Trabalho Final de Métodos Estatísticos 1

Consultores Responsáveis:
Felipe Bretas
Renan Marques
Tales Vaz
Victor Tavares

Requerente:
Maria Teresa

Sumário

1 Introdução

O presente estudo tem como objetivo analisar o desempenho dos estudantes do 5º ano e 9º ano do Ensino Fundamental e 3ª série do Ensino Médio na Prova Brasil de 2011. Para isso, foi utilizada uma amostra aleatória simples de 200 escolas, extraída do banco de dados dos resultados da avaliação. Além disso, foi selecionada uma subamostra de 50 escolas para comparação dos resultados obtidos.

A pesquisa busca descrever as características das escolas participantes e o desempenho médio dos alunos. Entre os principais focos da análise estão a estimativa da nota média em Língua Portuguesa e Matemática, a verificação da normalidade das distribuições das notas e a comparação da proficiência em Matemática segundo o local da escola (urbana ou rural). Além disso, o estudo examina a evolução do desempenho entre 2009 e 2011 e investiga possíveis associações entre variáveis como região, categoria administrativa, tamanho da escola e tamanho do município. Para os testes de hipóteses foi utilizado um nível de significância de 5%. Os dados foram disponibilizados pelo Inep e se referem às escolas que participaram da Prova Brasil 2011, uma avaliação aplicada pelo Ministério da Educação para diagnosticar o desempenho dos estudantes em redes municipais, estaduais e federais.

O software utilizado para análise estatística dos dados foi o R versão 4.4.2. 0 R é um software de programação gratuito largamente usado na área de estatística e visualização de dados que permite não só o manuseio e análise de bancos de dados, como também a confecção de gráficos.

2 Objetivos

2.1 Descrever as características das escolas e o desempenho de seus estudantes na Prova de Brasil em 2011.

2.2 Estimar a proporção de escolas que menos de 75% de seus estudantes participaram da Prova Brasil em 2011.

2.3 Estimar a proficiência média em Língua Portuguesa e em Matemática das escolas na Prova Brasil em 2011.

2.4 Verificar se houve melhora do resultado da Prova Brasil de 2009 para 2011. Na Prova Brasil realizada em 2009 a proficiência em Língua Portuguesa foi 184,3 e em Matemática foi 204,3.

2.5 Verificar se é possível afirmar que as notas em Língua Portuguesa e em Matemática são normalmente distribuídas.

2.6 Comparar a proficiência média em Matemática segundo o local da escola. Você diria que a proficiência em Matemática é maior em escolas urbanas?

2.7 Verificar se existe diferença significativa entre as notas de Língua Portuguesa e Matemática.

2.8 Comparar a proporção de escolas que menos de 75% de seus estudantes participaram da Prova Brasil em 2011 segundo:

2.8.1 Local da escola;

2.8.2 Região de localização da escola.

2.9 Verificar se:

2.9.1 Região e categoria administrativa estão associadas;

2.9.2 Tamanho da escola e tamanho do município estão associados.

2.10 Verificar se a nota em Língua Portuguesa é um bom indicador para predizer a nota existe em Matemática, ou seja se estão associadas.

3 Metodologia

4 Resultados

4.1 Análise 1

Nesta primeira análises, foi feito uma descritiva das variáveis a fim de compreender melhor as características da amostra.

Figura 1: Gráfico de setores do local da escola

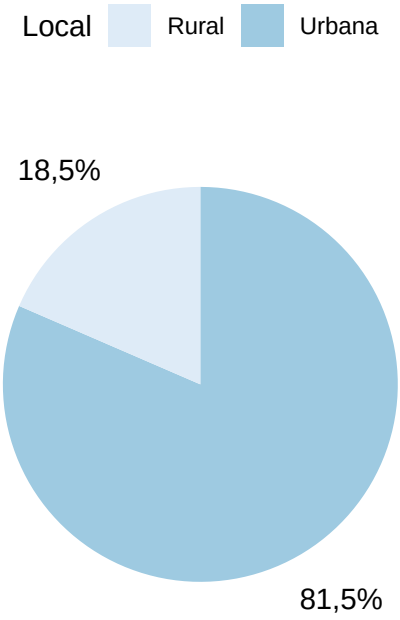


Figura 2: Gráfico de colunas da região da escola

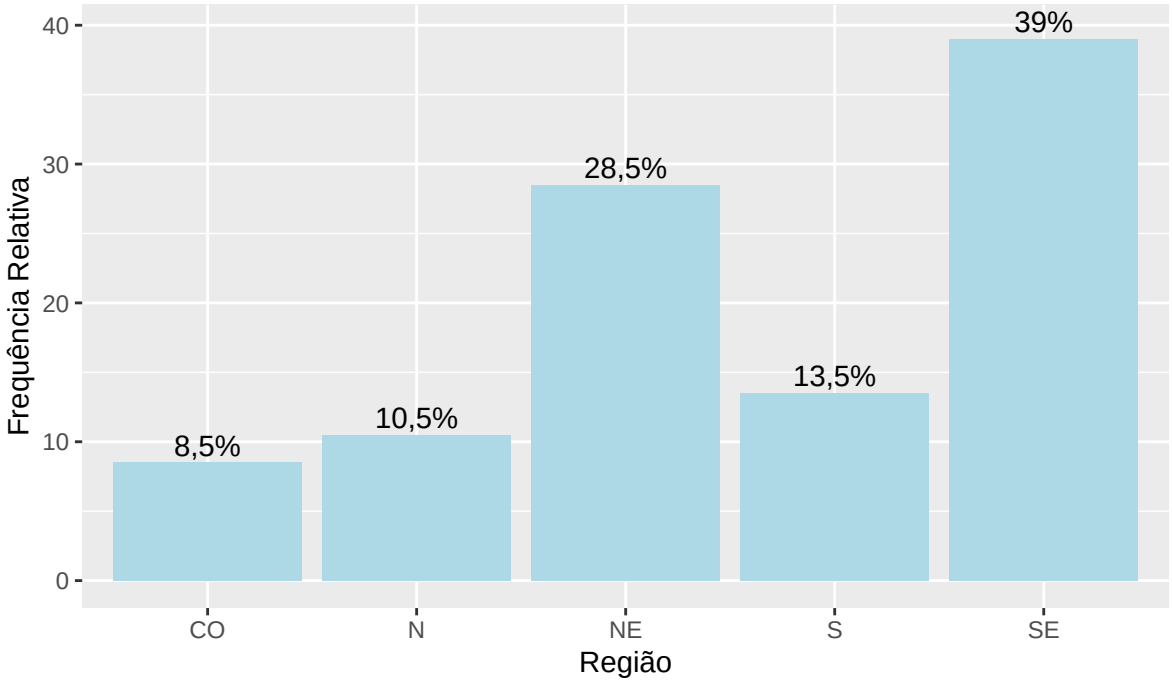


Figura 3: Gráfico de setores do tamanho do município

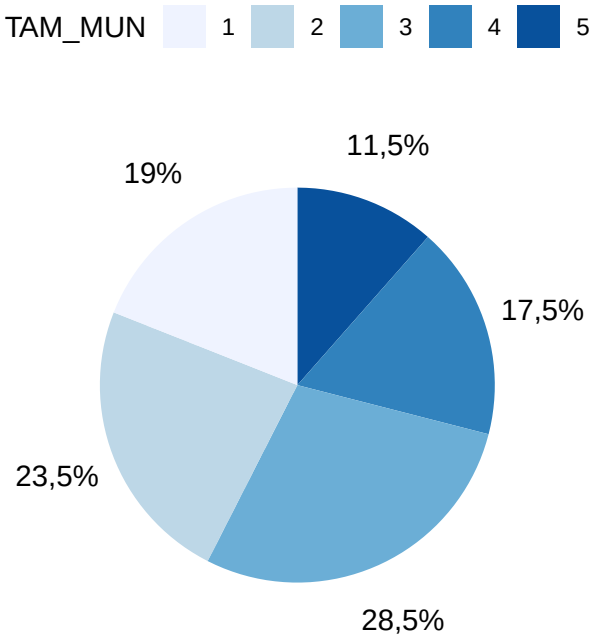


Figura 4: Gráfico de setores do tipo de administração da escola

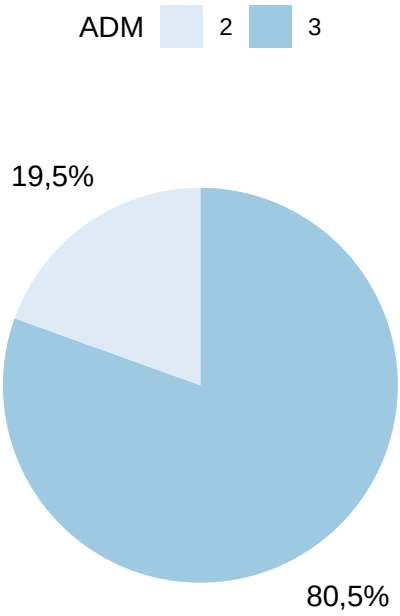


Figura 5: Gráfico de setores do tamanho da escola

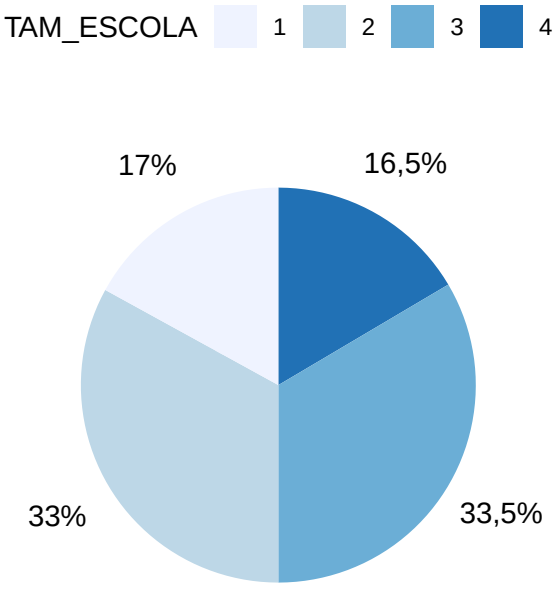


Figura 6: Boxplot do número de matriculados em cada escola

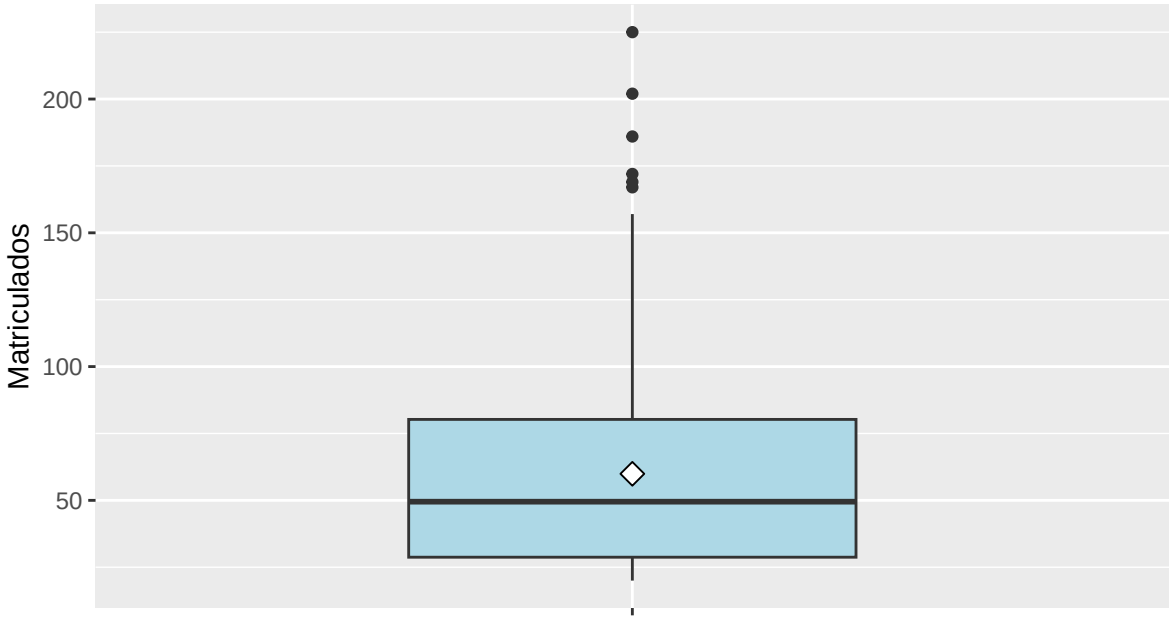


Figura 7: Gráfico de setores do local da escola(amostra de 50)

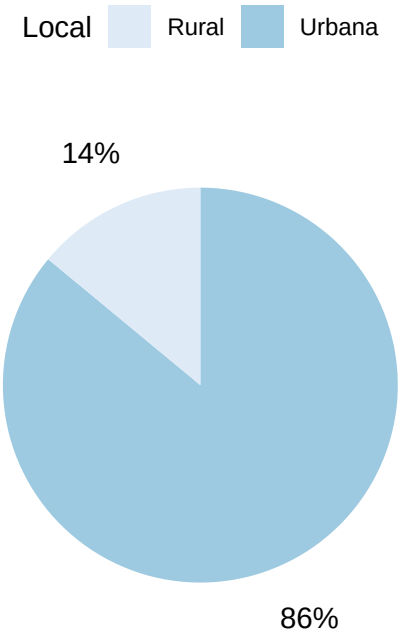


Figura 8: Gráfico de colunas da região da escola(amostra de 50)

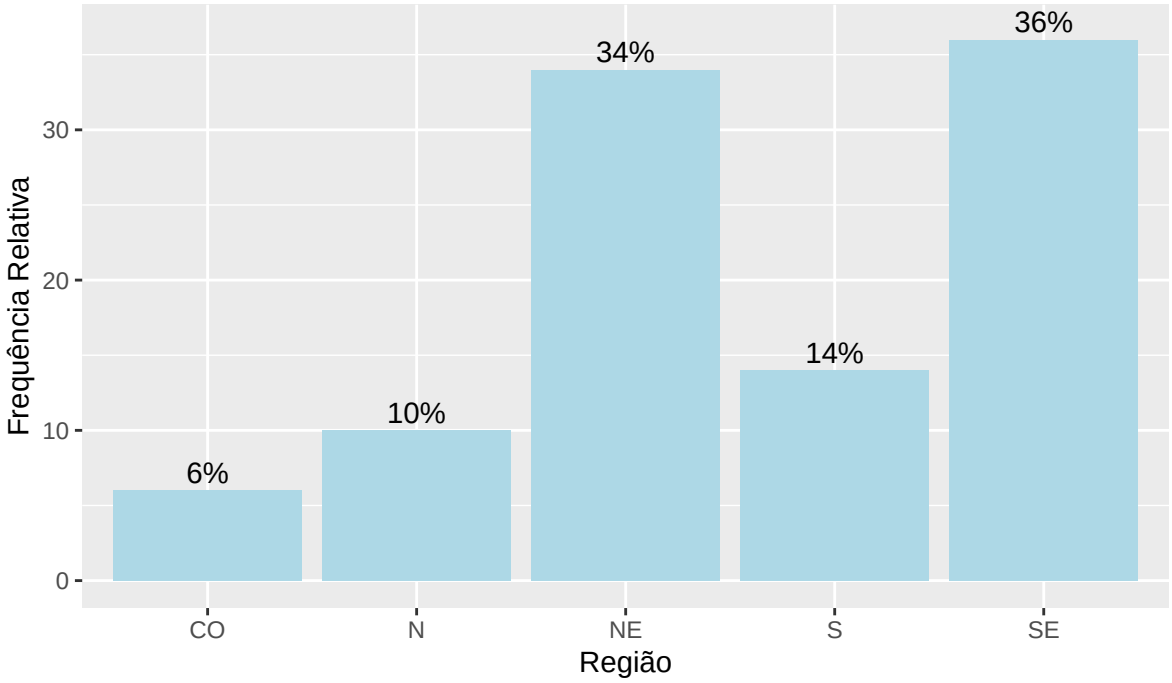


Figura 9: Gráfico de setores do tamanho do município(amostra de 50)

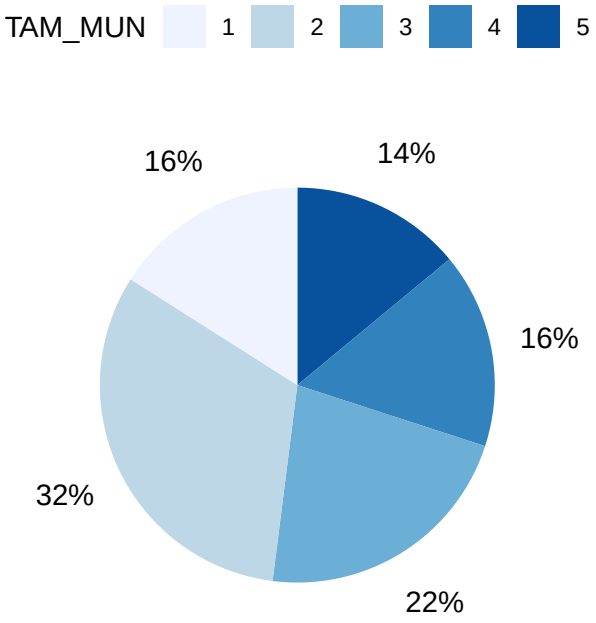


Figura 10: Gráfico de setores do tipo de administração da escola(amostra de 50)

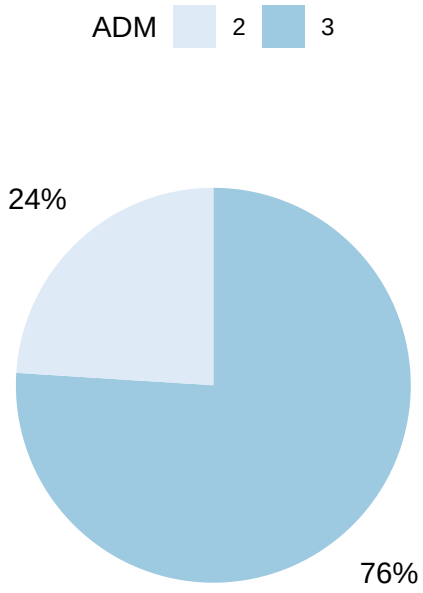


Figura 11: Gráfico de setores do tamanho da escola(amostra de 50)

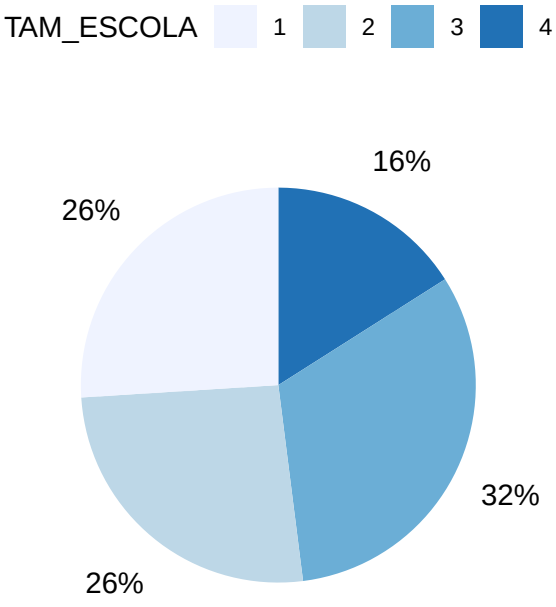
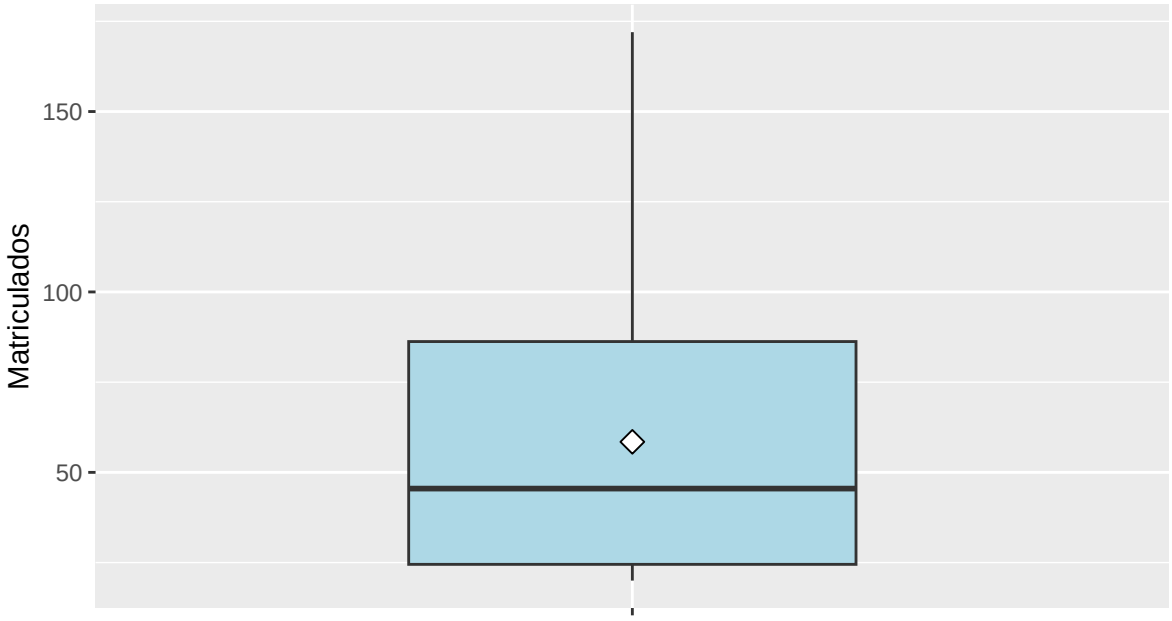


Figura 12: Boxplot do número de matriculados em cada escola(amostra de 50)



4.2 Análise 2

Esta análise tem como intenção estimar a proporção de todas as escolas que obtiveram uma participação média menor que 75% em todas as provas realizadas. Para isso foi utilizada a variável “Participação”, sendo esta quantitativa contínua. Para a construção do intervalo de confiança, foi suposto que ambas as notas seguem uma distribuição normal.

O intervalo de confiança de 95% obtido para a proporção média menor que 75% de participação foi $p \in$ (0,04; 0,11) para a amostra com 200 observações e $p \in$ (0; 0,16) para a amostra com 50 observações.

4.3 Análise 3

Está análise tem como objetivo estimar a média das notas de portugues e matemática na população do estudo. Para isso, foram utilizadas as variáveis “NOTA_LP” e “NOTA_MT”, ambas quantitativas contínuas. Para a construção do intervalo de confiança, foi suposto que ambas as notas seguem uma distribuição normal.

O intervalo de confiança de 95% obtido para a média da nota de português foi $mi_1 \in$ (183.76;190.39) e o para a média de matemática foi $mi_2 \in$ (202.90;210.65). Para a amostra de tamanho 50, por ser uma amostra menor, os intervalos de confiança de 95% para as notas são maiores: para a média da nota de português foi $mi_1 \in$

(179.48;194.46) e o para a média de matemática foi $mi_2 \in$ (199.82;216.64).

4.4 Análise 4

Nessa análise buscou-se verificar se houve melhora do resultado da Prova Brasil de 2009 para 2011. Na Prova de 2009 a média em língua portuguesa foi de 184.3 e a de matemática foi de 204.3. Para isso foram feitos dois testes de hipótese: 1°) A nota de língua portuguesa melhorou? 2°) A nota de matemática melhorou?

Hipóteses do primeiro teste:$H_0) mi_1 = 184,3$,$H_1) mi_1 > 184,3$.

Estatística do teste: $T = X - 184,3/23.9 \div \sqrt{200}$ tem distribuição de student com 199 graus de liberdade.

Região Crítica para a=5%: $mi_1 > 187.08$

Conclusão: como a média amostral foi de 187.07, não rejeitaremos H_0, entretanto, por estar bem próxima da região crítica, seria necessário estudar mais a fundo para afirmar esse resultado. O P-valor deu 0.0516, que é maior porém bem próximo de 0,05, o que contribui para não rejeitar H_0.

Hipóteses do segundo teste:$H_0) mi_1 = 204.3$,$H_1) mi_1 > 204.3$.

Estatística do teste: $T = X - 204.3/27.9 \div \sqrt{200}$ tem distribuição de student com 199 graus de liberdade.

Região Crítica para a=5%: $mi_1 > 207.55$

Conclusão: como a média amostral foi de 206.77, não rejeitaremos H_0. O P-valor deu 0.1056 que é maior que 0,05, o que contribui para não rejeitar H_0.

Para a Amostra de tamanho 50, o processo foi similar: testes de hipótese: 1°) A nota de língua portuguesa melhorou? 2°) A nota de matemática melhorou? Hipóteses do primeiro teste:$H_0) mi_1=184,3 ,H_1)mi_1>184,3$.

Estatística do teste: $T = X - 184,3/26.3 \div \sqrt{50}$ tem distribuição de student com 49 graus de liberdade.

Região Crítica para a=5%: $mi_1 > 190.55$

Conclusão: Como a média amostral foi de 186.96, não rejeitamos H_0. O P-valor deu aproximadamente 0,2 que é maior que 0,5, o que contribui para não rejeitar H_0.

Hipóteses do segundo teste:$H_0) mi_1 = 204.3$,$H_1) mi_1 > 204.3$.

Estatística do teste: $T = X - 204.3/29.5 \div \sqrt{50}$ tem distribuição de student com 49 graus de liberdade.

Região Critica para a=5%: $mi_1 > 211.31$

Conclusão: Como a média amostral foi de 208.23, não rejeitamos H_0.

Em suma, em ambas as amostras não houve melhora das notas. O P-valor deu aproximadamente 0,15 que é maior que 0,5, o que contribui para não rejeitar H_0.

4.5 Análise 5

Nessa análise tem o intuito descobrir se as notas de Língua Portuguesa e Matemática aderem ao modelo de distribuição normal. Foram utilizados para a confecção desta análise as variáveis “Nota Língua Portuguesa” e “Nota Matemática”, ambas quantitativas contínuas. Para isso foi necessário estimar a média e a desvio padrão de ambas as matérias, sendo estas, 187,08 de média e 23,91 de desvio padrão para Língua Portuguesa para a amostra de 200 obversavações; 186,97 de média e 26,37 de desvio padrão para Língua Portuguesa para a amostra de 50 observações; 206, 78 de média e 27,97 de desvio padrão para Matemática para a amostra de 200 observações; 208,23 de média e 29,60 de desvio padrão para Matemática para a amostra 50 observações. Para isso foi feito os testes de hipótese: 1°) A Nota de Língua Portuguesa adere a distribuição normal? 2°) A Nota de Matemática adere a distribuição normal?

Primeiramente serão analisadas as amostras com 200 observações.

Hipóteses primeiro teste: $H_0)$ Nota de Língua Portuguesa segue o modelo de distribuição normal; $H_1)$ Nota de Língua Portuguesa não segue o modelo de distribuição normal.

Estatística do Teste: $X^2 = \sum_{i=1}^{6}((O_i - E_i)^2) \div E_i$, Qui-Quadrado com 5 graus de liberdade.

Região Crítica para a=5%: $X_1^2 < 0,8312$ e $X_2^2 > 12,8325$

Conclusão: Como o valor obtido pela estatística do teste foi de $X^2 = 5,2896$, H_0 não será rejeitada. Utilizando o p-valor, a conclusão é a mesma, obtendo um valor de 0,3816, ainda não rejeitando H_0.

Hipóteses segundo teste: $H_0)$ Nota de Matemática segue o modelo de distribuição normal; $H_1)$ Nota de Matemática não segue o modelo de distribuição normal.

Estatística do Teste: $X^2 = \sum_{i=1}^{7}((O_i - E_i)^2) \div E_i$, Qui-Quadrado com 6 graus de liberdade.

Região Crítica para a=5%: $X_1^2 < 1,2373$ e $X_2^2 > 14,4494$

Conclusão: O valor obtido para estatística do teste foi de $X^2 = 4,1033$, H_0 não será rejeitada. O p-valor também aponta para a não rejeição de H_0, com o valor de 0,6627.

Será analisado de forma semelhante as amostras de 50 observações.

Hipóteses primeiro teste: $H_0)$ Nota de Língua Portuguesa segue o modelo de distribuição normal; $H_1)$ Nota de Língua Portuguesa não segue o modelo de distribuição normal.

Estatística do Teste: $X^2 = \sum_{i=1}^{5}((O_i - E_i)^2) \div E_i$, Qui-Quadrado com 4 graus de liberdade.

Região Crítica para α=5%: $X_1^2 < 0,4844$ e $X_2^2 > 11,1433$

Conclusão: O valor obtido para estatística do teste foi de $X^2 = 3,5847$, H_0 não será rejeitada. O p-valor também aponta para a não rejeição de H_0, com valor de 0,4651.

Hipóteses segundo teste: $H_0)$ Nota de Matemática segue o modelo de distribuição normal; $H_1)$ Nota de Matemática não segue o modelo de distribuição normal.

Estatística do Teste: $X^2 = \sum_{i=1}^{6}((O_i - E_i)^2) \div E_i$, Qui-Quadrado com 5 graus de liberdade.

Região Crítica para a=5%: $X_1^2 < 0,8312$ e $X_2^2 > 12,8325$

Conclusão: A estatística do teste obteve um valor de $X^2 = 1,6651$, assim não rejeitando H_0. O p-valor também aponta para a não rejeição de H_0, com um valor de 0,8933.

4.6 Análise 6

A análise teve como objetivo investigar se existem diferenças significativas no desempenho em Matemática entre alunos de escolas urbanas e rurais. Para isso, foi utilizado o teste para comparação de duas médias, adequado para amostras independentes, uma vez que as variâncias entre os grupos podem ser diferentes.

A hipótese nula H_0 propõe que não há diferença no desempenho entre alunos de escolas urbanas e rurais, enquanto a hipótese alternativa H_1 sugere que existe uma diferença significativa. A amostra total foi composta por 200 alunos, com n = 200. O teste gerou um valor t de 5,0852, com 56,626 graus de liberdade e um p-valor de 4.318e-06, que é significativamente menor que o nível de significância de 5%. Isso indica que existe uma diferença significativa no desempenho entre os dois grupos. Os alunos de escolas urbanas apresentaram uma média de notas de 211,11, enquanto os alunos de escolas rurais obtiveram uma média de 187,70, com uma diferença de 23,41 pontos, o que reforça a superioridade dos alunos urbanos. O intervalo de confiança de 95% (14.19;32.63) também confirma essa diferença significativa, evidenciando que os alunos urbanos tendem a ter um desempenho melhor. O boxplot da amostra demonstra que as notas de Matemática são mais concentradas em torno da mediana para os alunos urbanos, com menor dispersão, enquanto os alunos rurais apresentam maior variação em suas pontuações.

4.7 Análise 7

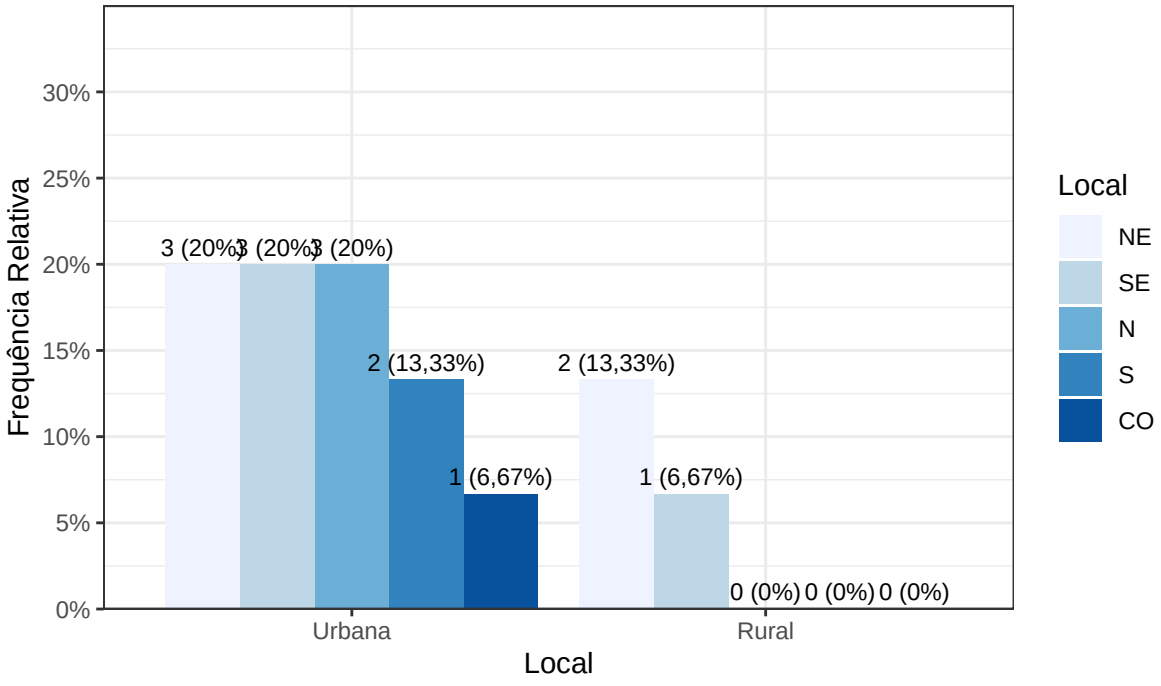
O objetivo desta análise foi avaliar se existe uma diferença estatisticamente significativa entre as notas de Língua Portuguesa e Matemática dos alunos. Como se tratam das mesmas pessoas avaliadas nas duas disciplinas, foi escolhido o teste para comparação de duas médias pareadas, adequado para amostras dependentes.

A hipótese nula H_0 sugere que não há diferença nas médias das notas entre as duas disciplinas, enquanto a hipótese alternativa H_1 indica que existe uma diferença significativa. O teste para a amostra de 200 alunos gerou um valor t de -28,28, com 199 graus de liberdade e um p-valor inferior a $2 * 10^{-16}$, o que é altamente significativo. A diferença média entre as notas foi de -19,70, com intervalo de confiança de 95% variando entre -21,07 e -18,32. Isso confirma que as notas em Língua Portuguesa são significativamente menores que as de Matemática. O boxplot reforça essa diferença, mostrando uma clara disparidade entre as distribuições das duas disciplinas. Para a amostra de 50 alunos, os resultados foram semelhantes, com t = -15,48, com 49 graus de liberdade e p-valor inferior a $2 * 10^{-16}$, evidenciando uma diferença média de -21,26, com intervalo de confiança entre -24,02 e -18,50. Ambos os testes indicam que a diferença entre as notas de Matemática e Língua Portuguesa é estatisticamente significativa.

4.8 Análise 8

Nesta análise comparou-se a proporção das escolas que menos de 75% de seus estudantes participaram da Prova Brasil em 2011 em relação ao local da escola e a região de localização da escola. Para isso foram utilizadas as variáveis “LOCAL” e “REG”, ambas qualitativas nominais. Espera-se que as proporções mantenham a proporção geral da amostra.

Figura 13: Gráfico de colunas do local pela região das escolas em que menos de 75% dos estudantes participaram da prova



Como pode ser observado, 80% dessas escolas são da região urbana e a maioria (60%) são do Nordeste ou do Sudeste. O que vai de acordo com a proporção geral, em que 81.5% são da região urbana e 67.5% são da região Nordeste ou Sudeste. Na amostra de 50, só em 4 escolas menos de 75% dos estudantes participaram da Prova Brasil, o que torna extremamente difícil verificar por meio dessa amostra se há uma relação entre a região/local da escola e a baixa adesão dos estudantes na prova.

4.9 Análise 9

Esta análise tem como intenção verificar se existe associação entre algumas variáveis presentes na amostra. Sendo assim, esta análise será segmentada em duas a fim de facilitar a compreensão e a leitura.

4.9.1 Análise 9a)

A fim de verificar a existência de uma associação entre as variáveis “Região”, qualitativa nominal, e “Categoria administrativa”, qualitativa nominal, foram feitos gráficos a seguir.

Figura 14: Gráfico de colunas da região pela categoria administrativa para amostra de 200

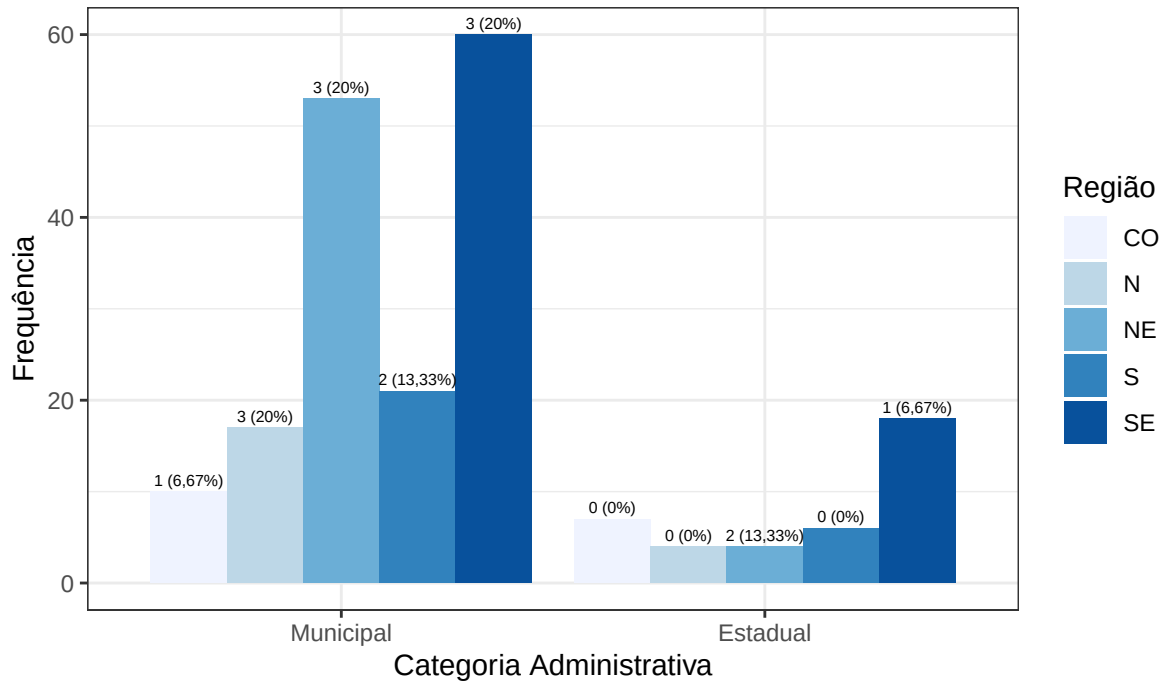
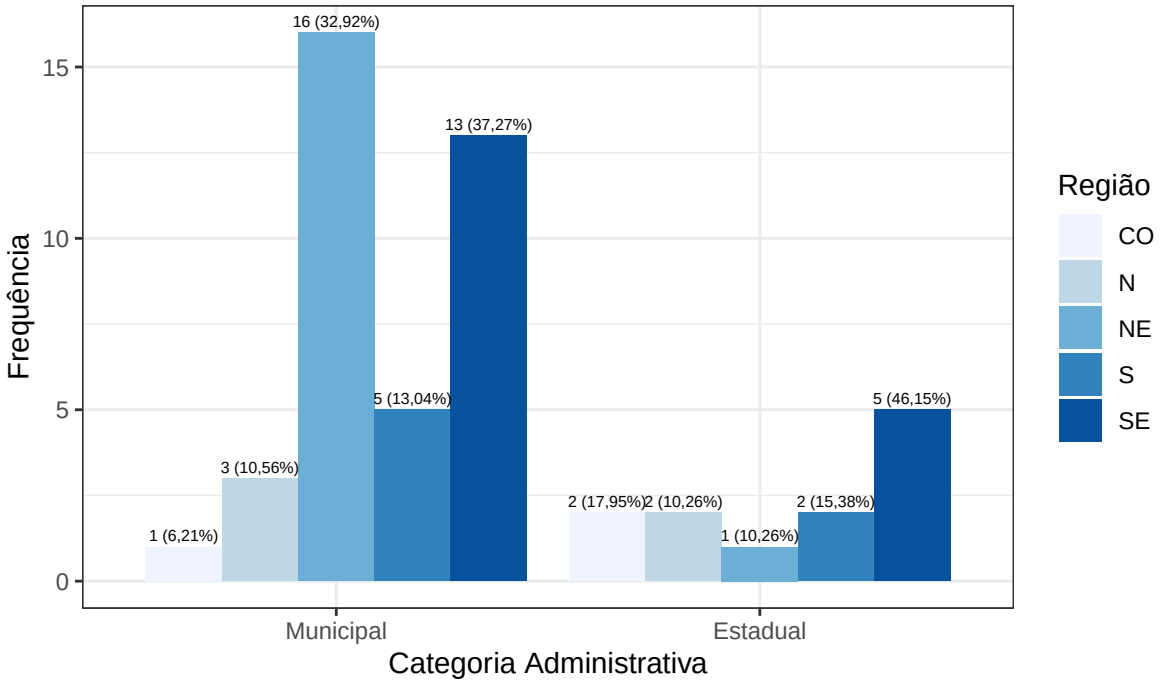


Figura 15: Gráfico de colunas da região pela categoria administrativa para amostra de 50



Como pode ser observado na **Figura 14** e **Figura 15**, existem mais observações presentes na categoria “Municipal”, totalizando 161 observações, mais que o quadruplo da categoria “Estadual”. Vale a pena ser ressaltado que ambas categorias, a região “Sudeste” é a que apresenta mais observações. Na categoria “Municipal”, pode-se observar uma grande quantidade de observações na região “Nordeste” com um total de 53. Para verificar a associação entre as variáveis foi utilizado o Coeficiente de Contingência Modificado, este que possui um valor entre 0 e 1 com valores próximos de 0 demonstrando uma associação fraca e valores próximos de 1 uma associação forte. Na análise em questão, obteve um valor de C^*=0,33 para a amostra de 200, apresentando uma associação de fraca a moderada entre as variáveis e C^*=0,26 para a amostra de 50, apresentando uma associação de fraca a moderada entre as variáveis, mas menor que a amostra de 200.

4.9.2 Análise 9b)

A fim de verificar a existência de uma associação entre as variáveis “Tamanho do Município”, qualitativa ordinal, e “Tamanho da Escola”, qualitativa ordinal, foram feitos os gráficos a seguir.

Figura 16: Gráfico de colunas do tamanho da escola pelo tamanho do município para amostra de 200

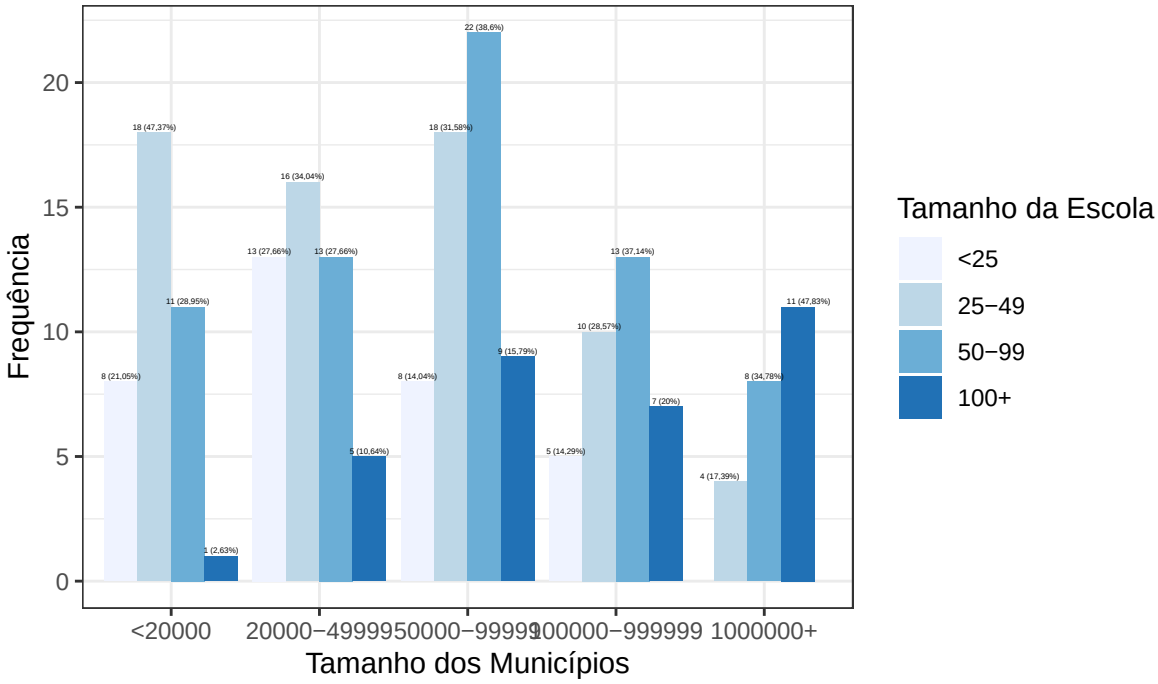
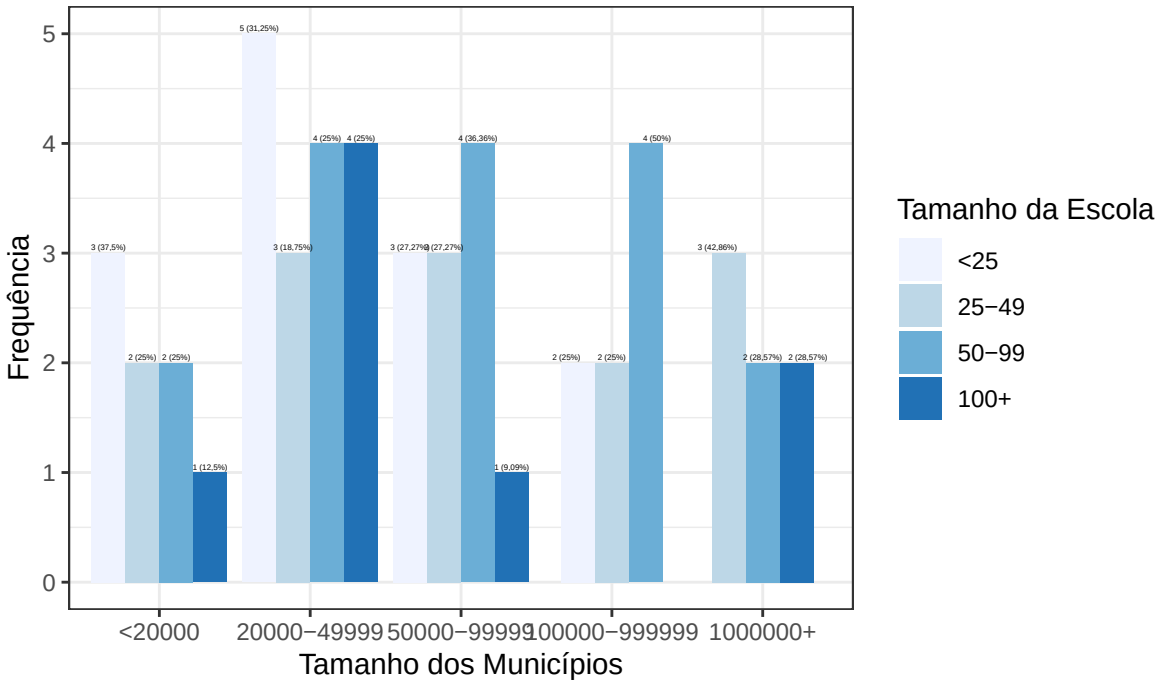


Figura 17: Gráfico de colunas do tamanho da escola pelo tamanho do município para amostra de 50



Observando a **Figura 16** e a **Figura 17** é possível observar existem mais escolas com tamanho “50000-99999” na amostra de 200, com um total 57 observações. Contudo, na amostra de 50, existem mais escolas com tamanho “20000-49999”,

com um valor de 16. Para verificar a associação entre as variáveis foi utilizado o Coeficiente de Contingência Modificado, este que possui um valor entre 0 e 1 com valores próximos de 0 demonstrando uma associação fraca e valores próximos de 1 uma associação forte. Na análise em questão, obteve um valor de C^*=0,44 para a amostra de 200, apresentando uma associação a moderada entre as variáveis e C^*=0,23 para a amostra de 50, apresentando uma associação de fraca a moderada entre as variáveis, sendo menor que a amostra de 200.

4.10 Análise 10

objetivo desta análise é investigar a relação entre o desempenho dos alunos nas disciplinas de Língua Portuguesa e Matemática. Para isso, foi utilizada a correlação de Pearson, que mede a força e a direção da associação linear entre as variáveis.

A hipótese nula H_0 propõe que não existe correlação entre as notas de Língua Portuguesa e Matemática, enquanto a hipótese alternativa H_1 sugere que há uma correlação significativa entre elas. A correlação de Pearson para a amostra de 200 alunos apresentou um coeficiente de 0.9398, com um valor t de 38,69 e 198 graus de liberdade, com p-valor inferior a $2 * 10^{-16}$. Isso indica uma forte associação positiva entre as notas das duas disciplinas. O gráfico de dispersão mostra um alinhamento claro entre as notas de Língua Portuguesa e Matemática, reforçando a forte correlação positiva. Para a amostra de 50 alunos, a correlação também foi alta, com coeficiente de 0.9067, t = 20,26 e p-valor inferior a $2 * 10^{-16}$, confirmando que a relação entre as notas das duas disciplinas é estatisticamente significativa, independentemente do tamanho da amostra.

5 Conclusões

Portanto, a partir das análises descritivas realizadas foi possível ter uma visão geral dos estudantes e das escolas integrantes da amostra. Na primeira análises, pode-se observar as medidas resumos da proficiência dos estudantes e gráficos que contribuíram para entender características pertinentes das escolas, não alterando de maneira significativa entre as amostras de tamanho 200 e de 50, sendo importante para uma boa visualização do panorama geral do projeto.

Já em relação a estimação da proporção de escolas com menos de 75% de participação, foi obtido para um intervalo de confiança de 95%, uma proporção de (0,04;0,11), ou 7,5% com erro 3,5% de erro, para as amostras de tamanho 200. Já para amostras de tamanho 50, teve a proporção de (0,00;0,16), ou 8% com margem de erro de 8%.

Ao se estimar a proficiência média dos alunos nas matérias de português, foi obtido para um intervalo de confiança de 95%, uma média de (183,76;190,30) para amostras de tamanho 200, e para amostras de tamanho 50, o intervalo foi de (179,48;194,46). Foi estimado também a proficiência média em matemática, onde foi obtido um intervalo de (202,90;210,65) para amostras de tamanho 200, e para amostras de tamanho 50, o intervalor foi de (199,82;216,64). Pode-se observar que as notas médias de português estimadas foram menor que as notas de matemática em ambas amostras e que, amostras de tamanho menor possuem um intervalo maior para um mesmo intervalo de confiança.