Planejamento e Análise de Experimentos (EEE933)
Estudo de Caso 2

Pedro Vinícius, Samara Silva e Savio Vieira

24 de Agosto de 2020

Introdução

O Índice de Massa Corporal (IMC) é uma medida de gordura corporal baseada na relação entre peso (em kg) e altura (em m) de um indivíduo e é comumente utilizado como uma ferramenta de triagem para indicar se uma pessoa está com um peso saudável para sua altura. Este índice é calculado conforme a Equação 1:

$$IMC = \frac{peso}{(altura)^2} \tag{1}$$

em kg/m^2. Cada faixa de IMC permite classificar o indivíduo em uma das seguintes categorias [2]:

- Baixo peso: $< 18,5$
- Peso normal: $18,5 - 25$
- Sobrepeso: $25 - 30$
- Obesidade: $30 - 35$
- Obesidade mórbida: > 40

Neste estudo de caso deseja-se comparar o IMC médio dos alunos de Pós-Graduação em Engenharia Elétrica (PPGEE) da Universidade Federal de Minas Gerais (UFMG) de dois semestres distintos: 2016/2 e 2017/2. Para tal fim, foram disponibilizadas duas amostras, contendo sexo, altura e peso de alguns alunos [3]. Assim, duas análises estatísticas independentes são propostas: (i) uma sobre o IMC médio dos alunos do sexo masculino e (ii) uma sobre o IMC médio dos alunos do sexo feminino. Para ambos os casos, a condução dos experimentos foram similares, no entanto, alguns testes tiveram que ser adaptados de acordo com as propriedades das distribuições amostrais investigadas.

Planejamento dos Experimentos

As hipóteses estatísticas foram definidas com o intuito de responder às questões propostas abaixo:

- Há evidências de que o IMC médio dos alunos do PPGEE/UFMG de 2016/2 é diferente do IMC médio dos alunos do PPGEE/UFMG de 2017/2 no que se refere ao sexo masculino?
- E quanto ao sexo feminino?

Em concordância com a proposta de comparação do IMC médio entre os alunos de semestres distintos, as hipóteses de teste podem ser formuladas sobre o parâmetro média:

$$\begin{cases} H_0 : \mu_{2016} = \mu_{2017} \\ H_1 : \mu_{2016} \neq \mu_{2017} \end{cases}$$

onde a hipótese nula H_0 implica na igualdade entre os IMCs médios dos alunos de 2016/2 e 2017/2 e a hipótese alternativa bilateral H_1 na diferença dos IMCs médios e, portanto, em uma potencial diferença dos estilos de vida dos alunos.

Os parâmetros experimentais para realização dos testes são:

- A probabilidade admissível de rejeição da hipótese nula quando ela é verdadeira é de apenas 5%, isto é, o nível de significância do teste é $\alpha = 0,05$;
- A potência do teste é de $\pi = 1 - \beta = 0,8$. Em outras palavras, deseja-se uma probabilidade de falha ao rejeitar a hipótese nula quando ela é falsa de 20%;
- O tamanho de efeito de mínima relevância prática foi definido em $\delta^* = 1$, ou seja, pretende-se detectar, a partir do teste de hipóteses, desvios de 1 kg/m^2.

Pré-Processamento dos Dados

Conforme mencionado anteriormente, as bases de dados imc_20162.csv e CS01_20172.csv foram disponibilizadas [3]. A amostra relativa ao semestre de 2016/2 dispõe dos atributos número de identificação do aluno, visto que a coleta manteve o sigilo dos estudantes, curso (graduação ou pós-graduação), gênero, peso (em kg) e altura (em m). A princípio, foi necessário extrair apenas as informações dos alunos cujo vínculo com a universidade era de discente da pós-graduação e, posteriormente, realizar a fragmentação por gênero, formando duas amostras independentes ($M2016$ e $F2016$). A amostra relativa ao semestre de 2017/2, por sua vez, compreendia os atributos peso (em kg), altura (em m), sexo e idade. Além disso, as observações eram somente de alunos da pós-graduação e, portanto, exigiu apenas a separação por gênero em duas outras amostras ($M2017$ e $F2017$).

A partir dos pesos e alturas disponíveis, os índices de massa corporal foram calculados para cada aluno, conforme a Equação 1. Por fim, as observações de interesse foram compiladas em uma única estrutura de dados. Os 10 primeiros IMCs de cada amostra podem ser visualizados abaixo, onde os valores "NA" presentes nas amostras femininas ($F2016$ e $F2017$) indicam que ambas possuem tamanho amostral $N < 10$, isto é, 7 e 4 observações, respectivamente. As amostras masculinas ($M2016$ e $M2017$), no entanto, apresentam 21 observações cada uma.

```
# 10 primeiras observações de cada amostra
show(IMCs[c(1:10),])
```

```
##       M2016    M2017    F2016    F2017
## 1  24.96801 29.73704 18.45917 17.36111
## 2  23.23346 26.95568 20.19509 20.83253
## 3  28.07504 29.06574 19.72318 17.84652
## 4  37.55102 30.42185 22.48133 17.74623
## 5  22.40879 20.76125 22.58955       NA
## 6  24.28098 24.38272 25.18079       NA
## 7  27.14304 23.74764 18.96193       NA
## 8  24.41928 22.49135       NA       NA
## 9  22.47121 24.89814       NA       NA
## 10 21.62630 22.40818       NA       NA
```

Análise Exploratória de Dados

Algumas primeiras propriedades das quatro amostras, como média, moda, mediana, valores extremos, variância e desvio podem ser obtidas de imediato.

```
# Estatísticas iniciais da amostra masculina de 2016
stats_M2016
```

```
##   Variância    Média     Moda  Mediana   Mínimo   Máximo   Desvio
## 1  18.69141 24.93595 24.96801 24.35542 17.57707 37.55102 4.323356
```

```
# Estatísticas iniciais da amostra masculina de 2017
stats_M2017
```

```
##   Variância    Média     Moda  Mediana   Mínimo   Máximo   Desvio
## 1  11.80097 24.28551 29.73704 23.74764 17.72212 30.42185 3.435254
```

```
# Estatísticas iniciais da amostra feminina de 2016
stats_F2016
```

```
##   Variância    Média     Moda  Mediana   Mínimo   Máximo   Desvio
## 1   5.83963 21.08443 18.45917 20.19509 18.45917 25.18079 2.416533
```

```
# Estatísticas iniciais da amostra feminina de 2017
stats_F2017
```

```
##   Variância   Média     Moda  Mediana   Mínimo   Máximo   Desvio
## 1  2.573854 18.4466 17.36111 17.79637 17.36111 20.83253 1.604324
```

A priori, é possível evidenciar que a diferença entre as médias amostrais masculinas ($\Delta\bar{x}_M = 0,6504$) é bem menos discrepante que a diferença entre as médias amostrais femininas ($\Delta\bar{x}_F = 2,6378$). No que se refere a diferença entre as variâncias amostrais, tanto a disparidade do gênero masculino quanto do gênero feminino são bastante expressivas, sendo $\Delta s_M = 6,8904$ e $\Delta s_F = 3,2657$, respectivamente. Outro fato interessante é que as observações coletadas retratam que, na média, os alunos entrevistados de ambos os sexos seguem um padrão de vida ideal ($\overline{IMC} \in [18,5;\ 25]$), onde os homens estão mais próximos do limite superior (sobrepeso) e as mulheres estão mais próximas do limite inferior (baixo peso).

A fim de compreender melhor os dados em estudo e, posteriormente, inferir sobre as populações de onde as amostras provêm, serão analisadas algumas representações gráficas. No que tange a distribuição de frequência das observações, pode-se constatar que as amostras masculinas apresentam um comportamento bastante similar ao de uma distribuição normal. No caso da amostra $M2016$, em específico, tal característica seria melhor assimilada caso não houvesse o *outlier* cujo IMC é 37.55 kg/m^2. Em relação às amostras femininas, não é possível identificar indícios de normalidade a partir de seus histogramas, uma vez que os seus respectivos tamanhos amostrais são muito pequenos.

```
# Histogram
ggplot(IMCs, aes(value)) + xlab(expression("IMC (kg/"*"m"^2*")")) +
    ylab("Frequência") + geom_histogram(bins = 20) +
    facet_wrap(~variable, scales = 'free')
```

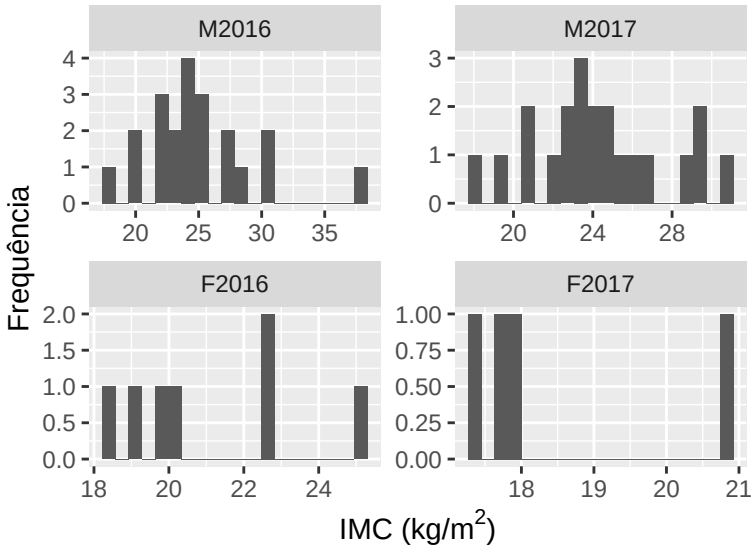

Figura 1: Histogramas.

Os diagramas de caixa, em princípio, corroboram algumas análises anteriores quanto às distribuições amostrais. O segundo quartil de $M2016$, em particular, está praticamente no centro da caixa, representando uma mediana próxima da média e, portanto, evidências de normalidade. O mesmo não ocorre para a amostra $M2017$, que visualmente apresenta maior assimetria do segundo quartil em relação ao centro da caixa. No entanto, a diferença entre a média e a mediana de $M2017$ ($\overline{x}_{2017}$ - $\overline{m}_{2017} = 0,5378$) é ainda menor que a mesma diferença para $M2016$ ($\overline{x}_{2016}$ - $\overline{m}_{2016} = 0,5805$), o que instiga análises ainda mais singulares. As assimetrias apresentadas para as distribuições amostrais do gênero feminino fortalecem os princípios de não-normalidade evidenciados anteriormente.

```
# Boxplot
ggplot(data = IMCs, aes(y = "", x = value)) + xlab(expression("IMC (kg/"*"m"^2*")")) +
    ylab("") + geom_boxplot(lwd = 0.3) +
    facet_wrap(~variable, scales = 'free') + coord_flip()
```

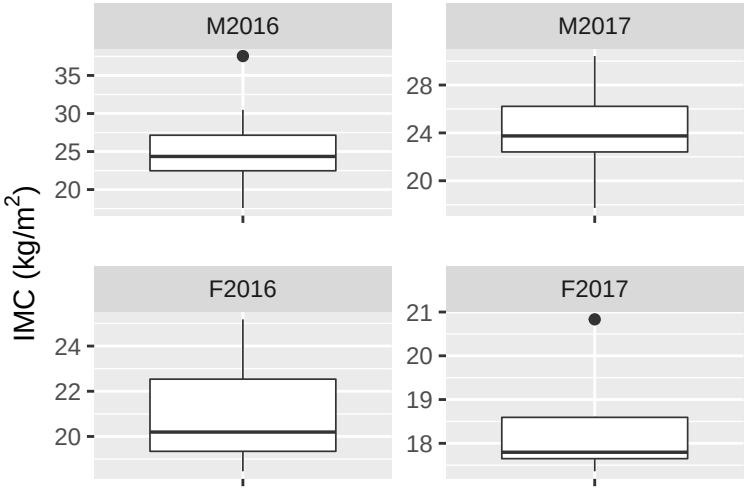

Figura 2: Boxplots.

Por fim, gráficos quantil-quantil foram utilizados para comparar as distribuições de probabilidade de cada uma das amostras (eixo das ordenadas) com uma distribuição normal (eixo das abcissas). Tal análise foi tomada para concluir sobre a normalidade das distribuições, corroborando ou refutando conclusões anteriores. Como esperado, o gráfico Q-Q da amostra masculina de 2016/2 sugere que os dados são normalmente distribuídos, uma vez que a reta se ajusta bem aos pontos (desconsiderando *outliers*). As amostras $M2017$ e $F2016$ também apresentaram bons ajustes do modelo aos dados e, consequentemente, também sugerem normalidade. Quanto à amostra feminina de 2017/2, há claros sinais de que os dados não seguem uma distribuição normal. Posteriormente, todas essas premissas serão validades a partir de testes estatísticos, como Kolmogorov-Smirnov e Shapiro-Wilk.

```
# QQ-Plots
ggplot(data = IMCs, aes(sample = value)) +
  facet_wrap(~variable, scales = "free") +
  stat_qq() + stat_qq_line() + scale_y_continuous(name = 'Quantis da Amostra') +
  scale_x_continuous(name = 'Quantis Teóricos Normais')
```

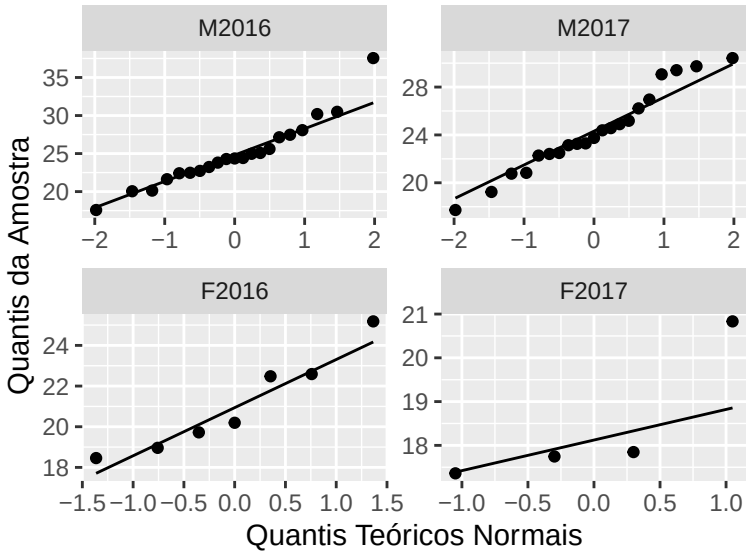

Figura 3: QQ-Plots

Tamanho de Efeito

Comumente, os estudos científicos apresentam a significância dos resultados alcançados, contudo, é aconselhável mensurar também a tamanho de efeito (importância real) concernente às diferenças encontradas em termos de média ou variância dos grupos avaliados [8]. Embora o valor p seja importante por trazer a informação sobre a existência ou não de diferenças entre os grupos avaliados, considerando um nível de significância α, ele não mensura o tamanho real deste efeito, ou seja, quão expressiva é a diferença encontrada. Por causa disso, recomenda-se que o tamanho de efeito esteja sempre presente em conjunto com o valor de p. [11]

A literatura aborda algumas metodologias para calculo do tamanho de efeito, como o Teste de Cohen, Teste de Glass, Teste de Hedges, Teste (ψ), dentre outros. Cada um desses testes têm diferentes maneiras de calcular o tamanho de efeito quanto a um determinado estimador, alguns fazendo uso da média (Cohen(d), Glass (Δ), Hedges(g), ψ) e outros usando as variâncias (Pearson, η^2, ω^2 e Cohen(f^2)). O teste de Cohen, por exemplo, calcula o tamanho de efeito d quanto à média, obtendo o quociente da diferença entre as médias dos grupos pelo desvio padrão agrupado, conforme mostrado nas Equações (2) e (3) [6].

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s} = \frac{\mu_1 - \mu_2}{s} \tag{2}$$

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \tag{3}$$

A medida mais usada para calcular o tamanho de efeito para um teste t de Student é o d de Cohen (Cohen 1998) [7]. A interpretação deste teste se dá a partir do d calculado pela Equação (2). O efeito é considerado muito pequeno quando d resulta entre $[0,01; 0,20[$ [10], pequeno, médio e grande quando $d \in [0,20; 0,50[$, $[0,50; 0,80]$ e $[0,80; 1,20[$, respectivamente [4], muito grande se $d \in [1,20; 2]$ [10] e enorme a partir de $2,0$ [10].

Para calcular o tamanho do efeito usando este método, algumas premissas assumidas: normalidade dos dados e homoscedasticidade. [9] Usando o Teste de Shapiro-Wilk é possível inferir sobre a normalidade das amostras, haja vista que a Hipótese Nula deste teste é de que os dados provêm de uma distribuição normal [12]. A função shapiro.test do pacote stats do R considera o nível de confiança de 90% ($\alpha = 0,1$) para avaliação das amostras fornecidas [1].

Tamanho do efeito para amostras masculinas

O Teste de Shapiro-Wilk evidenciou que a ambas as amostras do grupo masculino M2016 e M2017 apresentam características de normalidade, pois a hipótese nula desse teste não pôde ser refutada, com nível de confiança de 90%. Nesse caso, o valor de p resultou em $p = 0,1275$ e $p = 0,6206$, respectivamente.

Uma vez que a variância populacional não é conhecida, não se pode afirmar sobre a homoscedasticidade das amostras. Por isso, torna-se necessário a realização de um teste específico, como o Teste F para igualdade de variâncias. Esse teste tem como hipótese nula a proposição de que os grupos comparados têm a mesma variância, partindo do pressuposto de que os dados são normais. O teste estatístico F é calculado pelo quociente da variância do primeiro grupo pelo segundo. Se a resultante de F é maior que o valor crítico superior ou menor que o valor crítico inferior, a hipótese nula é rejeitada [5].

```
# Teste F: homocedasticidade das amostras masculinas (assume normalidade)
var.test(x = M2016, y = M2017, alternative = "two.sided")
```

```
##
##  F test to compare two variances
##
## data:  M2016 and M2017
## 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
```

A execução do teste F para o grupo de dados M2016 e M2017 resultou em $F = 1,5838$ e $p = 0,3119$, que reflete que a igualdade de variância pode ser assumida para os dados em questão. Com isso, pode-se executar o Cohen's d para computar o tamanho de efeito para as tais amostras.

Por meio do uso do pacote `effsize` do R [13] foi calculado o tamanho de efeito referente aos dados estudados, com uso da função `cohen.d`.

```
cohen.d(sort(M2016),sort(M2017))
```

```
##
## Cohen's d
##
## d estimate: 0.1665831 (negligible)
## 95 percent confidence interval:
##      lower      upper
## -0.4582151  0.7913813
```

```
cohen.d(sort(F2016),sort(F2017))
```

```
##
## Cohen's d
##
## d estimate: 1.21019 (large)
## 95 percent confidence interval:
##      lower      upper
## -0.3231255  2.7435055
```

Os resultados mostraram que, em relação à média do IMC do grupos masculinos, d resultou em $0,1665831$, que evidencia que não existem diferenças significativas entre as médias estudadas.

Tamanho do efeito para amostras femininas

Análise Estatística

Uma vez analisado a normalidade das amostras, pode-se definir quais testes são aplicáveis para comparação das médias. Para as observações do IMC masculino, tanto a amostra do segundo semestre de 2016 quando a do segundo semestre de 2017 são normais, sendo assim, pode-se aplicar o teste T de Welch para duas amostras. Porem, para as observações do IMC feminino, a amostra do segundo semestre de 2017 não é normal, sendo assim, para teste das médias das amostras do grupo feminino, deve-se utilizar um teste não-paramétrico. O teste T de Welch duas amostras tem como premissa que as amostras a serem testadas tenham comportamento normal, como já mencionado. Os parâmetros de entrada, além das duas amostras envolvidas, devem conter que tipo da hipótese alternativa, se bilateral, unilateral esquerda ou direita, neste experimento, deseja-se que as médias tenham valores diferentes, sendo assim, bilateral. Outro parâmetro de entrada é o valor da diferença das médias, neste caso, como a hipótese nula é que as médias das duas amostras são iguais, a diferença das médias $\mu_{2016} - \mu_{2017} = 0$, e por fim, o nível de confiança adotado para os testes foi de 95%.

```
(t_test <- t.test(x = M2016,
                  y = M2017,
                alternative = "two.sided",
                conf.level = 0.95))
```

```
##
##  Welch Two Sample t-test
##
## data:  M2016 and M2017
## 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
```

```
cat('Intervalo de confiança:', t_test$conf.int[1:2])
```

```
## Intervalo de confiança: -1.788823 3.089716
```

O resultado do teste T de Welch retornou um p-valor de $0,5925$, maior que o nível de significância de $0,05$, isso significa que com 95% de confiança, não deve-se rejeitar a hipótese nula H_0 de que as médias das duas amostras são iguais. Sendo a hipótese alternativa H_1 bilateral, o intervalo de confiança para a diferença das médias é [-1,788823 , 3,089716].

O teste de Wilcoxon foi apontado para análise das amostras do grupo feminino, por uma das amostras não ser normal. É um teste não-paramétrico, equivalente ao teste Mann-Whitney quando as amostras não são pareadas. A hipótese nula H_0 é de que as médias são iguais, sendo assim, o que nos parâmetro do teste de Wilcoxon chama-se de mudança de localização, o valor deve ser igual a zero, como, por padrão, já é. A hipótese alternativa H_1 é de que as médias são diferentes, um teste bilateral, e o nivel de confiança para o teste tambem é de 95%.

```
(wilcox_test <- wilcox.test(x = F2016,
                            y = F2017,
                            alternative = "two.sided",
                            conf.int = TRUE))
```

```
##
##  Wilcoxon rank sum test
```

```
##
## data:  F2016 and F2017
## W = 24, p-value = 0.07273
## alternative hypothesis: true location shift is not equal to 0
## 95 percent confidence interval:
##  -0.6374374  5.2284403
## sample estimates:
## difference in location
##               2.162763
```

```
cat('Intervalo de confiança:', wilcox_test$conf.int)
```

```
## Intervalo de confiança: -0.6374374 5.22844
```

Como o p-valor retornado foi maior que o nível de significância do experimento, $0,07273 > 0,05$, a hipótese nula H_0, com 95% de confiança, não pode ser rejeitada. O intervalo de confiança para a diferença das médias é $[-0,6374374,\ 5,22844]$. Para os testes tanto do grupo masculino quanto para o feminino, as hipóteses nulas H_0 foram aceitas, neste caso, hipóteses fracas que não foram rejeitadas. Não significa que são verdadeiras, mas que com os testes realizados a 95% de confiança, não pode-se rejeitá-las.

Validação de Premissas

```
shapiro.test(M2016)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  M2016
## W = 0.92833, p-value = 0.1275
```

```
shapiro.test(M2017)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  M2017
## W = 0.96494, p-value = 0.6206
```

```
shapiro.test(F2016)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  F2016
## W = 0.91974, p-value = 0.4674
```

```
shapiro.test(F2017)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  F2017
## W = 0.7475, p-value = 0.03659
```

```r
# Teste F: homocedasticidade das amostras masculinas (assume normalidade)
var.test(x = M2016, y = M2017, alternative = "two.sided")
```

```
##
##  F test to compare two variances
##
## data:  M2016 and M2017
## 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
```

Conclusões

Discussão de Melhorias

Atividades Desempenhadas

Referências

[1] Shapiro.test: Shapiro-Wilk Normality Test. https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/shapiro.test, 2020. From stats v3.6.2.

[2] Alexandra M. N. Borba, Juliane H. Wolff Wolff, and Rafaela Liberali. Avaliação do Perfil Antropométrico e Alimentar de Idosos Institucionalizados em Blumenau-Santa Catarina. *RBONE-Revista Brasileira de Obesidade, Nutrição e Emagrecimento*, 1(3), 2007.

[3] Felipe Campelo. Lecture Notes on Design and Analysis of Experiments. http://git.io/v3Kh8, 2018. Version 2.12; Creative Commons BY-NC-SA 4.0.

[4] Jacob Cohen. *Statistical power analysis for the behavioral sciences*. Academic press, 2013.

[5] Ji-Qian Fang. *Handbook of medical statistics*. World Scientific, 2018.

[6] Guanis de Barros Vilela Junior. Tamanho do efeito (effect size).

[7] Alboukadel Kassambara. T-test essentials: Definition, formula and calculation. https://www.datanovia.com/en/lessons/t-test-effect-size-using-cohens-d-measure/#:~:text=T%2Dtest%20conventional%20effect%20sizes,if%20it%20is%20statistically%20significant. Acesso: 15 ago.

[8] Ken Kelley and Kristopher J Preacher. On effect size. *Psychological methods*, 17(2):137, 2012.

[9] Chao-Ying Joanne Peng and Li-Ting Chen. Beyond cohen's d: Alternative effect size measures for between-subject designs. *The Journal of Experimental Education*, 82(1):22–50, 2014.

[10] Shlomo S Sawilowsky. New effect size rules of thumb. *Journal of Modern Applied Statistical Methods*, 8(2):26, 2009.

[11] Gail M Sullivan and Richard Feinn. Using effect size—or why the p value is not enough. *Journal of graduate medical education*, 4(3):279–282, 2012.

[12] R Development Core Team. Shapiro-Wilk Normality Test. https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/shapiro.test. Documentation reproduced from package stats, version 3.6.2, License: Part of R 3.6.2.

[13] Marco Torchiano and Maintainer Marco Torchiano. Package 'effsize'. 2020.