

UNIVERSIDADE FEDERAL DE JUIZ DE FORA
DEPARTAMENTO DE ESTATÍSTICA
CURSO DE ESTATÍSTICA

Iago Carvalho Cunha

**Análise de Dados Amostrais Complexos da Pesquisa
do PROALFA de Minas Gerais**

Juiz de Fora
2010

Iago Carvalho Cunha

*Análise de Dados Amostrais Complexos da Pesquisa do
PROALFA de Minas Gerais*

Monografia apresentada ao Curso de Estatística da
Universidade Federal de Juiz de Fora, como requi-
sito para a obtenção do grau de Bacharel em Es-
tatística.

Orientador: Marcel de Toledo Vieira

Doutor em Estatística - University of Southampton

Juiz de Fora

2010

Cunha, Iago

Análise de Dados Amostrais Complexos da Pesquisa do PROALFA
de Minas Gerais / Iago Cunha - 2010

37.p

1.Amostragem 2. Dados Amostrais Complexos. I.Título.

CDU N/A

Iago Carvalho Cunha

*Análise de Dados Amostrais Complexos da Pesquisa do
PROALFA de Minas Gerais*

Monografia apresentada ao Curso de Estatística da
Universidade Federal de Juiz de Fora, como requi-
sito para a obtenção do grau de Bacharel em Es-
tatística.

Aprovado em 06 de dezembro de 2010

BANCA EXAMINADORA

Marcel de Toledo Vieira

Doutor em Estatística - University of Southampton

Ronaldo Rocha Bastos

Doutor em Urban and Regional Planning - Liverpool University

Clécio da Silva Ferreira

Doutor em Estatística - Universidade de São Paulo

À minha mãe.

Resumo

As últimas décadas vêm testemunhando um aumento da frequência em que os levantamentos de dados são realizados, normalmente com o objetivo de auxiliar as tomadas de decisões em diversos setores da sociedade. Tal fato fez com que a importância da Amostragem, ramo da Estatística, se tornasse mais evidente. Na análise de dados amostrais complexos, a não consideração das características do planejamento amostral pode levar à ocorrência de vies. Este trabalho tem como objetivo principal observar e discutir alguns dos problemas encontrados quando tais características não são consideradas na estimação de alguns parâmetros e do erro padrão de seus estimadores. Além disso, será realizada uma comparação da eficiência dos diferentes planos amostrais utilizados nas pesquisas do PROALFA, realizadas nos anos de 2007, 2008 e 2009. Os resultados, em geral, mostram que o erro padrão das estimativas ficam subestimados quando as características do plano amostral não são consideradas para estimação. Isto faz com que os coeficientes estimados para modelos de regressão, por exemplo, sejam considerados significativos quando na verdade não são.

Palavras-chave: Dados amostrais complexos, PROALFA, Estimação.

Abstract

Recent decades have witnessed an increase in the frequency with which surveys are conducted, usually with the aim of assisting decision making in various sectors of society. This fact has made the importance of Survey Sampling to become more evident. In the analysis of complex survey data, not taking into account the characteristics of the survey design can lead to the occurrence of bias. This monograph's main objective is to observe and discuss some problems encountered when such characteristics are not considered in the estimation of some parameters and of standard errors estimators. In addition, there will be a comparison of the efficiency of different sampling schemes used in the PROALFA surveys conducted in 2007, 2008 and 2009. The results generally show that the standard error of the estimators are underestimated when the characteristics of the sampling design are not considered for estimation. As result, the estimated coefficients for regression models, for example, tend to be considered significant when in fact they are not.

Keywords: Complex survey data, PROALFA, Estimation.

Agradecimentos

Seria injusto dizer que esta monografia é resultado de uma caminhada que começou na Universidade Federal de Juiz de Fora e, por isto, agradeço desde já àqueles que passaram pela minha vida e de alguma forma contribuíram para o que sou hoje.

E agradeço, particularmente, a algumas pessoas que contribuíram diretamente na construção e finalização deste trabalho:

À minha mãe que nunca mediu esforços para me ajudar a chegar nesta etapa da minha vida. À minha irmã e aos meus sobrinhos que mesmo distante me enchem de alegria e força para seguir em frente.

Aos amigos e colegas, em especial, à Carolina que me acompanhou de perto durante quase toda minha caminhada na universidade, ao Bruno, ao Luís Gustavo, ao Samuel, à Laura, à Priscila, ao Roberto, ao Victor e ao Thiago pelo companheirismo, pelo apoio, pelos incentivos e por proporcionarem momentos de diversão tanto dentro quanto fora da UFJF.

Ao professor, orientador e amigo Marcel pelo apoio e ajuda sem os quais seria impossível finalizar este trabalho. Ao professor e coordenador do curso de Estatística Clécio pelo esforço em me ajudar a concluir o curso burocraticamente. E aos professores Ronaldo, Camila, André Hallack, Ana Paula Sobral, Henrique e Joaquim pelas belas aulas ministradas.

“Statistical thinking will one day be as necessary for efficient citizenship as the ability to read or write”.

H. G. Wells

Sumário

Lista de Figuras	7
Lista de Tabelas	8
1 Introdução	10
2 O PROALFA	13
2.1 População alvo e população amostrada	13
2.2 Planos amostrais	14
2.3 Plano amostral e população alvo do PROALFA 2007	14
2.4 Plano amostral e população alvo do PROALFA 2008	16
2.5 Plano amostral e população alvo do PROALFA 2009	17
3 Metodologia	19
3.1 Médias, erros padrão e intervalos de confiança	19
3.2 Modelos de regressão múltipla	21
4 Resultados e análise	25
4.1 Médias	25
4.2 Modelos de regressão	28
5 Conclusão	35
Referências Bibliográficas	36

Lista de Figuras

2.1	Níveis de estratificação	15
2.2	Níveis de estratificação para as SREs que possuem pelo menos um município com mais de 200 mil habitantes	15

Lista de Tabelas

4.1	Proficiências médias (em Língua Portuguesa) estimadas e seus respectivos erros padrão para rede Estadual de Minas Gerais	25
4.2	Proficiências médias em Língua Portuguesa estimadas e seus respectivos erros padrão para rede Municipal de Minas Gerais	26
4.3	Proficiências médias (em Língua Portuguesa) estimadas e seus respectivos erros padrão para rede Estadual em 5 SREs de Minas Gerais (considerando o Plano Amostral)	26
4.4	Proficiências médias (em Língua Portuguesa) estimadas e seus respectivos erros padrão para rede Municipal em 5 SREs de Minas Gerais (considerando o Plano Amostral)	27
4.5	Proficiências médias (em Língua Portuguesa) estimadas e seus respectivos erros padrão para rede Estadual em 5 SREs de Minas Gerais (não considerando o Plano Amostral)	27
4.6	Proficiências médias (em Língua Portuguesa) estimadas e seus respectivos erros padrão para rede Municipal em 5 SREs de Minas Gerais (não considerando o Plano Amostral)	28
4.7	Modelos iniciais estimados para explicação da proficiência em Língua Portuguesa da rede Estadual de 2007	29
4.8	Modelos finais estimados para explicação da proficiência em Língua Portuguesa da rede Estadual de 2007	29
4.9	Modelos iniciais estimados para explicação da proficiência em Língua Portuguesa da rede Municipal de 2007	30
4.10	Modelos finais estimados para explicação da proficiência em Língua Portuguesa da rede Municipal de 2007	30
4.11	Modelos iniciais estimados para explicação da proficiência em Língua Portuguesa da rede Estadual de 2008	31

4.12 Modelos finais estimados para explicação da proficiência em Língua Portuguesa da rede Estadual de 2008	31
4.13 Modelos iniciais estimados para explicação da proficiência em Língua Portuguesa da rede Municipal de 2008	32
4.14 Modelos finais estimados para explicação da proficiência em Língua Portuguesa da rede Municipal de 2008	32
4.15 Modelos iniciais estimados para explicação da proficiência em Língua Portuguesa da rede Estadual de 2009	33
4.16 Modelos finais estimados para explicação da proficiência em Língua Portuguesa da rede Estadual de 2009	33
4.17 Modelos iniciais estimados para explicação da proficiência em Língua Portuguesa da rede Municipal de 2009	34
4.18 Modelos finais estimados para explicação da proficiência em Língua Portuguesa da rede Municipal de 2009	34

1 Introdução

Os resultados das pesquisas de avaliação educacional em larga escala, que visam investigar o desempenho de alunos e escolas em relação às suas disciplinas curriculares (frequentemente Língua Portuguesa e Matemática), vem sendo amplamente utilizados para analisar a qualidade do sistema educacional no Brasil e para assistir os ministérios e secretarias de educação na definição de políticas públicas nesta área (FRANCO, BROOKE & ALVES, 2008).

Muitas destas avaliações, como a Avaliação Nacional da Educação Básica (ANEB), que é um dos programas que compõem o Sistema de Avaliação da Educação Básica (SAEB), e o Programa de Avaliação da Alfabetização (PROALFA), que é um dos programas que compõem o Sistema Mineiro de Avaliação da Educação Pública (SIMAVE), tem alguma(s) de suas etapas realizadas por amostragem através da adoção de planos amostrais complexos para a seleção da amostra de alunos a serem avaliados (VIEIRA, 2007; VIEIRA & SOUZA, 2008; VIEIRA & SOUZA, 2009; PESTANA, 1999). Os esquemas de seleção adotados são considerados complexos, pois são utilizados métodos de estratificação, conglomeração, seleção em múltiplos estágios e/ou probabilidades de seleção desiguais no sorteio da amostra.

Os motivos para se realizar essas avaliações por meio de planos amostrais complexos são, entre outros, limitações financeiras dos órgãos financiadores, facilitação logística da aplicação das provas e a possibilidade da produção de estimativas com níveis de precisão bastante satisfatórios. Além disso, a limitação dos recursos disponíveis para a realização das pesquisas poderia aumentar a possibilidade de ocorrência de erros não amostrais caso todas as pesquisas educacionais fossem realizadas de forma censitária (VICENTE et al, 2001, p. 20-21).

Os principais objetivos desta monografia são: (i) comparar a eficiência dos diferentes planos amostrais adotados para a seleção das amostras das pesquisas do PROALFA dos anos de 2007, 2008 e 2009, que buscaram avaliar o desempenho em Língua Portuguesa dos alunos dos 2º e 4º anos do ensino fundamental da rede pública de ensino de Minas Gerais dos respectivos anos; e (ii) avaliar os possíveis impactos da desconsideração do

plano amostral nos procedimentos de inferência a partir dos dados coletados, ou seja, considerando que o plano amostral utilizado tivesse sido amostragem aleatória simples em cada estrato.

Com estes objetivos, inicialmente foram estimados, para cada ano em que a pesquisa foi realizada, os parâmetros relativos à proficiência média dos alunos em Língua Portuguesa do 4º ano do ensino fundamental. Além disso, foram ajustados modelos de regressão múltipla tendo como variável resposta a proficiência dos alunos e como covariáveis algumas das outras variáveis que foram levantadas pela pesquisa. Como estratégia, para cada análise conduzida, considerou-se o plano amostral adotado para a seleção da amostra e, em seguida, desconsiderou-se algumas das características do plano amostral.

Sendo assim, as análises foram conduzidas de duas maneiras: (i) comparou-se a eficiência das estimativas obtidas para os três anos em que a pesquisa foi conduzida, através de métodos de estimação que consideram o plano amostral; (ii) avaliou-se os impactos dos planos amostrais complexos adotados, a partir da comparação de estimativas de variâncias produzidas desconsiderando algumas das características do plano amostral com estimativas produzidas considerando as características do plano amostral adotado.

Cabe ressaltar que os objetivos desta monografia são metodológicos. Desta forma, não buscou-se ajustar os melhores modelos aos dados amostrais analisados. Sendo assim, não foi objetivo deste trabalho o estudo com profundidade sobre como a proficiência em Língua Portuguesa se comporta em relação às variáveis explicativas consideradas nos modelos de regressão que foram ajustados. O principal objetivo no ajuste dos modelos de regressão foi observar o comportamento das estimativas dos coeficientes e dos erros padrão a fim de compará-las quando os estimamos considerando e não considerando os planos amostrais, e as principais conseqüências da desconsideração das características do plano amostral adotado, por exemplo, na seleção das covariáveis incluídas nos modelos finais.

Na Seção 2, é apresentada uma descrição das principais características do PROALFA, incluindo seus objetivos, população alvo e os planos amostrais utilizados. Em seguida, na Seção 3, é realizada uma descrição da metodologia de inferência para dados amostrais complexos que foi utilizada para o cálculo de todas as estimativas incluídas nesta monografia. Os resultados, bem como as análises dos mesmos são apresentados na

Seção 4, enquanto na Seção 5, são apresentadas conclusões.

2 O PROALFA

O PROALFA é um dos programas que compõem o SIMAVE e tem como principal objetivo a avaliação do desempenho em Língua Portuguesa dos alunos devidamente matriculados nos 2º, 3º e 4º anos do ensino fundamental das escolas públicas das redes municipal e estadual de ensino em Minas Gerais. Esta avaliação é composta de um teste de conhecimentos em Língua Portuguesa e um sucinto questionário contendo questões apenas sobre a idade do aluno, idade que entrou na escola, se fez pré-escola e quantas vezes foi reprovado durante a vida escolar. Nos anos de 2008 e 2009 foi incluída uma pergunta sobre o gênero do aluno.

A partir dos testes de conhecimento do PROALFA são construídas medidas com o auxílio da teoria da resposta ao item (TRI) que, neste contexto, são denominadas proficiências dos alunos (VIEIRA, 2001, p. 96; ANDRADE, 2000).

Para o 3º ano, o PROALFA tem caráter censitário, ou seja, todos os alunos são avaliados. Para os 2º e 4º anos, o PROALFA é realizado por amostragem, através da adoção de planos amostrais complexos. Como o foco desta monografia é o estudo dos efeitos de planos amostrais complexos, não consideraremos os dados coletados de forma censitária para o 3º ano. Desta forma, são descritos nesta seção os planos amostrais e a população alvo para as avaliações do 2º e 4º ano.

Cabe ressaltar que os dados utilizados nesta monografia foram fornecidos pelo Centro de Avaliação e Políticas Públicas da Educação (CAEd), instituição vinculada à Universidade Federal de Juiz de Fora (UFJF), que foi responsável pela realização das pesquisas do PROALFA entre os anos de 2006 a 2010.

2.1 População alvo e população amostrada

A população alvo é a população de interesse de uma pesquisa, ou seja, é a população para a qual se deseja inferir os resultados obtidos. Entretanto, nem sempre toda a população de interesse estará disponível e a amostra é colhida na parte da população que é acessível, também chamada de população amostrada ou população de referência (VICENTE, 2001;

BOLFARINE & BUSSAB, 2005).

2.2 Planos amostrais

Planos amostrais devem permitir a identificação do universo (conjunto de todos os elementos amostrais da população) a que se refere o estudo, a população de interesse (público alvo de onde será extraído a amostra) e os elementos amostrais (objetos de onde serão efetivamente retiradas as medidas para análise).

Levantamentos que têm como objetivo a realização de inferência estatística exigem a adoção de planos amostrais probabilísticos, ou seja, devem garantir que todos os elementos da população tenham uma probabilidade não nula de seleção e, conseqüentemente, garantam a possibilidade de associarmos uma probabilidade de seleção para todas as amostras possíveis (VICENTE, 2001; BOLFARINE & BUSSAB, 2005). Sendo assim, seja uma população de tamanho fixo denotada por $U = \{1, 2, \dots, N\}$ e uma amostra de U dada por $s = \{k_1, k_2, \dots, k_n\}$ com uma probabilidade de seleção denotada por $p(s)$. Segundo Bolfarine & Bussab (2005), a forma como essas probabilidades são associadas é que irá definir um planejamento amostral.

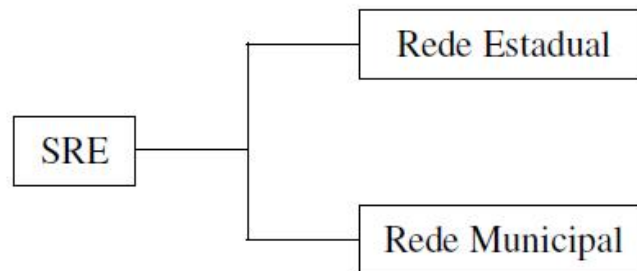
2.3 Plano amostral e população alvo do PROALFA 2007

Neste estudo, a população alvo considerada são os alunos do 2º e do 4º do ensino fundamental matriculados nas redes estadual e municipal do estado de Minas Gerais. A população amostrada, por sua vez, é constituída de alunos das escolas da rede estadual e de alunos das escolas da rede municipal (localizadas em municípios que aceitaram participar da pesquisa). Portanto, todas as inferências estatísticas realizadas poderão ser generalizadas para esta população amostrada.

O plano amostral utilizado no PROALFA 2007 considera uma amostragem aleatória estratificada por conglomerados em um estágio, onde as escolas são as unidades primárias de amostragem (UPAs) e, uma vez selecionada uma destas escolas, todos os alunos matriculados na série avaliada e presentes no dia da avaliação foram testados em

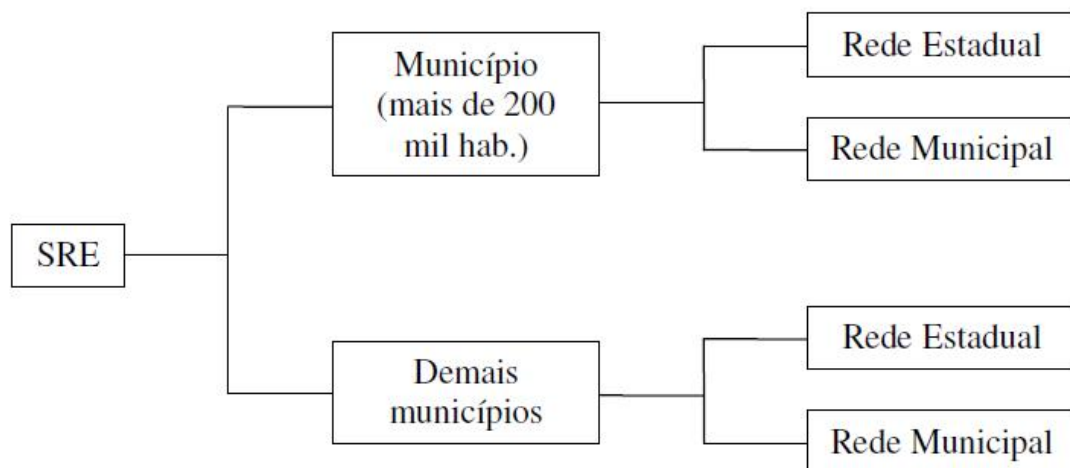
Língua Portuguesa. Além disso, foram definidos alguns domínios de interesse, ou seja, grupos de alunos para os quais desejava-se publicar resultados separadamente. A estratificação foi definida a partir das áreas de abrangência das Superintendências Regionais de Ensino (SREs) e de acordo com a rede de ensino à qual a escola pertence. Sendo assim, os níveis de estratificação podem ser definidos da seguinte maneira:

Figura 2.1: Níveis de estratificação



Além disso, a Secretaria Estadual de Educação de Minas Gerais (SEE/MG) determinou que o plano amostral do PROALFA 2007 deveria permitir a “produção de estimativas para todos os municípios com população (estimada em 2006) acima de 200 mil habitantes, garantindo, ainda, pelo menos dois municípios investigados por pólo de ensino” (VIEIRA, 2007). Sendo assim, os níveis de estratificação para as SREs que possuem pelo menos um município com mais de 200 mil habitantes podem ser representados da seguinte maneira:

Figura 2.2: Níveis de estratificação para as SREs que possuem pelo menos um município com mais de 200 mil habitantes



Desta forma, o total de estratos gerados foi de 130¹ e, para cada um destes estratos, o tamanho de amostra foi calculado de forma a permitir uma posterior inferência estatística e garantir um nível de precisão semelhante para a estimação de uma medida de proficiência média em Língua Portuguesa para cada um destes estratos (VIEIRA, 2007).

Estimativas de variâncias das médias das proficiências dos estratos considerados foram utilizadas para o cálculo dos tamanhos das amostras de cada estrato. Para o cálculo destas estimativas, foram utilizados os dados do 2º ano do PROALFA de 2006 e, nos casos em que a variância não pode ser estimada (estratos não contemplados em 2006 ou estratos com amostra de escolas com tamanho unitário), utilizaram-se estimativas de estratos semelhantes no que diz respeito ao tamanho, localização geográfica e rede de ensino. Para a determinação do tamanho da amostra, duas condições foram respeitadas: em cada estrato, pelo menos, duas escolas deveriam ser avaliadas e o tamanho final da amostra deveria ser de, aproximadamente, 30.000 alunos (devido à limitação do orçamento para a pesquisa). Sendo assim, ficou estabelecido o tamanho da amostra de conglomerados como 419 escolas.

Um procedimento de amostragem com probabilidades proporcionais ao tamanho (ppt) das escolas foi utilizado para selecioná-las em cada estrato de interesse, sendo que o número de alunos matriculados na escola no ano de interesse foi utilizado como variável *proxy* do tamanho da escola. Desta maneira, o tamanho total da amostra foi de 30.069 alunos para o 2º ano do ensino fundamental e 30.074 alunos para o 4º ano do ensino fundamental.

2.4 Plano amostral e população alvo do PROALFA 2008

No PROALFA de 2008, tanto a população alvo e a população amostrada quanto o plano amostral são muito semelhantes aos de 2007. A população alvo continua sendo a mesma, ou seja, alunos do 2º e 4º do ensino fundamental matriculados nas redes estadual e municipal do estado de Minas Gerais e a população amostrada pode ter sido alterada devido à entrada de escolas da rede municipal que não haviam participado ou à saída de escolas

¹Devido à recusa de alguns municípios em participar do PROALFA em 2007, o número final de estratos considerados foi de 127.

da rede municipal que haviam participado em 2007 (VIEIRA & SOUZA, 2008).

No que diz respeito ao plano amostral, uma das diferenças do realizado em 2008 para o de 2007 é que o número final de estratos considerados foi de 126, devido à recusa de alguns municípios em participar do projeto (caso a recusa não existisse, o número total de estratos teria sido 130).

Uma estimação da variância das médias das proficiências também foi realizada para o cálculo do tamanho das amostras de cada estrato e esta se deu através da base de dados do PROALFA 2007. Além disso, foi definido que cada estrato considerado deveria possuir, pelo menos, três escolas amostradas, com o objetivo de permitir posterior inferência estatística sobre cada estrato (VIEIRA & SOUZA, 2008).

Um procedimento de amostragem com probabilidades proporcionais (ppt) ao tamanho das escolas foi utilizado para selecioná-las em cada estrato de interesse, sendo que o número de turmas na escola do ano de interesse foi utilizado como variável *proxy* do tamanho da escola. Desta forma, ficou estabelecido o tamanho da amostra de conglomerados como 555 escolas resultando em uma amostra de 33.306 alunos para o 2º ano do ensino fundamental e 523 escolas resultando em uma amostra de 33.475 alunos para o 4º ano do ensino fundamental.

2.5 Plano amostral e população alvo do PROALFA 2009

Assim como no PROALFA 2008, tanto a população alvo quanto a população amostrada são muito semelhantes aos de 2007, ou seja, a população alvo continua a mesma (alunos do 2º e 4º anos do ensino fundamental matriculados nas redes estadual e municipal do estado de Minas Gerais) e, mais uma vez, a população amostrada pode ter sido alterada pelos motivos acima descritos. Porém, o plano amostral do PROALFA 2009 é bastante distinto dos planos utilizados nos PROALFAs de 2007 e 2008 (VIEIRA & SOUZA, 2009).

O plano amostral utilizado no PROALFA 2009 considera uma amostragem aleatória estratificada por conglomerados em dois estágios, onde as escolas são as UPAs e as turmas do ano de interesse são as unidades secundárias de amostragem (USAs). Todos os alunos de uma turma selecionada que estivessem presentes no dia da avaliação deveriam

ser avaliados.

Apesar do plano amostral utilizado no PROALFA 2009 ser distinto dos utilizados nos PROALFAs de 2007 e 2008, os níveis de estratificação permaneceram os mesmos. Desta forma, o total de estratos considerados foi de 128.

Uma estimação da variância das médias das proficiências para o cálculo do tamanho das amostras de cada estrato foi realizada e esta se deu através das bases de dados dos PROALFAs de 2007 e 2008. Além disso, cada estrato foi dividido em dois sub-estratos de tamanho: escolas com, no máximo, duas turmas nos anos de interesse e escolas com mais de duas turmas. No primeiro sub-estrato, apenas uma turma foi selecionada e no segundo, duas turmas foram selecionadas.

Um procedimento de amostragem sequencial de Poisson, que é um método de amostragem ppt, foi utilizado para selecionar as escolas em cada um dos sub-estratos de tamanho. Tal procedimento teve como base a seleção das escolas com probabilidades proporcionais ao tamanho da escola, sendo que o número de turmas na escola do ano de interesse foi utilizado como variável proxy do tamanho da escola. Além disso, diferentemente dos PROALFAs de 2007 e 2008, o tamanho total da amostra do PROALFA 2009 totalizou 2024 escolas para o 2º ano do ensino fundamental e 2002 escolas para o 4º ano, em ambos os casos, o total da amostra de alunos foi de, aproximadamente, 50 mil alunos.

Quadro 2.1 - Resumo das características dos planos amostrais dos PROALFAS de 2007, 2008 e 2009

Ano	Série	Características				
		Nº de alunos	Nº de escolas	Nº de estratos	Método de seleção	Nº de estágios
2007	2º Ano	30.069	419	127	ppt	1
	4º Ano	30.074	419			
2008	2º Ano	33.306	555	126	ppt	1
	4º Ano	33.475	523			
2009	2º Ano	50.000	2024	128	ppt	2
	4º Ano	50.000	2002			

3 Metodologia

Toda a metodologia que será apresentada nesta seção foi implementada no *software* livre R versão 2.10.1 (2009) com auxílio do pacote *survey* versão 3.22-1 (LUMLEY, 2010).

3.1 Médias, erros padrão e intervalos de confiança

Consideraremos inicialmente o caso em que a seleção da amostra é realizada por amostragem aleatória simples (AAS) dentro de cada estrato (ou seja, amostragem estratificada simples – AES) e que as médias populacionais das proficiências em cada estrato h são os nossos parâmetros de interesse.

Neste caso, tais parâmetros são estimados por

$$\bar{x}_h = \sum_{i=1}^{n_h} x_{hi}/n_h \quad (3.1)$$

onde \bar{x}_h aqui a média amostral da proficiência em Língua Portuguesa do h -ésimo estrato, x_{hi} a proficiência do i -ésimo aluno do h -ésimo estrato e n_h o número total de alunos avaliados no h -ésimo estrato (ou seja, o tamanho da amostra de alunos no estrato h).

Para o caso da AES, sobretudo para situações em que o tamanho da amostra é suficientemente grande, podemos dizer que (BOLFARINE & BUSSAB, 2005)

$$\bar{x}_h \sim N \left(\mu_h, \frac{\sigma_h^2}{n_h} \right) \quad (3.2)$$

onde μ_h representa a média populacional da proficiência do h -ésimo estrato e σ_h^2 a variância populacional da proficiência do h -ésimo estrato; portanto o erro padrão deste estimador é aproximadamente dado por σ_h/n_h e pode ser estimado por s_h/n_h , onde

$$s_h = \sqrt{\sum_{i=1}^{n_h} \frac{(x_{hi} - \bar{x}_h)^2}{n_h - 1}} \quad (3.3)$$

é um estimador de σ_h

A partir destas expressões, podemos construir os intervalos de confiança considerando AES para as médias das proficiências em cada estrato h , a partir de

$$\left(\bar{x}_h - z_{\alpha/2} \frac{s_h}{\sqrt{n_h}} < \mu_h < \bar{x}_h + z_{\alpha/2} \frac{s_h}{\sqrt{n_h}} \right) \quad (3.4)$$

onde onde $z_{\alpha/2}$ é obtido a partir da distribuição normal padrão e $(1 - \alpha)$ é o nível de confiança que, no nosso caso, será de 95%.

Em um segundo momento, para estimarmos as médias populacionais das proficiências em cada estrato h considerando qualquer plano amostral, podemos utilizar de forma mais geral o estimador de Horvitz-Thompson (HORVITZ & THOMPSON, 1952), dado por

$$\bar{x}_h^{HT} = \frac{1}{n_h} \sum_{i=1}^{n_h} \frac{x_{hi}}{\pi_{hi}} \quad (3.5)$$

onde π_{hi} representa a probabilidade de inclusão do i -ésimo aluno do h -ésimo estrato.

A seguir, apresentamos uma revisão sobre a estimação do erro padrão deste estimador, que é necessário para a construção de intervalos de confiança. Porém, para alguns planos amostrais complexos, não existe uma forma analítica direta para calcularmos este erro padrão. Desta forma, são utilizados normalmente outros métodos, tais como a linearização de Taylor (também chamado de método delta) e os métodos de replicação de Jackknife e bootstrap (WOLTER, 1985; VIEIRA, 2001). Neste trabalho, utilizamos o método da linearização de Taylor, pois este método está implementado no pacote *survey* do *software* R.

O método da linearização de Taylor vem sendo utilizado com frequência para se obter aproximações de estimadores não-lineares por estimadores lineares (VIEIRA, 2001). Este método tem como base expansão em séries de Taylor do estimador pontual do parâmetro de interesse, em torno do verdadeiro parâmetro e considera apenas as aproximações de primeira ordem, de acordo com Vieira (2009), onde maiores informações podem ser encontradas. Na subseção 3.2, será apresentada a expressão do estimador de linearização da variância para o caso em que o parâmetro de interesse é um coeficiente de

um modelo de regressão.

Intervalos de confiança podem ser construídos através de (VIEIRA, 2001)

$$(\bar{x}_h^{HT} - z_{\alpha/2}EP_L(\bar{x}_h^{HT}) < \mu_h < \bar{x}_h^{HT} + z_{\alpha/2}EP_L(\bar{x}_h^{HT})) \quad (3.6)$$

onde $EP_L(\bar{x}_h^{HT})$ é o erro padrão de \bar{x}_h^{HT} calculado a partir do método de linearização de Taylor.

3.2 Modelos de regressão múltipla

Outro objetivo deste trabalho é comparar a eficiência dos planos amostrais dos PROALFAs de 2007, 2008 e 2009 através da comparação de parâmetros de modelos de regressão múltipla estimados sem considerar o plano amostral e parâmetros dos mesmos modelos estimados levando em consideração as características do plano.

Um modelo de regressão múltipla é dado por:

$$y_i = \beta_0 + \sum_{j=1}^k \beta_j x_{ij} + \varepsilon_i, i = 1, 2, \dots, n \quad (3.7)$$

onde, β_0 representa o intercepto do modelo, β_j o j -ésimo coeficiente de regressão, x_{ij} denota a i -ésima observação da j -ésima variável explicativa, j varia de 1 até k , k denota o número de covariáveis consideradas, ε_i é chamado de termo de erro (ou resíduo) do i -ésimo aluno, i varia de 1 até n , e n denota o tamanho da amostra de alunos.

Esta equação pode ser escrita em sua forma matricial como:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (3.8)$$

$$\text{onde } \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} \text{ e } \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

Para que este modelo possa ser ajustado, alguns pressupostos devem ser atendidos, entre os quais devemos destacar que os resíduos devem ser independentes e identicamente distribuídos segundo uma distribuição normal com média 0 (zero) e variância constante (σ^2), o que implica que as observações y_i são normalmente e independentemente distribuídas com média dada por $\beta_0 + \sum_{j=1}^k \beta_j x_{ij}$ e variância σ^2 , além disso, não deve existir colinearidade perfeita entre quaisquer das variáveis independentes (KMENTA, 1988).

Para estimarmos os valores de β quando não levamos em consideração o plano amostral, podemos utilizar o estimador de mínimos quadrados ordinários (KMENTA, 1988). Tal estimador, em sua forma matricial, é dado por:

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \quad (3.9)$$

onde \mathbf{X}' é a matriz transposta de \mathbf{X} e $(\mathbf{X}'\mathbf{X})^{-1}$ é a matriz inversa de $(\mathbf{X}'\mathbf{X})$. O erro padrão de $\hat{\beta}$, neste caso, é dado pela raiz quadrada dos termos da diagonal da matriz dada por $(\mathbf{X}'\mathbf{X})^{-1}\hat{\sigma}^2$, sendo $\hat{\sigma}^2$ um estimador de σ^2 (KMENTA, 1988).

Quando os dados a serem analisados foram coletados a partir de planos amostrais complexos, o pressuposto de independência entre as observações é normalmente violado. Sendo assim, para a estimação dos parâmetros de interesse devemos adotar métodos que não levem em consideração tal pressuposto e que considerem as características do plano amostral. Por exemplo, quando desejamos estimar os valores de β considerando todas as características do plano amostral, o método de estimação mais frequentemente adotado é o método da máxima pseudo-verossimilhança (MPV) que tem como base incorporação dos pesos amostrais w_i de cada indivíduo i , definidos como o inverso das probabilidades de inclusão π_i . O método de MPV está implementado no pacote *survey* do R e sua expressão para o estimador de β é dada por

$$\hat{\beta}_{MPV} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{y} \quad (3.10)$$

onde $\mathbf{W} = \text{diag}[(w_1, \dots, w_n)]$.

Para maiores informações sobre o método de MPV, pode-se consultar Vieira (2009), por exemplo. No contexto de dados amostrais complexos, para estimar o erro padrão deste estimador estaremos utilizando o método da linearização de Taylor. Através

deste método, obtemos a seguinte expressão da variância assintótica do estimador de MPV de β (VIEIRA, 2001)

$$V_L(\hat{\beta}_{MPV}) = (\mathbf{X}'\mathbf{X})^{-1}V\left(\sum_{i=1}^n w_i x_i \varepsilon_i\right) (\mathbf{X}'\mathbf{X})^{-1} \quad (3.11)$$

onde $V\left(\sum_{i=1}^n w_i x_i \varepsilon_i\right) = \sum_{i=1}^n \sum_{j=1}^n \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \varepsilon_i x_i x_j' \varepsilon_j$, com π_{ij} representando a probabilidade de inclusão conjunta dos alunos i e j .

Esta variância pode ser estimada, por sua vez, da seguinte maneira (VIEIRA, 2001)

$$\hat{V}_L(\hat{\beta}_{MPV}) = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\hat{V}\left(\sum_{i=1}^n w_i x_i \varepsilon_i\right) (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1} \quad (3.12)$$

onde $V\left(\sum_{i=1}^n w_i x_i \varepsilon_i\right) = \sum_{i=1}^n \sum_{j=1}^n (w_i w_j - w_{ij}) \hat{\varepsilon}_i x_i x_j' \hat{\varepsilon}_j$.

Em ambos os casos – considerando e não considerando o plano amostral – utilizamos um método de construção dos modelos denominado *backward* que consiste em inserir no modelo todas as variáveis disponíveis e, em seguida, retirar, uma por uma, as variáveis que não apresentarem significância. Um método de construção automático foi escolhido pois é importante para compararmos os modelos, uma vez que pesquisadores diferentes poderiam chegar a modelos diferentes caso um método automático não fosse utilizado.

Neste trabalho, o nível de significância adotado é de 5%. Sendo assim, denominamos os modelos com todas as variáveis disponíveis inseridas no modelo como sendo modelos iniciais e os modelos com apenas as variáveis significativas como sendo modelos finais.

As variáveis disponíveis para construção dos modelos dos PROALFAs de 2007 encontram-se descritas no quadro 3.2.

Quadro 3.1 - Descrição das variáveis disponíveis no PROALFA 2007 para construção dos modelos de regressão

Variáveis	Descrição
Proficiência	Proficiência em Língua Portuguesa
Idade	Idade do aluno no dia em que respondeu ao questionário.
Idade Escola	Idade do aluno quando entrou na escola.
Repetiu	Número de vezes que o aluno repetiu de ano.
Pré-Escola	Variável <i>dummy</i> . Indica se o aluno fez pré-escola. Codificação: 0 - não; 1 - sim.

As variáveis disponíveis para construção dos modelos dos PROALFAs de 2008 e 2009, por sua vez, encontram-se descritas no quadro 3.2.

Quadro 3.2 - Descrição das variáveis disponíveis nos PROALFAs 2008 e 2009 para construção dos modelos de regressão

Variáveis	Descrição
Proficiência	Proficiência em Língua Portuguesa
Gênero Masculino	Variável <i>dummy</i> . Gênero do aluno. Codificação: 0 - feminino; 1 - masculino.
Idade	Idade do aluno no dia em que respondeu ao questionário.
Idade Escola	Idade do aluno quando entrou na escola.
Repetiu	Número de vezes que o aluno repetiu de ano.
Pré-Escola	Variável <i>dummy</i> . Indica se o aluno fez pré-escola. Codificação: 0 - não; 1 - sim.

4 Resultados e análise

4.1 Médias

Nas tabelas 4.1 e 4.2, abaixo, encontram-se as estimativas das proficiências médias (em Língua Portuguesa) e seus respectivos erros padrão e intervalos de confiança de 95% para as redes Estadual e Municipal de Minas Gerais. Analisando-as, percebemos que quando o plano amostral não é considerado em sua forma integral para o cálculo das estimativas, o erro padrão é sempre subestimado. Além disso, os erros padrão das estimativas foram sempre menores nos resultados de 2009 quando considerado o plano amostral, fato que era esperado, pois, teoricamente, o plano amostral realizado em 2009 é mais eficiente que os de 2007 e 2008.

Tabela 4.1: Proficiências médias (em Língua Portuguesa) estimadas e seus respectivos erros padrão para rede Estadual de Minas Gerais

Ano	Considerando o Plano Amostral				Não Considerando o Plano Amostral			
	Média	Erro Padrão	IC - LI*	IC - LS*	Média	Erro Padrão	IC - LI*	IC - LS*
2007	541,4	3,3	534,9	547,9	545,1	0,7	543,8	546,4
2008	574,2	4,2	565,9	582,5	578,0	0,7	576,7	579,4
2009	565,3	2,2	561,1	569,6	566,4	0,6	565,3	567,5

*IC - LI e IC - LS representam o limite inferior e superior do intervalo de confiança, respectivamente.

Tabela 4.2: Proficiências médias em Língua Portuguesa estimadas e seus respectivos erros padrão para rede Municipal de Minas Gerais

Ano	Considerando o Plano Amostral				Não Considerando o Plano Amostral			
	Média	Erro Padrão	IC - LI*	IC - LS*	Média	Erro Padrão	IC - LI*	IC - LS*
2007	532,8	4,8	523,3	542,2	531,8	0,9	530,1	533,5
2008	544,0	3,1	538,0	550,0	543,0	0,7	541,7	544,4
2009	542,5	1,6	539,4	545,6	543,7	0,6	542,5	544,8

*IC - LI e IC - LS representam o limite inferior e superior do intervalo de confiança, respectivamente.

Já nas tabelas 4.3, 4.4, 4.5 e 4.6, a seguir, encontram-se as estimativas das proficiências médias (em Língua Portuguesa) e seus respectivos erros padrão e intervalos de confiança de 95% para 5 SREs divididas por rede do estado de Minas Gerais. Nelas podemos perceber que, de maneira geral, o efeito de subestimação dos erros padrão das estimativas quando não considerado o plano amostral e a diminuição dos erros padrão ao longo do tempo quando considerado o plano amostral também acontecem.

Tabela 4.3: Proficiências médias (em Língua Portuguesa) estimadas e seus respectivos erros padrão para rede Estadual em 5 SREs de Minas Gerais (considerando o Plano Amostral)

SRE	2007		2008		2009	
	Média	Erro Padrão	Média	Erro Padrão	Média	Erro Padrão
1	544,7	26,0	605,6	23,7	551,0	7,2
2	554,4	30,6	566,7	23,4	560,4	10,8
3	539,4	14,1	604,8	13,0	581,4	8,1
4	565,6	23,7	582,8	44,7	567,0	28,8
5	513,5	17,2	553,4	19,3	557,4	11,6

Tabela 4.4: Proficiências médias (em Língua Portuguesa) estimadas e seus respectivos erros padrão para rede Municipal em 5 SREs de Minas Gerais (considerando o Plano Amostral)

SRE	2007		2008		2009	
	Média	Erro Padrão	Média	Erro Padrão	Média	Erro Padrão
1	512,1	4,3	553,7	15,4	505,9	11,2
2	447,3	30,4	533,1	35,0	540,8	16,2
3	520,4	5,0	558,6	13,8	560,0	6,8
4	566,7	19,2	570,7	26,2	555,9	10,8
5	509,6	51,1	566,2	33,2	551,4	7,2

Tabela 4.5: Proficiências médias (em Língua Portuguesa) estimadas e seus respectivos erros padrão para rede Estadual em 5 SREs de Minas Gerais (não considerando o Plano Amostral)

SRE	2007		2008		2009	
	Média	Erro Padrão	Média	Erro Padrão	Média	Erro Padrão
1	528,3	5,9	609,0	6,6	549,6	4,1
2	541,4	4,9	571,5	5,3	562,5	4,3
3	543,9	3,3	599,5	3,4	580,8	3,1
4	575,7	5,0	604,1	7,0	565,5	5,8
5	543,9	7,2	549,2	5,5	557,8	4,7

Tabela 4.6: Proficiências médias (em Língua Portuguesa) estimadas e seus respectivos erros padrão para rede Municipal em 5 SREs de Minas Gerais (não considerando o Plano Amostral)

SRE	2007		2008		2009	
	Média	Erro Padrão	Média	Erro Padrão	Média	Erro Padrão
1	515,6	4,7	571,8	5,4	506,4	4,9
2	435,4	11,4	529,1	6,7	533,3	5,8
3	512,4	9,1	552,1	4,0	560,6	3,1
4	571,4	5,0	580,0	4,3	555,1	4,9
5	535,1	20,0	547,6	11,1	549,9	4,0

4.2 Modelos de regressão

Assim como nas estimativas das médias, os erros padrão dos coeficientes dos modelos de regressão tanto iniciais quanto finais de todos os anos e redes de ensino também apresentam uma característica de subestimação quando não considerados o plano amostral. Porém, em todos os modelos, os coeficientes estimados quando considerado e não considerado o plano amostral apresentam o mesmo sinal (ver tabelas 4.7 a 4.18).

Nos modelos da rede Estadual de 2007, representados nas tabelas 4.7 e 4.8, percebemos que os modelos finais para os dois tipos de estimação são iguais, ou seja, neste caso o fato de os erros padrão serem maiores quando considerado o plano amostral não influenciou na significância dos coeficientes.

Tabela 4.7: Modelos iniciais estimados para explicação da proficiência em Língua Portuguesa da rede Estadual de 2007

Covariáveis	Considerando o Plano Amostral			Não Considerando o Plano Amostral		
	Coefficiente	Erro Padrão	p-valor	Coefficiente	Erro Padrão	p-valor
Intercepto	599,93	27,97	<0,01	603,00	11,78	<0,01
Idade	-1,59	2,87	0,58	-0,85	1,22	0,49
Idade Escola	-8,13	1,87	<0,01	-9,68	0,79	<0,01
Repetiu	-24,80	3,22	<0,01	-29,63	1,55	<0,01
Pré-Escola	16,51	3,47	<0,01	19,04	1,77	<0,01

Tabela 4.8: Modelos finais estimados para explicação da proficiência em Língua Portuguesa da rede Estadual de 2007

Covariáveis	Considerando o Plano Amostral			Não Considerando o Plano Amostral		
	Coefficiente	Erro Padrão	p-valor	Coefficiente	Erro Padrão	p-valor
Intercepto	585,49	12,06	<0,01	594,82	5,03	<0,01
Idade	-	-	-	-	-	-
Idade Escola	-8,21	1,85	<0,01	-9,66	0,78	<0,01
Repetiu	-26,28	2,75	<0,01	-30,55	1,12	<0,01
Pré-Escola	16,86	3,44	<0,01	19,34	1,75	<0,01

Já nos modelos da rede Municipal de 2007, representados nas tabelas 4.9 e 4.10, percebemos que os modelos finais para os dois tipos de estimação não são iguais, ou seja, neste caso o fato de os erros padrão serem maiores quando considerado o plano amostral influenciou na significância dos coeficientes. A variável Pré-Escola foi significativa no modelo que não considera o plano amostral e não foi significativa no modelo que considera o plano amostral.

Tabela 4.9: Modelos iniciais estimados para explicação da proficiência em Língua Portuguesa da rede Municipal de 2007

Covariáveis	Considerando o Plano Amostral			Não Considerando o Plano Amostral		
	Coefficiente	Erro Padrão	p-valor	Coefficiente	Erro Padrão	p-valor
Intercepto	606,42	22,34	<0,01	594,95	14,43	<0,01
Idade	-0,55	2,31	0,81	-0,94	1,44	0,51
Idade Escola	-9,04	2,63	<0,01	-8,57	1,14	<0,01
Repetiu	-22,69	3,07	<0,01	-27,43	1,82	<0,01
Pré-Escola	-5,69	8,11	0,48	10,07	2,29	<0,01

Tabela 4.10: Modelos finais estimados para explicação da proficiência em Língua Portuguesa da rede Municipal de 2007

Covariáveis	Considerando o Plano Amostral			Não Considerando o Plano Amostral		
	Coefficiente	Erro Padrão	p-valor	Coefficiente	Erro Padrão	p-valor
Intercepto	589,78	16,99	<0,01	586,83	7,25	<0,01
Idade	-	-	-	-	-	-
Idade Escola	-7,78	2,77	0,01	-8,68	1,13	<0,01
Repetiu	-23,02	2,24	<0,01	-28,29	1,25	<0,01
Pré-Escola	-	-	-	9,92	2,28	<0,01

Um efeito parecido com o que ocorreu nos modelos para a rede Municipal de 2007 ocorreu nos modelos para a rede Estadual e Municipal de 2008, ou seja, uma

variável que foi significante quando o plano amostral não é considerado passou a não ser quando consideramos o plano amostral na estimação. No modelo da rede Estadual de 2008 (tabelas 4.11 e 4.12), notamos que a variável Idade passou a ser não significante quando considerado o plano amostral.

Tabela 4.11: Modelos iniciais estimados para explicação da proficiência em Língua Portuguesa da rede Estadual de 2008

Covariáveis	Considerando o Plano Amostral			Não Considerando o Plano Amostral		
	Coefficiente	Erro Padrão	p-valor	Coefficiente	Erro Padrão	p-valor
Intercepto	664,91	32,39	<0,01	681,29	12,53	<0,01
Gênero Masculino	-10,80	2,04	<0,01	-10,75	1,49	<0,01
Idade	-4,84	3,57	0,18	-7,53	1,33	<0,01
Idade Escola	-7,29	3,66	0,05	-5,48	0,99	<0,01
Repetiu	-17,91	5,65	<0,01	-12,75	1,72	<0,01
Pré-Escola	14,14	5,98	0,02	12,72	2,11	<0,01

Tabela 4.12: Modelos finais estimados para explicação da proficiência em Língua Portuguesa da rede Estadual de 2008

Covariáveis	Considerando o Plano Amostral			Não Considerando o Plano Amostral		
	Coefficiente	Erro Padrão	p-valor	Coefficiente	Erro Padrão	p-valor
Intercepto	625,21	21,15	<0,01	681,29	12,53	<0,01
Gênero Masculino	-10,89	2,06	<0,01	-10,75	1,49	<0,01
Idade	-	-	-	-7,53	1,33	<0,01
Idade Escola	-8,06	3,53	0,02	-5,48	0,99	<0,01
Repetiu	-21,87	4,10	<0,01	-12,75	1,72	<0,01
Pré-Escola	14,98	6,03	0,01	12,72	2,11	<0,01

Já nos modelos para a rede Municipal de 2008 (tabelas 4.13 e 4.14) a variável que deixou de ser significativa quando considerado o plano amostral foi Pré-Escola. A variável Idade não apresentou significância tanto quando consideramos quando não consideramos o plano amostral.

Tabela 4.13: Modelos iniciais estimados para explicação da proficiência em Língua Portuguesa da rede Municipal de 2008

Covariáveis	Considerando o Plano Amostral			Não Considerando o Plano Amostral		
	Coeficiente	Erro Padrão	p-valor	Coeficiente	Erro Padrão	p-valor
Intercepto	579,27	32,24	<0,01	583,11	10,42	<0,01
Gênero Masculino	-14,70	1,90	<0,01	-15,59	1,46	<0,01
Idade	-0,66	3,55	0,85	-0,98	1,09	0,37
Idade Escola	-4,64	2,23	0,04	-4,17	0,86	<0,01
Repetiu	-14,14	3,78	<0,01	-12,25	1,41	<0,01
Pré-Escola	13,41	8,10	0,10	7,00	1,97	<0,01

Tabela 4.14: Modelos finais estimados para explicação da proficiência em Língua Portuguesa da rede Municipal de 2008

Covariáveis	Considerando o Plano Amostral			Não Considerando o Plano Amostral		
	Coeficiente	Erro Padrão	p-valor	Coeficiente	Erro Padrão	p-valor
Intercepto	591,49	12,68	<0,01	575,46	5,63	<0,01
Gênero Masculino	-15,06	1,82	<0,01	-15,75	1,45	<0,01
Idade	-	-	-	-	-	-
Idade Escola	-5,71	2,25	0,01	-4,34	0,85	<0,01
Repetiu	-15,46	2,26	<0,01	-13,12	1,01	<0,01
Pré-Escola	-	-	-	6,88	1,96	<0,01

Nos modelos estimados para a rede Estadual de 2009 (tabelas 4.15 e 4.16), o fato de os erros padrão dos coeficientes serem maiores quando consideramos o plano amostral não influenciou na significância dos mesmos, ou seja, todos os coeficientes foram significantes em ambos os modelos.

Tabela 4.15: Modelos iniciais estimados para explicação da proficiência em Língua Portuguesa da rede Estadual de 2009

Covariáveis	Considerando o Plano Amostral			Não Considerando o Plano Amostral		
	Coeficiente	Erro Padrão	p-valor	Coeficiente	Erro Padrão	p-valor
Intercepto	693,13	20,65	<0,01	697,09	12,38	<0,01
Gênero Masculino	-14,76	1,58	<0,01	-15,62	1,28	<0,01
Idade	-9,42	1,94	<0,01	-8,78	1,28	<0,01
Idade Escola	-6,81	2,28	<0,01	-7,73	1,00	<0,01
Repetiu	-21,61	2,92	<0,01	-22,25	1,84	<0,01
Pré-Escola	14,43	4,60	<0,01	11,00	1,86	<0,01

Tabela 4.16: Modelos finais estimados para explicação da proficiência em Língua Portuguesa da rede Estadual de 2009

Covariáveis	Considerando o Plano Amostral			Não Considerando o Plano Amostral		
	Coeficiente	Erro Padrão	p-valor	Coeficiente	Erro Padrão	p-valor
Intercepto	693,13	20,65	<0,01	697,09	12,38	<0,01
Gênero Masculino	-14,76	1,58	<0,01	-15,62	1,28	<0,01
Idade	-9,42	1,94	<0,01	-8,78	1,28	<0,01
Idade Escola	-6,81	2,28	<0,01	-7,73	1,00	<0,01
Repetiu	-21,61	2,92	<0,01	-22,25	1,84	<0,01
Pré-Escola	14,43	4,60	<0,01	11,00	1,86	<0,01

Já nos modelos da rede Municipal de 2009 representados nas tabelas 4.17 e 4.18, percebemos que a variável Idade passou a ser não significativa quando consideramos o plano amostral.

Tabela 4.17: Modelos iniciais estimados para explicação da proficiência em Língua Portuguesa da rede Municipal de 2009

Covariáveis	Considerando o Plano Amostral			Não Considerando o Plano Amostral		
	Coefficiente	Erro Padrão	p-valor	Coefficiente	Erro Padrão	p-valor
Intercepto	624,46	19,17	<0,01	623,48	11,51	<0,01
Gênero Masculino	-13,59	1,48	<0,01	-14,43	1,31	<0,01
Idade	-2,94	2,00	0,14	-3,44	1,18	<0,01
Idade Escola	-7,81	2,14	<0,01	-6,44	0,82	<0,01
Repetiu	-27,50	2,45	<0,01	-25,95	1,51	<0,01
Pré-Escola	14,41	3,17	<0,01	12,71	1,74	<0,01

Tabela 4.18: Modelos finais estimados para explicação da proficiência em Língua Portuguesa da rede Municipal de 2009

Covariáveis	Considerando o Plano Amostral			Não Considerando o Plano Amostral		
	Coefficiente	Erro Padrão	p-valor	Coefficiente	Erro Padrão	p-valor
Intercepto	598,43	13,03	<0,01	623,48	11,51	<0,01
Gênero Masculino	-13,59	1,46	<0,01	-14,43	1,31	<0,01
Idade	-	-	-	-3,44	1,18	<0,01
Idade Escola	-7,96	2,07	<0,01	-6,44	0,82	<0,01
Repetiu	-30,18	1,67	<0,01	-25,95	1,51	<0,01
Pré-Escola	14,38	3,16	<0,01	12,71	1,74	<0,01

5 Conclusão

Como os nossos objetivos principais eram de ordem metodológica, cabe ressaltar que não procuramos interpretar os resultados produzidos e apresentados nesta monografia sob o ponto de vista substantivo da Educação.

Através dos resultados apresentados neste trabalho, vimos que problemas podem ocorrer quando não consideramos todas as características do plano amostral nas análises estatísticas. No caso das estimativas de médias, a amplitude dos intervalos de confiança pode ser considerada menor do que realmente deveria ser e no caso dos modelos de regressão, em alguns casos, coeficientes que não deveriam ser considerados significantes o são. Em estudos onde o objetivo é auxiliar as tomadas de decisão de órgãos públicos na criação de políticas públicas, por exemplo, ou, até mesmo, empresas privadas, corre-se o risco da decisão ser tomada com base nestes resultados mal elaborados.

Vale ressaltar que os erros padrão de dados provenientes de planejamentos amostrais que possuem conglomeração em alguma(s) de suas etapas podem ser maiores caso a variação da característica de interesse dentro dos conglomerados for muito pequena. Isto ocorre nos dados utilizados neste trabalho, pois dentro de escolas e turmas os alunos tendem a obter notas parecidas entre si quando comparadas com alunos fora dos conglomerados que percentem.

Além disso, vimos que os erros padrão das estimativas pontuais quando consideramos todas as características do plano amostral, de um modo geral, foram menores no ano de 2009 do que em 2007 e 2008, ou seja, a precisão dos resultados melhorou ao longo do tempo. Este resultado era esperado, pois o plano amostral elaborado para 2009 é, teoricamente, melhor que os demais.

Referências Bibliográficas

- [1] ANDRADE, D. F. et al. *Teoria da Resposta ao Item: Conceitos e Aplicações*. São Paulo: Associação Brasileira de Estatística (ABE), 2000.
- [2] BOLFARINE, H; BUSSAB, W. O. *Elementos de Amostragem*. 1 ed. São Paulo: Blucher, 2005.
- [3] FRANCO, C; BROOKE, N; ALVES, F. Estudo longitudinal sobre qualidade e equidade no ensino fundamental brasileiro: GERES 2005. *Ensaio: Avaliação e Políticas Públicas em Educação*, v. 16, n.61, p. 625-637, dez. 2008.
- [4] HORVITZ, D. G; THOMPSON, D. J. A Generalization of Sampling Without Replacement from a Finite Universe. *Journal of the American Statistical Association*, v. 47, n. 260, p. 663-685, dez. 1952.
- [5] KMENTA, J. *Elementos de Econometria*. 1 ed. São Paulo: Atlas, v. 2, 1988.
- [6] LUMLEY, T. *survey: analysis of complex survey samples*. R package version 3.22-1, 2010. Disponível em: <<http://www.R-project.org>>.
- [7] PESTANA, M. I. O Sistema de Avaliação Brasileiro. *Revista Brasileira de Estudos Pedagógicos*, v. 79, n.191.
- [8] R Development Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing: Vienna, 2009. Disponível em <<http://www.R-project.org>>.
- [9] VICENTE, P. et al. *Sondagens. A amostragem como factor decisivo de qualidade*. 2 ed. rev. e corrigida. Lisboa: Sílabo, 2001.
- [10] VIEIRA, M. D. T. *Um Estudo Comparativo das Metodologias de Modelagem de Dados Amostrais Complexos - Uma Aplicação ao SAEB 99*. Dissertação de Mestrado Pontifícia Universidade Católica do Rio ed Janeiro, Rio de Janeiro, 2001.
- [11] ----- *Plano Amostral da Pesquisa do PROALFA de 2007*. Relatório Técnico. Juiz de Fora: Departamento de Estatística, UFJF, 2007.

-
- [12] ----- . *Analysis of Longitudinal Survey Data*. 1 ed. Saarbrücken: VDM Verlag Dr. Müller, 2009.
- [13] VIEIRA, M. D. T; SOUZA, A. C. *Plano Amostral da Pesquisa do PROALFA de 2008*. Relatório Técnico. Juiz de Fora: Departamento de Estatística, UFJF, 2008.
- [14] ----- . *Plano Amostral da Pesquisa do PROALFA de 2009*. Relatório Técnico. Juiz de Fora: Departamento de Estatística, UFJF, 2009.
- [15] WOLTER, K. M. *Introduction to Variance Estimation*. New York: Springer-Verlag, 1985.