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 estatísticos tiveram que ser adaptados de acordo com as propriedades das distribuições amostrais em investigação, que foram validadas previamente.

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?
- Há evidências de que a mediana do IMC dos alunos do PPGEE/UFMG de 2016/2 é diferente da mediana do IMC dos alunos do PPGEE/UFMG de 2017/2 no que se refere ao sexo feminino?

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

$$\begin{cases} H_0 : \mu_{M2016} = \mu_{M2017} \\ H_1 : \mu_{M2016} \neq \mu_{M2017} \end{cases} \quad \begin{cases} H_0 : m_{F2016} = m_{F2017} \\ H_1 : m_{F2016} \neq m_{F2017} \end{cases}$$

onde a hipótese nula H_0 implica na igualdade entre os IMCs médios ou medianos dos alunos de 2016/2 e 2017/2 e a hipótese alternativa bilateral H_1 na diferença dos IMCs médios ou medianos 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%.

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 8 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 $N < 8$, isto é, 7 e 4 observações, respectivamente. As amostras masculinas ($M2016$ e $M2017$), no entanto, apresentam 21 observações cada uma.

```
##      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
```

Análise Exploratória de Dados

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

```
##       Variância    Média     Moda  Mediana   Mínimo   Máximo   Desvio
## M2016 18.691408 24.93595 24.96801 24.35542 17.57707 37.55102 4.323356
## M2017 11.800968 24.28551 29.73704 23.74764 17.72212 30.42185 3.435254
## F2016  5.839630 21.08443 18.45917 20.19509 18.45917 25.18079 2.416533
## F2017  2.573854 18.44660 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.

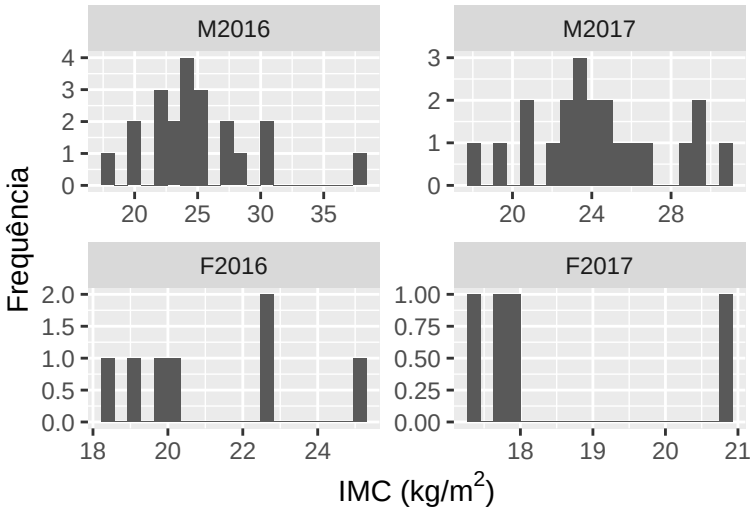

Figura 1: Histogramas.

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.

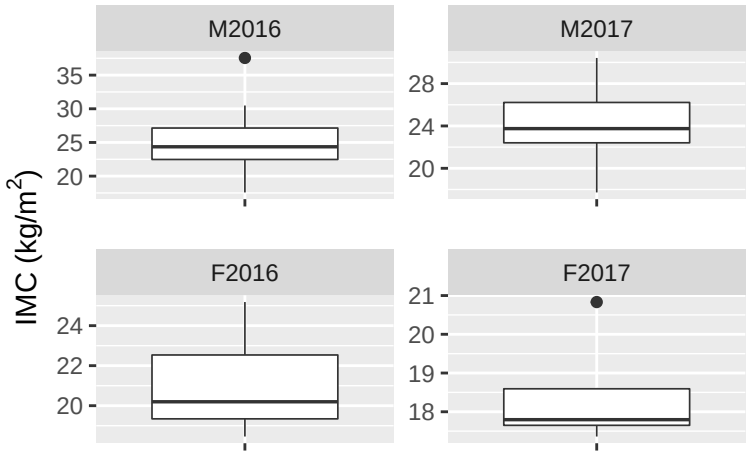

Figura 2: Boxplots.

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.

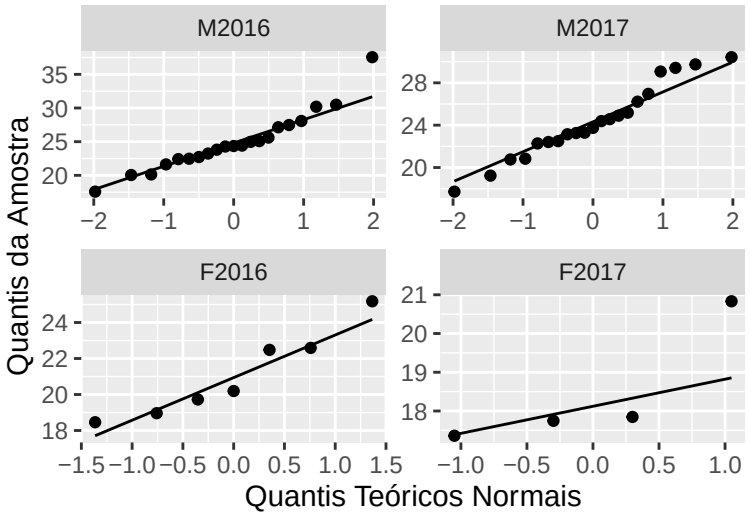

Figura 3: QQ-Plots

Parte 1: Amostras Masculinas

Tamanho de Efeito

Geralmente, os estudos científicos relatam a significância dos resultados alcançados. Contudo, é aconselhável mensurar também o tamanho de efeito (importância real) pertinente às diferenças encontradas em termos de média ou variância dos grupos avaliados [6]. A literatura aborda algumas metodologias para isso, como o Teste de Cohen, Teste de Glass, Teste de Hedges, 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 a partir das variâncias (Pearson, η^2, ω^2 e Cohen(f^2)). O primeiro teste de Cohen mencionado, 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 a Equação (2) [13]:

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

onde $\overline{x}_1$ e $\overline{x}_2$ são, respectivamente, as médias das amostras 1 e 2 e s é o desvio padrão agrupado dado pela Equação (3):

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

em que n_1 e n_2 são, nessa ordem, os tamanhos amostrais dos grupos 1 e 2, bem como s_1 e s_2 são os seus respectivos desvios amostrais.

A medida mais usada para calcular o tamanho de efeito para um teste t de Student é o d de Cohen [5]. A interpretação deste teste se dá a partir do d calculado pela Equação (2) conforme a seguinte classificação [9]:

- Muito pequeno: $0,01 \leq d < 0,20$
- Pequeno: $0,20 \leq d < 0,50$
- Médio: $0,50 \leq d < 0,80$
- Grande: $0,80 \leq d < 1,20$
- Muito grande: $1,20 \leq d < 2$
- Enorme: $d \geq 2,0$

Para calcular o tamanho de efeito usando este método, algumas premissas devem ser assumidas sobre os dados: normalidade e homocedasticidade [7]. Usando o Teste de Shapiro-Wilk é possível inferir sobre a normalidade das amostras, tendo em vista que a hipótese nula H_0 deste teste é de que os dados provêm de uma distribuição normal [11]. 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]. A validação da homocedasticidade será realizada a posteriori, uma vez que a premissa de alguns dos testes capazes de verificar tal premissa assumem normalidade.

O Teste de Shapiro-Wilk evidenciou que as amostras do grupo masculino $M2016$ e $M2017$ provêm de uma distribuição populacional com caráter normal ao nível de confiança de 90%, pois a hipótese nula desse teste não pôde ser refutada. Nesse caso, o valor de p resultou em $0,1275$ e $0,6206$, respectivamente, ambos superiores ao nível de significância $\alpha = 0,1$.

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

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

Uma vez que as variâncias populacionais não são conhecidas, nada se pode afirmar a cerca da homocedasticidade (igualdade de variâncias). 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 resultam de populações com 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 [4].

```
##
##  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 calcular o d de Cohen como uma boa estimativa de tamanho de efeito para as amostras masculinas. O pacote `effsize` do R [12] permite tal cálculo a partir da função `cohen.d`.

```
##
## Cohen's d
```

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

O tamanho de efeito reportado é $d = 0,1665$.

Poder do Teste

No entanto, o cálculo anterior não leva em consideração a potência desejada ($\pi = 0,80$) e nem o nível de significância estipulado ($\alpha = 0,05$). É bem provável que, com essas especificações e com os tamanhos amostrais consideravelmente pequenos ($n_M = n_{M2016} = n_{M2017} = 21$), não seja possível detectar desvios no IMC médio masculino de $0,1665$ kg/m^2 a partir das hipóteses de teste, conforme sugere o cálculo do d de Cohen.

Supondo o desvio padrão conjugado $s_M = 3,9046$ calculado a partir da Equação (3), o nível de significância $\alpha = 0,05$, o tamanho amostral n_M e o d de Cohen calculado anteriormente, a potência obtida pelo método `power.t.test` é de apenas $3,4\%$. Tal valor é cerca de $76,6$ p.p menor que a potência desejada e, portanto, não é aceitável para realização dos experimentos.

```
##
##      Two-sample t test power calculation
##
##               n = 21
##           delta = 0.1665831
##              sd = 3.904637
##       sig.level = 0.05
##           power = 0.03400076
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

Para que fosse possível detectar efeitos iguais ou maiores que $d = 0,1665$ de forma a preservar os parâmetros experimentais desejados, seriam necessárias cerca de 8626 amostras para cada um dos semestres.

```
##
##      Two-sample t test power calculation
##
##               n = 8625.523
##           delta = 0.1665831
##              sd = 3.904637
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

Como não existem possibilidades de se obter mais amostras para os experimentos, é possível detectar, a um nível de significância $\alpha = 0,05$ e potência $\pi = 0,80$, um tamanho de efeito de mínima relevância prática de $d = 3,4596$.

```
##
##      Two-sample t test power calculation
##
##               n = 21
```

```
##           delta = 3.459679
##              sd = 3.904637
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

Análise Estatística

Uma vez que as premissas de normalidade e homocedasticidade foram validadas, o teste estatístico t de Student pode ser efetuado para comparação das médias. A hipótese nula é que a média da variável estudada (IMC) é igual nas duas populações, ou seja, a diferença entre a média das duas populações é igual a zero. Já a hipótese alternativa indica que há diferenças nas médias das populações, ou seja, $\mu_{2016} - \mu_{2017} \neq 0$. O nível de confiança adotado para os testes foi de 95%.

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

O resultado do teste t de Student retornou um p-valor de $0,5923$, maior que o nível de significância de $0,05$. Assim, com 95% de confiança não é possível rejeitar a hipótese nula H_0 de que as médias das duas populações são iguais. Sendo a hipótese alternativa H_1 bilateral, o intervalo de confiança para a diferença das médias é $[-1,784943;\ 3,085836]$. É interessante ressaltar que o resultado para o teste t de Welch foi bastante similar em termos de p-valor e intervalo de confiança e, portanto, também falhou em rejeitar a hipótese nula.

Parte 2: Amostras Femininas

Análise Estatística

De forma a verificar as premissas de normalidade para as amostras femininas, foram realizados testes de Shapiro-Wilk para ambas amostras com um nível de significância de $\alpha = 0,10$ [11].

```
##
##  Shapiro-Wilk normality test
##
## data:  na.omit(F2016)
## W = 0.91974, p-value = 0.4674

##
##  Shapiro-Wilk normality test
##
## data:  na.omit(F2017)
## W = 0.7475, p-value = 0.03659
```

No que se refere à amostra feminina do semestre 2016/2, cujo tamanho amostral é $n_{F2016} = 7$, não é possível rejeitar a hipótese nula de que a amostra proveio de uma população com distribuição normal ao nível

de confiança de 90% ($p > \alpha$). Quanto a amostra feminina de 2017/2, cujo tamanho amostral é $n_{F2017} = 4$, não se pode afirmar o mesmo, uma vez que a hipótese nula é rejeitada ao nível de confiança de 90% ($p < \alpha$). Em outras palavras, pode-se afirmar com 90% de confiança que $F2016$ provém de uma população normal e $F2017$ não.

Com o intuito de averiguar também a premissa de homocedasticidade, utilizou-se o teste de Fligner-Killeen [10], que não pressupõe amostras provenientes de população com distribuição normal. A hipótese nula desse teste é de que as variâncias das duas populações, das quais foram retiradas as amostras testadas, são iguais. Já a hipótese alternativa é de que essas varianças não são iguais.

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  value by variable
## Fligner-Killeen:med chi-squared = 0.71101, df = 1, p-value = 0.3991
```

Assim, a partir do p-valor substancialmente maior que o nível de significância, pode-se afirmar que não é possível refutar a hipótese nula ao nível de confiança de 95% e, consequentemente, as amostras provêm de populações homocedásticas.

Por fim, o teste de Wilcoxon para amostras independentes foi definido para análise das populações femininas, uma vez que ele não pressupõe normalidade das amostras. Tal teste é não-paramétrico, equivalente ao teste Mann-Whitney quando as amostras não são pareadas, e sua hipótese nula é de que as medianas das populações são iguais. A hipótese alternativa é de que as medianas populacionais são diferentes.

```
##
##  Wilcoxon rank sum test
##
## data:  na.omit(F2016) and na.omit(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
```

```
## 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 não pôde ser rejeitada ao nível de confiança de 95%. O intervalo de confiança bilateral para a diferença das medianas é $[-0,6374374;\ 5,22844]$.

Tamanho de Efeito

$$A = \frac{n_1 n_2 - U}{n_1 n_2} \tag{4}$$

onde $W = n_1 n_2 - U = 24$, reportado pela função `wilcox.test` em R [8].

```
##
## Vargha and Delaney A
##
## A estimate: 0.8571429 (large)
```

Poder do Teste

Dado que os tamanhos amostrais são distintos ($n_{F2016} \neq n_{F2017}$) e extremamente pequenos, é bem provável que a potência para efeitos maiores ou iguais a $A = 0,8571$ seja substancialmente menor que a desejada ($\pi = 0,80$).

Conclusões

Discussão de Melhorias

Atividades Desempenhadas

As hipóteses de teste para ambos experimentos foram definidas em concordância com os três autores. Em relação à primeira parte do estudo de caso, tanto o pré-processamento dos dados, quanto a análise exploratória e o poder do teste foram conduzidos pelo Pedro. A Samara realizou o cálculo do tamanho de efeito e a validação das premissas assumidas para os testes estatísticos. A análise estatística, por sua vez, foi feita pelo Sávio. A segunda parte do estudo de caso, referente às inferências sobre a distribuição feminina, foi realizada após sucessivas investigações. A análise estatística foi novamente realizada pelo Sávio e o cálculo do tamanho de efeito, bem como a validação das premissas inerentes, foram conduzidas em conjunto. Por fim, o Pedro realizou a análise de potência do teste e a Samara conclui o trabalho com as considerações finais e discussões de melhoria.

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

[5] Alboukadel Kassambara. T-test Essentials: Definition, Formula and d Calculation. https://www.datanovia.com/en/lessons/t-test-effect-size-using-cohens-d-measure/. Acesso em 15 Agosto de 2020.

[6] Ken Kelley and Kristopher J Preacher. On Effect Size. *Psychological methods*, 17(2):137, 2012.

[7] 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.

[8] John Ruscio. A Probability-Based Measure of Effect Size: Robustness to Base Rates and Other Factors. *Psychological methods*, 13(1):19, 2008.

[9] Shlomo S Sawilowsky. New Effect Size Rules of Thumb. *Journal of Modern Applied Statistical Methods*, 8(2):26, 2009.

[10] R Development Core Team. Fligner-Killeen Test Of Homogeneity Of Variances. https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/fligner.test. Documentation reproduced from package stats, version 3.6.2, License: Part of R 3.6.2.

[11] 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.

[12] Marco Torchiano. Package effsize, 2020. Version 0.8.0.

[13] Sonia VIEIRA. Bioestatística: Tópicos Avançados. *Campus, Rio de Janeiro*, 2003.