UNIVERSIDADE DE SÃO PAULO – ESCOLA DE ARTES, CIÊNCIAS E HUMANIDADES
SISTEMAS DE INFORMAÇÃO

APLICAÇÃO DE MÉTODOS ESTATÍSTICOS NA ANÁLISE DE TENDÊNCIAS DE AVALIAÇÃO ENTRE GRUPOS DE ALUNOS

SÃO PAULO
2021

DIMITRI OLIVEIRA DABKIEWICZ, nUSP 12543008
GUILHERME DE ABREU BARRETO, nUSP 12543033
KAYO RICARDO LIMA DA SILVA, nUSP 12543029
LINCOLN MORALE BRAGA, nUSP 12608705
LUCAS PANTA DE MOURA, nUSP 12608650

APLICAÇÃO DE MÉTODOS ESTATÍSTICOS NA ANÁLISE DE TENDÊNCIAS DE AVALIAÇÃO ENTRE GRUPOS DE ALUNOS

Relatório apresentado para aprovação na disciplina de Tratamento e Análise de Dados, do curso de Sistemas de Informação da Universidade de São Paulo.

Orientador: Prof⁰ Dr. Regis Rossi Alves Faria

RESUMO

O presente relatório descreve medidas resumo e coeficientes de associação ou correlação para então demonstrar a aplicação destes recursos na análise de um conjunto de dados composto por notas de duas turmas de uma dada disciplina. Baseados nestes, são construídos gráficos e tabelas para, por fim, inferir tendências comuns e distintas entre as turmas.

Palavras-chave:

Estatística, medidas de posição central, medidas de dispersão, coeficientes de associação ou correlação, aplicação prática

SUMÁRIO

1 INTRODUÇÃO.....6

2 ANÁLISES INICIAIS.....7

2.1 MEDIDAS-RESUMO.....7

2.1.1 Medidas de posição.....8

2.1.1.1 Média aritmética simples.....8

2.1.1.2 Mediana.....8

2.1.1.3 Quartil.....8

2.1.1.4 Moda.....9

2.1.2 Medidas de dispersão.....9

2.1.2.1 Variância.....9

2.1.2.2 Desvio padrão.....9

2.2 GRÁFICOS.....10

2.2.1 Histograma.....10

2.2.2 Boxplot.....11

2.2.3 Gráfico de barras.....12

2.2.4 Gráfico de setores.....13

3 ANÁLISES CONJUNTAS.....15

4 CONCLUSÃO.....17

5 REFERÊNCIAS.....18

APÊNDICE A – A DIVISÃO DE TAREFAS DO GRUPO

1 INTRODUÇÃO

Em diversas situações de análise quantitativa busca-se, para fazer sentido de uma série de dados (amostra), resumi-los apresentando um ou alguns poucos valores que sejam representativos do conjunto. Tais valores são denominados “medidas-resumo”. No mais, havendo mais de uma amostra, é possível, ainda, identificar relações entre as medidas-resumo destas calculando-se coeficientes de associação ou correlação. É o objetivo deste relatório definir as medidas-resumo e os coeficientes de associação para demonstrar o uso destas na realização de inferências sobre uma situação cotidiana: a avaliação do resultado de diferentes turmas de alunos em uma dada disciplina. Para tal, utiliza-se dos softwares Google Planilhas (2021), localizado para português brasileiro, mais a linguagem de programação Python (versão 3.9.6), com as bibliotecas Pandas e Seaborn (versões 1.2.5-1 e 0.11.1-1), disponibilizadas na plataforma Google Colaboratory (2021), para a confecção de gráficos, tabelas, e a realização dos cálculos estatísticos necessários. Por fim, pretende-se demonstrar a utilidade de medidas tais para a obtenção de um panorama geral de um número de informações coletadas.

2 ANÁLISES INICIAIS

Neste estudo consideramos duas turmas de alunos de uma mesma matéria, denominadas A e B, sendo a primeira composta por 53 alunos que realizaram um curso introdutório (Curso I) e a segunda composta por 39 alunos que realizaram um curso avançado (Curso II). As notas obtidas pelos alunos da turma A e B foram, respectivamente:

$$
\begin{array}{ccccccccc}
08,75 & 10,00 & 10,00 & 10,00 & 07,50 & 10,00 & 07,50 & 07,50 & 07,50 \\
07,50 & 02,50 & 07,50 & 10,00 & 07,50 & 07,50 & 07,50 & 10,00 & 10,00 \\
02,50 & 05,00 & 10,00 & 10,00 & 10,00 & 10,00 & 07,50 & 02,50 & 00,00 \\
07,50 & 10,00 & 10,00 & 10,00 & 02,50 & 05,00 & 10,00 & 10,00 & 07,50 \\
07,50 & 07,50 & 07,50 & 02,50 & 02,50 & 07,50 & 10,00 & 10,00 & 10,00 \\
10,00 & 10,00 & 07,50 & 05,00 & 07,50 & 10,00 & 10,00 & 10,00 &
\end{array}
\tag{1}
$$

$$
\begin{array}{ccccccccc}
08,00 & 07,60 & 07,20 & 09,20 & 08,40 & 09,60 & 04,67 & 10,00 & 08,00 \\
05,47 & 09,60 & 03,47 & 10,00 & 01,60 & 09,60 & 05,20 & 07,20 & 07,20 \\
07,20 & 09,20 & 09,60 & 09,20 & 07,20 & 09,20 & 09,60 & 07,20 & 09,60 \\
06,00 & 06,00 & 09,60 & 07,87 & 04,27 & 06,80 & 06,27 & 10,00 & 06,80 \\
08,67 & 09,20 & 07,47 & & & & & &
\end{array}
\tag{2}
$$

Tais dados foram listados em duas colunas de uma planilha eletrônica criada usando o Google Planilhas, ocupando, assim, os índices **(A2:A54)** e **(B2:B40)**, respectivamente.

2.1 MEDIDAS-RESUMO

O uso de medidas-resumo nos permite realizar inferências generalizadas sobre os valores encontrados em uma amostra. Por um lado, medidas de posição nos permitem formular índices representativos de uma tendência observada na amostra como um todo (BUSSAB; MORENTTIN, 2017, p. 35), por outro, medidas de dispersão indicam a ocorrência de valores mais ou menos discrepantes com relação a esta tendência geral (Ibid., p.38).

As medidas de posição utilizadas em nosso estudo para a elaboração de gráficos e tabelas foram as medidas de tendência central média aritmética simples, mediana e moda; e as medidas de variabilidade: desvio padrão e variância. Embora o cálculo destas seja aqui explicado, este foi feito automaticamente neste estudo utilizando fórmulas do *Google Planilhas* que, não obstante, produzem resultados correspondentes.

2.1.1 Medidas de posição

2.1.1.1 Média aritmética simples

O valor (M) resultante da soma dos valores de uma amostra dividida pelo número de elementos nesta (Ibid., p. 35). Este pode ser expresso algebricamente por:

$$\mathrm{M} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{3}$$

Onde *n* é o número de elementos da amostra e x_i é o valor do elemento de índice *i*.

A fórmula “=MÉDIA” permite o cálculo automatizado deste índice. Assim, aplicando-a como “=MÉDIA(A2:A54)” e “=MÉDIA(B2:B40)” obtemos que as médias das notas das turmas A e B são **7,81** e **7,67**, respectivamente.

2.1.1.2 Mediana

O valor central, ou a média de dois valores centrais, (Md) de uma amostra quando esta é disposta sequencialmente (Ibid.). Este pode ser expresso algebricamente por:

$$\mathrm{Md} = \begin{cases} x\left(\dfrac{n+1}{2}\right), \text{ se n for ímpar;} \\ \dfrac{x\left(\frac{n}{2}\right) + x\left(\frac{n+1}{2}\right)}{2}, \text{ se n for par.} \end{cases} \tag{4}$$

A fórmula “=MED” permite o cálculo automatizado deste índice. Assim, aplicando-a como “=MED(A2:A54)” e “=MED(B2:B40)” obtemos que as medianas das notas das turmas A e B são **7,50** e **7,87**, respectivamente.

2.1.1.3 Quartil

O valor central, ou a média de dois valores centrais, que dividem a quarta parte de uma amostra do restante desta, quando esta é disposta sequencialmente (Ibid., p.42). Uma amostra qualquer, passível de ser dividida em quatro partes,

possui, portanto, três quartis (q1, q2 e q3), sendo o segundo destes a própria mediana.

Dispondo as notas de ambas as turmas sequencialmente em ordem crescente e contando os valores aqueles que encontram-se nas quartas partes, identificamos que estes são

a) **q1 = 7,5** e **q3 = 10** para a turma A, as médias dos dos 13º e 14º elementos, e 40º e 41º elementos, respectivamente;

b) **q1 = 6,8** e **q3 = 9,6** para a turma B, as notas dos 10º e 30º elementos, respectivamente.

2.1.1.4 Moda

O valor mais frequente em uma dada amostra, havendo pelo menos dois valores iguais. As séries em que existe mais de um valor mais frequente são denominadas bimodal, trimodal, etc. (Ibid., 35).

A fórmula “=MODO” permite a identificação automática deste índice. Assim, aplicando-a como “=MODO(A2:A54)” e “=MODO(B2:B40)” obtemos que as modas das notas das turmas A e B são **10,00** e **9,60**, respectivamente.

2.1.2 Medidas de dispersão

2.1.2.1 Variância

A razão (V) entre a somatória do quadrado das diferenças entre cada valor e a média aritmética, e o número de elementos na amostra (Ibid.). Pode ser expressa algebricamente por:

$$\mathrm{V} = \frac{\sum_{i=1}^{n}(x_i - M)^2}{n} \tag{5}$$

A fórmula “=VAR.S” permite o cálculo automatizado deste índice. Assim, aplicando-a como “=VAR.S(A2:A54)” e “=VAR.S(B2:B40)” obtemos que as variâncias das notas das turmas A e B são, aproximadamente, **7,15** e **4,00**, respectivamente.

2.1.2.2 Desvio padrão

Sendo a variância uma medida de dimensão igual ao quadrado da diferença entre os dados, sua proporção pode levar a problemas de interpretação. Assim o

sendo, o mais comum é apresentarmos o desvio padrão (Dp) enquanto medida de dispersão, esta que é definida enquanto a raiz quadrada positiva da variância (Ibid., p.38-39). Logo, esta pode ser expressa algebricamente por:

$$\mathrm{Dp} = \sqrt{V} \tag{6}$$

A fórmula “=DESVPAD” permite o cálculo automatizado deste índice. Assim, aplicando-a como “=DESVPAD(A2:A54)” e “=DESVPAD(B2:B40)” obtemos que os desvios padrão das notas das turmas A e B são, aproximadamente, **2,67** e **2,00**, respectivamente.

2.2 GRÁFICOS

Amparados pelas medidas-resumo descritas anteriormente, é possível sumarizar esses conjuntos de dados e produzir gráficos que ilustram as tendências observadas.

2.2.1 Histograma

Conforme descrevem Bussab e Morenttin (2017, p. 18), um histograma trata-se de um tipo de gráfico de barras contíguas onde os elementos de uma amostra são agrupados em classes as quais cada qual

a) corresponde a um conjunto de valores distintos, mas de igual amplitude;

b) é representada por uma barra cuja altura indica a frequência de elementos contidos nesta.

O histograma destaca-se enquanto recurso gráfico para sumarizar a ocorrência de valores contínuos, tal qual são as notas observadas, de acordo com uma escala qualquer.

Em nosso estudo estabelecemos que a amplitude amplitude (A) para as classes seria proporcional aos valores máximo ($v_{máx}$), mínimo (v_{min}), e número de elementos (n) observados na amostra, conforme a seguinte razão:

$$\mathrm{A} = \frac{v_{máx} - v_{min}}{\sqrt{n}} \tag{7}$$

Aplicando a fórmula “=(MÁXIMO(A2:A54) - MÍNIMO(A2:A54)) / RAIZ(CONT.NÚM(A2:A54))” para a turma A e a mesma fórmula com índices

correspondentes à turma B, obtêm-se, assim, os valores aproximados de 1,43 e 1,34 os quais foram igualados em **1,5** para fins comparativos.

Na sequência, os índices (A2:A54) e (B2:B40) foram selecionados na criação de gráficos do tipo histograma, que tiveram as amplitudes das barras definidas para 1,5. Os resultados foram os seguintes:

Gráfico 1: Turma A (Curso I)

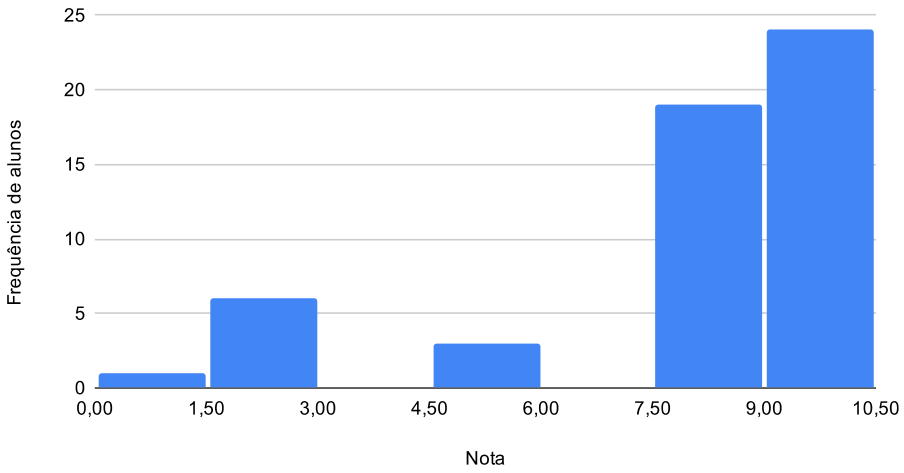
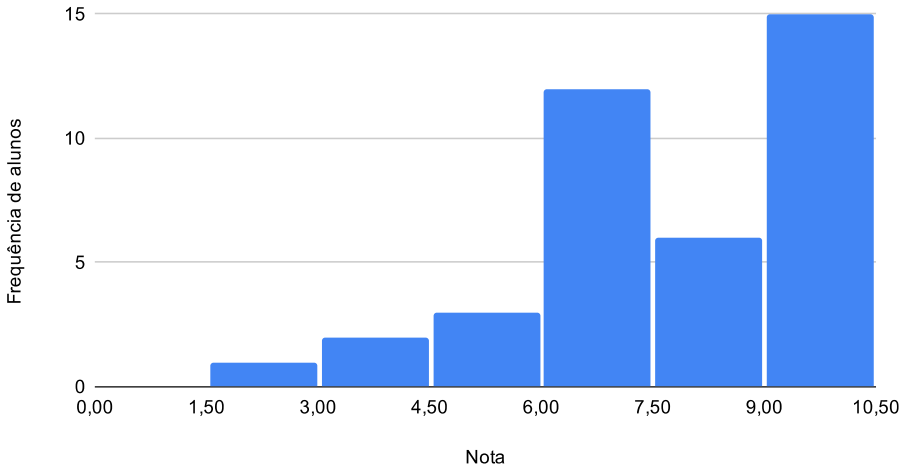


Gráfico 2: Turma B (Curso II)



Fonte: Autores

2.2.2 Boxplot

O *Boxplot*, ou "caixa-de-bigodes", trata-se de um diagrama que destaca a amplitude dos valores encontrados em uma amostra (Ibid., p. 48). Este possui os seguintes indicadores:

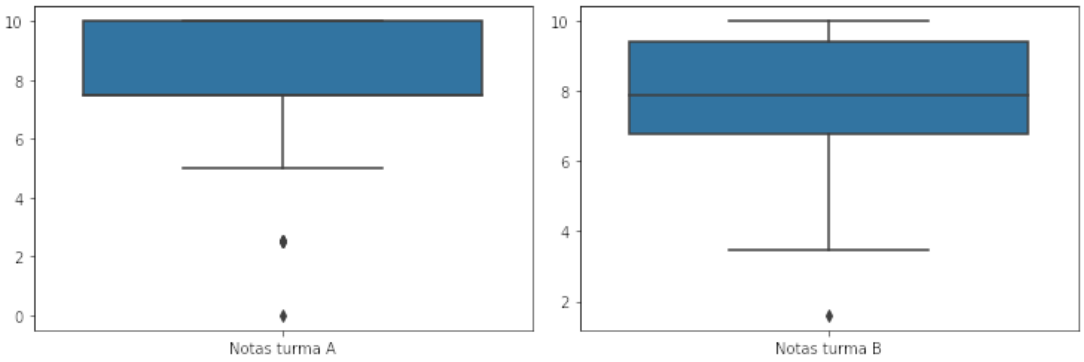
a) uma caixa, que indica a posição dos valores entre o primeiro e terceiro quartis;

b) uma linha no interior desta caixa, que indica a posição da mediana;

c) prolongamentos superior e inferior(os bigodes) que marcam a posição dos valores mais remotos que não ultrapassam os limites superior (LS = q3 + 1,5 (q3 - q1)) ou inferior (LI = q1 - 1,5 (q3 - q1));

d) e finalmente pontos exteriores que denotam observações destoantes das demais, denominadas valores atípicos ou *outliers*.

Para criarmos boxplots, utilizamo-nos dos dados anteriormente descritos, as posições dos quartis anteriormente obtidas, a linguagem de programação *Python* e suas bibliotecas *Pandas* (para leitura dos dados) e *Seaborn* (para criação do gráfico). Os resultados foram os seguintes:



Fonte: Autores

2.2.3 Gráfico de barras

Um gráfico de barras com intervalos unitários pode ser aplicado para se saber com precisão o número de alunos que obtiveram uma nota menor ou igual a uma determinada nota máxima, em intervalos regulares. Para criarmos um destes, primeiro geramos uma sequência de 1 à 10 nos índices (C2:C11), para então aplicar a fórmula “=FREQUÊNCIA” nos índices D2 e E2 com os parâmetros (A2:A54) e (B2:B40), respectivamente. O resultado da aplicação desta fórmula foi a criação de séries numéricas que quantificam o número de alunos que obtiveram notas correspondentes a cada intervalo unitário da série (C2:C11). Selecionados os índices em (C2:C11) mais, respectivamente, as sequências (D2:D11) e (E2:E11), geramos os seguintes gráficos:

Gráfico 3:Turma A (Curso I)

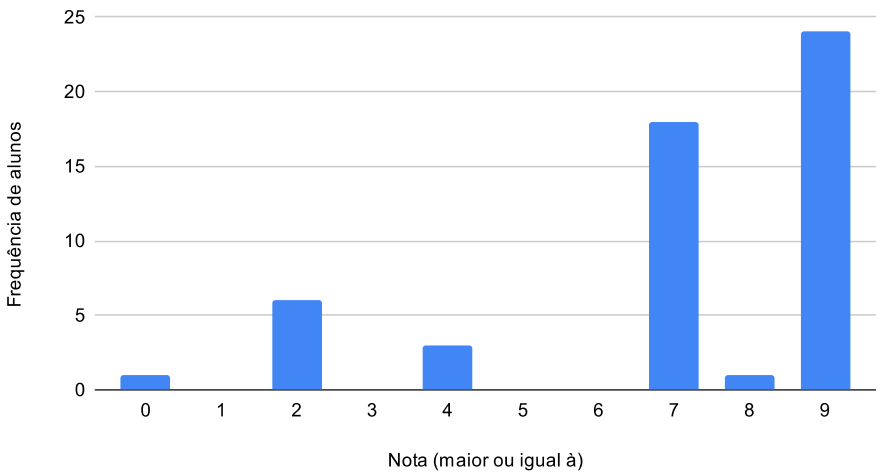
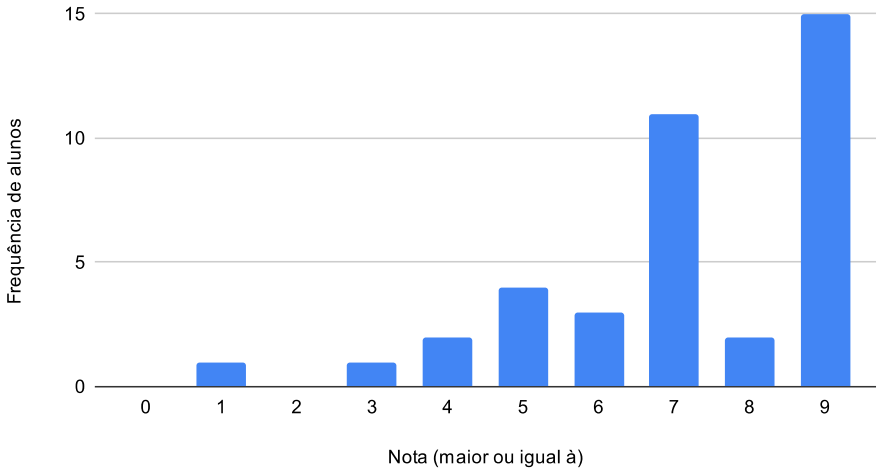


Gráfico 4: Turma B (Curso II)



Fonte: Autores

2.2.4 Gráfico de setores

Por vez, um gráfico de setores, ou “gráfico de pizza”, permite visualizar com maior precisão as frações do total de alunos que obtiveram uma nota menor ou igual a determinadas notas máximas. Utilizando-se dos mesmos parâmetros dos dois gráficos anteriores, obtivemos as seguintes imagens:

Gráfico 5: Turma A (Curso I)

Percentual de alunos versus nota (maior ou igual à)

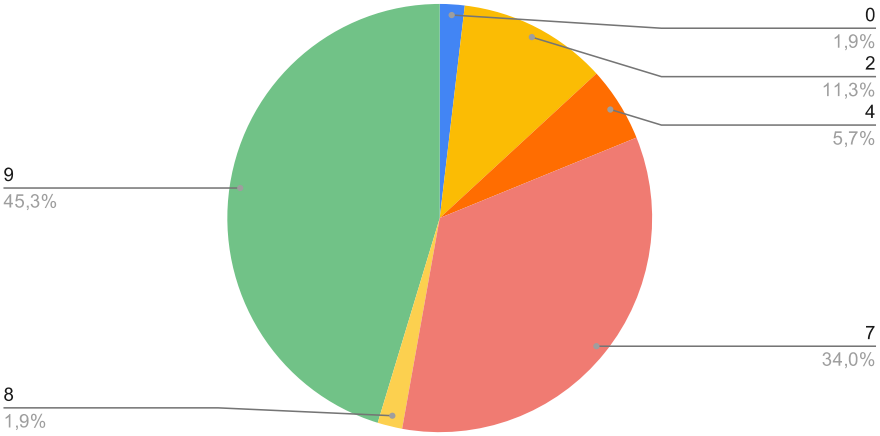
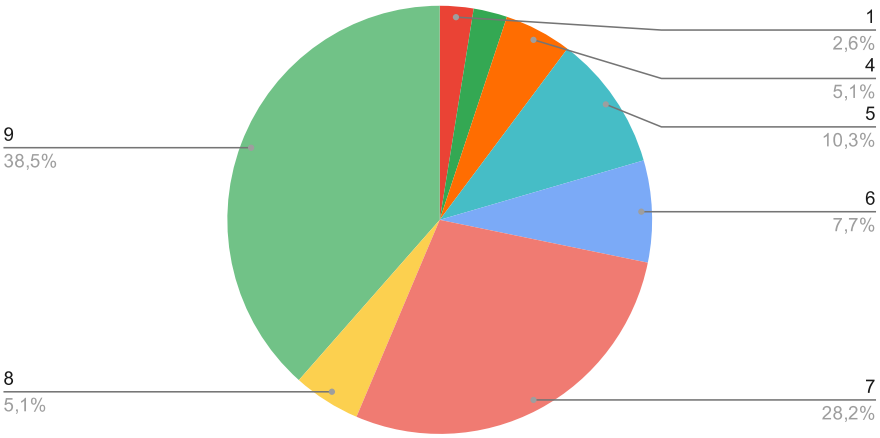


Gráfico 6: Turma B (Curso II)

Percentual de alunos versus nota (maior ou igual à)



Fonte: Autores

3 ANÁLISES CONJUNTAS

Obtidas as tendências gerais observadas em ambas as turmas, é possível aferir se, entre elas, há uma forma de associação. Consideremos um índice de aprovação dos alunos, estabelecido em uma nota maior ou igual a 5. Existe maior ou menor chance de um dado aluno ser reprovado em função deste estar cursando o curso introdutório (curso I) ou avançado (curso II)? De acordo com as unidades expostas nos gráficos 3 à 4 e percentuais expostos nos gráficos 5 à 6, temos:

Valores observados	Notas >= 5	Notas < 5	Total
Turma A (curso I)	46 (88,8%)	7 (13,2%)	53 (100%)
Turma B (curso II)	35 (89,75%)	4 (10,25%)	39 (100%)
Total	81 (88%)	11 (12%)	92 (100%)

Fonte: Autores

Fossem os cursos e turmas inconsequentes sobre a taxa de aprovações, teríamos:

Valores esperados	Notas >= 5	Notas < 5	Total
Turma A (curso I)	47 (≅ 88%)	6 (≅ 12%)	53 (100%)
Turma B (curso II)	34 (≅ 88%)	5 (≅ 12%)	39 (100%)
Total	81 (88%)	11 (12%)	92 (100%)

Fonte: Autores

Mas o que se observa é um ligeiro desvio:

Desvio	Notas >= 5	Notas < 5	Total
Turma A (curso I)	- 1 (≅ - 0,020%)	+ 1 (≅ + 0,020%)	0 (0%)
Turma B (curso II)	+ 1 (≅ + 0,025%)	- 1 (≅ - 0,025%)	0 (0%)
Total	0 (≅ + 0,005%)	0 (≅ - 0,005%)	0 (0%)

Fonte: Autores

Tal impacto sobre as notas em função da variação de turmas ou cursos pode ser aferida por índices denominados coeficientes de associação ou correlação: “medidas que descrevem, por meio de um único número, a associação entre duas variáveis" (Ibid., p. 76). Aqui, utilizaremos o coeficiente de contingência “T” proposto por Tschuprov (apud., Ibid.), que pode ser expresso algebricamente como:

$$\chi^2 = \sum_{i=1}^{r}\sum_{i=1}^{s}\frac{(n_{ij} - n_{ij}*)^2}{n_{ij}*};\ T = \sqrt{\frac{\chi^2}{n\sqrt{(r-1)(s-1)}}} \quad (8)$$

Onde:

a) n_{ij} é o valor do elemento na linha *i* e coluna *j*, e n_{ij}* é o valor esperado nessa mesma posição;

b) *n* é o número total de elementos (92);

c) *r* e *s* são o número máximo das linhas e colunas que descrevem os elementos (*r* = *s* = *2*).

O resultado da aplicação deste coeficiente para nossa amostra é de aproximadamente **0,067**, quando este índice pode variar entre 0, indicando a ausência de associação, e 1, indicando forte associação.

4 CONCLUSÃO

Apesar das turmas possuírem números diferentes de alunos, o que dificultou nossa análise pois impossibilitou a utilização de determinadas técnicas de comparação (tais quais a construção de gráficos *quantis x quantis*), conseguimos compará-las nos utilizando dadas técnicas pautadas pelo uso de medidas de posição e dispersão.

Primeiramente, identificamos tendências comuns entre estas pela aplicação de medidas de posição. Por meio destas pudemos observar que o desempenho geral de ambas as turmas é bastante similar (variação de apenas 0,14 na nota média destas, aproximadamente) e positivo (com moda igual à nota máxima e próxima desta em menos de uma unidade, respectivamente, e todos os quantis localizados acima da nota 5). Todavia, a aplicação de medidas de dispersão revela que as notas obtidas pelos alunos da turma A foram significativamente mais discrepantes entre si do que aquilo que foi observado entre os alunos da turma B. De fato, os gráficos 1 e 2 ilustram uma descontinuidade nas notas obtidas entre grupos de alunos da turma A, enquanto a turma B apresenta continuidade. No mais, os diagramas de boxplot revelam que isso se deve à maior presença de valores atípicos (*outliers*) na turma A. Ainda, os gráficos 3 e 4 nos permitem quantificar com precisão o número de ocorrências destes e outros valores, enquanto os gráficos 5 e 6 nos mostram a fração que estes representam no total de observações.

Finalmente, estabelecendo uma nota de corte em 5,0, e utilizando-se dos dados expostos nos gráficos 3 à 6, pudemos avaliar por meio de tabelas quocientes de aprovações em ambas as turmas e compará-los, fixando um coeficiente de contingência que revelou haver pouca, mas não nula, associação entre as turmas ou cursos com suas respectivas taxas de aprovação. Sendo assim, demonstrou-se que o curso avançado foi ligeiramente mais favorável à ocorrência de aprovações quando comparado ao curso introdutório.

5 REFERÊNCIAS

BUSSAB, W.; MORENTTIN, P. **Estatística Básica**. 9. ed. São Paulo: Saraiva, 2017.

Google Colaboratory. Disponível em: <https://colab.research.google.com/notebooks/intro.ipynb>. Acesso em: 25 jul. 2021.

Google Planilhas. Disponível em: <https://docs.google.com/spreadsheets/>. Acesso em: 17 jul. 2021.

Apêndice A – A divisão de tarefas do grupo

Para realização deste trabalho, os seguintes autores ficaram responsáveis pelas seguintes tarefas, especificamente:

a) Análises iniciais: realizadas por Kayo da Silva e Lucas de Moura;

b) Análises conjuntas: realizadas por Dimitri Dabkiewicz e Lincoln Braga;

c) Revisão, redação final e diagramação: realizados por Guilherme Barreto.