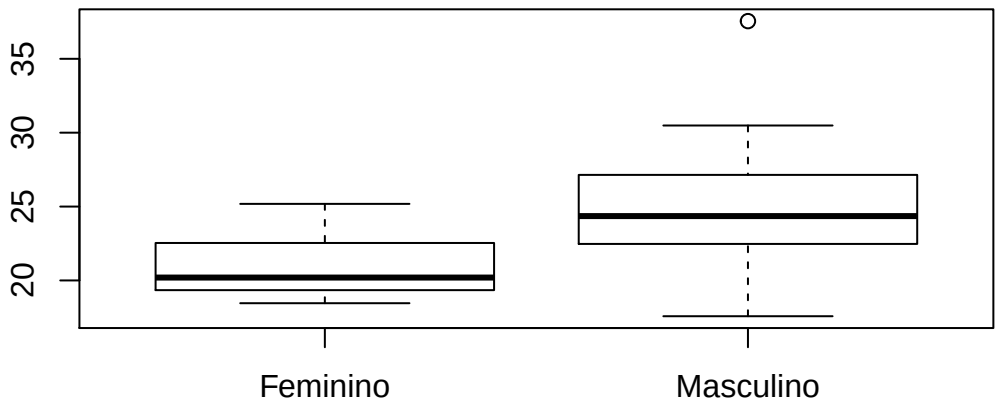


Figure 1: BoxPlot da Amostra das Sub-Populações de 2016

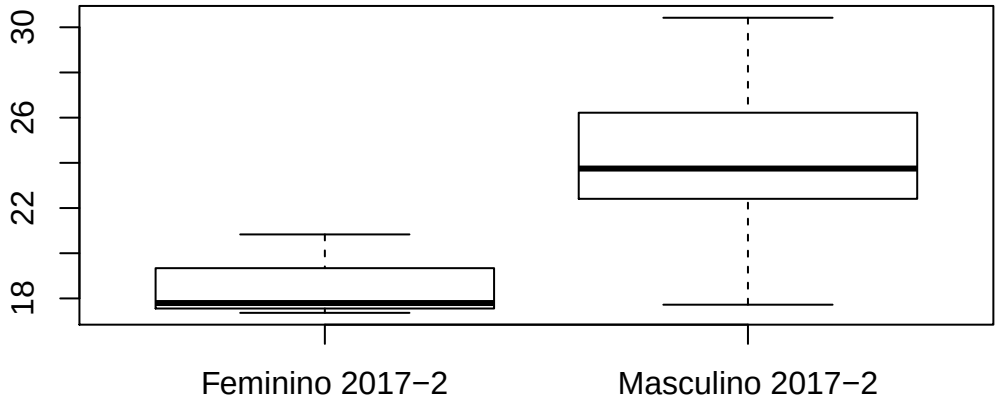


Figure 2: BoxPlot das Amostra das Sub-Populações 2017

amostras. Na verdade, o *boxplot* não é um gráfico adequado para se fazer inferências com esse número de amostras.

Análise dos Dados Experimentais Femininos

Teste de Hipótese

Conforme dito anteriormente, ao dividir a população em sub-populações, masculino e feminino, fica clara a necessidade de considerar uma hipótese nula para cada gênero e comparar os conjuntos de dados. Não há motivos para acreditar que exista uma diferença específica entre as duas populações, e o que deseja-se verificar é a existência de alguma diferença considerável entre as médias de IMC entre os anos. As hipóteses nulas foram definidas como uma diferença entre as médias das populações igual a 0.

Considerando os argumentos acima, a hipótese, para o caso feminino, está descrita a seguir:

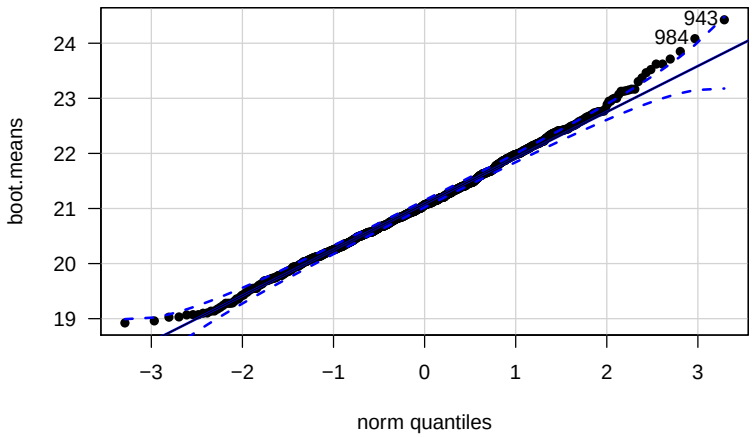
$$\begin{cases} H_{f0} : \mu_{f,2016} - \mu_{f,2017} = 0 \\ H_{f1} : \mu_{f,2016} - \mu_{f,2017} \neq 0 \end{cases}$$

Com a finalidade de averiguar a normalidade de cada uma delas, tem-se:

```
#Feminino - 2016
K <- 999
boot.means <- numeric(K)
for (i in seq(K)){
  boot.sample_2016 <- sample(IMC_Feminino_2016, replace = TRUE) # sample with replacement
  boot.means[i] <- mean(boot.sample_2016)
}

qqPlot(boot.means, las = 1, pch = 16)
```

```
## [1] 943 984
```

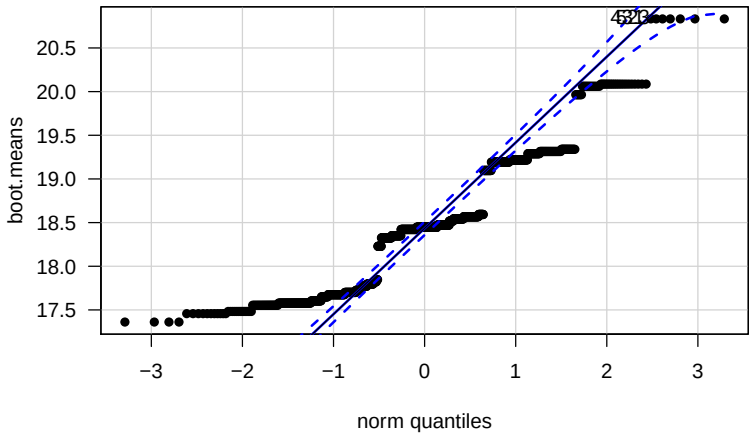
```
qqline(boot.means)
```



```r
# Feminin0 - 2017
for (i in seq(K)){
  boot.sample_2017 <- sample(IMC_Feminino_2017, replace = TRUE) # sample with replacement
  boot.means[i] <- mean(boot.sample_2017)
}

qqPlot(boot.means, las = 1, pch = 16)
```

```
## [1] 431 523
```

```r
qqline(boot.means)
```



Nota-se que a segunda amostra do dados femininos possui uma forte tendência a não normalidade. Aplica-se o teste Shapiro-Wilk para a certificação:

```r
(shapiro.test(IMC_Feminino_2016))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  IMC_Feminino_2016
## W = 0.91974, p-value = 0.4674
```

```r
(shapiro.test(IMC_Feminino_2017))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  IMC_Feminino_2017
## W = 0.7475, p-value = 0.03659
```

A amostra de 2016, claramente, segue uma distribuição normal. Assim, pode-se utilizar um teste T para análise das hipóteses. Já as amostras de 2017 não seguem uma normal.

Segundo (Montgomery and Runger 2012), a análise do intervalo de confiança foi relizado pela técnica de *bootstrap*. O método cria um intervalo de confiança que é verificado e, a partir daí, conclui sobre a rejeição ou não da hipótese nula. Se o intervalo de confiança contiver o valor da variável de interesse sob H_0 ($\mu_1 - \mu_2 = 0$), não tem rejeição da hipótese nula.

```
alpha <- 0.05
IntConf <- two.boot(IMC_Feminino_2016, IMC_Feminino_2017, mean, R = 999)
boot.ci(IntConf, conf = 1-alpha, type = "bca")
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 999 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = IntConf, conf = 1 - alpha, type = "bca")
##
## Intervals :
## Level       BCa
## 95%   ( 0.429,  4.812 )
## Calculations and Intervals on Original Scale
```

Acima, pode-se observar o resultado do cálculo do intervalo de confiança utilizando a técnica de *bootstrap*. Uma vez que o zero não está contido no intervalo de confiança da média, a hipótese nula é rejeitada. Nas amostras femininas, os IMCs médios são diferentes, $\mu_1 \neq \mu_2$.

Após analisar as médias relacionadas da população feminina, realiza-se os testes de hipótese com a população masculinas.Assume-se a seguinte hipótese nula:

$$\begin{cases} H_{h0} : \mu_{h,2016} - \mu_{h,2017} = 0 \\ H_{h1} : \mu_{h,2016} - \mu_{h,2017} \neq 0 \end{cases}$$

Primeiramente testamos rejeição ou não da hipótese nula quanto aos dados tenderem a normalidade. Assim utilizamos o teste de Shapiro Wilk:

```
(shapiro.test(IMC_masculino_2016))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  IMC_masculino_2016
## W = 0.92833, p-value = 0.1275
```

```
(shapiro.test(IMC_masculino_2017))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  IMC_masculino_2017
## W = 0.96494, p-value = 0.6206
```

Dado o p-valor do teste, pode-se assumir que não podemos rejeitar a hipótese nula de não normalidade. Dessa forma, assumindo que os dados de ambas as populações masculinas tendem a normalidade e são independentes, realizamos o teste T para comparar a média das populações:

```
t.test(IMC_masculino_2016,IMC_masculino_2017, alternative='two.sided',
       mu=0, paired=FALSE,conf.level = 0.95);
```

```
##
##  Welch Two Sample t-test
##
## data:  IMC_masculino_2016 and IMC_masculino_2017
## t = 0.53979, df = 38.057, p-value = 0.5925
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.788823  3.089716
## sample estimates:
## mean of x mean of y
##  24.93595  24.28551
```

Tendo em vista a proximidade das médias(IMC_masculino_2016=24.93 e IMC_masculino_2017=24.28), o tamanho do efeito(df=38.05) e o p-valor=0.59(maior que a significncia de 0.05), não se pode rejeitar a hipótese nula de que as populações tem diferença de médias igual à 0. No caso, vale destacar que o valor estimado de intervalo de confiança foi de -1.78 até 3.08.

Diante desses resultados, faz-se nescessário o cálculo da potência, estimando-se o erro de tipo II. Utilizamos o efeito d de cohen para o cálculo do efeito, já que estamos comparando duas populações.Assim, temos que:

```
d_C_masculino=cohen.d(IMC_masculino_2016,IMC_masculino_2017)
```

Caculando a potência por meio do teste T, temos que:

```
power.t.test(n=length(IMC_masculino_2016+IMC_masculino_2017),
             delta = d_C_masculino$estimate ,
    sd=sqrt(0.18),sig.level =0.05, type="two.sample",alternative ="two.sided")
```

```
##
##      Two-sample t test power calculation
##
##               n = 21
##           delta = 0.1665831
##              sd = 0.4242641
##       sig.level = 0.05
##           power = 0.2363497
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

Nesse caso a potência encontrada é de 0.23.

Discussão para Melhoria do Experimento

De acordo com (Ministério da Saúde 2017), o cálculo de IMC é feito para adultos em geral sem diferenciação de sexo. A primeira sugestão seria a realização da coleta dos dados em outros períodos, ou seja, a coleta de

amostras muito maiores, aumentando, assim, a potência dos testes. Os valores de altura e peso em média são distintos para homens e mulheres, porém com a vantagem de amostras maiores, a análise geral poderia ser feita.

Além disso, para que os dados coletados apresentem melhor veracidade, aconselha-se a utilização de instrumentos adequados para medição de altura e peso das sub-populações. Assim, pode-se minimizar os erros de coleta.

Uma outra sugestão seria a realização de análises levando em consideração a idade, já que, de acordo com essa característica, há uma grande tendência ao aumento de peso. Além disso, informações como prática de exercícios e alimentação saudável seriam também relevantes para um estudo mais detalhado de sub-populações.

Conclusões

Conclusões sobre a Análise Feminina

A ideia do estudo era comparar duas amostras diferentes. Situações com observações não normais e com poucas amostras foram notadas. De acordo com gráficos e testes, concluiu-se que uma das amostras femininas não pode ser considerada normal e houve a necessidade de uma análise alternativa. O maior problema, aparentemente, das amostras femininas foi o baixo número de observações, dificultando a conclusão das características da distribuição populacional. Com mais amostras, a análise de 2017 poderia apresentar uma tendência a assemelhar-se com uma normal, uma vez que as amostras Masculinas e Feminina-2016 apresentaram características de normalidade. O teste de potencia também é interferido com o baixo tamanho amostral.

Como a hipótese nula foi rejeitada, podemos admitir a hipótese alternativa de que houve, sim, uma mudança da qualidade de vida dos estudantes entre os dois períodos analisados.

Conclusões sobre a Análise Masculina

Quanto as amostras masculinas podemos concluir que não houve variação significativa no IMC da população de 2016 para 2017, dado que as amostras possuiam indicativo de normalidade e o teste T não rejeitou a hipótese nula.

Referências

"How Often Is B.M.I. Misleading?" 2015. https://www.nytimes.com/interactive/projects/cp/summer-of-science-2015/latest/how-often-is-bmi-misleading.

Ministério da Saúde. 2017. "IMC Em Adultos." http://portalms.saude.gov.br/component/content/article/804-imc/40509-imc-em-adultos.

Montgomery, Douglas, and George Runger. 2012. *Estatística Aplicada E Probabilidade Para Engenheiros*. LTC.

"The Health Risk of Obesity—Better Metrics Imperative." 2013. https://science.sciencemag.org/content/341/6148/856.summary.