CIÊNCIA e NATURA

UFSM

Ci. e Nat., Santa Maria v.42
Commemorative Edition: Statistic, e50, 2020
DOI: 10.5902/2179460X40424
ISSN 2179-460X

Statistics

Submissão 11/10/19 aprovação 06/03/19 publicação 30/11/20

# Análise do rendimento de estudantes da área da saúde frente a uma metodologia ativa de ensino para testes de hipóteses

Score analysis for health area student's of an active teaching methodology for hypothesis testing content

**Rodrigo Fioravanti Pereira[I] ,Fernando de Jesus Moreira Junior[II], Ileana Maria Greca[III], Jesus Meneses Villagra[IV]**

## RESUMO

Apresenta-se uma análise do rendimento acerca do conteúdo de testes de hipóteses para a média populacional de alunos da área da saúde que participaram da metodologia ativa de ensino chamada Problem Based Learning adaptada ao contexto de ensino e amparada pela Teoria dos Campos Conceituais. Este rendimento foi comparado com o de alunos que tiveram aulas tradicionais por meio de um instrumento criado para este fim e procurou replicar questões típicas deste conteúdo. O instrumento foi aplicado a 99 estudantes de uma instituição de ensino superior do sul do Brasil, pelo ambiente Moodle, presencialmente com uso dos softwares Excel e Bioestat. Os rendimentos foram analisados pela Análise Exploratória dos Dados, a Teoria Clássica dos Testes e Testes de Hipóteses e foram maiores para os alunos que não participaram do PBL. Possíveis causas são a melhor adequação do instrumento ao sistema tradicional de ensino, limitações contextuais para a aplicação da metodologia ativa sem descaracteriza-la como a baixa carga horária e a resistência a novas metodologias. Conclui-se que a diferença de rendimentos não é grande a ponto de desestimular o esforço de adaptar metodologias ativas de ensino à Bioestatística, dadas suas outras vantagens para além do rendimento.

**Palavras-chave:** Ensino de Estatística; Análise de Testes; Metodologia Ativa de Ensino.

## ABSTRACT

An analysis of the performance of the hypothesis testing content for the population average of health students who participated in the active teaching methodology called Problem Based Learning adapted to the teaching context and supported by the Conceptual Fields Theory is presented. This performance was compared with that of students who took traditional classes through an instrument created for this purpose that sought to replicate typical questions of this content. The instrument was applied to 99 students from a higher education institution in southern Brazil, by the Moodle environment, in person and using Excel and Bioestat software. Scores were analyzed by Exploratory Data Analysis, the Classical Test Theory and Hypothesis Tests and were higher for students who did not participate in the PBL, but the differences were not significant at $\alpha = 5\%$. Possible causes are the better adaptation of the instrument to the traditional education system, contextual limitations to the application of the active methodology such as low workload and resistance to new methodologies. It is concluded that the difference in scores is not large enough to discourage the effort to adapt active methodologies to biostatistics, given its advantages other than yield.

**Keywords:** Statistics Teaching; Test Analysis; Active Teaching Methodology.

[I] Universidade Franciscana, Santa Maria, Brasil. E-mail: prof.rodrigopereira@gmail.com.
[II] Universidade Federal de Santa Maria, Santa Maria, Brasil. E-mail: fmjunior@smail.ufsm.br.
[III] Universidad de Burgos, Burgos, Espanha. E-mail: imgreca@ubu.es.
[IV] Universidad de Burgos, Burgos, Espanha. E-mail: meneses@ubu.es.

## 1 INTRODUÇÃO

Em março de 2019 a revista Nature publicou um artigo intitulado *"Scientists rise up against statistical significance"*, que termina de forma contundente,

> Nosso chamado para retirar a significância estatística e usar intervalos de confiança como intervalos de compatibilidade não é uma panaceia. Embora elimine muitas práticas ruins, poderia muito bem introduzir novas. Assim, monitorar a literatura quanto a abusos estatísticos deve ser uma prioridade permanente para a comunidade científica. Porém, a erradicação da categorização ajudará a interromper reivindicações superconfiantes, declarações injustificadas de 'sem diferença' e declarações absurdas sobre 'falha de replicação' quando os resultados dos estudos originais e de replicação forem altamente compatíveis. O uso indevido da significância estatística causou muitos danos à comunidade científica e àqueles que dependem de aconselhamento científico. P valores, intervalos e outras medidas estatísticas têm seu lugar, mas é hora da significância estatística desaparecer. (AMRHEIN; GREENLAND; MCSHANE, 2019, p. 307)

O artigo expõe problemas referentes à aplicação e interpretação da estatística. Quanto ao p-valor, trabalhos como os de Cohen (2011) e Panagiotakos (2008) ressaltam problemas com sua má interpretação. O problema ganhou tamanha proporção que a *American Statistics Assosiation* (ASA), preocupada com o mau uso do p-valor, lançou o *"The ASA's Statement on p-Values: Context, Process, and Purpose"* (WASSERSTEIN; LAZAR, 2016), para tentar regular o seu uso.

Neste cenário, a formação básica em estatística precisa de atenção desde o primeiro contato com a disciplina, o que ocorre, normalmente, nos primeiros semestres da graduação. Neste período da formação, não é raro que o sentimento do aluno frente a estatística seja negativo, principalmente ocasionado pela relação com a matemática e/ou do sentimento de pouca aplicabilidade na área, entre outras razões, discutidas, para o caso da área da saúde, em Pereira, Dufranc e Villagra (2019). Particularmente no que se refere aos testes estatísticos, Post e Van Duijn (2014) afirmam que é preciso trabalhar para o entendimento dos princípios por trás dos testes de hipóteses, assim, mudanças no ensino dos testes de hipóteses precisam ser feitas.

No âmbito de uma pesquisa mais ampla e ainda em desenvolvimento, no campo da didática da estatística e que pretende contribuir para a melhora do quadro discutido acima, utilizou-se um instrumento, composto por dez itens fechados considerados típicos para o conteúdo de Testes de Hipóteses para a Média Populacional (THMP), com

o objetivo de medir o desempenho de alunos de graduação da área da saúde após experimentarem a metodologia ativa conhecida por *Problem Based Learning* (PBL), amparada pela Teoria dos Campos Conceituais (TCC) e modificada levando-se em conta um contexto próprio de ensino e comparar este desempenho com os alunos que tiveram aulas tradicionais, adicionalmente, compara-se os rendimentos entre os gêneros e cursos dos alunos. A TCC é uma teoria de aprendizagem desenvolvida por Gerárd Vergnaud e está melhor descrita em Moreira (2002).

O presente artigo trata da pesquisa parcial que descreve as características de uma determinada população quanto ao nível de rendimento em um instrumento do tipo questionário, logo, pela visão de Gil (2002), esta pesquisa classifica-se como descritiva. Para tal, usou-se a abordagem quantitativa da Análise Exploratória dos Dados (AED) e a Teoria Clássica dos Testes (TCT), entendidas como mostrado a seguir.

## 2 TÉCNICAS ESTATÍSTICAS PARA ANÁLISE DOS ITENS

Tukey (1977) pondera que a AED refere-se ao olhar para os dados e ver o que eles parecem dizer, concentrando-se em aritmética simples e figuras fáceis de produzir. É preciso levar em conta que as aparências são descrições parciais que servem para se extrair novos *insights*, não se preocupando com a confirmação.

Batanero, Estepa e Godino (1991) lembram que antes de Tukey (1977) introduzir a AED, a análise se baseava fundamentalmente nos cálculos, trazendo duas consequências: a diminuição da importância visual da representação dos dados e a equiparação da análise a um modelo comprobatório preestabelecido, reduzindo a análise a um teste de hipóteses que barra a extração de qualquer outra informação que se possa deduzir dos dados. Neste sentido, Tukey (1977) apresenta o seguinte princípio: “É importante entender o que você pode fazer antes de aprender a medir o quão bem você parece ter feito”.

Para Behrens (1997), cada pesquisa desenvolve sua própria AED, segundo suas necessidades. Considerando este trabalho, a AED contou com visualização de dados, tabelas de frequências e medidas resumo. Para além da AED, existe a TCT, que segundo Soares, Amorim e Silva (2018), é uma das vertentes da psicometria moderna, ela se preocupa em explicar o resultado final total (escore) do teste. Neste trabalho

apresenta-se a análise das medidas típicas encontradas na TCT para um instrumento (teste) específico, que são o índice de dificuldade, o índice de discriminação, coeficiente bisserial e o alpha de Cronbach.

O índice de dificuldade é dado pela proporção de respondentes que acertaram o item e mede o grau de dificuldade do item. Quanto maior for o índice de dificuldade, mais fácil é o item. O índice de discriminação é calculado por meio da diferença entre a proporção de acertos dos 27% do escores mais extremos (escores maiores menos escores menores) e mede a capacidade do item de diferenciar respondentes de maior e menor habilidade. Quanto maior for o resultado do índice, maior será sua discriminação. Já o coeficiente bisserial mede o grau de associação entre a proporção de acertos do item e o valor do escore na prova, ajudando a responder a seguinte questão: acertar um determinado item tem relação com escores altos na prova?

O coeficiente bisserial, pela notação de Borgatto e Andrade (2012), é dado por:

$$r_{bis} = \frac{M^+ - M^-}{S} \cdot \frac{p(1-p)}{h(p)}$$

Onde,

$M^+ \rightarrow$ média do escore para os que acertaram o item;

$M^- \rightarrow$ média do escore para os que erraram o item;

$S \rightarrow$ desvio padrão do escore de todos os respondentes;

$p \rightarrow$ proporção de acerto do item;

$h(p) \rightarrow$ valor da densidade da distribuição normal com média 0 e variância 1 no ponto em que a área da curva à esquerda deste ponto é igual a p.

Para o caso de não se poder presumir que os escores assumam uma distribuição normal, usa-se o coeficiente de correlação ponto-bisserial, dado a seguir:

$$r_{pbis} = \frac{M^+ - M^-}{S} \cdot \sqrt{\frac{p}{1-p}}$$

A consistência interna do instrumento será auferida pelo alpha de Cronbach (CRONBACH, 1951), que é dado por

$$\alpha = \left(\frac{k}{k-1}\right) \cdot \left(1 - \frac{\sum_{i=1}^{k} {S_i}^2}{{S_t}^2}\right)$$

Onde,

$k \rightarrow$ quantidade de itens do questionário;

$S_i^2 \rightarrow$ variância de cada item;

$S_t^2 \rightarrow$ variância total do escore.

O $\alpha$ de Cronbach mede a correlação entre as respostas do instrumento, que tem a mesma escala de medição, por meio da análise do perfil das respostas dadas pelos estudantes. Ele mede a fiabilidade do instrumento e quanto mais próximo de 1, mais fiável é o instrumento, entretanto, não há consenso sobre o valor de $\alpha$ a partir do qual diz-se que o instrumento é altamente fiável (MAROCO; GARCIA-MARQUES, 2006).

A estimação não paramétrica de densidade por núcleo (kernel) com função de núcleo gaussiana será utilizada no contexto da AED e trata-se da estimação do formato da distribuição populacional por meio de uma suavização do histograma amostral buscando uma aproximação da curva normal que permite uma análise visual da distribuição dos dados. (ESTATCAMP, 2019) e (ORANGE DATA MINING, 2019)

Ainda na AED, o gráfico cartesiano que relaciona o escore com a proporção de acertos por item, permite verificar a evolução da proporção de acertos em cada item a medida em que o rendimento dos estudantes aumenta. O esforço dedicado à AED cria o ambiente propício para os testes de hipóteses, no sentido de cumprir com seus pré-requisitos e corroborar com seus resultados. Assim, não se corre o risco de basear-se a análise estatística somente nos testes de hipóteses e resultados do tipo “há significância” ou “não há significância”, que empobrece o trabalho. Com estas ideias, passa-se a descrever a metodologia utilizada.

## 3 METODOLOGIA

Com o objetivo de analisar o rendimento (escore) de alunos da área da saúde que responderam a um instrumento fechado contendo itens típicos sobre os THMP, estratificados por gênero (feminino ou masculino), PBL (participou ou não participou) e curso (Biomedicina, Nutrição ou Psicologia), desenvolveu-se uma AED seguida da análise de medidas típicas da TCT e testes de hipóteses, descritas acima.

O instrumento utilizado possui dez questões típicas, de múltipla escolha, tal como utilizadas em livros e materiais didáticos da área. Entretanto, mesmo questões típicas podem variar bastante entre professores e instituições, logo, o instrumento levou em conta tanto questões teóricas quanto práticas e as realidades de duas instituições de ensino superior (IES).

Ele foi construído com o auxílio da Sigma Jr, empresa júnior do curso de Estatística da Universidade Federal de Santa Maria (UFSM) a qual gerou sua primeira versão consultando professores de estatística e bioestatística daquela universidade. Em seguida, o questionário passou por avaliação de outros quatro professores da Universidade Franciscana, gerando algumas alterações antes de sua versão final, com 10 questões de múltipla escolha.

Os dados foram coletados nos dois semestres de 2018 e no primeiro semestre de 2019, nas turmas o Professor Pesquisador foi ministrante da disciplina de Bioestatística, perfazendo as quantidades de alunos da tabela 1:

Tabela 1 – Quantidades de alunos por curso e participação no PBL

| | | Curso | | | |
|---|---|---|---|---|---|
| | | BMD | NUT | PSC | Total |
| PBL | Não | 12 | 0 | 26 | 38 |
| | Sim | 0 | 10 | 51 | 61 |
| Total | | 12 | 10 | 77 | 99 |

O instrumento foi aplicado a 99 estudantes de uma instituição de ensino superior da cidade de Santa Maria, RS, por meio do ambiente virtual de aprendizagem Moodle, de maneira presencial e individual, em laboratório de informática, durante duas horas/aula (100min) cada aplicação. Os estudantes tinham acesso à internet e eram familiarizados com os *softwares* Excel e Bioestat.

Os dados coletados foram exportados do Moodle para uma planilha Excel onde foram devidamente limpos, restando as colunas Identificador, Gênero, PBL, Curso, e a resposta certa (1) ou errada (0) em cada um dos 10 itens.

Ainda na planilha Excel foi calculado o índice de discriminação. A AED teve dois momentos distintos, um com foco no instrumento e no rendimento geral dos estudantes e o outro nos escores dentro das variáveis estratificadoras.

No primeiro momento lançou-se mão das medidas resumo dos escores proporcionadas pelo software IBM SPSS[1], que também gerou o histograma com a curva normal subjacente, com visualização apoiada pelo boxplot feito pelo software Orange Canvas[2]. Em seguida, procedeu-se ao cálculo das medidas da TCT, por meio do pacote 'ltm' do software R, que gera também o gráfico dos escores vs proporção de acertos no item.

No segundo momento, cada comparação dentro das variáveis estratificadoras contou com boxplots comparativos, seguindo para a estimação de núcleo gaussiano, criados no Orange Canvas. Adiante, tem-se a comparação das médias dos grupos por meio de um gráfico de colunas gerado no software Tableau[3] e testes de hipóteses.

## 4 RESULTADOS E DISCUSSÕES

O instrumento foi aplicado a 99 alunos, dos quais, 61 (61,6%) participaram da PBL, 76 (76,8%) eram do gênero feminino, 12 (12,1%) eram do curso de Biomedicina, 10 (10,1%) do curso de Nutrição e 77 (77,8%) de Psicologia.

O rendimento dos estudantes (escore) tem as seguintes medidas resumo:

Tabela 2 – Medidas resumo para a variável escore (gerado por SPSS)

| Escores | | Estatística |
|---|---|---|
| Média | | 6,56 |
| 95% Intervalo de Confiança para Média | Limite inferior | 6,16 |
| | Limite superior | 6,95 |
| 5% da média aparada | | 6,58 |
| Mediana | | 6,00 |
| Variância | | 3,841 |
| Desvio Padrão | | 1,960 |
| Mínimo | | 2 |
| Máximo | | 10 |
| Intervalo | | 8 |
| Amplitude interquartil | | 3 |
| Assimetria | | 0,106 |
| Curtose | | -0,360 |

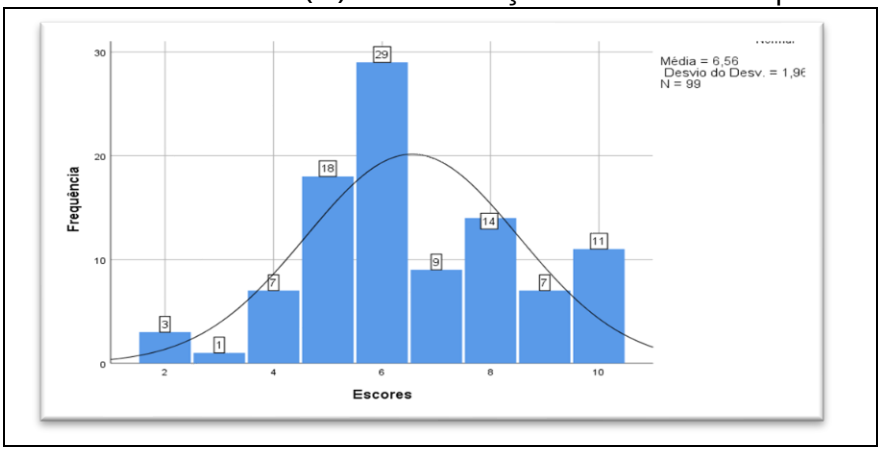
[1] v. 25
[2] v. 3.21.0
[3] v. 2019.2.3

De forma geral, o rendimento dos estudantes é considerado razoável se levarmos em conta a média, que é suficiente para aprovação na instituição onde a pesquisa está sendo desenvolvida e o desvio padrão que determina um coeficiente de variação de 30% que é alto, porém comum para esta instituição[4].

### 4.1 Análise do Instrumento

A distribuição dos escores dos 99 alunos, com a respectiva curva normal, é mostrada na figura 1:

Figura 1 – Distribuição de frequências de escore, média, desvio padrão, quantidade de elementos (N) e distribuição normal correspondente (gerado no SPSS)

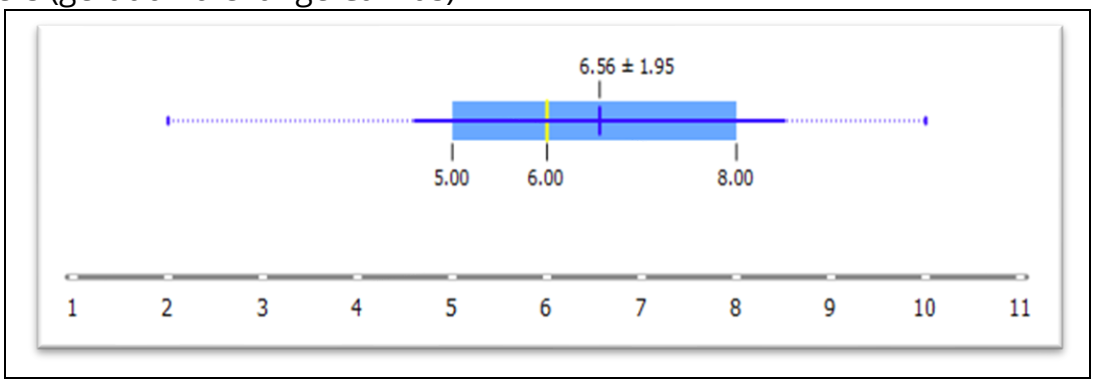



Percebe-se, pela figura 1, que os itens 3 e 7 possuem frequências bem menores do que as esperadas para a distribuição normal subjacente, enquanto que os itens 6 e 10 possuem frequências bem maiores. Neste caso, a imagem parece mostrar que não há adequação da distribuição dos escores com a curva normal correspondente.

A figura 2 apresenta o boxplot dos escores, com os quartis, a média (6,56) e o desvio padrão (1,95), o escore mínimo (2) e o máximo (10) e indica uma assimetria à direita desta distribuição se considerarmos que a mediana (6,0) é menor do que a média, depondo contra a normalidade dos dados.

[4] Embora não haja estudo que permita a comparação destes resultados, acredita-se, pela experiência prévia, que estes resultados sejam, de fato, típicos desta instituição.

Figura 2 – Boxplot para a variável escore com média, desvio padrão, quartis e whiskers (gerado no Orange Canvas)




O teste de aderência de qui-quadrado para normalidade trouxe um valor calculado igual a 16,03, contra um valor tabelado de 12,59, a um nível de 5% de significância, levando à rejeição da hipótese de normalidade dos dados para a variável escore. O teste também indicou que os maiores valores calculados ocorreram nos itens Q6, Q7 e Q10, corroborando com a inspeção visual anterior que indicava estes itens com forte descompasso com a curva normal subjacente, a menos do item Q3, que o teste qui-quadrado não identificou como discrepante à curva normal. Também o teste de normalidade de Shapiro-Wilk, ao mesmo nível de significância de 5%, não permitiu pressupor que os dados provenham de população normalmente distribuída ($p < 0,000$).

Por todos estes motivos contra a pressuposição de normalidade dos dados de escore, a correlação bisserial será substituída pela correlação ponto-bisserial, gerando a tabela 1, abaixo.

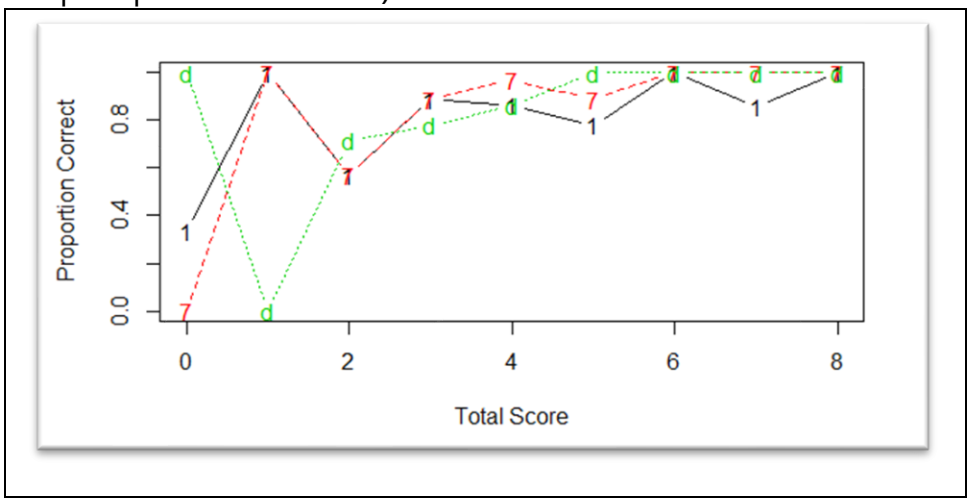
Tabela 3 - Medidas da TCT por item (gerado por Excel e R, pacote ltm)

| Item | Índice de dificuldade | Índice de discriminação | Coef. Ponto Bisserial | Alfa de Cronbach (geral: 0,5437) Excluindo os itens |
|---|---|---|---|---|
| Q1 | 86% | 0,21 | 0,28 | 0,54 |
| Q2 | 40% | 0,35 | 0,41 | 0,54 |
| Q3 | 62% | 0,62 | 0,53 | 0,49 |
| Q4 | 45% | 0,54 | 0,52 | 0,50 |
| Q5 | 51% | 0,50 | 0,48 | 0,51 |
| Q6 | 71% | 0,46 | 0,41 | 0,53 |
| Q7 | 90% | 0,28 | 0,42 | 0,51 |
| Q8 | 73% | 0,63 | 0,47 | 0,51 |
| Q9 | 49% | 0,70 | 0,57 | 0,48 |
| Q10 | 89% | 0,24 | 0,28 | 0,54 |

O item Q2 teve a menor proporção de acertos (40%), mesmo assim, considera-se o item acessível aos estudantes, visto que não trouxe um resultado do tipo "ninguém ou pouquíssimos alunos acertaram o item". Ao mesmo tempo, os itens Q1, Q7 e Q10 alcançaram mais de 86% de acertos, sendo considerados os muito fáceis.

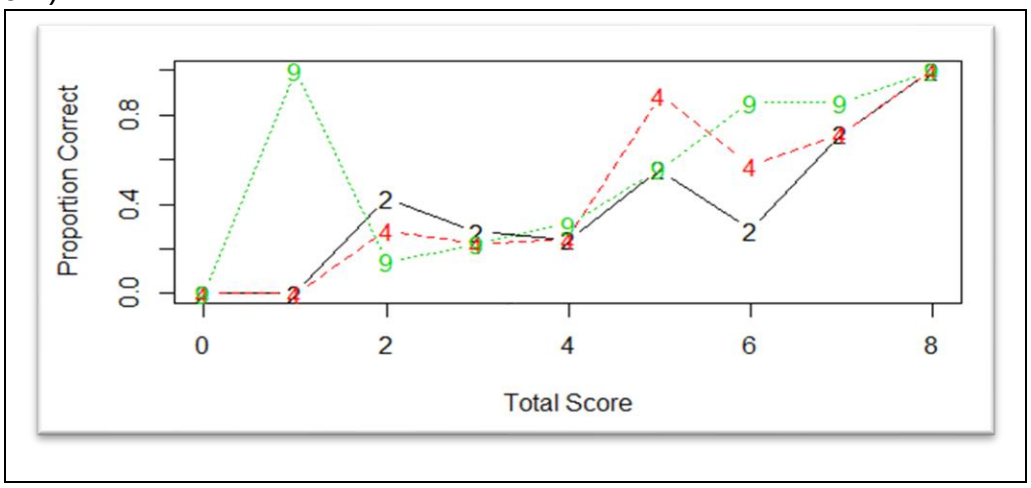
A figura 3 mostra o gráfico com o escore total em relação com a proporção de acertos dos itens Q1, Q7 e Q10, revelando que mesmo os estudantes com baixos escores já alcançam altas proporções de acertos para estes itens, indicando que eles não são capazes de distinguir candidatos com alto rendimento dos com baixo rendimento.

Figura 3 – Gráfico do escore versus a proporção de acertos por item. O número 1 refere-se ao item Q1, o 2 ao Q2 e a letra d indica o item Q10. O pacote do R fez o seguinte ajuste: Escore = Total Score + 2, pois nenhum estudante obteve escores 0 ou 1 (gerado pelo pacote 'ltm' do R)




Por outro lado, os itens Q2, Q4 e Q9 (menores índices de dificuldades, por isto, mais difíceis) só tiveram níveis altos de acertos para alunos com um rendimento maior do que sete (figura 4), fazendo parecer que os itens são capazes de separar os candidatos com alto rendimento dos demais. Resultado reforçado pelos índices de discriminação altos dos itens Q4 e Q9 (0,54 e 0,70, respectivamente), ao contrário do item Q2 (0,35). Também o coeficiente ponto bisserial de Q4 e Q9 foram altos (0,52 e 0,57), ao contrário do item Q2 (0,41). Assim, parece que Q2, embora seja o mais difícil, acerta-lo não indica necessariamente bom rendimento no instrumento, enfraquecendo a capacidade do item de separar baixos rendimentos de altos rendimentos, ao contrário dos itens Q4 e Q9, também considerados difíceis.

Figura 4 – Gráfico do escore versus a proporção de acertos por item. O número2 refere-se ao item Q2, o 4 ao Q4 e o 9 ao Q9. O pacote do R fez o seguinte ajuste: Escore = Total Score + 2, pois nenhum estudante obteve escores 0 ou 1 (gerado pelo pacote 'ltm' do R)

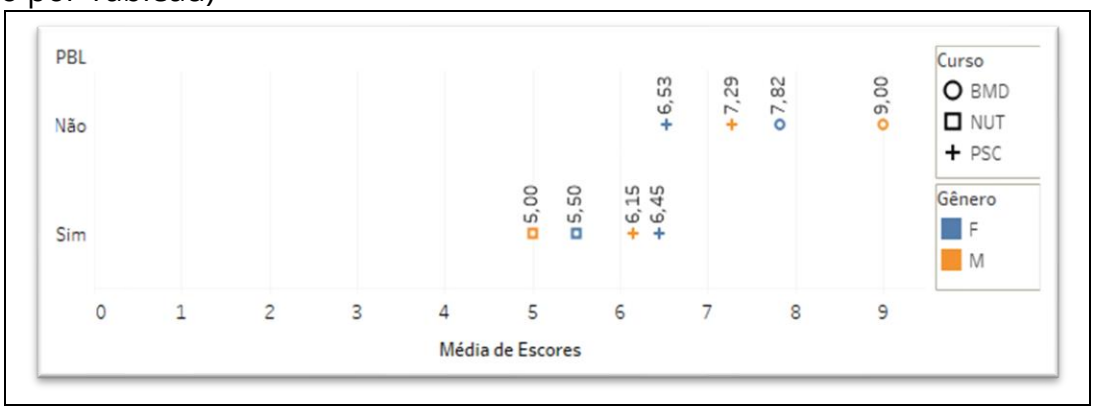



O α de Cronbach geral (0,5437) não é alto, mesmo que o valor 0,6 seja considerável aceitável para alguns cenários de pesquisa, levando-se em conta que instrumentos grandes tendem a ter valores de α maiores (MAROCO; GARCIA-MARQUES, 2006), o que não é o caso do instrumento aqui analisado. A exclusão do item também não gerou uma elevação na confiabilidade, sendo desnecessária a redução do número de itens do instrumento.

## 4.2 Análise dos escores com estratificações

Feitas as ponderações acerca do instrumento, passa-se a comparar os resultados dos escores por gênero, curso e participação no PBL, primeiramente, apresentando um resumo das médias considerando as variáveis estratificadoras, na figura 5.

Figura 5 – Médias dos escores separadas por participação no PBL, curso e gênero (gerado por Tableau)



As médias mais altas foram do curso de Biomedicina (nenhum dos 12 alunos participou do PBL), de ambos os gêneros, enquanto que as menores médias foram do curso de Nutrição (todos os 10 alunos participaram do PBL). O curso de Psicologia obteve médias intermediárias (51 alunos participaram e 26 não participaram do PBL). Assim, as médias mais altas foram dos que não participaram do PBL.

### 4.2.1 Estratificação dos escores por gênero

As médias dos escores para o gênero feminino (F) e masculino (M) foram muito parecidas, mas suas dispersões não (2,1 e 1,31 respectivamente), como se vê nas figuras 6 e 7.

Figura 6 – Boxplots de escore separado por gênero (M: masculino e F: feminino) (gerado por Orange Canvas)

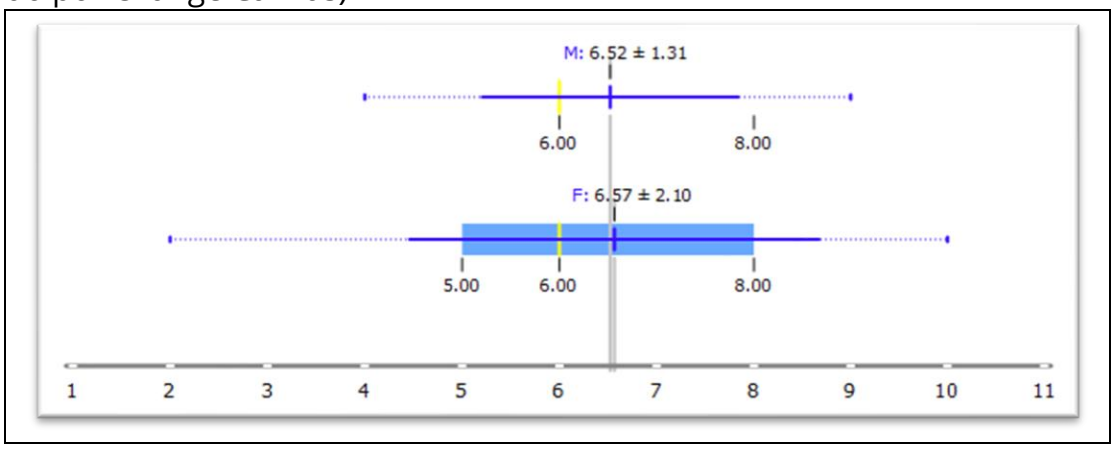
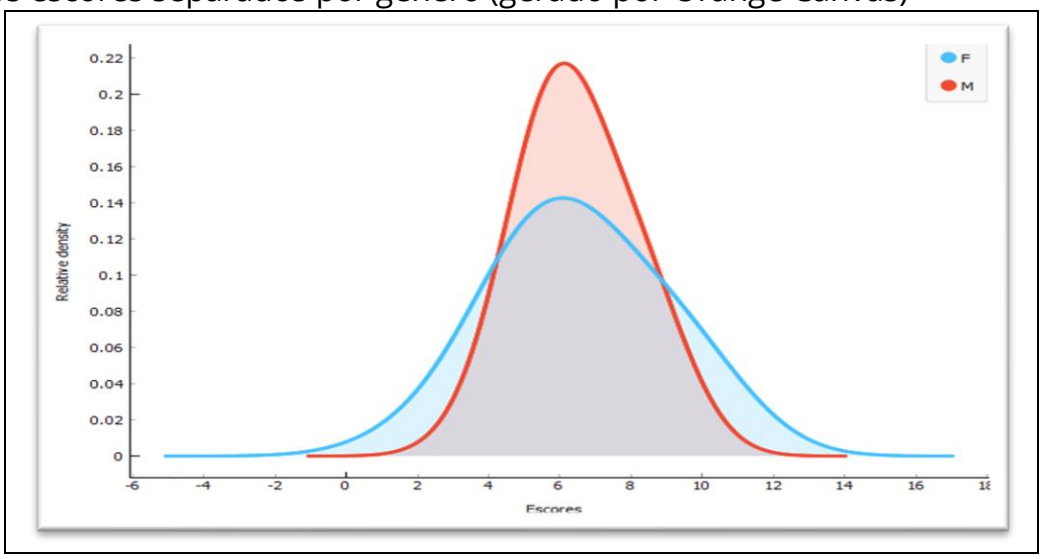
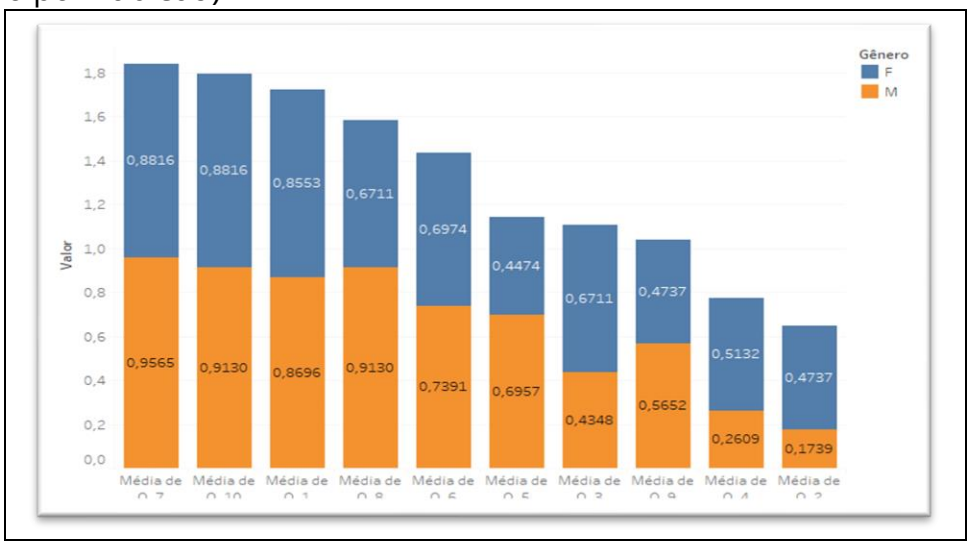



Figura 7 – Curva de estimação da densidade de probabilidade núcleo gaussiano para os escores separados por gênero (gerado por Orange Canvas)



Pela figura 8, vê-se que em sete itens (Q1, Q5, Q6, Q7, Q8, Q9 e Q10), a média de masculino foi maior do que a de feminino. Em seguida, procura-se saber se esta diferença é significativa.

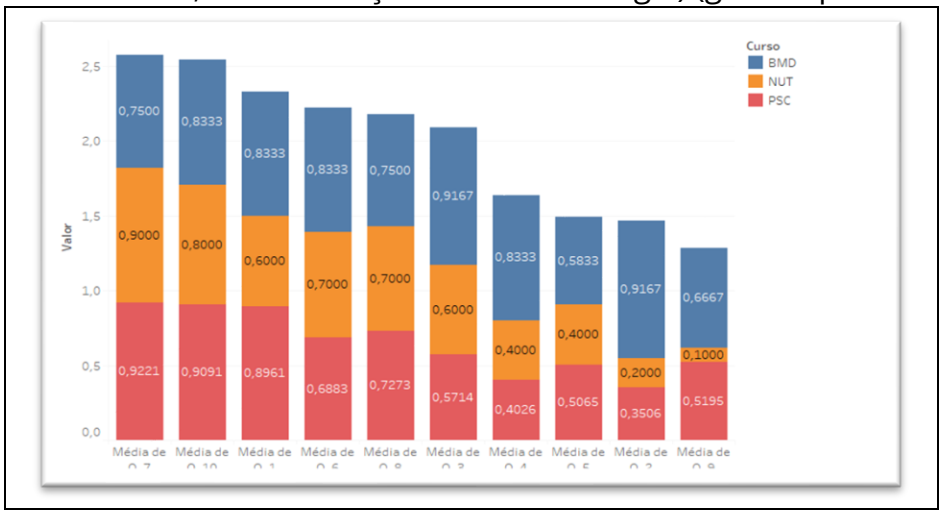
Figura 8 – Distribuição dos escores e média dos escores separados por gênero (gerado por Tableau)




O teste de normalidade de Shapiro-Wilk indicou que é preciso comparar os escores dos diferentes gêneros por um teste não paramétrico (p = 0,001). Utilizou-se o teste U de Mann-Whitney para amostras independentes que corroborou com a hipótese de não haver diferença significativa entre as populações da variável escore separada pelo gênero (p = 0,039), a um nível de significância de 5%, corroborando com a análise visual.

### 4.2.2 Estratificação dos escores por curso

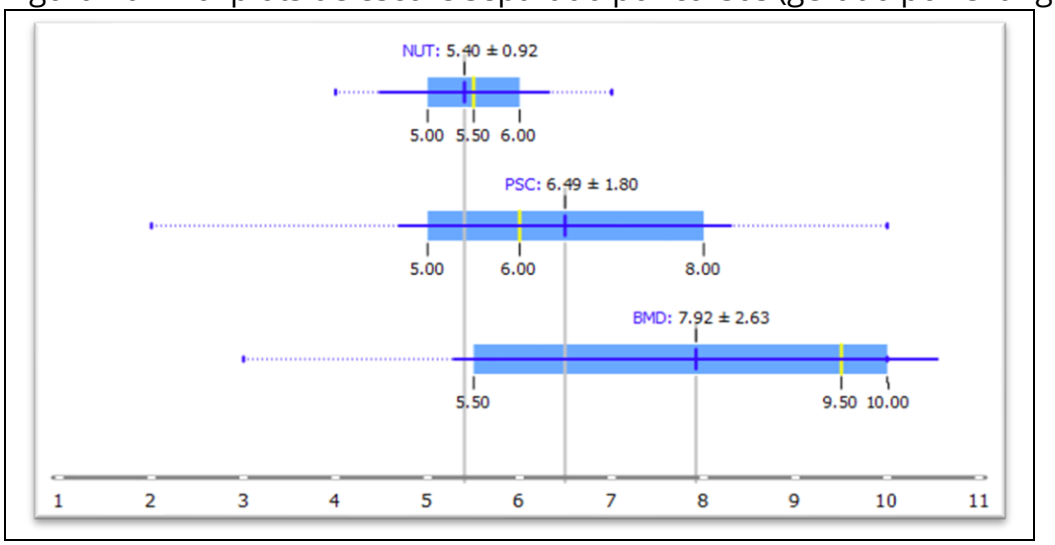
Figura 9 – Distribuição dos escores e média dos escores separados por curso (BMD: Biomedicina; NUT: Nutrição e PSC: Psicologia) (gerado por Tableau)



Alguns itens geraram escores muito diferentes entre os três cursos. O item Q.4 teve um rendimento médio no curso de Biomedicina praticamente o dobro que os outros dois. O item Q2 obteve valores ainda mais extremos do que o Q4.

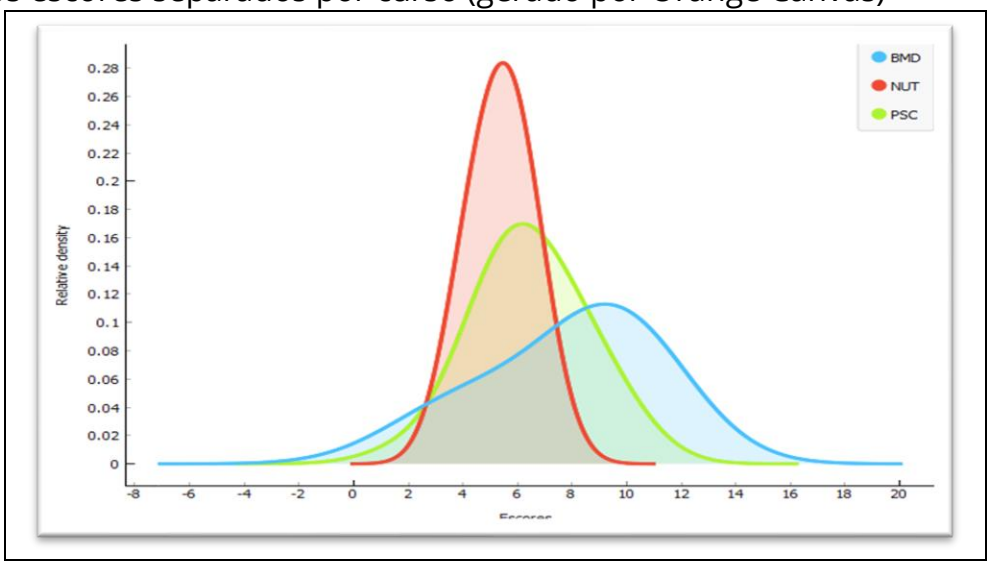
O curso de Biomedicina obteve rendimento médio superior em quase todos os itens, exceto nos itens Q7, Q10 e Q1. O rendimento da Nutrição nos itens Q.9 e Q.2 foi bastante inferior do que os demais, como se vê na figura 9.

Figura 10 – Boxplots de escore separado por cursos (gerado por Orange Canvas)




Os boxplots da figura 10 apresentam grande distanciamento entre as médias dos escores de cada curso, porém não há uma separação total de seus intervalos interquartílicos, indicando que, em sua variabilidade, as notas não apresentam uma diferença que distingue o rendimento dos cursos.

Figura 11 – Curva de estimação da densidade de probabilidade núcleo gaussiano para os escores separados por curso (gerado por Orange Canvas)



O curso de nutrição mostrou, pela análise da figura 11, um escore médio mais baixo do que os demais e suas notas não variaram tanto quanto os outros. O curso de Biomedicina obteve uma média e variação maior nas notas.

O teste de normalidade de Shapiro-Wilk aponta que a variável escore provém de população normalmente distribuída somente no curso de Nutrição (p = 0,245). Neste caso, optou-se por uma comparação entre os escores dos cursos por meio do teste de Kruskal-Wallis para amostras independentes que indicou que não há diferença significativa entre os escores dos cursos (p = 0,06), embora as amostras indiquem que o escore médio do curso de Biomedicina supere a do curso de Psicologia e esse supera o escore médio do curso de Nutrição.

### 4.2.3 Estratificação dos escores por participação na PBL

Figura 12 – Boxplots de escore separado por participação no PBL (gerado por Orange Canvas)

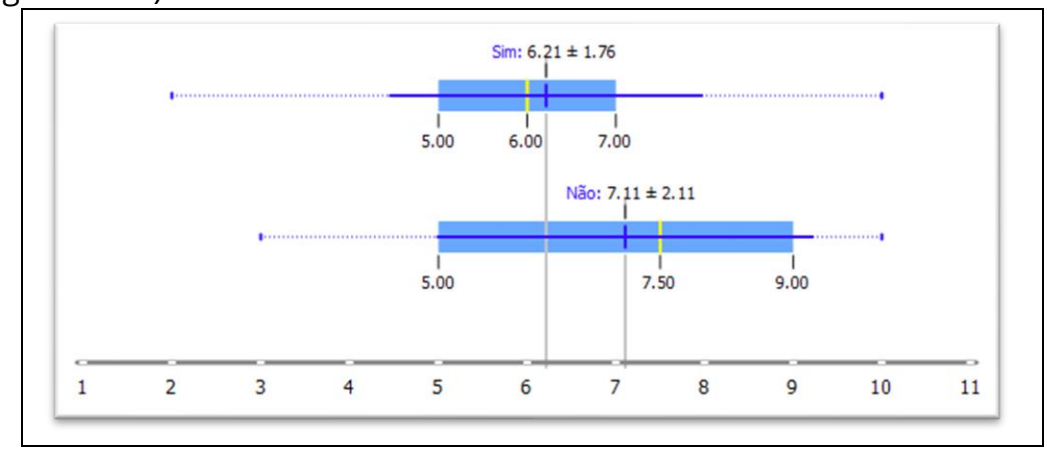
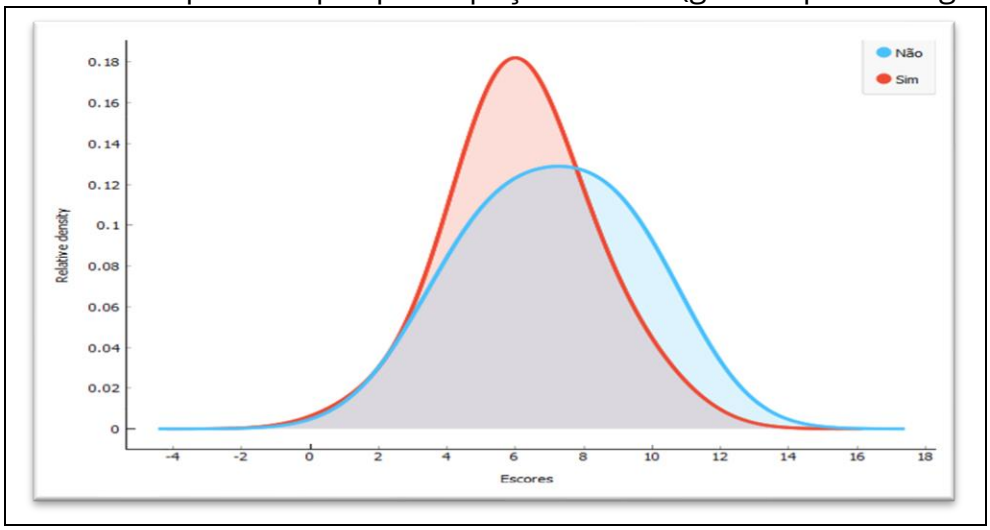
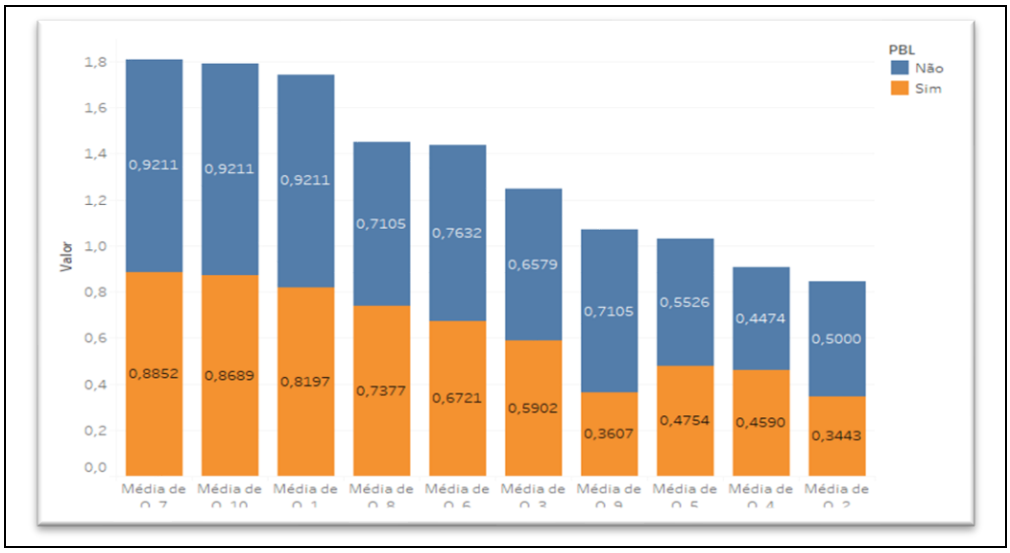



Figura 13 – Curva de estimação da densidade de probabilidade núcleo gaussiano para os escores separados por participação no PBL (gerado por Orange Canvas)



Pela observação das figuras 12 e 13, a média dos escores de quem participou da PBL foi menor do que os demais, mas houve maior dispersão dos escores dos que não participaram.

Figura 14 – Distribuição dos escores e média dos escores separados por participação no PBL (gerado por Tableau)




Pela figura 14, percebe-se que a média dos escores de quem não participou da PBL só não foi maior do que a dos que participaram nos itens Q8 e Q4 e mesmo nestes itens, a diferença foi pequena. Isto aponta para uma consistência de melhor rendimento dos alunos “Não” para os alunos “Sim”.

O teste de normalidade de Shapiro-Wilk recomenda que os escores não provêm de população normalmente distribuída quando se considera a participação ou não no PBL (Sim: p = 0,02; Não: p =0,008), indicando o uso de teste não paramétrico para a comparação dos dois grupos. Considerando o teste U de Mann-Whitney para amostras independentes, não rejeitamos a hipótese de igualdade entre os escores de quem participou do PBL e dos que não participaram da metodologia (p = 0,059).

Não há diferença significativa entre as médias dos escores quando estratificados pela variável PBL, embora as amostras indiquem que o escore médio dos que não participaram da PBL supere a dos que participaram.

## 5 CONCLUSÕES

O instrumento criado procura reproduzir questões de uma prova típica sobre o conteúdo de THMP, comumente aplicada em sala de aula. Esta forma de construção não gerou uma fiabilidade (alfa de Cronbach) alta, fazendo pensar que as atividades "livrescas", mesmo quando consideram itens práticos e teóricos, não possuem grande fiabilidade para a avaliação do conhecimento dos THMP.

A metodologia ativa PBL, no contexto desta pesquisa, ateve-se ao processo investigativo inerente ao fazer estatístico, o que diminui o foco do aluno na prova. O menor rendimento do PBL, aqui apresentado, pode dever-se também a este fato.

Por outro lado, as diferenças de rendimento encontradas não permitem discartar sumariamente o PBL. Os rendimentos levemente menores reforçam a ideia de que o esforço da adaptação do PBL para um contexto historicamente desenvolvido para adequar-se ao modelo tradicional pode gerar resultados, no mínimo, semelhantes. Outros ganhos esperados desta metodologia ativa de ensino são bem vindos, como a inserção do estudante em ambiente de pesquisa, o trabalho cooperativo, o protagonismo do aluno, a construção colaborativa do conhecimento, o autodidatismo entre outras.

Também é preciso considerar que as mudanças no PBL, no âmbito desta pesquisa, ainda estão se estabelecendo e, além disto, até mesmo um professor experiente sente dificuldades com as metodologias ativas tanto quanto os próprios alunos, já versados no modelo tradicional, no qual estabeleceram sua zona de conforto.

**REFERÊNCIAS**

AMRHEIN, V.; GREENLAND, S.; MCSHANE, B. Scientists rise up against statistical significance. **Nature**, v. 567, n. 7748, p. 305–307, 20 mar. 2019.

BATANERO, C.; ESTEPA, A.; GODINO, J. D. **Análisis Exploratorio de Datos: Sus Posibilidades en la Enseñanza Secundaria**. Suma, v. 9, p. 25–31, 1991.

BEHRENS, J. T. Principles and Procedures of Exploratory Data Analysis. **Psychological Methods**, v. 2, n. 2, p. 131–160, 1997.

BORGATTO, A. F.; ANDRADE, D. F. DE. Análise clássica de testes com diferentes graus de dificuldade. **Estudos em Avaliação Educacional**, v. 23, n. 52, p. 146, 30 ago. 2012.

COHEN, H. W. P Values: Use and Misuse in Medical Literature. **American Journal of Hypertension**, v. 24, n. 1, p. 18–23, 1 jan. 2011.

CRONBACH, L. J. **Coefficient Alpha And The Internal Structure Of Tests**. PSYCHOMETRIKA, v. 16, n. 3, p. 297–334, 1951.

ESTATCAMP. **Estimação não paramétrica de densidades: método do núcleo - Análise de Capacidade**. Disponível em: http://www.portalaction.com.br/analise-de-capacidade/431-estimacao-nao-parametrica-de-densidades-metodo-do-nucleo. Acesso em: 25 set. 2019.

GIL, A. C. **Como Elaborar Projetos de Pesquisa/Antonio Carlos Gil**. 3a ed. São Paulo: Atlas, 2002.

MAROCO, J.; GARCIA-MARQUES, T. Qual a fiabilidade do alfa de Cronbach? Questões antigas e soluções modernas? **Laboratório de Psicologia**, v. 4, n. 1, p. 65–90, 2006.

MOREIRA, M. A. A Teoria dos Campos Conceituais de Vergnaud, O Ensino de Ciências e a Pesquisa nesta Área. **Investigações em Ensino de Ciências**, n. 1, p. 7–29, 2002.

ORANGE DATA MINING. **Distributions**. Disponível em: https://docs.biolab.si//3/visual-programming/widgets/visualize/distributions.html. Acesso em: 26 set. 2019.

PANAGIOTAKOS, D. B. Value of p-value in biomedical research. **The open cardiovascular medicine journal**, v. 2, p. 97–9, 2008.

PEREIRA, R. F.; GRECA, I. M..; VILLAGRA, J. A. M. Caminhos do ensino de estatística para a área da saúde. **Revista Latinoamericana de Investigación en Matemática Educativa**, v. 22, n. 1, p. 67–96, 31 mar. 2019.

POST, W. J.; VAN DUIJN, M. A. J. Teaching Hypothesis Testing: a Necessary Challenge. **The 9th International Conference on Teaching Statistics**. 2014.

SOARES, J. A. R.; AMORIM, A. F.; SILVA, C. R. DA. Avaliação Educacional em Larga Escala e Algumas Considerações Sobre a TCT e a TRI. **Revista Ciências Exatas e Naturais**, v. 20, n. 1, p. 119–125, 2018.

TUKEY, J. W. **Exploratory Data Analysis**. [s.l.] Addison-Wesley, 1977.

WASSERSTEIN, R. L.; LAZAR, N. A. The ASA's Statement on p-Values: Context, Process, and Purpose. **The American Statistician**, 2016.