Universidade de São Paulo - USP
Escola Politécnica
PRO3200 - Estatística

Impacto da pandemia de COVID-19 na educação pública do Estado de São Paulo e a disparidade entre suas regiões

Integrantes:
Beatriz Cristine Almeida dos Santos 11808860
Camille Peixoto Almeida 12702259
Isabela Belapetravicius 11801971
Natália Elloá Ramos Oliveira 12555871
Victor Villela dos Santos 12553754

1 Introdução

Neste relatório, estuda-se a influência da pandemia do coronavírus na educação pública do estado de São Paulo, tópico que se tornou pauta de discussão com a queda dos casos de COVID e a subsequente volta das atividades presenciais.

Para esclarecer este cenário, serão analisados os dados referentes aos anos "pré-pandemia" e "pós-pandemia", disponibilizados pelo governo de São Paulo por meio do site *Dados Abertos da Educação* [1].

Ainda, surgem alguns questionamentos a respeito da disparidade econômica e social entre as regiões do Estado e de como isso afeta a educação pública, principlmente ao considerar o período posterior à pandemia.

Por isso, também será feita uma análise da influência do investimento em educação no rendimento escolar por região de São Paulo, a partir de dados do *Portal da Transparência* [2].

2 Objetivos

Tem-se como objetivo principal estudar as consequências da pandemia na educação, ao comparar o rendimento acadêmico dos alunos da rede pública nos períodos pré e pós-pandêmico.

Dessa forma, é possível avaliar se os efeitos da pandemia foram, no geral, negativos ou positivos, e se esses impactos ainda perduram nos estudantes do estado de São Paulo.

Em uma segunda análise, busca-se entender a heterogeneidade do rendimento escolar de acordo com a região do Estado – se ela é significativa e, em caso afirmativo, como ela pode ser justificada.

3 Descrição de Dados

As informações aqui analisadas foram obtidas com base no banco de dados do SARESP (Sistema de Avaliação de Rendimento Escolar), avaliação anual aplicada pela Secretaria da Eduacação do Estado de São Paulo que visa diagnosticar e acompanhar a evolução da educação básica paulista.

Com o intuito de avaliar o aprendizado e o rendimento dos alunos da rede pública, a análise será feita a partir das notas de proficiência das turmas de 9^o ano do fundamental e 3^o ano do médio nas grandes áreas de Língua Portuguesa e Matemática, considerando os períodos pré (2018 e 2019) e pós (2021 e 2022) pandemia.

Agora, para o estudo da relação entre investimento na educação e rendimento acadêmico, foram utilizados as informações disponibilizadas pelo Tesouro Nacional acerca do FUNDEB (Fundo de Manutenção e Desenvolvimento da Educação Básica e de Valorização dos Profissionais da Educação), sendo esta a principal fonte de financiamento da educação básica pública do país.

Por fim, para uma melhor interpretação dos dados, o estudo será desenvolvido separadamente para cada grande região, detalhadas na seção seguinte.

4 Metodologia

4.1 Regiões de São Paulo

Em um primeiro momento, questiona-se a existência de cenários educacionais muito distintos, tendo em vista a grande quantidade de municípios pertencentes ao estados de São Paulo e sua consequente diversidade econômica e social.

Por esse motivo, a base de dados utilizada foi dividida em sete grandes regiões, de acordo com o tamanho da população apresentado a seguir:

Tabela 1: Estimativa populacional 2021 (Referência)

Região	**População**
São Paulo	12.396.372
Vale do Paraíba + Litoral Norte	2.950.000
Campinas	1.223.237
Riberão Preto	720.116
Sorocaba	695.328
Baixada Santista	433.991

O restante dos municípios paulistas foi agrupado em uma única classe, denominada "Interior".

4.2 Nível de Proeficiência

Depois, para avaliar o rendimento acadêmico dos alunos de cada região nas disciplinas de Língua Portuguesa e Matemática, calculam-se as medidas de posição central (média, moda e mediana) e de dispersão (desvio padrão, variância) dos níveis de proficiência.

O nível de proficiência é um dos critérios de avaliação utilizados pelo SARESP e consiste em uma classificação de 0 a 1000 pontos a partir das expectativas anuais de aprendizagem dos alunos, estabelecidas para cada série escolar.

De acordo com a nota de proficiência, são definidas quatro categorias – insuficiente, suficiente básico, suficiente adequado e avançado – referentes ao nível de domínio dos conteúdos, competências e habilidades necessários para cada ano escolar em determinada área do conhecimento.

As categorias para as áreas de Língua Portuguesa e Matemática estão exibidas nas tabelas abaixo:

Tabela 2: Língua Portuguesa

Nível	3°EF	**5°EF**	**7°EF**	**9°EF**	**3°EM**
Insuficiente	<125	<150	<175	<200	<250
Suficiente: Básico	[125, 174]	[150, 199]	[175, 224]	[200, 274]	[250, 299]
Suficiente: Adequado	[175, 254]	[200, 249]	[225, 274]	[275, 324]	[300, 374]
Avançado	≥ 225	≥ 250	≥ 275	≥ 325	≥ 375

Tabela 3: Matemática

Nível	3°EF	**5°EF**	**7°EF**	**9°EF**	**3°EM**
Insuficiente	<150	<175	<200	<225	<275
Suficiente: Básico	[150, 199]	[175, 224]	[200, 249]	[225, 299]	[275, 349]
Suficiente: Adequado	[200, 249]	[225, 274]	[250, 299]	[300, 349]	[350, 399]
Avançado	≥ 50	≥ 275	≥ 300	≥ 350	≥ 400

4.3 Intervalo de Confiança e Testes de Hipóteses

A partir das medidas amostrais obtidas anteriormente, faz-se interessante encontrar uma estimativa intervalar para as medidas populacionais. Para isso, será construído o intervalo de confiança para a média, considerando um nível de confiança de 95%.

Além disso, também se tem o interesse de testar, a 5% de significância, hipóteses de comparação entre duas médias, objetivando-se assim a formulação de conclusões sobre o rendimento escolar em cada região de São Paulo ao longo dos anos, principalmente no período pré e pós pandêmico.

4.4 ANOVA e Regressão Linear

Por fim, busca-se avaliar se a diferença entre as médias de proficiência para cada região é significativa. Para tal, será feita uma análise de variâncias (ANOVA), empregada em comparações múltiplas – neste caso, de médias.

Ainda, caso seja possível afirmar que existe uma variação expressiva do rendimento escolar entre as diferentes regiões de São Paulo, será utilizado o conceito de regressão linear a fim de se estudar a relação entre investimento em educação e rendimento acadêmico. Assim, é possível avaliar se o ensino sofre influência do aspecto econômico de um município.

5 Modelos Estatísticos

5.1 Intervalo de Confiança para a Média

A primeira técnica estatística utilizada neste estudo será a contrução de intervalo de confiança para estimar medidas populacionais (no caso, a média) a partir de valores amostrais presentes no banco de dados.

As expressões analíticas dos intervalos de confiança estão apresentadas a seguir e serão utilizadas para a obtenção dos resultados exibidos na próxima seção.

$$\overline{x} - t_{n-1;\frac{\alpha}{2}} \frac{s}{\sqrt{n}} < \mu < \overline{x} + t_{n-1;\frac{\alpha}{2}} \frac{s}{\sqrt{n}} \tag{1}$$

Sendo:

1. $\overline{x}$: a média amostral
2. $t_{\frac{\alpha}{2}}$: os valores da distribuição de probabilidades t-student
3. n: tamanho da amostra
4. s: desvio padrão amostral

Vale destacar que o nível de confiança, denotado por $(1 - \alpha)$, representa a probabilidade de o parâmetro populacional estar contido no intervalo de confiança gerado.

5.2 Teste de Hipóteses para Duas Médias, com σ's Desconhecidos e Diferentes

A segunda técnica estatística empregada será a formulação de testes de hipóteses, com o intuito de comparar o rendimento acadêmico nas diferentes regiões de São Paulo e tirar conclusões sobre o impacto da pandemia na educação pública do estado.

A seguir, exibem-se as expressões analíticas de teste de hipóteses, que serão utilizadas para a obtenção dos resultados apresentados na próxima seção.

Inicialmente, definem-se a hipótese nula e a hipótese a ser testada:

$$H_0\colon \mu_1 = \mu_2$$
$$H_1\colon \mu_1 > \mu_2$$

Depois, encontra-se o valor crítico, dado por:

$$(x_1 - x_2)_{critico} = t_{\alpha;\nu} \times \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \tag{2}$$

Em que

$$w_1 = \frac{s_1^2}{n_1} \quad w_2 = \frac{s_2^2}{n_2} \tag{3}$$

$$\nu = \frac{(w_1 + w_2)^2}{\frac{w_1^2}{n_1+1} + \frac{w_2^2}{n_2+1}} - 2 \tag{4}$$

Sendo:

1. $\overline{x_i}$: médias amostrais
2. $t_{\alpha,\nu}$: os valores da distribuição de probabilidades t-student
3. ν: graus de liberdade
4. n_i: tamanho de cada amostra
5. s_i^2: variâncias amostrais

Se

$$\overline{x_1} - \overline{x_2} > +(x_1 - x_2)_{critico} \Rightarrow \textit{Rejeita-se } H_0$$

Vale destacar que o nível de significância, denotado por α, representa a probabilidade de rejeitar a hipótese nula quando esta é verdadeira.

5.3 ANOVA de Um Fator

Outra técnica abordada neste relatório será a construção de uma tabela de análise de variâncias, mais conhecida como ANOVA, utilizada em comparações de múltiplas médias.

As expressões analíticas da tabela ANOVA, essenciais para a obtenção dos resultados subsequentes, apresentam-se abaixo:

Primeiro, definem-se a hipótese nula e a hipótese a ser testada:

H_0*: Todas as médias são iguais.*
H_1*: Ao menos uma média é diferente.*

Depois, monta-se a tabela ANOVA:

Fonte	*SQ*	*GL*	*QM*	F_{calc}
Entre	*SQE*	GL_E	s_E^2	$\frac{s_E^2}{s_R^2}$
Residual	*SQR*	GL_R	s_R^2	
Total	*SQT*	GL_T	s_T^2	

Em que:

1. *SQ*: soma de quadrados
2. *GL*: graus de liberdade
3. *QM*: quadrado médio, $\frac{SQ}{GL}$

Ainda, o valor crítico é dado por:

$$F_{crit} = F_{GL_E;GL_R;\alpha} \tag{5}$$

Se

$$F_{calc} > F_{critico} \Rightarrow \textit{Rejeita-se } H_0$$

5.4 Regressão Linear

Finalmente, a última técnica tratada nesta análise será o ajuste de um modelo linear da proficiência média em função do investimento em educação para cada município de São Paulo. Aqui, o intuito é avaliar se o fator econômico é um dos responsáveis pela disparidade entre o rendimento acadêmico das diferentes regiões.

Em um modelo de regressão linear simples, têm-se o parâmetro do coeficiente angular (representado por β_1) e o parâmetro do coeficiente linear (representado por β_0), ou seja:

$$y = \beta_0 + \beta_1 x + e \tag{6}$$

Para determinar se há ou não correlação linear entre as variáveis analisadas, pode-se realizar um teste de hipóteses, em que:

- H_0*:* $\beta_0 = 0$ *e* $\beta_1 = 0$ *(Não há correlação)*
- H_1*:* $\beta_0 \neq 0$ *e* $\beta_1 \neq 0$ *(Há correlação)*

Desta vez, a tabela ANOVA é tal que:

Fonte	*SQ*	*GL*	*QM*	F_{calc}
Regrassão	*SQReg*	GL_{Reg}	s^2_{Reg}	$\frac{s^2_E}{s^2_R}$
Residual	*SQR*	GL_R	s^2_R	
Total	*SQT*	GL_T	s^2_T	

De novo, o valor crítico é dado por:

$$F_{crit} = F_{GL_E;GL_R;\alpha} \tag{7}$$

E se

$$F_{calc} > F_{critico} \Rightarrow \textit{Rejeita-se } H_0$$

6 Análise dos resultados

6.1 Medidas Estatísticas

A seguir estão apresentadas as medidas de posição central e de dispersão para cada região do estado de São Paulo durante o período definido anteriormente como pré e pós pandemia:

- *Período pré pandemia*

Ano 2018							
Região	Série	Disciplina	Média	Mediana	Moda	Desvio Padrão	Variância
Interior	9 EF	LP	215,35	242,50	0	96,61	9332,75
		MAT	226,18	250,00	0	98,83	9766,88
	3 EM	LP	220,08	266,40	0	123,09	15151,40
		MAT	233,20	272,50	0	118,18	13967,25
Baixada Santista	9 EF	LP	212,11	239,85	0	96,35	9284,12
		MAT	217,58	241,80	0	94,92	9009,90
	3 EM	LP	224,32	267,90	0	119,46	14271,06
		MAP	232,24	267,90	0	111,09	12340,05
Campinas	9 EF	LP	218,96	245,40	0	94,78	8982,88
		MAT	229,47	251,20	0	94,92	9009,86
	3 EM	LP	224,24	269,60	0	121,59	14784,19
		MAT	235,88	273,60	0	115,33	13301,49
Ribeirão Preto	9 EF	LP	217,85	241,10	0	90,59	8205,77
		MAT	227,35	247,60	0	93,58	8757,80
	3 EM	LP	229,72	269,20	0	115,73	13394,12
		MAT	238,11	273,40	0	111,76	12491,07
São Paulo	9 EF	LP	212,34	240,05	0	96,35	9282,96
		MAT	217,89	241,30	0	94,28	8888,29
	3 EM	LP	218,83	261,90	0	119,85	14362,96
		MAT	224,16	260,10	0	112,12	12570,42
Sorocaba	9 EF	LP	219,68	246,60	0	94,96	9017,58
		MAT	228,30	250,90	0	95,36	9093,07
	3 EM	LP	221,35	266,90	0	121,04	14650,56
		MAT	228,82	268,00	0	116,40	13549,07
Vale do Paraíba e Litoral Norte	9 EF	LP	213,48	241,50	0	98,11	9625,93
		MAT	222,60	247,00	0	100,69	10137,87
	3 EM	LP	231,22	271,80	0	116,23	13509,76
		MAT	239,65	275,60	0	113,42	12865,20

Tabela 4: Medidas estatísticas para 2018.

Ano 2019							
Região	Série	Disciplina	Média	Mediana	Moda	Desvio Padrão	Variância
Interior	9 EF	LP	218,34	242,50	0	95,34	9090,54
		MAT	232,18	256,10	0	100,84	10168,89
	3 EM	LP	225,29	264,70	0	117,39	13780,63
		MAT	229,52	269,00	0	120,62	14548,23
Baixada Santista	9 EF	LP	213,86	238,30	0	94,93	9012,26
		MAT	229,08	249,05	0	92,46	8548,99
	3 EM	LP	223,46	260,60	0	114,79	13175,71
		MAP	226,59	261,50	0	112,53	12663,50
Campinas	9 EF	LP	220,77	243,90	0	93,11	8670,17
		MAT	236,58	253,30	0	89,76	8057,57
	3 EM	LP	223,80	265,50	0	120,38	14490,55
		MAT	229,08	268,10	0	119,68	14322,34
Ribeirão Preto	9 EF	LP	216,98	238,80	0	91,09	8296,73
		MAT	231,47	250,80	0	92,76	8605,22
	3 EM	LP	229,45	266,80	0	115,33	13300,15
		MAT	240,83	269,70	0	106,70	11384,27
São Paulo	9 EF	LP	210,92	237,70	0	98,18	9638,60
		MAT	227,20	246,20	0	90,38	8168,37
	3 EM	LP	214,11	253,90	0	119,00	14161,29
		MAT	214,99	253,00	0	116,51	13575,43
Sorocaba	9 EF	LP	221,33	245,10	0	93,79	8795,73
		MAT	236,40	255,90	0	93,55	8751,03
	3 EM	LP	223,52	262,10	0	115,79	13408,21
		MAT	225,98	264,60	0	117,81	13879,58
Vale do Paraíba e Litoral Norte	9 EF	LP	215,02	240,20	0	96,54	9319,35
		MAT	234,56	255,40	0	96,33	9279,42
	3 EM	LP	228,41	268,50	0	117,13	13719,13
		MAT	242,00	272,70	0	109,58	12006,88

Tabela 5: Medidas estatísticas para 2019.

- *Período pós pandemia*

Ano 2021							
Região	Série	Disciplina	Média	Mediana	Moda	Desvio Padrão	Variância
Interior	9 EF	LP	202,88	230,50	0	97,97	9598,80
		MAT	209,78	236,95	0	102,09	10421,97
	3 EM	LP	180,73	229,80	0	127,74	16317,20
		MAT	184,36	234,50	0	129,72	16827,18
Baixada Santista	9 EF	LP	199,04	227,20	0	97,81	9566,93
		MAT	203,24	229,80	0	100,10	10020,12
	3 EM	LP	180,02	229,30	0	127,98	16378,50
		MAP	179,80	230,30	0	127,07	16146,02
Campinas	9 EF	LP	205,25	232,70	0	98,28	9659,30
		MAT	208,94	235,30	0	100,82	10164,27
	3 EM	LP	186,61	238,80	0	129,12	16670,69
		MAT	187,90	238,65	0	129,58	16791,11
Ribeirão Preto	9 EF	LP	193,18	223,50	0	100,10	10020,16
		MAT	197,51	226,10	0	103,32	10675,68
	3 EM	LP	183,48	232,90	0	126,33	15959,02
		MAT	185,69	235,05	0	127,43	16239,66
São Paulo	9 EF	LP	200,26	230,30	0	99,96	9991,81
		MAT	202,84	231,20	0	101,99	10401,55
	3 EM	LP	180,39	233,80	0	130,09	16923,58
		MAT	178,52	230,20	0	128,26	16451,86
Sorocaba	9 EF	LP	202,76	232,30	0	99,77	9954,97
		MAT	207,05	233,70	0	102,85	10577,37
	3 EM	LP	179,49	232,90	0	129,84	16859,70
		MAT	181,11	233,70	0	130,61	17058,03
Vale do Paraíba e Litoral Norte	9 EF	LP	208,87	233,60	0	94,59	8947,26
		MAT	216,71	239,70	0	99,15	9831,42
	3 EM	LP	191,77	240,40	0	125,15	15661,73
		MAT	194,39	241,10	0	126,42	15981,81

Tabela 6: Medidas estatísticas para 2021.

Ano 2022							
Região	Série	Disciplina	Média	Mediana	Moda	Desvio Padrão	Variância
Interior	9 EF	LP	219,09	240,10	0	89,77	8058,30
		MAT	224,98	241,90	0	91,75	8418,18
	3 EM	LP	217,59	253,10	0	114,76	13170,20
		MAT	220,64	253,80	0	116,24	13511,01
Baixada Santista	9 EF	LP	209,69	231,30	0	90,60	8207,67
		MAT	212,90	231,10	0	90,57	8202,91
	3 EM	LP	218,59	250,80	0	112,03	12549,69
		MAP	218,07	246,80	0	111,27	12380,64
Campinas	9 EF	LP	220,37	242,30	0	90,66	8219,71
		MAT	224,61	242,60	0	92,15	8490,86
	3 EM	LP	216,56	251,90	0	115,36	13308,18
		MAT	218,58	250,70	0	116,28	13521,26
Ribeirão Preto	9 EF	LP	204,58	227,80	0	92,71	8595,52
		MAT	209,49	229,50	0	93,94	8823,84
	3 EM	LP	216,47	249,90	0	111,72	12480,51
		MAT	217,90	248,00	0	112,29	12608,82
São Paulo	9 EF	LP	214,46	235,50	0	88,92	7907,19
		MAT	217,10	234,20	0	88,89	7901,92
	3 EM	LP	211,49	246,50	0	114,88	13196,98
		MAT	210,13	241,80	0	113,57	12899,16
Sorocaba	9 EF	LP	217,13	239,20	0	92,09	8479,89
		MAT	220,47	238,70	0	93,10	8666,87
	3 EM	LP	215,42	251,80	0	115,18	13266,55
		MAT	217,31	250,50	0	116,17	13495,05
Vale do Paraíba e Litoral Norte	9 EF	LP	220,54	241,10	0	89,79	8062,36
		MAT	226,39	242,10	0	91,57	8384,84
	3 EM	LP	218,45	253,80	0	113,77	12944,20
		MAT	220,18	252,80	0	114,46	13100,62

Tabela 7: Medidas estatísticas para 2022.

6.2 Representações Gráficas

Para melhor visualização das medidas estatísticas obtidas e apresentadas na seção anterior, estas foram representadas em gráficos de barras, expostos mais abaixo.

- *Análise de todas as regiões em 2019 e 2021*

Em um primeiro momento, analisam-se as médias de proficiência de todas as regiões paulistas no ano imediatamente anterior à pandemia – 2019 – e no ano de volta gradual das atividades presenciais –2021:

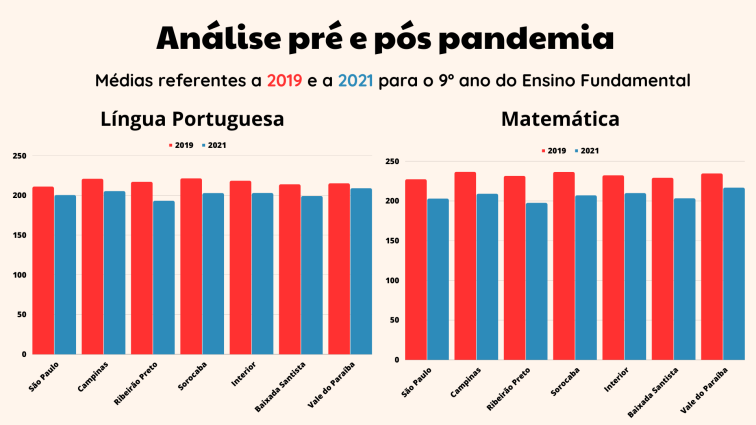

Figura 1: Média do nível de proficiência em 2019 e 2021para o ano de conclusão do ensino fundamental.

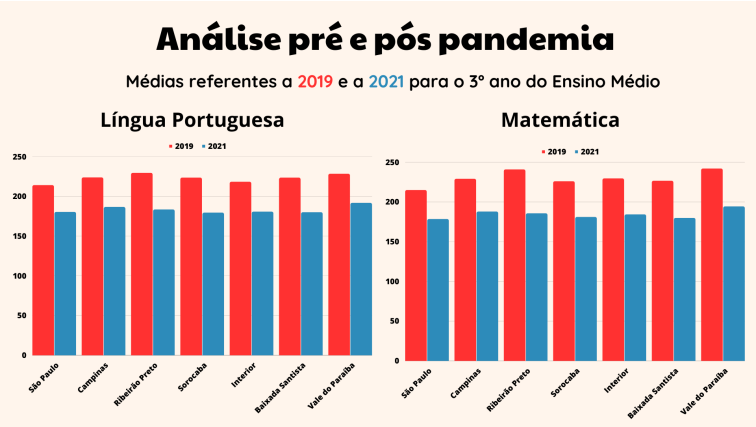

Figura 2: Média do nível de proficiência em 2019 e 2021para o ano de conclusão do ensino médio.

A partir dos gráficos, é possível perceber que, para todas as regiões analisadas, o rendimento acadêmico diminuiu após a suspensão das aulas presenciais devido ao cenário de pandemia.

Além disso, nota-se que para o 3^o ano do ensino médio a queda foi visualmente mais significativa.

- *Análise da região metropolitana de São paulo ao longo dos anos*

Agora, para analisar as médias de proficiência ao longo dos últimos dez anos, escolhe-se como exemplo a região metropolitana de São Paulo, a mais populosa do estado e do Brasil:

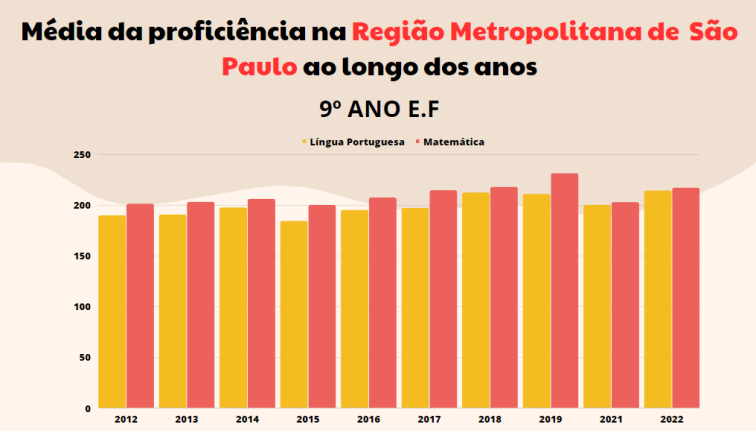

Figura 3: Média do nível de proficiência para a região metropolitana de São Paulo no ano de conclusão do ensino fundamental.

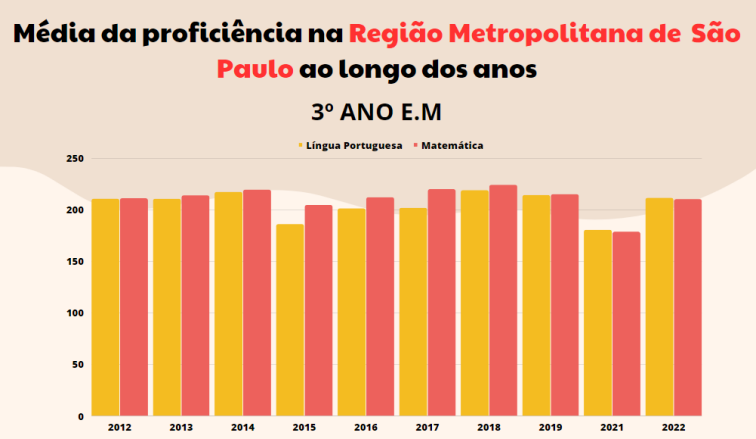

Figura 4: Média do nível de proficiência para a região metropolitana de São Paulo no ano de conclusão do ensino médio.

A partir dos gráficos, é possível observar que, ao longo dos anos, o rendimento acadêmico sofreu pequenas variações, sendo queda em 2021 a mais significativa delas, principalmente ao analisar o 3^o ano do ensino médio.

Além disso, nota-se que as médias de proeficiência de 2022 aumentaram, o que aponta para uma possível retomada do ritmo de estudos que havia no período anterior à pandemia.

6.3 Intervalos de Confiança

Como o número amostral é superior a 1000 graus de liberdade para todas as regiões analisadas, o valor da distribuição t-Student coincide até a terceira casa decimal com o valor da distribuição normal.

Como exemplo, são construídos os intervalos de confiança para os anos de 2018, 2019, 2021 e 2022 para a **região metropolitana de São Paulo**, considerando um nível de confiança de 95% (logo, $\alpha = 0.05$):

Anos	**Intervalos de Confiança**	
Séries	**9°EF**	**3°EM**
2018	211,94 - 212,81	218,40 - 219,35
2019	210,53 - 211,39	213,67 - 214,64
2021	199,67 - 200,97	179,70 - 181,21
2022	214,09 - 214,89	212,00 - 211,07

Tabela 8: Intervalo de confiança para a média de notas de proeficiência em Língua Portuguesa com nível de confiança de 95%, graus de liberdade = ∞, e $t_{\frac{\alpha}{2}} = 1{,}96$.

Anos	**Intervalos de Confiança**	
Séries	**9°EF**	**3°EM**
2018	217,50 - 218,35	223,76 - 224,65
2019	226,85 - 227,63	214,56 - 215,51
2021	202,24 - 203,57	177,84 - 179,34
2022	216,73 - 217,53	209,71 - 210,63

Tabela 9: Intervalo de confiança para a média de notas de proeficiência em Matemática com nível de confiança de 95%, graus de liberdade = ∞, e $t_{\frac{\alpha}{2}} = 1{,}96$.

Ao analisar os tamanhos de intervalo, nota-se que o principal fator para torna-lo maior é o desvio padrão, ou seja, quanto maior o desvio, maior o tamanho do intervalo de confiança.

Isto porque o número de elementos de cada amostra, que seria outro fator determinante para o tamanho do intervalo, não varia de forma significativa, considerando que a análise é feita para a mesma região.

6.4 Teste de Hipóteses

Considera-se inicialmente que as médias de proficiência entre 2019 e 2021 são iguais, e se quer provar que, na verdade, o rendimento acadêmico em 2021 sofreu uma queda,isto é:

$$H_0: \mu_{2019} = \mu_{2021}$$
$$H_1: \mu_{2019} > \mu_{2021}$$

De novo, como exemplo, os testes de hipóteses são realizados para a **região metropolitana de São Paulo**, considerando um nível de significânica de 5%:

Teste de Hipóteses			
9º EF		**3º EM**	
Amostral	**Crítico**	**Amostral**	**Crítico**
10,66	0,71	33,72	0,82

Tabela 10: Teste de hipóteses para as médias de notas de proeficiência de 2019 e 2021 em Língua Portuguesa com nível de significância de 5%.

Teste de Hipóteses			
9º EF		**3º EM**	
Amostral	**Crítico**	**Amostral**	**Crítico**
24,36	0,71	36,47	0,81

Tabela 11: Teste de hipóteses para as médias de notas de proeficiência de 2019 e 2021 em Matemática com nível de significância de 5%.

A partir dos valores críticos obtidos, nota-se que em todos os casos $\overline{x}_{2019} - \overline{x}_{2021} >>> (x_{2019} - x_{2021})_{critico}$, logo a **hipótese nula é rejeitada**.

Além disso, vale observar que os testes de hipóteses foram realizados também para as demais regiões de São Paulo, que resultaram nas mesmas conclusões.

Portanto, é possível afirmar que o rendimento escolar na rede pública do estado de São paulo sofreu uma queda de 2019 para 2021, sendo este um possível impacto da pandemia de COVID-19.

6.5 ANOVA

Aqui, considera-se inicialmente que as médias de proficiência das diferentes regiões de São Paulo são iguais para todos os anos, e se quer provar que, na verdade, há pelo menos uma das médias que destoa das demais, isto é:

H_0: *Todas as médias são iguais.*
H_1: *Ao menos uma média é diferente.*

Abaixo, apresenta-se a tabela ANOVA obtida:

6.6 Regressão Linear

7 Conclusão

Foi possível observar que a pandemia afetou negativamente o rendimento escolar dos estudantes do 9° ano do Ensino Fundamental e do 3° ano do Ensino Médio – os anos de conclusão dos ciclos escolares – para a região de maior importância populacional, econômica e cultural do estado de São Paulo, uma vez que, tanto para a disciplina de Língua Portuguesa como para a disciplina de Matemática, o nível de proeficiência diminuiu da passagem do ano de 2019 para o ano de 2021.

Assim, conclui-se que o cenário da pandemia de 2020 foi prejudicial à educação dos estudantes que não se adaptaram ao ensino remoto por diversos motivos, tais como:

- **Mudança repentina para o ensino remoto**: com o fechamento das escolas e medidas de distanciamento social, muitas escolas tiveram que adotar o ensino remoto de maneira rápida e improvisada. Professores e alunos tiveram que se adaptar rapidamente a novas plataformas e tecnologias de ensino, o que não foi fácil principalmente para alunos muito de séries mais novas epara professores de idade mais avançada. A transição abrupta para o ensino online pode ter resultado em dificuldades de comunicação, falta de acesso a recursos adequados e problemas de conectividade.

- **Desigualdade de acesso**: Nem todos os alunos têm acesso igualitário à internet de alta velocidade, dispositivos eletrônicos adequados ou ambientes propícios para o aprendizado em casa. Isso criou uma divisão digital entre aqueles que tinham acesso adequado à tecnologia e recursos e aqueles que não tinham, aprofundando as desigualdades educacionais. Alunos de famílias economicamente desfavorecidas ou de áreas rurais podem ter sido especialmente afetados.

- **Falta de interação presencial**: O ensino remoto privou os alunos da interação presencial com professores e colegas de classe. A interação social desempenha um papel importante no processo educacional, ajudando os alunos a se engajarem, a trocarem ideias e a esclarecerem dúvidas. A falta de interação face a face pode ter afetado negativamente a motivação e o envolvimento dos alunos.

- **Sobrecarga de trabalho para professores e alunos**: Tanto os professores quanto os alunos enfrentaram uma carga de trabalho adicional durante a pandemia. Os professores tiveram que adaptar seus métodos de ensino, criar materiais online e lidar com os desafios técnicos do ensino remoto. Os alunos, por sua vez, tiveram que se ajustar a novas rotinas de estudo, gerenciar seu tempo e aprender de forma independente, o que pode ter sido desafiador para muitos.

- **Impacto na saúde mental**: A pandemia e o isolamento social tiveram um impacto significativo na saúde mental de estudantes e educadores. O estresse, a ansiedade e a falta de motivação podem ter afetado negativamente o desempenho acadêmico e a capacidade de concentração dos alunos.

É importante ressaltar que, embora a educação durante a pandemia tenha enfrentado muitos desafios, também houve esforços significativos para mitigar esses problemas. Muitos professores e instituições de ensino se adaptaram rapidamente e desenvolveram estratégias de ensino online mais eficazes. Além disso, medidas foram tomadas para fornecer dispositivos eletrônicos e acesso à internet para alunos de famílias desfavorecidas. No entanto, ainda há muito a ser feito para superar os obstáculos e melhorar a qualidade da educação durante esse período desafiador.

8 Bibliografia

1. Dados Abertos da Educação. Microdados de alunos do Sistema de Avaliação de Rendimento Escolar do Estado de São Paulo (SARESP): Disponível em: `https://dados.educacao.sp.gov.br/dataset/microdados-de-alunos-do-sistema-de-C3%A7%C3%A3o-de-rendimento-escolar-do-estado-de-s%C3%Ao-paulo`. Acesso em: 02/05/2023

2. Portal da Transparência. TesouroNacional. O Fundo de Manutenção e Desenvolvimento da Educação Básica e de Valorização dos Profissionais da Educação (FUNDEB). Disponível em: `https://sisweb.tesouro.gov.br/apex/f?p=2600:1::IR_962295:NO`. Acesso em: 14/06/2023

3. HO, Linda Lee; RIBEIRO, Celma de Oliveira. Intervalo de confiança.PRO3200 - Estatística, Departamento de Engenharia e Produção, Universidade de São Paulo.2022.

4. RAMOS, Alberto. Apostila de Estatística-PRO3200. Escola Politécnica da Universidade de São Paulo, Departamento de Engenharia de Produção, São Paulo.2021

5. Lista de municípios de São Paulo por população (Estado de São Paulo). Wikipedia. Disponível em: `https://pt.wikipedia.org/wiki/Lista_de_munic%C3%ADpios_de_S%C3%A3o_Paulo_por_popula%C3%A7%C3%A3o` Acesso em: 25/05/2023

6. IBGE: Vale do Paraíba se aproxima de 2,6 milhões de habitantes. Band.com.br. Disponível em: https://www.band.uol.com.br/band-vale/noticias/ibge-vale-do-paraiba-se-aproxima-de-26-milhoes-de-habitantes-veja-por-cidade-16368529/ Acesso em: 25/05/2023

7. População do Litoral Norte de SP ultrapassa 345 mil habitantes em 2021, estima IBGE. Tamoios News.

Disponível em: https://www.tamoiosnews.com.br/noticias/populacao-do-litoral-norte-de-sp-em-2021-ibge/ Acesso em: 26/05/2023

8. População paulista cresceu 20% de 2001 a 2021. SEADE-Fundação Sistema Estadual de Análise de dados.

   Disponível em: https://www.seade.gov.br/populacao-paulista-cresceu-20-de-2001-a-2021/: :text=AAcesso em: 26/05/2023

9. Dados abertos da Educação. SARESP.

   Disponível em: https://dados.educacao.sp.gov.br/story/saresp Acesso em: 24/05/2023

9 Apêndice - Script do projeto em linguagem R

9.1 Carregamento dos dados

```
#Carrega 10 anos de dados da SARESP
#Os dados de serie nao estao regularizados 2011 pra tras(pelo menos o padrao
mudou desde essa epoca, se soubermos como era antes vai ser possivel usar esses dados)
carrega_todos_dados_educacao <- function(){
  df2022 <- read.csv("MICRODADOS_SARESP_2022_0.csv", sep = ";", header = TRUE,
  stringsAsFactors = FALSE, encoding = "UTF-8")
  df2021 <- read.csv("MICRODADOS_SARESP_2021.csv", stringsAsFactors = FALSE,
  encoding = "UTF-8")
  df2019 <- read.csv("MICRODADOS_SARESP_2019.csv",   sep = ";", header = TRUE,
  stringsAsFactors = FALSE, encoding = "UTF-8")
  df2018 <- read.csv("MICRODADOS_SARESP_2018.csv",   sep = ";", header = TRUE,
  stringsAsFactors = FALSE, encoding = "UTF-8")
  df2017 <- read.csv("MICRODADOS_SARESP_2017.csv",   sep = ";", header = TRUE,
  stringsAsFactors = FALSE, encoding = "UTF-8")
  df2016 <- read.csv("MICRODADOS_SARESP_2016.csv",   sep = ";", header = TRUE,
  stringsAsFactors = FALSE, encoding = "UTF-8")
  df2015 <- read.csv("MICRODADOS_SARESP_2015.csv",   sep = ";", header = TRUE,
  stringsAsFactors = FALSE, encoding = "UTF-8")
  df2014 <- read.csv("MICRODADOS_SARESP_2014.csv",   sep = ";", header = TRUE,
  stringsAsFactors = FALSE, encoding = "UTF-8")
  df2013 <- read.csv("MICRODADOS_SARESP_2013.csv",   sep = ";", header = TRUE,
  stringsAsFactors = FALSE, encoding = "UTF-8")
  df2012 <- read.csv("MICRODADOS_SARESP_2012.csv",   sep = ";", header = TRUE,
  stringsAsFactors = FALSE, encoding = "UTF-8")

  #Altera alguns nomes de colunas para padronizacao
  colnames(df2021)[45] <- "VALIDADE"
  colnames(df2022)[45] <- "VALIDADE"
  colnames(df2022)[11] <- "SERIE"
```

```
  colnames(df2021)[11] <- "SERIE"

  #Remove dos dados da prova de ciencia
  df2022 <- df2022[,-c(26, 29, 32, 35, 38, 41, 44)]
  df2021 <- df2021[,-c(26, 29, 32, 35, 38, 41, 44)]

  #Adiciona da coluna "Ano"
  df2022$Ano_Estudo <- 2022
  df2021$Ano_Estudo <- 2021
  df2019$Ano_Estudo <- 2019
  df2018$Ano_Estudo <- 2018
  df2017$Ano_Estudo <- 2017
  df2016$Ano_Estudo <- 2016
  df2015$Ano_Estudo <- 2015
  df2014$Ano_Estudo <- 2014
  df2013$Ano_Estudo <- 2013
  df2012$Ano_Estudo <- 2012

  df2022 <- subset(df2022, df2022$VALIDADE != 0)
  df2021 <- subset(df2021, df2021$VALIDADE != 0)
  df2019 <- subset(df2019, df2019$VALIDADE != 0)
  df2018 <- subset(df2018, df2018$VALIDADE != 0)
  df2017 <- subset(df2017, df2017$VALIDADE != 0)
  df2016 <- subset(df2016, df2016$VALIDADE != 0)
  df2015 <- subset(df2015, df2015$VALIDADE != 0)
  df2014 <- subset(df2014, df2014$VALIDADE != 0)
  df2013 <- subset(df2013, df2013$VALIDADE != 0)
  df2012 <- subset(df2012, df2012$VALIDADE != 0)
  return(list(df2022, df2021, df2019, df2018, df2017, df2016,
  df2015, df2014, df2013, df2012))
}


#Carrega somente 4 anos de dados da SARESP. Implementado para agilizar testes.
carrega_poucos_dados_educacao <- function(){
  df2022 <- read.csv("MICRODADOS_SARESP_2022_0.csv", sep = ";", header = TRUE,
  stringsAsFactors = FALSE, encoding = "UTF-8")
  df2021 <- read.csv("MICRODADOS_SARESP_2021.csv", stringsAsFactors = FALSE,
  encoding = "UTF-8")
  df2019 <- read.csv("MICRODADOS_SARESP_2019.csv",   sep = ";", header = TRUE,
  stringsAsFactors = FALSE, encoding = "UTF-8")
  df2018 <- read.csv("MICRODADOS_SARESP_2018.csv",   sep = ";", header = TRUE,
  stringsAsFactors = FALSE, encoding = "UTF-8")

  #Altera alguns nomes de colunas para padronizacao
  colnames(df2021)[45] <- "VALIDADE"
```

```
  colnames(df2022)[45] <- "VALIDADE"
  colnames(df2022)[11] <- "SERIE"
  colnames(df2021)[11] <- "SERIE"

  #Remove dos dados da prova de ciencia
  df2022 <- df2022[,-c(26, 29, 32, 35, 38, 41, 44)]
  df2021 <- df2021[,-c(26, 29, 32, 35, 38, 41, 44)]

  #Adiciona da coluna "Ano"
  df2022$Ano_Estudo <- 2022
  df2021$Ano_Estudo <- 2021
  df2019$Ano_Estudo <- 2019
  df2018$Ano_Estudo <- 2018

  df2022 <- subset(df2022, df2022$VALIDADE != 0)
  df2021 <- subset(df2021, df2021$VALIDADE != 0)
  df2019 <- subset(df2019, df2019$VALIDADE != 0)
  df2018 <- subset(df2018, df2018$VALIDADE != 0)
  return(list(df2022, df2021, df2019, df2018))
}


#Carrega somente um ano de dados. Ideal para testar funcoes que so dependam de
uma lista de dados
carrega_2022_educacao <- function(){
  df2022 <- read.csv("MICRODADOS_SARESP_2022_0.csv", sep = ";", header = TRUE,
  stringsAsFactors = FALSE, encoding = "UTF-8")
  colnames(df2022)[45] <- "VALIDADE"
  colnames(df2022)[11] <- "SERIE"
  df2022 <- df2022[,-c(26, 29, 32, 35, 38, 41, 44)]
  df2022$Ano_Estudo <- 2022
  df2022 <- subset(df2022, df2022$VALIDADE != 0)
  return(df2022)
}
```

9.2 Código principal

```
    #Carrega funcoes de outros scripts.
library(dplyr)
source("funcoes.R")
source("carrega_dados.R")

#Carrega dados
lista_dataframes_educacao <- carrega_todos_dados_educacao()
```

```
#Cria dataframe de resultados
Resultados <- data.frame(
    Disciplina =NA,
    Media =NA,
    Moda =NA,
    Variancia =NA,
    Desvpad =NA,
    NumeroAmostras =NA,
    Minimo =NA,
    Q1 =NA,
    Mediana =NA,
    Q3 =NA,
    Maximo =NA,
    NumeroSubBasico =NA,
    NumeroBasico =NA,
    NumeroAdequado =NA,
    NumeroAvancado =NA
    )[numeric(0), ]

iteradores <- c("Ano_Estudo", "SERIE", "RegiaoMetropolitana", "PERIODO")

#Cria colunas no df Resultados
for (iter in seq_along(iteradores)){
  Resultados[[iteradores[iter]]] <- vector(length = 0)
}

subsets <- list()
#Separa os dados baseado na lista de iteradores
for (d in seq_along(lista_dataframes_educacao)){
    data <- lista_dataframes_educacao[[d]]
    subsets <- append(subsets, split(data, do.call(interaction, data[iteradores])))
}
subsets <- Filter(function(df) nrow(df) > 0, subsets)


for (i in seq_along(subsets)){
    df <- subsets[[i]]
    informacoes <- c()
    for (iter in iteradores){
        informacoes <- c(informacoes, df[[iter]][1])
    }

    print(informacoes)

    informacoes_lp <- c("Lingua Portuguesa", medidas_descritivas(df$profic_lp),
        as.vector(table(df$nivel_profic_lp)[1:4]), informacoes)
```

```
    informacoes_mat <- c("Matematica", medidas_descritivas(df$profic_mat),
        as.vector(table(df$nivel_profic_mat)[1:4]), informacoes)
    print(length(informacoes_lp))
    Resultados <- rbind(Resultados, informacoes_lp)
    Resultados <- rbind(Resultados, informacoes_mat)
}

#Esse trecho de código existe puramente pq a formatação das colunas se perde no
meio do for loop. Não gosto dessa solução mas é o que tem pra hoje
colnames(Resultados) <- c(c("Disciplina", "Media", "Moda", "Variancia", "Desvpad",
  "NumeroAmostras", "Minimo", "Q1", "Mediana", "Q3", "Maximo", "NumeroSubBasico",
  "NumeroBasico", "NumeroAdequado", "NumeroAvancado"), iteradores)
```

9.3 Funções

```
    #Funcao Moda
mode <- function(x) {
  unique_vals <- unique(x)
  counts <- tabulate(match(x, unique_vals))
  max_count <- max(counts)
  modes <- unique_vals[counts == max_count]

  return(mean(modes))
}


#Amostragem dos dados
#QUEBRADO
amostra <- function(dados, n){
  z <- sample(1:nrow(dados), size = n, replace = TRUE)
  return(dados[z[1:n],])
}


#Remove linhas NULL ou NA, e converte os valores para numeros
formata_lista <- function(lista){
  lista <- subset(lista, !is.na(lista) & lista != "NULL")
  lista <- gsub(",", ".", lista)
  lista <- as.numeric(lista)
  return(lista)
}


#Retorna medidas descritivas de uma lista de dados
medidas_descritivas <- function(lista){
```

```
  lista <- formata_lista(lista)

  media <- mean(lista)
  moda <- mode(lista)
  variancia <- var(lista)
  desvpad <- sd(lista)
  n <- length(lista)
  minimo <- min(lista)
  q1 <- as.numeric(quantile(lista, 0.25))
  mediana <- median(lista)
  q3 <-as.numeric(quantile(lista, 0.75))
  maximo <- max(lista)

  return(c(media, moda, variancia, desvpad, n, minimo, q1, mediana, q3, maximo))
}


#Intervalo de confianca para uma media usando T-Student; Alfa = significancia
#A variavel "medidas descritivas" deve ser obtida pela funcao de mesmo nome
intervalo_confianca_media <- function(medidas_descritivas, alfa, n){
  media <- medidas_descritivas[1]
   desvpad <- medidas_descritivas[4]
   n <- medidas_descritivas[5]

   erro <- qt(alfa, n-1) * desvpad/sqrt(n)
   return(c(media-erro, media+erro))
}
```