Estudo de Caso 02

Marcus e Paulo

15 de abril de 2016

O Experimento

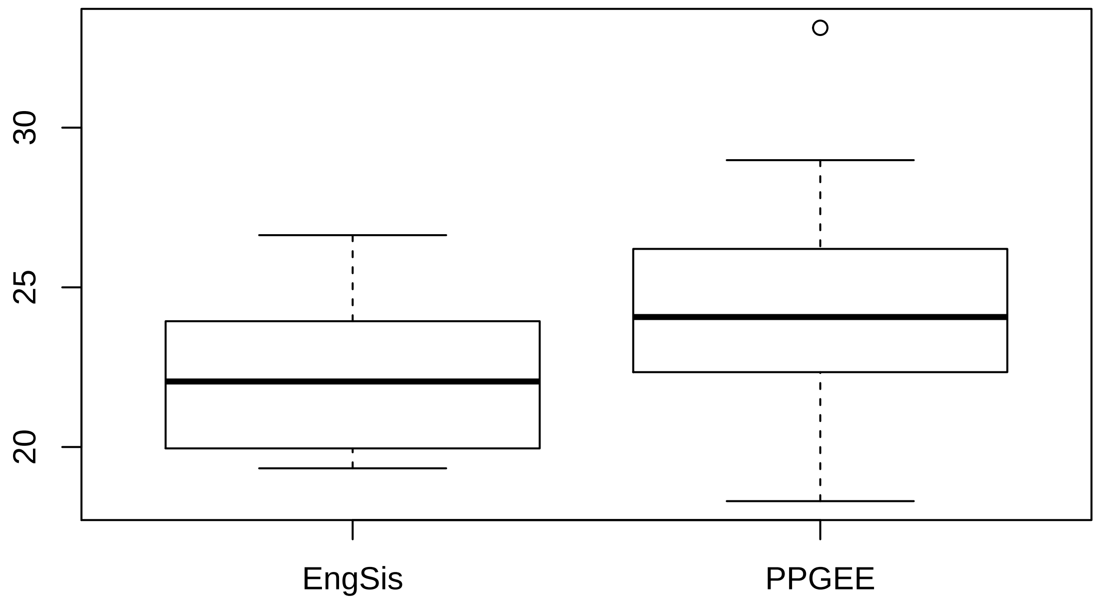
Neste experimento, comparamos o BMI (Body Mass Index ou Índice de Massa Corporal) de duas populações de estudantes, sendo uma delas, contendo alunos do curso de graduação em Engenharia de Sistemas, cujo tamanho do espaço amostral será definido como n_1, e a outra com estudantes da pós-graduação em Engenharia Elétrica (PPGEE), com espaço amostral de tamanho n_2.

Os dados de massa e altura coletados para o cálculo do BMI foram retirados do arquivo Data.csv, onde temos uma amostra com $n_1 = 13$ observações e outra com $n_2 = 28$ observações.

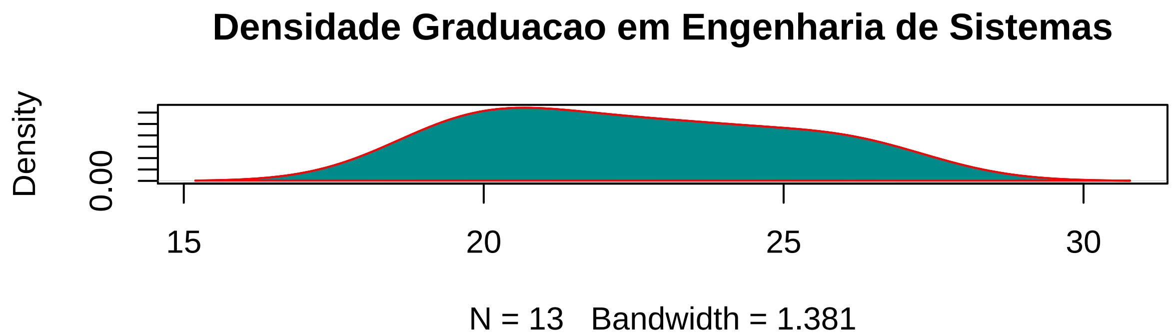
O BMI, neste caso, está sendo utilizado como um valor, que sozinho, já é suficiente para determinar o estilo de vida dos alunos de ambas as populações (apesar de termos certas limitações em seu uso). Com esta consideração, podemos relacionar a forma física dos estudantes com o seu curso, o que nos leva à questão científica de interesse para este estudo de caso: **"Qual o efeito do curso na forma física dos alunos?"**.

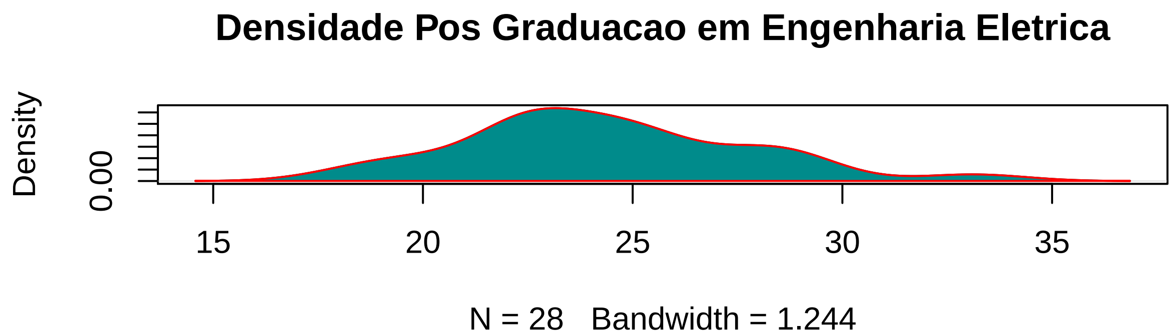
Análise exploratória dos dados

```r
##Leitura dos Dados
Data <- read.csv("Data.csv")

##Separacao das Posicoes de cada curso
posGrad <- (Data$Course == "EngSis")
posPosGrad <- !(posGrad)

#Calculo BMI
Data <- cbind(Data,
              BMI = (Data$Weight.kg) / (Data$Height.cm/100)^2 )

##Box Plot
par(mfrow = c(1, 1))
boxplot(BMI ~ Course, data = Data)
```




```
##Distribuicao de Densidade
par(mfrow = c(2, 1))
plot(density(Data$BMI [posGrad] ),
     main = "Densidade Graduacao em Engenharia de Sistemas", xlab = NULL)
polygon(density(Data$BMI [posGrad]), col="darkcyan", border = "red")

plot(density(Data$BMI [posPosGrad]),
     main = "Densidade Pos Graduacao em Engenharia Eletrica", xlab = NULL)
polygon(density(Data$BMI [posPosGrad]), col="darkcyan", border = "red")
```





Na figura com a função de densidade da população correspondente aos alunos de gradução em Engenharia de Sistemas, vemos que a função não tem uma forma exatamente normal, devido ao seu tamanho amostral, que é pequeno.

Como o tamanho amostral da população dos estudantes de pós-graduação em Engenharia Elétrica é maior, a curva teve um aspecto mais parecido com o de uma distribuição normal em relação à figura anterior, mas ainda precisa de mais observações. Note que no intervalo entre 30 e 35 no eixo x, temos uma protuberância incomum que não foi vista na outra figura, isso é o efeito causado por único ponto que é um “outlier” nesta população. Podemos observar que ele causou uma distorção considerável na distribuição de densidade do PPGEE, uma vez que temos poucas amostras.

No gráfico PPGEE gerado pela função “boxplot”, podemos perceber que os dados seguem uma distribuição muito bem comportada, que é aproximadamente simétrica em relação à mediana, sendo este fator, um ótimo indicador de estarmos trabalhando com uma distribuição realmente simétrica, isto é, não tem uma cauda muito longa nem para um lado e nem para o outro. Isso é concluído sem considerarmos a observação acima do ponto de máximo valor do gráfico, ao levá-la em conta, vemos que ela influencia um pouco na cauda direita da função de densidade desta população, ou seja, essa observação é o “outlier” citado anteriormente, mas ela não está distante o suficiente do gráfico para influenciar demais na média, uma vez que temos quase 30 observações nesta amostra.

Ao comparar o gráfico PPGEE com o gráfico EngSis, nota-se que a mediana deste último se encontra num ponto com menor valor que o ponto onde a mediana de PPGEE está localizada, em outras palavras, a distribuição amostral de EngSis está centrada em um valor de BMI menor do que o valor no centro da distribuição amostral de PPGEE. Com essa informação, temos um leve **indício** de que a média dos BMIs da população de alunos do curso de graduação em Engenharia de Sistemas seja menor que a média dos BMIs dos estudantes do PPGEE. Mas temos que realizar um teste de hipótese, porque o gráfico de EngSis é assimétrico e o ponto de mínimo valor da “caixa” é um pouco maior que o valor deste mesmo ponto do gráfico PPGEE, ou seja, temos características no gráfico que podem levar ao erro.

Escolha dos parâmetros necessários

Escolhemos os parâmetros como segue:

- Significância $\alpha = 0.05$
- Potência (1 - β) = 0.80
- Efeito de Interesse Mínimo $\delta^* = 1.26$

O valor do efeito de interesse mínimo foi definido como a diferença que um BMI teria se o aluno errasse em 2kg o seu peso (para menos) e em 2cm a sua altura (para mais), por ficar muito tempo sem se pesar e/ou se medir, por exemplo.

Definição das hipóteses do teste

Para este estudo de caso, desejamos inferir algo sobre a diferença entre os valores das médias dos BMIs de cada população, uma vez que estamos interessados em compará-las para saber o efeito do curso na forma física dos alunos. Tendo isto em mente, as hipóteses foram definidas da seguinte forma:

- $H_0 : \mu_1 - \mu_2 = 0$
- $H_1 : \mu_1 - \mu_2 \neq 0$

A hipótese alternativa foi definida como bilateral, pois não estamos interessados em saber se a média dos alunos de graduação em Engenharia de Sistemas é a maior das duas, ou se a média dos alunos da pós-graduação em Engenharia Elétrica é a maior, mas sim em usar a informação de qual é maior, ou menor (depois de realizados os testes e tiradas as conclusões) para tentar entender o efeito do curso sobre os estudantes.

O teste utilizado foi o teste T de duas amostras (two sample t-test), porque este tipo de teste é usado quando não conhecemos a variância real do processo. Como temos duas populações diferentes e foram tiradas uma amostra de cada uma, utilizamos a versão do teste para duas amostras.

Premissas do teste

Antes da realização do teste temos que apresentar as seguintes premissas com relação aos dados e com relação ao mesmo.

Dados:

1. Os alunos reportaram suas massas e alturas corretamente, sem viés.

Teste:

1. Os dados formam uma distribuição normal com média μ e variância σ^2;
2. O desvio padrão de cada população é desconhecido;
3. As duas amostras são independentes, ou seja, não há nenhum relacionamento entre os indivíduos de uma amostra com os da outra;

Relação entre as Variâncias

Quando temos um problema com duas amostras, antes de testar a hipótese das seções anteriores, é preciso fazer um teste, que, através dele, podemos assumir similaridade entre as variâncias de ambas as populações ou não. O teste F foi o escolhido, por ser responsável em detectar **igualdade** entre as variâncias.

As hipóteses do teste F são formuladas da mesma maneira da seção anterior, como feito abaixo:

- $H_0 : \sigma_1^2 = \sigma_2^2$
- $H_1 : \sigma_1^2 \neq \sigma_2^2$

A variável de teste do teste F é o quociente entre as variâncias amostrais, que serão chamadas de s_1^2 (cujo tamanho de sua amostra associada é n_1) e s_2^2 (cujo tamanho de sua amostra associada é n_2):

$$F = \frac{s_1^2}{s_2^2}$$

```
##F-test
var.test(Data$BMI [posGrad], Data$BMI [posPosGrad], conf.level = 1 - 0.05)
```

```
##
##  F test to compare two variances
##
## data:  Data$BMI[posGrad] and Data$BMI[posPosGrad]
## F = 0.58221, num df = 12, denom df = 27, p-value = 0.3256
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.2358306 1.7396886
## sample estimates:
## ratio of variances
##          0.5822072
```

O resultado da função "var.test" nos mostra que o valor 1 está dentro do intervalo de confiança, o que significa que não podemos rejeitar a hipótese nula deste teste F. Observando os valores extremos do intervalo de confiança, para o nosso caso, podemos assumir igualdade entre as variâncias, uma vez que o mesmo inclui o 1 (que é o valor atingido por F quando as variâncias amostrais são iguais) e pelo fato do seu tamanho não influenciar muito neste estudo de caso, ou seja, seus valores extremos não estão tão longe um do outro a ponto de interferir na nossa análise se assumirmos igualdade entre as variâncias.

Teste de hipótese

```
with(Data,
     t.test(BMI~Course,
            alternative = "two.sided",
            mu = 0,
            conf.level = 0.95,
            var.equal = TRUE))
```

```
##
##  Two Sample t-test
##
## data:  BMI by Course
## t = -1.7259, df = 39, p-value = 0.09228
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.9443195  0.3122583
## sample estimates:
## mean in group EngSis  mean in group PPGEE
##             22.52300             24.33903
```

O resultado da função "t.test" do R nos mostra um valor negativo de t_0 (-1.7259), que indica que a diferença entre as médias de cada amostra é menor que a hipótese nula. O valor-p é maior do que o nível de significância α, isto é, não temos evidência para a rejeição da hipótese nula.

Pelo intervalo de confiança indicado na resposta da função vemos que o 0 (valor da hipótese nula) está entre seus valores extremos (-3.9443195 e 0.3122583).

Para analisar o valor para o qual os dados das amostras tendem, consideramos o valor de δ^* igual a 1.26 como definido anteriormente. O módulo da diferença entre a hipótese nula e o valor extremo da direita no intervalo de confiança é menor que o valor desse parâmetro ($0.3122583 < 1.26$), e o módulo da diferença entre o valor extremo da esquerda do intervalo e a hipótese nula é maior que este valor ($3.9443195 > 1.26$), ou seja, o intervalo [-3.9443195,0] é mais significativo para a análise, pois supera o tamanho de efeito mínimo δ^*.

Com essas informações, percebemos que os dados estão sugerindo que a população composta pelos alunos do curso de graduação em Engenharia de Sistemas tem uma média de BMIs μ_1, menor do que a da pós-graduação em Engenharia Elétrica μ_2, uma vez que pelo resultado do teste, parece ter uma chance maior da diferença entre elas estar entre 0 e -4, aproximadamente. Isso faz sentido de se supor ao ver as médias de cada amostra na última linha do resultado da função.

Potência do teste

```r
require("pwr");
```

```
## Loading required package: pwr
```

```r
potenciaTest <- pwr.t2n.test(n1 = 13,
                             n2 = 28,
                             d = 1.26,
                             sig.level = 0.05,
                             alternative = "two.sided");
print( paste("Potencia do teste", potenciaTest$power, sep = " = "))
```

```
## [1] "Potencia do teste = 0.955470519247858"
```

Estamos interessados em saber a potência do nosso teste pra encontrar uma diferença de BMI de $\pm$ 1.26 kg m^{-2} entre as médias das amostras dos alunos de graduação em Engenharia de Sistemas e de pós-graduação em Engenharia Elétrica.

O valor desta potência retornado pela função é de aproximadamente 0.96, ou 96%, isto é, foi apresentado um valor de potência alto. Isso significa que a probabilidade de se cometer um erro do tipo II é baixa e a **chance** é grande da diferença entre as médias das populações ser algo em torno do tamanho de efeito mínimo, baseado nos resultados utilizando as duas amostras.

Para obter uma potência de aproximadamente 80% ou 0.8, como definido nas seções anteriores, nós podemos diminuir o número de observações de 28 para 10, uma redução muito útil, deixando a coleta de dados mais simples, se uma potência de 80% for considerada suficiente.

Verificação das premissas

Devem ser verificadas as premissas do teste consideradas:

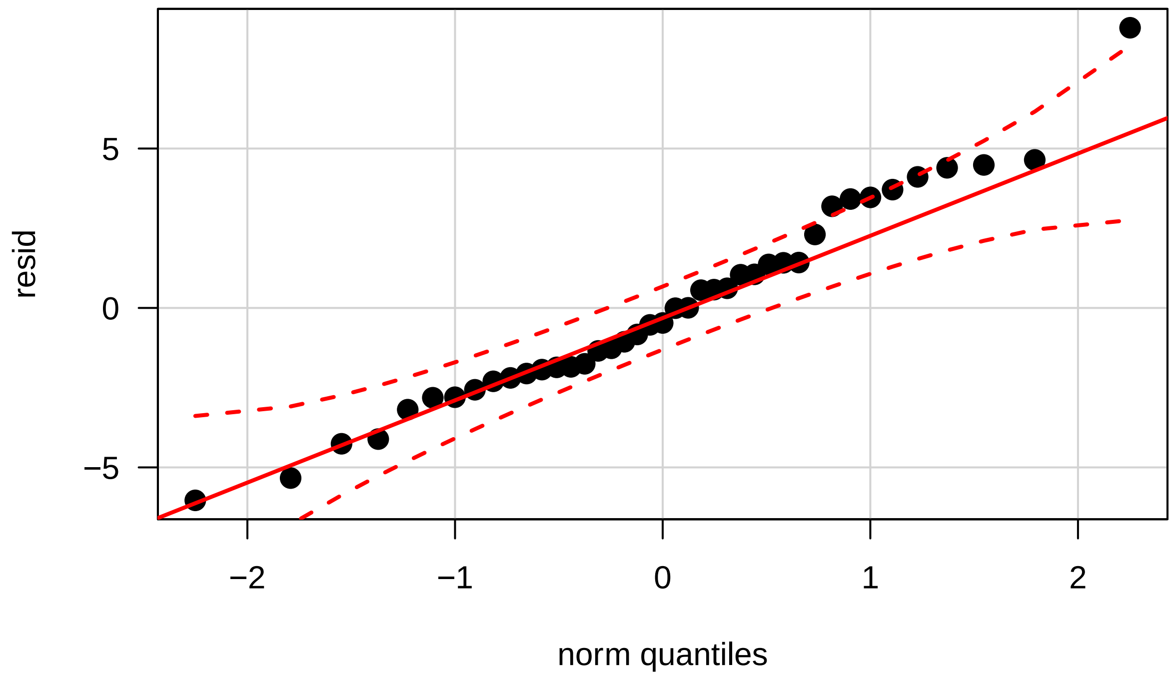
- Normalidade dos resíduos

```
means <- aggregate(BMI ~ Course, data = Data, mean)
Means <- numeric()
Means[posGrad] <- means$BMI[1]
Means[posPosGrad] <- means$BMI[2]
Data <- cbind(Data, Means)
resid <- Data$BMI - Data$Means
```

```
require("car")
```

```
## Loading required package: car
```

```
par(mfrow = c(1, 1))
 qqPlot(resid,
        pch=16,
        cex=1.5,
        las=1)
```



No gráfico plotado pelo teste Shapiro-Wilk, nota-se que a grande maioria dos dados está localizada próxima à reta, entre as curvas pontilhadas, indicando que estávamos certos ao assumir que os dados poderiam ser modelados por uma distribuição normal. Apenas alguns poucos valores estão fora da faixa definida pelas curvas, sendo que somente um deles está consideravelmente longe da reta, então podemos desprezar seus efeitos neste caso.

- Independência dos resíduos

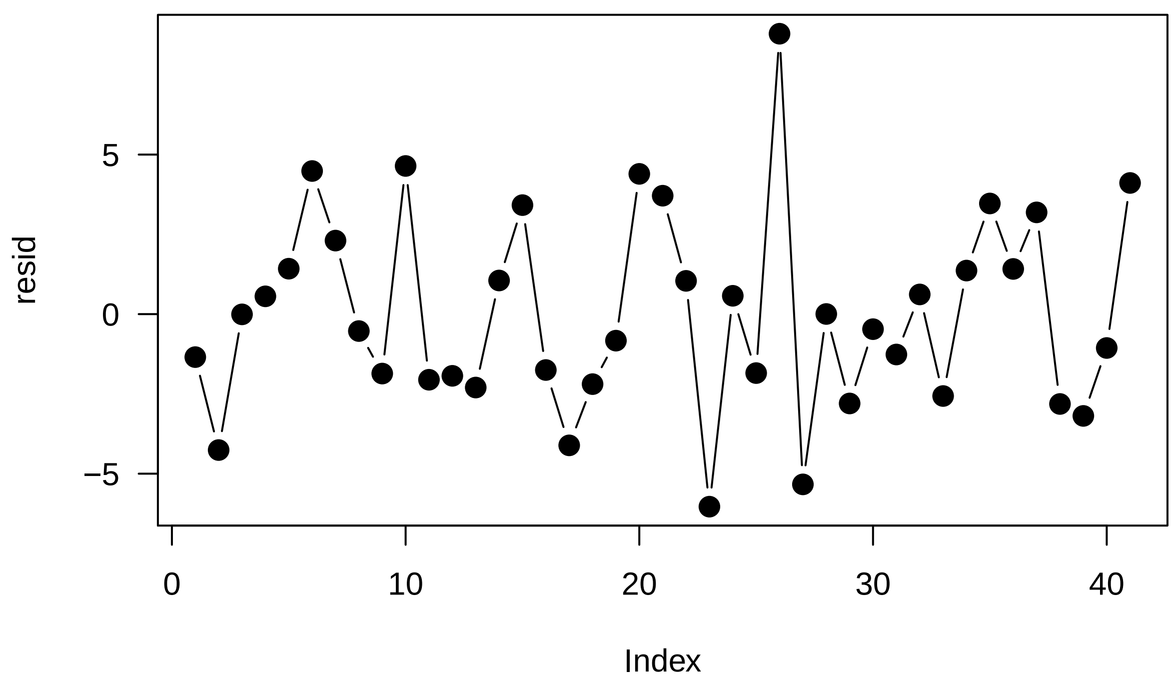
```
require(lmtest)
```

```
## Loading required package: lmtest

## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
with(Data,
     dwtest(BMI~Course))
```

```
##
##  Durbin-Watson test
##
## data:  BMI ~ Course
## DW = 2.0644, p-value = 0.5197
## alternative hypothesis: true autocorrelation is greater than 0
```

```r
plot(resid,
     pch=16,
     cex=1.5,
     type="b",
     las=1)
```



Antes de fazer o experimento, já não tínhamos nenhuma razão em supor que existe autocorrelação serial, isso só seria verdade, se o professor tivesse coletado os dados de forma diferente.

Se os mesmos tivessem sido coletados em forma de pergunta, de modo sequencial e em voz alta, era esperado que o que cada um dissesse, seria influenciado pelo que cada um pensa que deveria ser o mais adequado, por exemplo, se um determinado aluno achar que está muito magro, o mesmo vai tentar dizer um valor de peso um pouco maior do que realmente é, por causa da expectativa de grupo.

Mas como o professor pediu que todos escrevessem no papel seu peso e altura, de forma anônima, ele tentou bloquear essa expectativa. Ao fazer isso, a mesma foi diminuída. Se a coleta dos dados fosse de forma mais precisa, utilizando uma balança e uma fita métrica, o efeito dessa expectativa diminuiria mais ainda, a ponto de não existir, ou seja, não teríamos que nos preocupar com isso e nem com o viés. Então, o teste desta seção foi feito mais para ilustrar o que já sabíamos, que não temos autocorrelação e, consequentemente, os dados são independentes.

Executando o teste de Durbin-Watson, obteve-se um valor-p de 0.5197, aproximadamente 10 vezes maior do que o nível de significância, ou seja, não temos nenhum indício de uma autocorrrelação diferente de zero para nenhum dos dois lados. Portanto, para este caso, presumimos corretamente a independência dos dados.

Conclusões e Recomendações

Concluímos com este experimento que não temos evidência de que a média dos BMIs dos alunos do curso de Engenharia de Sistemas seja igual à média dos BMIs dos estudantes da PPGEE. Não só não temos evidência disso, como os dados parecem sugerir (pelo que foi visto e analisado nas seções anteriores com mais detalhes), dando um nível de confiança de 95%, que a média dos BMIs dos estudantes de Engenharia de Sistemas é menor do que a dos alunos do PPGEE.

Levando em consideração o que os dados parecem mostrar, podemos levantar algumas informações sobre o efeito do curso na forma física dos alunos, procurando responder à questão de interesse deste experimento.

Os alunos de Engenharia de Sistemas aparentarem ter um BMI menor em média, pode ter a ver com o estilo de vida dos mesmos não permitir que eles tenham horários regulares para se alimentar, isto é, não têm tempo para comer. Geralmente os alunos da pós são mais velhos do que os alunos da graduação, alguns até muito mais velhos, então o fator "idade" pode influenciar, uma vez que pessoas na faixa de 35 a 40 anos tendem a adquirir mais peso do que pessoas da faixa de 20 a 30 anos. A ansiedade que pode se perpetuar durante um curso de pós graduação associada a esse metabolismo mais lento também pode estar relacionada com esta média um pouco mais alta, porque a ansiedade costuma fazer algumas pessoas comerem mais.

Algumas recomendações que poderiam ser seguidas pelos alunos, baseadas nas conclusões anteriores seriam: uma melhor organização do tempo, procurar comer de três em três horas um sanduíche natural e uma fruta, por exemplo; praticar exercícios para aliviar os efeitos de ansiedade, assim como para facilitar a queima de gordura, e tomar vitaminas de forma a complementar a alimentação.

Possíveis formas de melhorar o experimento

Para melhorar a precisão das conclusões do experimento, seria interessante, aumentar o tamanho do espaço amostral de cada população o máximo possível, desse modo, diminuimos o intervalo de confiança tanto do teste T, quanto do teste F, o que leva a conclusões mais confiáveis ainda, porque uma vez que temos o resultado do teste F mais preciso, mostrando de forma ainda mais clara se as variâncias são iguais, teremos um teste T que considera melhor a realidade e também apresentará um resultado mais próximo da situação real.

Outra forma de melhorar este experimento seria coletar os dados dos estudantes de cada curso através da escolha de alguns alunos selecionados, ou mesmo o professor, que usarão instrumentos como uma balança e uma fita métrica para que os dados tenham um nível de exatidão maior.