UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
FACULDADE DE MEDICINA
PROGRAMA DE PÓS-GRADUAÇÃO EM EPIDEMIOLOGIA

TESE DE DOUTORADO

ANÁLISE DE DADOS EPIDEMIOLÓGICOS INCORPORANDO PLANOS AMOSTRAIS COMPLEXOS

Iara Denise Endruweit Battisti

Orientador: Profa. Dra. Jandyra Maria Guimarães Fachel

Co-orientador: Prof. Dr. João Riboldi

Porto Alegre, julho de 2008.

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
FACULDADE DE MEDICINA
PROGRAMA DE PÓS-GRADUAÇÃO EM EPIDEMIOLOGIA

TESE DE DOUTORADO

ANÁLISE DE DADOS EPIDEMIOLÓGICOS INCORPORANDO PLANOS AMOSTRAIS COMPLEXOS

Iara Denise Endruweit Battisti

Orientador: Profa. Dra. Jandyra Maria Guimarães Fachel
Co-orientador: Prof. Dr. João Riboldi

A apresentação desta tese é exigência do Programa de Pós-graduação em Epidemiologia, Universidade Federal do Rio Grande do Sul, para obtenção do título de Doutor.

Porto Alegre, Brasil.
2008

BANCA EXAMINADORA

Profa. Dra. Maria Inês Reinert Azambuja, Programa de Pós-Graduação em Epidemiologia, Faculdade de Medicina, Universidade Federal do Rio Grande do Sul

Profa. Dra. Roselaine Ruviaro Zanini, Departamento de Estatística, Centro de Ciências Naturais e Exatas, Universidade Federal de Santa Maria

Profa. Dra. Sídia Maria Callegari Jacques, Departamento de Estatística, Instituto de Matemática, Universidade Federal do Rio Grande do Sul

AGRADECIMENTOS

Gostaria de expressar todo meu carinho e meu agradecimento a tantas pessoas que me acompanharam nessa caminhada e que foram muito importantes para a conclusão desta tese. Meu agradecimento, meu carinho por cada um é enorme, mesmo não conseguindo expressar tão bem pelas palavras.

À Deus que me iluminou nessa caminhada.

Ao meu marido Gerson, meu companheiro, pelo seu incentivo, pelo seu amor, por iluminar a minha vida. Te amo!

Aos meus pais Salo e Teresinha, os primeiros a incentivarem meus estudos, por acreditarem sempre em mim, pelo carinho, por estarem sempre presentes me apoiando. Amo vocês.

A minha irmã Leila, por sua alegria contagiante, pelo apoio. Te amo maninha!

Ao meu vô Antônio, à minha *oma* Hilda, tios, tias, primos, primas, sogro e sogra pelo apoio e à minha vó Gema que somente acompanhou o início desta caminhada.

À minha orientadora, professora Jandyra, pela valiosa sugestão do tema de pesquisa, por ter me aceitado como orientanda, pela orientação, compreensão e amizade. Muito obrigada!

Ao meu co-orientador, professor Riboldi, por ter me sugerido esse curso em uma conversa num simpósio em que participávamos e desde então foi meu incentivador, pela orientação, valiosas contribuições, compreensão e amizade. Muito obrigada!

À professora Elsa Mundstock, pelas valiosas contribuições, pela sua dedicação e paciência em vários momentos de discussão e estudo. Muito obrigada!

Aos professores do Programa de Pós-Graduação em Epidemiologia pelos valiosos ensinamentos, especialmente ao professor Bruce Duncan, à professora Maria Inês Schmidt e ao professor Sotero Mengue por terem aceitado o meu ingresso na banca de seleção deste curso.

A Universidade Regional do Noroeste do Estado do Rio Grande do Sul - UNIJUÍ, ao Departamento de Física, Estatística e Matemática – DeFEM e ao Mestrado em Modelagem Matemática pelo apoio para realizar meu doutorado.

À banca examinadora externa Maria Inês Azambuja, Roselaine Zanini e Sídia Callegari Jacques por aceitarem a fazer parte da banca.

À banca examinadora interna Álvaro Vigo, Maria Inês Azambuja e Suzi Camey por aceitarem a fazer parte da banca e pelas valiosas contribuições.

Às colegas do curso Anaelena, Andréia, Angela Isabel, Eliana, Juliana, Maria Inês, Luciana, Roselaine e Stela pelos momentos de estudo e pelos momentos de descontração em Porto Alegre ou pelo *msn*.

À colega e amiga Silvania Bottaro pelo incentivo a iniciar essa caminhada.

Aos colegas e às colegas da UNIJUI, aos amigos e às amigas, são tantos, prefiro não citar nomes, todos que colaboraram sabem quem são e o quanto foram importantes para mim nesse caminhada.

Também, não poderia deixar de citar um companheiro incondicional neste último ano de caminhada, nas longas horas de estudos e dedicação para conclusão desta tese: *my dog Ingo.* E, *my dog Julia* que foi companheira nos primeiros anos dessa caminhada.

SUMÁRIO

ABREVIATURAS E SIGLAS

AAS	Amostragem aleatória simples
AC	Amostragem complexa
ACO	Amostragem por conglomerado
AE	Amostragem estratificada
AS	Amostragem sistemática
BRR	Replicação repetida balanceada
CCI	Coeficiente de correlação intra-conglomerado
CDC	Centers for Disease Prevention and Control
CNDDM	Campanha Nacional de Detecção de Diabetes Mellitus
DEFF	*Design effect*
EP	Erro padrão
EPA	Efeito do plano amostral
EQM	Erro quadrático médio
ERG	Ensaio aleatorizado em grupo
ESCA	*La Encuesta de Salud de Catalunya*
FCF	Fator de correção para população finita
IBGE	Instituto Brasileiro de Geografia e Estatística
IGLS	*Iterative Generalized Least Squares*
IID	Independente identicamente distribuído
IMC	Índice de Massa Corporal
INPE	Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira
MCBS	*Medicare Current Beneficiay Survey*
MEPS	*Medical Expenditure Panel Survey*
ML	*Maximum Likelihood*
MPV	Máxima Pseudo-Verossimilhança
NCHS	*National Center for Health Statistics*
NHANES	*National Health and Nutrition Examination Survey*

NHIS	*National Health Interview Survey*
PAQUID	*Personnes Agees Quid*
PNAD	Pesquisa Nacional por Amostra de Domicílios
PNDS	Pesquisa Nacional sobre Demografia e Saúde
PPT	Probabilidade proporcional ao tamanho
PPV	Pesquisa de Padrões de Vida
REML	*Residual Maximum Likelihood*
RIGLS	*Restricted/Reweighted Iterative Generalized Least Squares*
SAEB	Sistema Nacional de Avaliação da Educação Básica
SUGI	*SAS Users Group International*
UBS	Unidade Básica de Saúde
UFA	Unidade Final de Amostragem
UPA	Unidade primária de amostragem
UPAs	Unidades primárias de amostragem
USA	Unidade secundária de amostragem

RESUMO

Introdução: Muitos estudos epidemiológicos utilizam amostragem complexa para coleta de dados. A amostragem complexa pode ter uma ou mais das seguintes características: estratos, conglomerados e probabilidades desiguais de seleção. Se estas características não forem incorporadas na análise de dados, as estimativas pontuais e erros-padrões são incorretos. Assim é necessário ampliar a compreensão do impacto de cada característica nos resultados para incentivar os pesquisadores a utilizarem metodologias adequadas para análise dos dados, obtendo conclusões válidas para a população de onde provém a amostra. Para tratar as estruturas complexas do plano amostral existem duas principais metodologias: abordagem da amostragem complexa e abordagem de modelos multinível.

Objetivos: Descrever e comparar métodos para tratamento de dados provindos de planos amostrais complexos através de duas abordagens: amostragem complexa e modelos multinível, utilizando dados de dois estudos epidemiológicos.

Métodos: Para avaliar o impacto do plano amostral complexo, assim como de cada característica do plano amostral nas estimativas de média, proporção, coeficientes da regressão de Poisson e seus correspondentes erros padrões utilizaram-se os dados da busca ativa domiciliar dos participantes na Campanha Nacional de Detecção de Diabetes Mellitus – CNDDM de 2001, obtidos por amostragem estratificada com conglomerado em três estágios. Para comparar a abordagem da amostragem complexa e a abordagem de modelos multinível ajustaram-se modelos de regressão linear com e sem pesos amostrais utilizando os dados de um estudo do desempenho das crianças na avaliação de conhecimento, percepções e crenças sobre aleitamento materno, realizado com escolares da quinta série do ensino fundamental, no município de Ijuí/RS, estudo aleatorizado, com amostra estratificada por conglomerados.

Resultados: As estimativas pontuais de média e proporção são semelhantes comparando-se amostragem complexa e amostragem aleatória simples, porém observou-se grande diferença nos erros padrões. O mesmo foi observado nas estimativas dos coeficientes da regressão de Poisson com menor efeito do plano amostral. Na comparação da abordagem da amostragem complexa com modelos

multinível observou-se diferença nos erros padrões dos coeficientes da regressão entre as duas abordagens, sendo que os mesmos são maiores na amostragem complexa. Também, na análise não ponderada, as significâncias dos coeficientes no modelo final foram semelhantes entre as duas abordagens, porém houve diferença na análise ponderada para um dos coeficientes.

Conclusões: Os resultados encontrados a partir dos dois estudos evidenciaram a necessidade de incorporar a complexidade do plano amostral na análise dos dados. A questão de pesquisa poderá ser um fator importante na escolha entre a abordagem da amostragem complexa e a abordagem de modelos multinível.

Palavras-chave: amostragem complexa; efeito do plano amostral; modelos multinível.

ABSTRACT

Introduction: Many epidemiological studies use complex samples for data collection. Complex sampling may have one or more of the following characteristics: stratification, clustering and unequal selection probabilities. If these characteristics are not incorporated into data analysis, point estimates and standard errors are incorrect. Greater understanding of the effect of each characteristic on results should stimulate researchers to use adequate methods for data analysis and, therefore, to reach conclusions that are valid for the population that generated the sample. Two major methods are used to deal with complex sampling designs: the complex sample approach and the multilevel model approach.

Objective: To describe and compare methods to deal with data in complex sampling designs using complex sample and multilevel model approaches in two epidemiological studies.

Method: Data retrieved from a house-to-house survey of participants in the 2001 Brazilian Diabetes Detection Campaign (Campanha Nacional de Detecção de Diabetes Melitus - CNDDM) and collected by stratified clustering sampling in three stages were used to evaluate the impact of complex sampling designs, as well as of each of their characteristics, on the estimates of means, proportions, Poisson regression coefficients and their corresponding standard errors. To compare the complex sample and the multilevel model approaches, linear regression models were adjusted with and without sample weights using data from a random study that used stratified cluster sampling and investigated the performance of children in the evaluation of knowledge, perceptions and beliefs about maternal breastfeeding conducted with fifth grade students in Ijuí, Brazil.

Results: Mean and proportion point estimates were similar when complex sampling and simple random sampling were compared, but there was a great difference in standard errors. The same was found for estimates of Poisson regression coefficients that were less affected by sampling design. The complex sample approach showed significantly greater standard errors of the regression coefficients than the multilevel model approach. Also, unweighted analysis showed that the significance of

coefficients in the final models was similar in the two approaches, but there was a difference in one of the coefficients in weighted analysis.

Conclusions: Results of the two studies showed that sampling design complexity should be incorporated into data analysis. Research questions seem to be a determinant factor in the choice of either a complex sample or a multilevel model approach.

Key words: Complex sample; design effect; multilevel model.

LISTA DE QUADROS

ARTIGO 1

ARTIGO 2

ANEXO A

ANEXO C

LISTA DE TABELAS

ARTIGO 1

ARTIGO 2

LISTA DE FIGURAS

ARTIGO 1

ARTIGO 2

1. APRESENTAÇÃO

Este trabalho consiste na tese de doutorado intitulada “Análise de dados epidemiológicos incorporando planos amostrais complexos”, apresentada ao Programa de Pós-Graduação em Epidemiologia da Universidade Federal do Rio Grande do Sul. O trabalho é apresentado em três partes, na ordem que segue:

1. Introdução, Revisão da Literatura e Objetivos
2. Artigos
3. Conclusões e Considerações Finais.

Documentos de apoio, incluindo o Projeto de Pesquisa, estão apresentados nos anexos.

2. INTRODUÇÃO

Muitos estudos epidemiológicos utilizam amostragem probabilística para coleta de dados. Os métodos de amostragem probabilística freqüentemente utilizados são: amostragem aleatória simples (AAS), amostragem estratificada (AE), amostragem sistemática (AS) e amostragem por conglomerado (ACO). É muito comum, em grandes inquéritos para estudo da saúde de indivíduos, a aplicação de duas ou mais destas formas de amostragem ao mesmo tempo e, no caso de amostragem por conglomerados, a utilização de dois ou mais estágios de seleção das unidades amostrais. A amostragem complexa pode ter uma ou mais das seguintes características: estratificação, conglomeração e probabilidades desiguais de seleção.

Nos Estados Unidos, organizações como o Centro Nacional para Estatísticas de Saúde (*National Center for Health Statistics*) e o Departamento de Censo Americano (*United States Bureau of the Census*) disponibilizam dados secundários sobre saúde da população obtidos com amostragem complexa. No Brasil, dados secundários importantes, provindos de amostragem complexa, são disponibilizados pelo IBGE (Instituto Brasileiro de Geografia e Estatística), como, por exemplo, os dados da Pesquisa Nacional por Amostra de Domicílios (PNAD) (Silva et al., 2002).

A maioria dos aplicativos estatísticos contempla, em seu módulo básico, técnicas de análise de dados com metodologias apropriadas para amostragem aleatória simples, mas não apropriadas para os demais tipos de amostragem probabilística. Mais recentemente, os aplicativos estatísticos introduziram rotinas com metodologias para análise de dados obtidos com amostragem complexa. Assim,

torna-se mais fácil a aplicação das técnicas de análise incorporando adequadamente as diversas características de planos amostrais complexos, tanto na estimação de medidas descritivas, fornecendo a precisão dessas estimativas, como no ajuste de modelos (Silva et al., 2002).

Cabe ressaltar que, métodos de análise de dados amostrais complexos já existiam teoricamente há bastante tempo, porém, devido às dificuldades computacionais antes do advento das rotinas especiais, não eram utilizados. As rotinas para tratamento de dados de amostragem complexa são ainda pouco utilizadas nas diversas áreas de aplicação da Estatística e também na análise de dados epidemiológicos, apontando-se alguns fatores como: a) pouca disponibilidade de referências metodológicas sobre amostragem complexa com aplicações em Epidemiologia; b) muito recente incorporação destas técnicas nos principais aplicativos estatísticos; c) pouca divulgação metodológica das conseqüências em ignorar o plano amostral na análise de dados.

Se o plano amostral for complexo e os dados forem analisados como provenientes de uma amostragem aleatória simples, isto é, ignorando as características de estratificação, conglomeração e probabilidades desiguais de seleção, os resultados podem fornecer estimativas incorretas das variâncias dos estimadores.

Pretende-se com este estudo contribuir para a divulgação da necessidade de considerar o plano amostral complexo na análise de dados epidemiológicos. Também aponta-se a magnitude de erros nas estimativas e/ou diferenças nas precisões destas estimativas ao ignorar a complexidade do delineamento amostral na análise de dados.

Por fim, compara-se a metodologia de amostragem complexa com uma metodologia alternativa em certas aplicações – modelagem.

3. REVISÃO DE LITERATURA

Esta revisão de literatura engloba uma abordagem aplicada sobre amostragem complexa, também referenciada e indexada como *complex survey*. Primeiramente, apresenta-se o conceito de amostragem e tipos de amostragem. Na seqüência, os métodos de estimação de parâmetros para amostragem complexa e também os efeitos das características do plano amostral nas diversas estimativas, a partir da revisão de vários estudos já realizados. Na seção seguinte, os aplicativos para análise de dados de amostragem complexa são apresentados e, na última seção, uma revisão de modelos multinível.

3.1 AMOSTRAGEM

Amostragem é o processo pelo qual se obtém uma ou mais amostras de uma população de interesse. Na amostragem seleciona-se parte de uma população e observa-a com a finalidade de estimar parâmetros populacionais (características populacionais). É importante que os diferentes procedimentos amostrais satisfaçam aos seguintes critérios (Mundstock, 2005):

1) que as amostras sejam representativas da população;
2) que forneçam estimativas precisas das características da população, podendo medir sua confiabilidade;
3) que tenham pequeno custo para selecionar a amostra.

Na amostragem probabilística é possível calcular, com antecedência, a probabilidade de obter-se cada uma das amostras possíveis, sendo que todas as unidades da população (podendo ser indivíduos, residências, hospitais, entre outras) têm probabilidade maior que zero de participar da amostra. É importante observar que a aleatoriedade da amostra depende do processo pelo qual ela é obtida. Nas amostras probabilísticas é possível estimar, com uma determinada probabilidade, os erros de amostragem ou as discrepâncias entre as estimativas amostrais e os valores populacionais que seriam obtidos observando todas as unidades da população (Cochran, 1977).

Ao contrário, a amostragem não-probabilística é um procedimento pelo qual não se pode associar probabilidade de seleção às unidades e, conseqüentemente, não é possível determinar a confiabilidade dos resultados da amostra em termos probabilísticos.

Existem diferentes métodos de obtenção de uma amostra probabilística de uma população, os quais são brevemente descritos a seguir. Para maior detalhamento consulta-se Bolfarine e Bussab (2005) e Cochran (1977).

3.1.1 Tipos de amostragem aleatória

Amostragem aleatória simples

A amostragem aleatória simples (AAS) consiste na seleção de n unidades de uma população de tamanho N, de maneira que cada uma das amostras possíveis tenha a mesma probabilidade de ser selecionada.

As unidades da população são numeradas de 1 a N e depois são obtidos números aleatórios da tabela ou do computador. Uma amostra aleatória simples é selecionada extraindo-se uma unidade de cada vez. As unidades correspondentes aos n números sorteados constituem a amostra.

O sorteio de unidades amostrais pode ser realizado com ou sem reposição. A amostragem é dita sem reposição quando um elemento selecionado em uma extração é excluído da população para as extrações subseqüentes. A amostragem é dita com reposição quando os N elementos da população permanecem em todas as extrações, isto é, uma unidade selecionada em uma extração é reposta e pode ser extraída novamente. Assim, um elemento pode estar repetido na mesma amostra.

A probabilidade de inclusão da unidade i na amostra com reposição é dada por (Bolfarine e Bussab, 2005; Figueiredo, 2004):

$$\pi_i = 1 - (1 - \frac{1}{N})^n, i = 1,..., N$$

E, na amostra sem reposição é:

$$\pi_i = \frac{n}{N}, i = 1,..., N$$

Os resultados da amostra com e sem reposição são similares quando o N tende a ser grande em relação ao tamanho da amostra (Bolfarine e Bussab, 2005).

Amostragem estratificada

Na amostragem estratificada (AE), a população é dividida em sub-populações mutuamente exclusivas chamadas de estratos. Este tipo de amostragem consiste em selecionar uma amostra em cada estrato e combinar estas amostras numa única

amostra para estimar parâmetros da população. Tem como vantagem o aumento da precisão das estimativas, possibilidade de obtenção de informações em nível de estrato e facilidade na coleta de dados, por razões físicas ou administrativas. Silva (1998) aponta as seguintes razões para estratificar:

1. Deseja-se aumentar a precisão da estimativa global, partindo-se do conhecimento de que a variabilidade da característica estudada é grande;
2. Necessidade de obter-se estimativas para diversos segmentos da população;
3. Deseja-se que a amostra mantenha a composição da população segundo algumas características básicas (sexo, idade,...);
4. Conveniência administrativa ou operacional;
5. Deseja-se controlar o efeito de alguma característica na distribuição da característica que está sendo avaliada. Por exemplo, o efeito da escolaridade sobre o estado nutricional de crianças menores de cinco anos pode ser controlado pela composição de uma amostra que contenha os diversos níveis de escolaridade dos chefes de família da população estudada.

Amostragem sistemática

A amostragem sistemática (AS) é indicada quando a população está organizada em alguma ordem. Consiste em selecionar, aleatoriamente, uma unidade amostral entre as k primeiras unidades populacionais e, a partir daí, selecionar as restantes a intervalos fixos em cada k unidades.

Amostragem por conglomerados

Na amostragem por conglomerados (ACO), a população é dividida em M grupos ou conglomerados que servem como unidades primárias de amostragem (UPA), de maneira que cada unidade da população é associada com um e somente um conglomerado. Cada conglomerado é formado por N_i unidades, chamadas unidades secundárias de amostragem (USA). Das M unidades primárias (conglomerados) na população é selecionada uma amostra de tamanho m. A amostragem por conglomerados pode ser realizada em etapa única ou em mais etapas, como segue:

- Amostragem em etapa única: todas as unidades do conglomerado selecionado são incluídas na amostra;
- Amostragem em duas etapas (bietápica): nos conglomerados selecionados são extraídas amostras de n_i unidades secundárias. Neste caso, em cada etapa (estágio) têm-se diferentes unidades amostrais, definidas por unidade primária de amostragem (UPA) e unidade secundária de amostragem (USA), respectivamente para o primeiro e segundo estágios;
- Amostragem em várias etapas (multietápica): o processo pode ser estendido a várias etapas de amostragem, não sendo necessário aplicar o mesmo método de seleção das unidades em todos os níveis.

Segundo Bolfarine e Bussab (2005), a amostragem por conglomerado é muito usada em populações humanas, onde freqüentemente são sorteadas cidades, depois os bairros, os quarteirões, os domicílios e, finalmente, os moradores. Os autores

citam que uma das inconveniências para o uso da amostragem de conglomerado prende-se ao fato de que as unidades, dentro de um mesmo conglomerado, tendem a ter valores parecidos em relação às variáveis que estão sendo pesquisadas, e isso torna este tipo de amostragem menos eficiente.

Comparando-se AAS com a ACO de mesmo tamanho, a amostra por conglomerados tende a ter: custo menor por elemento, maior variância e maior complexidade para análises estatísticas (Bolfarine e Bussab, 2005; Ray, 1983).

Amostragem com probabilidades variáveis

Em alguns procedimentos amostrais, algumas unidades da população são "mais importantes" por terem uma contribuição maior no valor do parâmetro, neste caso estabelecem-se probabilidades desiguais de seleção às diferentes unidades da população (Natarajan et al., 2008).

Nos casos em que a probabilidade de seleção é proporcional à uma medida de tamanho da população, o procedimento amostral é definido como amostragem com probabilidade proporcional ao tamanho (PPT). A vantagem em selecionar a unidade de amostragem com PPT é obter uma amostra mais representativa da população e assim, aumentar a precisão dos estimadores quando comparados à AAS (Mundstock, 2005).

Amostragem complexa

As características freqüentemente presentes em um plano amostral complexo são:

- Estratificação;
- Conglomeração;
- Probabilidades variáveis.

Vieira (2007) aponta os seguintes motivos para a não adoção de AAS em pesquisas de grande porte:

- Quando a população está geograficamente dispersa, pode ser extremamente custoso chegar às unidades selecionadas;
- Um cadastro de boa qualidade para a população alvo pode não estar prontamente disponível;
- Pode ser um procedimento ineficiente quando a população não é homogênea e quando subgrupos apresentam diferenças de tamanho.

Alguns exemplos de amostragem complexa realizados por instituições oficiais:

- Pesquisa Nacional por Amostra de Domicílios (PNAD) – IBGE;
- Sistema Nacional de Avaliação da Educação Básica (Saeb) – INPE (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira).

3.1.2 Erro amostral

Um dos objetivos numa pesquisa por amostragem é estimar parâmetros, por exemplo, média, proporção, variância, coeficientes de modelos de regressão, entre outros. E, como já referenciado anteriormente, a um processo de amostragem está associado um erro amostral.

Segundo Bolfarine e Bussab (2005), o estudo do erro amostral consiste, basicamente, em verificar o comportamento da diferença entre o valor observado na amostra e o parâmetro, quando este valor amostral é observado em todas as possíveis amostras que poderiam ser formadas através do plano amostral escolhido. O valor esperado do quadrado dessa diferença (erro quadrático médio - EQM) informa sobre a precisão do estimador. No caso de estimador não viesado (valor da diferença igual a zero), o EQM é a variância do estimador, calculada em relação à distribuição amostral do estimador; extraindo-se a raiz quadrada tem-se o desvio padrão. Neste caso, o desvio padrão do estimador recebe o nome de erro padrão e indica o erro médio esperado pelo uso do estimador e do plano amostral especificado.

Bolfarine e Bussab (2005) preferem chamar os erros amostrais de desvios devido ao plano amostral, os quais tendem a desaparecer com o crescimento do tamanho da amostra. Complementam que, para alguns planos amostrais bastante complexos, não há expressões explícitas para estes desvios, sendo necessário o recurso de técnicas aproximadas. Às vezes, por facilidade de cálculo, emprestam-se fórmulas de um plano mais simples para o cálculo de erro padrão de outros planos amostrais mais complexos, praticando-se um “erro-técnico”. O objeto da presente pesquisa é justamente as conseqüências do “erro-técnico”.

A diferença entre o valor amostral e o parâmetro pode ser afetada por outro tipo de erro, o erro não amostral, devido a outros fatores que não o plano amostral, como por exemplo, inadequações de mensuração, entrevistas, codificação, entre outras. Nas pesquisas tenta-se minimizar esses erros, os quais podem ocorrer até na pesquisa de todas as unidades populacionais, isto é, não por processo de amostragem.

A seguir, apresentam-se algumas possíveis ocorrências de erros não amostrais (Bolfarine e Bussab, 2005):

A. Unidades perdidas (falta de resposta)
    i. Falta de resposta total:
        a. Falta de contato com a unidade;
        b. Recusa;
        c. Abandono durante a pesquisa;
        d. Incapacidade em responder;
        e. Perda de documento;
    ii. Falta de resposta parcial:
        a. Recusa em questões sensíveis (ex: renda);
        b. Incompreensão;
        c. Dados incoerentes;

B. Falhas na definição e administração
    i. Sistemas de referência:
        a. Erros de omissão (cobertura incompleta), exclusão de elementos de interesse. Resulta de diferenças entre as diversas populações;
        b. Inclusão de elementos não sorteados ou de outras populações

ii. Efeito do entrevistador;

iii. Insuficiência do questionário-redação;

iv. Erros de codificação e digitação;

C. Avaliação das conseqüências

i. Comparação com resultados de outras pesquisas;

ii. Efeito do processo de imputação, caso tenha sido usado;

iii. Programas de consistência de dados;

iv. Volume de não respondentes;

v. Diferença de perfil de respondentes e não respondentes.

O trabalho de Vasconcellos et al. (2005) aponta que o uso da amostragem inversa, como utilizada na Pesquisa Mundial de Saúde no Brasil, tem a principal vantagem de eliminar as correções de não-resposta.

3.2 ESTIMAÇÃO DE PARÂMETROS CONSIDERANDO AMOSTRAGEM COMPLEXA

Nesta seção, apresentam-se os métodos para estimação do erro padrão para amostragem complexa e na seqüência, os estudos realizados sobre amostragem complexa.

3.2.1 Métodos de Estimação em Amostragem Complexa

Os métodos exatos para estimação de parâmetros considerando AAS, AS, AE e ACO são bem conhecidos e referenciados em livros clássicos sobre amostragem, como Cochran (1977) e Kish (1965). Porém, na amostragem complexa, quase sempre são necessários métodos aproximados para estimar os parâmetros, os quais são detalhados em livros mais recentes, podendo-se citar Lehtonen e Pahkinen (2004) e Chambers e Skinner (2003).

Os estimadores do total, freqüentemente utilizados são estimadores lineares ponderados. Um caso particular é o estimador de Horvitz-Thompson. Outros estimadores usados são do tipo razão ou regressão, os quais são não-lineares (Pessoa e Silva, 1998). Os estimadores de razão e regressão são viciados para pequenas amostras, porém o vício é desprezível para amostras grandes e existem expressões aproximadas para as variâncias de aleatorização.

A estimação do erro padrão no caso de estimadores não lineares requer o uso de técnicas especiais, tais como: Linearização de Taylor e Re-amostragem (Kreuter e Valliant, 2007; Lehtonen e Pahkinen, 2004; Chambers e Skinner, 2003), que produzem resultados semelhantes (Rodgers-Farmer e Davis, 2001).

Estimação do Total

Apresenta-se a seguir, a metodologia para estimação do total para um delineamento estratificado por conglomerados em estágio único, onde os conglomerados (UPA) são selecionados, aleatoriamente, com ou sem reposição

dentro de cada estrato e todos os elementos do conglomerado são pesquisados. Os pesos amostrais são incorporados na análise e a correção para população finita (FCF) é aplicada no estimador da variância nos casos em que a UPA for selecionada sem reposição (StataCorp, 2005):

Sejam $h = 1,...,L$ estratos, $i = 1,...,N_h$ UPAs no estrato h e $j = 1,...,M_{hi}$ indivíduos na UPA i e no estrato h, então:

$$M = \sum_{h=1}^{L} \sum_{i=1}^{N_h} M_{hi}$$

é o número de indivíduos na população.

Seja Y_{hij} o valor da variável Y do indivíduo j, na UPA i e no estrato h, então o total populacional de Y é:

$$Y = \sum_{h=1}^{L} \sum_{i=1}^{N_h} \sum_{j=1}^{M_{hi}} Y_{hij}$$

O número de indivíduos na amostra é:

$$m = \sum_{h=1}^{L} \sum_{i=1}^{n_h} m_{hi}$$

Seja y_{hij} a variável de pesquisa do indivíduo da amostra, em que $h = 1,...,L$; $i = 1,...,n_h$; e $j = 1,...,m_{hi}$. O estimador de Y é:

$$\hat{Y} = \sum_{h=1}^{L} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hij} y_{hij}$$

onde w_{hij} é o peso amostral definido por $w_{hij} = \frac{N_h}{n_h}$.

O estimador para o tamanho da população é:

$$\hat{M} = \sum_{h=1}^{L} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hij}$$

O estimador da variância de $\hat{Y}$ é:

$$\hat{V}(\hat{Y}) = \sum_{h=1}^{l} (1 - f_h) \frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2$$

em que y_{hi} é o total ponderado da UPA i no estrato h:

$$y_{hi} = \sum_{j=1}^{m_{hi}} w_{hij} y_{hij}$$

e $\bar{y}_h$ é a média do total das UPAs para o estrato h:

$$\bar{y}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} y_{hi}$$

O fator $(1 - f_h)$ é o FCF para o estrato h e f_h é a razão de amostragem para o estrato h definida como $f_h = \frac{n_h}{N_h}$.

Ressalta-se que a estimação dos totais é a base sobre a qual assenta-se a estimação de médias, razões, taxas e proporções.

Muitos modelos paramétricos podem ser ajustados empregando o método da Máxima Pseudo-Verossimilhança (MPV) para estimar os parâmetros, com dados obtidos através de diferentes planos amostrais. Os estimadores de MPV não serão

únicos, já que existem diversas maneiras de definir os pesos w_j correspondentes a diferentes estimadores de totais. Os pesos mais usados são do estimador simples para totais-estimador.

O procedimento de MPV proporciona estimativas consistentes e é razoavelmente simples de calcular, tanto para os estimadores como para as variâncias dos estimadores dos parâmetros. Este procedimento é a base para o desenvolvimento de rotinas capazes de incorporar adequadamente os efeitos de planos amostrais complexos disponíveis no STATA, SAS e outros aplicativos.

Detalhes sobre procedimento de estimação de parâmetros encontram-se em StataCorp (2005), Silva et al. (2002) e Pessoa e Silva (1998).

Métodos de re-amostragem

No método de re-amostragem são retiradas repetidas amostras de tamanho menor que a amostra original. A estimativa é calculada para cada sub-amostra e a variância é calculada entre as estimativas das sub-amostras. As sub-amostras são geradas por diferentes métodos de re-amostragem (Kreuter e Valliant, 2007): grupos aleatórios, *jackknife*, re-amostragem repetida balanceada – *balanced repeated resampling* (BRR) e *bootstrap*.

3.2.2 Efeitos da incorporação de diferentes características do delineamento amostral complexo nas estimativas pontuais e erros padrões

A análise de dados provindos de amostragem complexa apresenta dois desafios: (1) obter estimativa pontual correta e (2) calcular corretamente variância e erros padrões. As três características – estratos, conglomerados e pesos - possíveis de estarem presentes numa amostragem complexa têm diferentes efeitos na estimativa pontual e na variância desta estimativa (Kreuter e Valliant, 2007).

Segundo Lemeshow e Cook (1999), na época do artigo era comum encontrar resultados de análise onde o delineamento complexo havia sido ignorado. Essa abordagem seria mais freqüente em países em desenvolvimento, onde utilizavam-se aplicativos estatísticos que não possuíam módulos específicos para amostragem complexa (Sousa e Silva, 2003). Um dos motivos apontados por Sousa e Silva (2003) seria o desconhecimento do impacto nas variâncias dos estimadores quando se ignora o delineamento complexo e, outro, a falta de disponibilidade das rotinas adequadas nos aplicativos estatísticos.

No entanto, atualmente encontram-se vários estudos que incorporam o delineamento amostral complexo na análise de dados e apresentam o efeito do plano amostral (EPA), como os estudos de Oliveira (2007), Giatti e Barreto (2006), Teixeira et al. (2006), Egede (2003, 2004), Bueno et al. (2003), Hernández et al. (2003), Barros e Bertoldi (2002), entre outros.

É interessante salientar que institutos governamentais, como o Instituto Brasileiro de Geografia e Estatística (IBGE), utilizam há bastante tempo métodos de análise apropriados para dados provindos de delineamentos amostrais complexos

também disponibilizando em seus bancos de dados as informações das características do plano de amostragem.

Vasconcellos e Portela (2001) apontam que planos amostrais complexos são utilizados pelos institutos oficias de estatística para obter amostras de forma operacionalmente otimizada e, portanto, com o menor custo possível. Isso implica na utilização de planos de amostragem com probabilidades desiguais de seleção, estratificação e tratamentos para não-resposta, que resultam, de um modo geral, em fatores de expansão ou pesos amostrais de alta variabilidade e dificultam a aplicação das técnicas tradicionais da inferência estatística clássica.

Efeito do conglomerado

Conglomerados geográficos são freqüentemente utilizados para reduzir custos administrativos (Carlson, 2003), permitindo um custo menor por indivíduo amostrado comparada a uma AAS, porém têm a desvantagem de que a análise estatística torna-se mais complexa e, geralmente, produz incrementos nas variâncias dos estimadores (Cordeiro, 2001). Desconsiderar o efeito do conglomerado na análise de dados pode ocasionar pouco impacto nas estimativas pontuais, mas pode resultar em subestimação da variabilidade, isto é, erros padrões subestimados e intervalos de confiança menores (Horton e Fitzmaurice, 2004).

Os conglomerados devem ser heterogêneos dentro de si e homogêneos entre si, isto é, os conglomerados devem ser bastante semelhantes entre si em relação à variável de interesse e os elementos dentro de cada conglomerado devem ser menos parecidos (Barata, 2005). Na prática, isso dificilmente ocorre, observando-se bastante

homogeneidade dentro dos conglomerados. Essa homogeneidade em relação às variáveis que estão sendo pesquisadas torna o plano amostral por conglomerado menos eficiente (Bolfarine e Bussab, 2005), pois resultam erros padrões maiores quando comparados com AAS de igual tamanho (Kreuter e Valliant, 2007).

Segundo Pérez et al. (2004), se dados provindos de amostragem complexa forem analisados como sendo de uma AAS, as variabilidades serão subestimadas em magnitude que depende do coeficiente de correlação intra-conglomerado (CCI) da variável que está sendo analisada. Isto é, quando a correlação é grande, a variabilidade é subestimada em maior magnitude.

Quando a população é dividida em grupos (conglomerados, por exemplo), a variância da população é uma função da variância entre conglomerados e dentro de conglomerados, no caso de amostragem em etapa única, assim definida (Mundstock, 2005, Cochran, 1977):

$$\sigma^2 = \sigma_E^2 + \sigma_D^2$$

em que:

σ_E^2: variância entre conglomerados

σ_D^2: variância dentro dos conglomerados

No caso de amostragem multietápica, cada etapa contribui para a variância do estimador.

O CCI indica a homogeneidade dos conglomerados, sendo calculado da seguinte forma (Bolfarine e Bussab, 2005):

$$CCI = \frac{S_E^2 - \frac{S_D^2}{m-1}}{S_E^2 + S_D^2}$$

S_E^2 = variância entre conglomerados

S_D^2 = variância dentro dos conglomerados

m = tamanho dos conglomerados

O intervalo de variação do CCI é de $\frac{-1}{m-1}$ a 1. No caso de máxima homogeneidade, isto é, dentro dos conglomerados todas as observação são iguais entre si, o CCI assume valor igual a 1. No outro extremo de heterogeneidade o CCI assume valor igual a $\frac{-1}{m-1}$.

O efeito do plano amostral por conglomerados de tamanho igual é dado por (Bolfarine e Bussab, 2005):

$$EPA = 1 + CCI(m-1)$$

Assim, a eficiência dependerá do tipo de conglomeração. Usualmente, CCI é positivo, então a conglomeração resulta em perda de eficiência em relação à AAS.

Efeito do estrato

Ao contrário dos conglomerados, os estratos devem ser heterogêneos entre si e homogêneos dentro de si, isto é, devem ser diferentes em relação à variável de interesse e muito parecidos dentro de cada estrato. Uma correta estratificação aumenta a precisão da amostra estratificada em relação a AAS. O efeito dependerá da eficácia da variável de estratificação para separar grupos compostos por elementos

semelhantes quanto à variável que está sendo pesquisada, e, ao mesmo tempo, diferentes entre si (Silva, 1998) em relação a outras variáveis. Por exemplo, se a variável de interesse é o IMC (índice de massa corporal) os estratos podem ser definidos pelas diferentes faixas de IMC, porém diferentes entre si em relação a outras variáveis.

Segundo Horton e Fitzmaurice (2004), a estratificação proporciona facilidade logística na fase de coleta de dados. Desconsiderar o efeito da estratificação nesta fase pode ter pequeno impacto nas estimativas dos parâmetros, mas superestima a variabilidade, isto é, os erros padrões serão superestimados e intervalos de confiança serão maiores no caso dos estratos serem homogêneos em relação à variável de estudo.

Efeito do peso amostral

Para compensar as probabilidades desiguais de seleção, são atribuídas ponderações diferenciadas aos elementos da amostra. No presente estudo, serão elas chamadas de pesos amostrais, correspondendo ao inverso do produto das probabilidades de inclusão nos diversos estágios de seleção (Szwarcwald e Damacena, 2007; Korn e Graubard, 1995), sob pena de obter resultados incorretos (Pfeffermann, 1996). Por exemplo, no caso de delineamento em dois estágios, o peso é calculado como $w_1.w_2$, respectivamente o peso no primeiro estágio, e no segundo estágio do delineamento. Em muitos planos amostrais, a soma dos pesos será igual ao tamanho da população (UCLA, 2005).

Hernández et al. (2003), Brogan (2003) e Korn e Graubard (1995) definem o peso como um fator de expansão, sendo que o valor do peso indica o número de indivíduos na população que cada observação na amostra representa. Outros tipos de ponderações são ajustes para não resposta e outros fatores, como pós-estratificação (Carlson, 2003; Korn e Graubard, 1991).

As estimativas de parâmetros da população são influenciadas por pesos distintos das observações e as estimativas de variância são influenciadas pela conglomeração, estratificação e pesos (Brogan, 2003; Leite e Silva, 2002; Silva et al., 2002; Pessoa e Silva, 1998). O cálculo das estimativas de variância desempenha papel essencial na realização da inferência analítica, permitindo a avaliação da precisão das estimativas e a formulação de testes de hipóteses sobre os parâmetros dos modelos. Assim, é muito importante que estimativas sejam obtidas corretamente.

Ciol et al. (2006) comentam que se os pesos amostrais não forem considerados na análise de dados, o erro padrão da média pode ser superestimado ou subestimado, dependendo da variabilidade do estrato.

Desconsiderar o peso amostral, segundo Horton e Fitzmaurice (2004) e Guillén et al. (2000) resultará em diferença nas estimativas e em subestimação da variabilidade, isto é, erros padrões subestimados e intervalos de confiança mais estreitos.

Kreuter e Valliant (2007) comentam que o aumento que pode ocorrer no erro padrão devido ao uso de ponderação é, algumas vezes, usado como argumento contra a ponderação. Porém, apontam que as estimativas obtidas quando se desconsideram os pesos não são válidas para toda a população.

A magnitude da diferença entre estimativas ponderadas e não ponderadas dependerá da variabilidade dos dados. Quando os pesos amostrais têm pouca variabilidade, as estimativas pontuais, considerando AAS, são similares às obtidas ao considerar ponderação (Korn e Graubard, 1991).

Outro fator que contribui na diferença entre a estimativa da análise ponderada e não ponderada é a relação entre o valor dos pesos e a variável de análise (Pérez et al., 2004). Por exemplo, se um grupo for sobreamostrado e a prevalência da variável resposta (desfecho) neste grupo for grande, então, a estimativa do desfecho será superestimada na análise não ponderada.

Kreuter e Valliant (2007) apresentam o ajuste do peso final para não resposta e pós-estratificação para o indivíduo i da amostra como $w_i f_{NRi} f_{PSi}$, em que w_i é o peso devido ao delineamento amostral e f_{NRi} e f_{PSi} são o ajustamento para não resposta e pós-estratificação, respectivamente.

Efeito do plano amostral (EPA)

O efeito do plano amostral (EPA) ou *design effect* (DEFF) é utilizado para medir o efeito do plano amostral sobre a variância de um estimador (Pessoa e Silva, 1998). O EPA foi proposto por Kish (1965), sendo definido como a razão entre a variância do estimador para o plano amostral complexo (verdadeiro) e a variância do estimador para AAS, para o mesmo tamanho da amostra n. Portanto, o EPA de Kish para um estimador $\hat{\theta}$ é:

$$\text{EPA}\ (\hat{\theta}) = \frac{V_{VERD}(\hat{\theta})}{V_{AAS}(\hat{\theta})}$$

Segundo Leite e Silva (1999), valores elevados do EPA destacam a importância da consideração do plano amostral efetivamente utilizado ao estimar as variâncias associadas às estimativas dos parâmetros.

O EPA é importante para avaliar subestimativas, ou até superestimativas dos erros padrões, utilizando-se as diferentes características do delineamento amostral e diferentes métodos de estimação (Sousa e Silva, 2003).

A interpretação do EPA é apresentada no Quadro 1 (Leite e Silva, 2002), sendo que o valor do EPA poderá variar muito entre as diferentes variáveis da pesquisa (Sturgis, 2004).

Quadro 1 – Interpretação dos resultados do efeito do plano amostral (EPA)

EPA	**Variância sob AAS**
< 1	Superestimada
= 1	Não há diferença entre as estimativas
> 1	Subestimada

Efeitos detectados em diferentes estudos

Nesta seção, apresenta-se uma revisão de estudos na literatura com abordagem metodológica na análise de dados provenientes de amostragem complexa, iniciando por estudos realizados no Brasil e apresentando após, estudos em outros países.

O impacto do plano por conglomerados e o efeito de ponderação foi avaliado por Sousa e Silva (2003) com dados de 1.355 indivíduos na PNDS96 (Pesquisa Nacional sobre Demografia e Saúde em 1996) constituindo uma amostragem estratificada em dois estágios, através do Epi Info 6.04b (CSAMPLE). No banco de dados, variáveis definem conglomerados, estratos e peso global, este último obtido pelo produto do peso devido ao plano de amostragem pelo peso devido à ausência de resposta e, também o peso padronizado, com o objetivo de obter o total ponderado da amostra igual ao total não ponderado.

Sousa e Silva (2003) consideraram quatro estratégias de análise: (1) conglomerado, sem ponderação, (2) ponderação devido ao plano de amostragem, (3) ponderação devido à ausência de resposta e (4) ponderação global. Para todas as estratégias foram obtidos: estimativas de prevalência, erros padrões, intervalos de confiança, EPA e vícios das estimativas. Concluíram que, a ponderação não aumentou a precisão das estimativas, além da precisão já incluída pelo plano por conglomerados. Apontam que os conglomerados influenciaram a precisão das estimativas, para duas das seis variáveis estudadas, com EPA superiores a 1,5, indicando a importância de considerar os conglomerados na análise. Sugerem a possível existência de heterogeneidade dentro dos conglomerados para outras variáveis em que EPA foi inferior a 1. Recomendam que o cálculo dos efeitos do delineamento e sua publicação devem tornar-se prática usual nas pesquisas.

Kneipp e Yarandi (2002) realizaram estudo sobre questões inerentes a delineamentos amostrais de grandes pesquisas nacionais, explicando a estimação da variância, quando há ponderação e, ainda justificando a necessidade de usar aplicativos que incluem amostragem complexa. Compararam os resultados utilizando

ponderação apenas para amostra, realizada no SPSS, e a ponderação para amostra e variância, realizada no STATA. Analisando dados de 9.482 indivíduos do MEPS (*Medical Expenditure Panel Survey*), com amostragem estratificada em três estágios, obtiveram intervalo de confiança mais amplo incorporando ponderação na análise realizada no STATA. Na análise realizada no SPSS, o intervalo de confiança é mais estreito, devido à redução da variabilidade. Observaram ainda, que o teste *t* de *Student* e a regressão linear foram menos afetados que o teste qui-quadrado, quando a ponderação não é considerada no STATA.

O estudo de Heeringa e Liu (1997), realizado com quatro bases de dados referente à saúde mental, mostrou que o efeito do delineamento é maior que 1 para quase todas as estimativas de prevalência, perfazendo 92% das estimativas obtidas. No mesmo estudo aplicaram-se modelos de regressão logística para verificar o efeito do delineamento, utilizando três diferentes estratégias: (a) não fazendo ajustamento nos dados (SAS), (b) ajustando os dados com ponderações (SAS) e (c) ajustando os dados com ponderações, estratificação e conglomerado, usando o SUDAAN. Os resultados da análise aplicando regressão logística mostraram diferença na estimativa pontual dos coeficientes entre análise ponderada e análise não-ponderada usando o SAS. Também houve diferença de erros padrões, sendo eles maiores na análise ponderada. Os coeficientes obtidos no SUDAAN não diferiram da análise ponderada do SAS, porém os erros padrões foram maiores. Concluem que é importante a incorporação das características do delineamento amostral na inferência univariada e multivariada.

Pérez et al. (2004) analisaram as particularidades de métodos de estimação de parâmetros, como a média, o total e o percentual e seus respectivos erros padrões, e

compararam modelos de regressão logística, para dados provenientes de amostragem complexa, considerando três estratégias de análise: (a) AAS, (b) incorporando pesos e (c) incorporando o delineamento complexo, em dados da *Segunda Encuesta Nacional de Factores de Riesgo y Afecciones Crônicas No Trasmisibles* realizado em Cuba nos anos de 2000 e 2001, com 22.851 indivíduos. Obtiveram prevalências e médias estimadas similares para as três estratégias de análise devido à pequena variabilidade das ponderações amostrais. Porém, a precisão diferiu, sendo menores para AAS e bem menores na análise ponderada do que no delineamento complexo. Os erros padrões foram bem menores na estratégia (b) porque neste caso o tamanho da amostra é igual à soma de todos os pesos amostrais. Da mesma forma, os resultados da regressão logística mostraram intervalos de confiança mais amplos para a análise incorporando o delineamento complexo.

Lemeshow et al. (1998) analisaram os dados de 3.777 indivíduos referentes ao projeto PAQUID (*Personnes Agees Quid)* considerando o delineamento complexo da amostragem, já que, originalmente, a análise foi realizada considerando uma AAS. Os autores utilizaram o STATA, definindo a variável de estratificação, de conglomeração e de ponderação. O valor de ponderação associado a cada indivíduo definiu-se em duas etapas, pois o processo de amostragem englobou estratificação e pós-estratificação. Na primeira etapa, a ponderação devido aos estratos e conglomerados foi definida como o inverso da probabilidade de seleção de cada indivíduo e, na segunda etapa, a ponderação devido a pós-estratificação foi definida pela razão entre a proporção populacional e proporção amostral nas categorias estratificadas posteriormente. O resultado desta segunda etapa foi utilizado para corrigir o peso da primeira etapa, obtendo-se assim, o peso final associado a cada

indivíduo. Foram encontradas variâncias (para médias) subestimadas quando se considerou AAS. Já os coeficientes de regressão e seus erros padrões não diferiram tanto quanto para a média, considerando amostragem complexa e AAS.

No estudo de Lemeshow e Cook (1999), considerando dados da NHANES III (*National Health and Nutrition Examination Survey III*, EUA, 1988-1994) com 19.683 indivíduos obtidos por amostragem estratificada em quatro estágios e os dados do PAQUID (*Personnes Agees Quid*, França, 1988), já abordados em Lemeshow et al. (1998), as estratégias de análise foram: amostragem complexa e AAS. Comparando as duas abordagens observaram-se maiores diferenças para estimativas pontuais de média e seus erros padrões do que para as estimativas de coeficientes de regressão e razão de chances. Os autores concluem que as diferenças encontradas nos resultados entre as duas estratégias de análise e ainda a disponibilidade de aplicativos, como o STATA e SUDAAN, demonstram a necessidade do uso das técnicas apropriadas para a análise de dados provindos de delineamento amostral complexo.

Guillén et al. (2000) ilustram como incorporar o plano amostral complexo na análise de dados, utilizando o STATA para obter estimativas corretas de média, proporção, erro padrão e coeficientes de regressão logística, com dados de 15.000 indivíduos da ESCA (*La Encuesta de Salud de Catalunya* - 1994*)* provenientes de um delineamento estratificado em dois estágios. Utilizam três estratégias de análise: (a) considerando AAS, (b) considerando somente pesos e (c) considerando o delineamento complexo. Concluiram que, ignorar o delineamento amostral resulta em estimativas viesadas dos parâmetros, sendo que a análise ponderada produz estimativas pontuais não viesadas, porém, os erros padrões são muito menores.

Dados de 19.127 indivíduos da NHIS (*National Health Interview Survey* - 1994) foram utilizados por Rodgers-Farmer e Davis (2001) para apontar estimativas pontuais viesadas, erros padrões subestimados e testes de significância com decisões errôneas quando utilizam-se rotinas tradicionais nos aplicativos tradicionais para analisar dados sob amostragem complexa. Ajustaram modelo de regressão múltipla em cinco estratégias: (1) assumindo AAS, (2) assumindo AAS e pesos amostrais, (3) assumindo AAS e pesos normalizados (4) assumindo peso amostral e um único estrato com reposição, isto é, somente o efeito dos conglomerados, e (5) considerando peso amostral e delineamento complexo (com reposição). Os três primeiros casos foram executados no SPSS e os dois últimos no SUDAAN.

Comparando as duas primeiras estratégias, obteve-se diferença entre os coeficientes e erros padrões: os erros padrões foram menores considerando pesos, pois as rotinas tradicionais subestimam as variâncias das estatísticas ponderadas. Na terceira análise, resultou uma suave redução no erro da variância estimada, comparada com a estratégia 1. A quarta análise apresentou erros padrões alterados, valor de p alterado e coeficientes iguais, comparando com a estratégia 3, com EPA maiores que 1 para todos os coeficientes, mostrando que os erros padrões foram subestimados quando se considerou AAS. Ainda, os autores afirmam que desconsiderar o delineamento quando o efeito do delineamento é maior que 2 resulta em uma significativa subestimação do erro padrão. Na estratégia 5, comparada à estratégia 4, os coeficientes não são diferentes, mas os erros padrões são e, conseqüentemente, o valor de p. Neste caso, uma variável independente mostrou-se não significativa no modelo, sendo que nas outras quatro estratégias esta variável foi

significativa. Por fim, os autores recomendam a incorporação das características do delineamento amostral.

Korn e Graubard (1995) analisaram resultados de regressão linear simples, diferença entre proporção, regressão logística e diferença entre médias obtidas para os dados da *National Maternal and Infant Health Survey – 1998* no aplicativo SUDAAN. Os resultados obtidos mostraram estimativas diferentes para as análises ponderada e não ponderada. Comentam que estimadores ponderados têm a desvantagem de maior variabilidade que estimadores não-ponderados.

Dois exemplos são apresentados por Wang et al. (1997), considerando amostragem estratificada em dois estágios. Os resultados demonstraram que os erros padrões das estimativas de média, desconsiderando o delineamento complexo (estratificação, conglomerado e peso), foram menores que os correspondentes para estimativas ajustadas ao delineamento. Na estimativa de proporção também houve diferença, e seus erros padrões foram menores que os correspondentes considerando a ponderação. Concluem que, desconsiderar ponderação na estimação de média e proporção, quando a probabilidade não for a mesma para todos os elementos da amostra, resulta em estimativas não corretas.

Dados da *Behavioral Risk Factor Surveillance System* do CDC (*Centers for Disease Prevention and Control*) foram analisados por Brogan (2003), no SUDAAN, considerando o delineamento complexo, e no SAS, considerando duas situações: análise não ponderada (atribuindo peso 1 para todos os valores da variável peso) e análise ponderada. Os resultados do estudo mostram que a prevalência é superestimada em 10% não incorporando ponderação no SAS (rotina tradicional), comparado ao SUDAAN. A prevalência também foi superestimada, mas em menor

grandeza, na análise ponderada no SAS. O erro padrão é superestimado utilizando o SUDAAN em 35% comparando com a análise não ponderada do SAS e quase a mesma diferença para a análise ponderada do SAS. Ainda, os resultados do teste qui-quadrado mostram que a análise não ponderada do SAS resulta em valor de qui-quadrado maior, gerando valor p mais significativo, comparada à análise do SUDAAN.

Ciol et al. (2006) encontraram diferença nos resultados da regressão logística entre a análise ponderada e não ponderada com dados da *Medicare Current Beneficiay Survey* (MCBS) realizada com a população dos Estados Unidos. No estudo, concluíram que a maior diferença entre os coeficientes dos modelos de regressão logística considerando os pesos e desconsiderando-os foi devido à sobreamostragem de um estrato e à alta proporção da categoria de referência da variável explicativa nesse estrato.

Silva et al. (2002), utilizando os dados da PNAD, descrevem como podem ser considerados os diversos aspectos do plano amostral complexo pois, muitas vezes, as análises feitas por analistas que trabalham fora da agência produtora dos dados utilizam modelagem supondo dados obtidos através de AAS com reposição. Descrevem como podem ser considerados na análise dos dados da PNAD e os diversos aspectos de seu plano amostral complexo: estratificação, conglomeração, probabilidades desiguais de seleção e ajustes dos pesos para calibração.

Lehtonen et al. (2002) discutem a análise de dados amostrais complexos através de técnicas de modelagem multivariada, demonstrando os métodos de *design-based* (amostragem complexa) e *model-based* (modelos multinível), sendo o interesse maior no efeito da conglomeração.

3.3 APLICATIVOS PARA ANÁLISE DE DADOS PROVINDOS DE AMOSTRAGEM COMPLEXA

Atualmente, existe grande disponibilidade de aplicativos para análise de dados provindos de amostragem complexa, porém eles ainda são pouco utilizados. Assim, nesta seção são apresentados aplicativos disponíveis para tratamento de dados com delineamento amostral complexo com o intuito de uma maior divulgação dos mesmos. Na seqüência, apresentam-se alguns estudos que abordaram o uso e comparação de aplicativos para incorporar o delineamento complexo.

Aplicativos para análise de amostragem complexa

Aplicativos disponíveis para análise de dados provindos de amostragem complexa são apresentados no quadro 2. Esses aplicativos variam quanto a recursos disponíveis e preços, sendo que há disponibilidade de aplicativo livre.

Para que seja possível incorporar o plano amostral utilizado na pesquisa, é necessário que estejam disponíveis no banco de dados tanto as variáveis da pesquisa quanto os identificadores das características do plano amostral, como conglomerados e estratos, nos diferentes níveis de amostragem e também o peso final.

No anexo A constam os comandos para medidas descritivas, regressão linear, regressão logística e regressão de Poisson considerando um delineamento estratificado por conglomerado e pesos amostrais.

STATA v.9

No presente estudo foram utilizados o STATA v.8 e o STATA v.9. Nas duas versões do programa, é usado o prefixo *svy* nos comandos referentes ao tratamento de dados de amostragem complexa, permitindo incorporar características como estratificação, conglomeração e peso amostral. O STATA v.8 permite analisar somente o primeiro estágio da amostragem porém, o STATA v.9 permite analisar vários estágios.

Quadro 2 – Aplicativos estatísticos para análise de dados com delineamento amostral complexo

Aplicativo	**Versão atual**	**Mantenedora**	**Página na internet**
STATA	v. 10	Stata Corporation	http://www.stata.com/
SAS	v. 9.1.3	SAS Institute Inc.	http://www.sas.com/technologies/analytics/statistics/index.html
SPSS	v. 16	SPSS Inc.	http://www.spss.com/complex_samples/
SUDDAN	v. 9.0.3	Research Triangle Institute	http://www.rti.org/sudaan/
WesVar	v. 5.1	WESTAT	http://www.westat.com/wesvar/
EPI INFO	v. 3.4.3	Center for Disease Control and Prevention	http://www.cdc.gov/epiinfo/
R	v. 2.6.1 v.3.6-13[a]	R Foundation	http://www.r-project.org/

[a] versão do pacote de amostragem complexa

No anexo B apresenta-se um fluxograma mostrando como é realizada a definição do plano amostral no aplicativo STATA.

Archer e Lemeshow (2006) desenvolveram um teste de ajuste para modelo de regressão logística para amostragem complexa no STATA, implementando um comando, denominado – *svylogitgof* e que está disponível na seguinte página da internet: http://www.stata-journal.com/software/sj6-1/st0099/. Este comando é usado logo após o comando *svy: logistic*, que ajusta o modelo de regressão logística.

Estudos abordando o uso de aplicativos apropriados para análise de amostragem complexa

Nesta seção, apresenta-se uma abordagem histórica de trabalhos anteriores que comparam aplicativos disponíveis nas respectivas épocas. Com o desenvolvimento rápido dos aplicativos, mudanças praticamente anuais de versões e a inclusão contínua de um maior número de técnicas estatísticas programadas para amostragem complexa, muitos destes estudos comparativos referem-se ao aplicativo na versão atualizada da época. Atualmente, com a incorporação de novas técnicas estatísticas no módulo de *complex survey* e de novas opções para diferentes tipos de delineamentos amostrais, os aplicativos estão igualando-se, diferenciando-se apenas na maior ou na menor interatividade com o usuário.

Pessoa e Silva (1998) apontaram que a revolução da informática havia criado condições extremamente favoráveis à utilização de dados estatísticos produzidos por órgãos como o IBGE. Porém, discutiram que certos cuidados precisam ser tomados

para a utilização correta dos dados de pesquisas como as que o IBGE produz. Ressaltaram que estes dados provêm de população finita, envolvendo diferentes características, como probabilidades distintas de seleção, estratificação e conglomeração das unidades, ajustes para compensar não-resposta e outros ajustes. Alertam que aplicativos tradicionais de análise ignoram estes aspectos, podendo resultar em estimativas incorretas.

Wang (2001) alertou para o uso incorreto de procedimentos tradicionais na análise de dados de amostragem complexa, indicando que aplicativos estatísticos como o SUDAAN, STATA, SAS e WesVar foram os pioneiros a introduzir rotinas para análise apropriada destes dados.

STATA v.5.0, SUDAAN v.7.0 e WesVarPC v.2.02 foram comparados por Cohen (1997), em relação à facilidade, eficiência (tempo de processamento) e diversificação de técnicas de análise disponíveis para amostragem complexa. Concluíram que não era possível indicar um melhor ou pior aplicativo, sugerindo que o usuário escolhesse a partir de suas necessidades específicas.

Para Chantala e Tabor (1999) as estimativas e erros padrões foram diferentes na análise dos dados do *Add Health Data* no SUDAAN e no STATA, quando as características do delineamento amostral foram incorporadas. Os autores discutem que estimativas pontuais de média, proporções e parâmetros de regressão são afetadas pelos pesos e que as variâncias estimadas são afetadas por conglomerado, estratificação e peso. Também comentam que na análise de sub-populações, o uso somente dos dados da sub-população resulta em estimativa pontual correta, porém o erro padrão poderá não ser correto, já que a estrutura do delineamento não está

disponível. Entretanto, nos aplicativos estatísticos para análise de delineamento complexo são disponibilizadas análises de sub-populações.

Lee (2000) avaliando os mesmos aplicativos recomendava que a facilidade de uso, custo e os recursos disponíveis de interesse de cada usuário são considerações importantes que o usuário deve avaliar na escolha do aplicativo estatístico.

Sousa e Silva (2000) chegaram à mesma conclusão que Cohen (1997), avaliando facilidade de aplicação, eficiência computacional e exatidão dos resultados do módulo CSAMPLE do Epi Info v.6.04, do STATA v.5 e do aplicativo WesVarPC v.2.12. Utilizaram dados da PNDS (Pesquisa Nacional sobre Demografia e Saúde, 1996), para analisar média e proporção utilizando os aplicativos acima citados. Ainda, concluíram que o Epi-Info é mais limitado na disponibilidade de técnicas de análise, porém seu uso é simples e gratuito. O STATA e WesVarPC têm diversidade de técnicas de análise, porém têm custo de aquisição.

A facilidade de analisar dados considerando a amostragem complexa, utilizando o STATA foi demonstrada por Lemeshow e Cook (1999) e Guillén et al. (2000). Os autores apontam que isto deve encorajar os pesquisadores a adotar essa estratégia de análise.

Schaefer et al. (2003) comparando SUDAAN v.8, STATA v.8, SAS v.8 e WesVar v.4 concluíram que cada usuário deve avaliar suas necessidades na seleção do aplicativo estatístico. Argumentam que estes aplicativos variam em relação às técnicas de análise disponível.

A necessidade de incorporar o plano amostral foi assunto de estudos apresentados em diversas edições da conferência anual do SUGI (*SAS Users Group International*). Iniciando por An e Watts (1998), mostrando os procedimentos de

amostras complexas incorporados no SAS v.7, como SURVEYSELECT, SURVEYMEANS E SURVEYREG. Tompkins e Siller (2000) compararam médias, proporções e seus erros padrões no SAS v.6.12 (PROC MEANS, rotina que não considera o delineamento complexo), SAS v.7 (PROC SURVEYMEANS, rotina que considera o delineamento complexo) e SUDAAN. Os resultados mostraram erros padrões maiores em análises incorporando o delineamento complexo no SAS v.7 e no SUDAAN. Os autores reforçam, então, a necessidade do uso de aplicativos que considerem o plano amostral para evitar inferências incorretas como as que estariam sendo produzidas pelos procedimentos simples.

No trabalho de Gosset et al. (2002) há algoritmos para calcular variâncias pelo método da reamostragem, já que dados públicos muitas vezes disponibilizam somente os pesos. Concluíram que os resultados são equivalentes aos obtidos pelo SUDAAN. Berglund e Arbor (2002) demonstraram o uso dos comandos no PROC SURVEYMEANS E PROC SURVEYREG no SAS. Além destes procedimentos, Cassel e Rousey (2003) demonstraram o uso do SURVEYSELECT.

Mais recentemente, Siller e Tompkins (2005) compararam estimativas de média, proporção e erros padrões produzidas por quatro aplicativos – SAS, SPSS, STATA e SUDAAN usando dois bancos de dados do NCHS (*National Center for Health Statistics*). Esses dados foram obtidos através de amostragem complexa incluindo conglomerados, estratificação e diferentes probabilidades de seleção. Apontam a necessidade de utilizar um aplicativo que considere tal complexidade da amostra na análise dos dados. Os quatro aplicativos analisados utilizam linearização por série de Taylor como método de estimação. Concluem que os resultados são idênticos entre os quatro aplicativos analisados.

Na recente conferência do SUGI, Gosset et al. (2006) chamam a atenção para a necessidade no uso de um aplicativo estatístico que considere o plano amostral para analisar dados oficiais como NHANES (*National Health and Nutrition Examination Survey*). Demonstraram o uso das rotinas SURVEYMEANS, SURVEYFREQ, SURVEYREG e SURVEYLOGISTIC do SAS com dados do NHANES.

Figueiredo (2004) realizou estudo com a biblioteca ADAC do aplicativo R comparando-a com o SUDDAN. Apontou vantagens da utilização da biblioteca ADAC, a linguagem aberta e gratuita, tornando-a de fácil acesso, a qual suporta bancos gerados de qualquer outro programa e a disponibilidade para acesso à estrutura de cada função. Como desvantagens, descreveu que algumas rotinas não estão ainda otimizadas ocupando muita memória, aceita apenas alguns tipos de planos amostrais e sua estrutura de análise para regressão se restringe à regressão normal e logística. Para o SUDAAN apontou como vantagens a exigência de pouca memória, disponibilidade de mais técnicas implementadas e, como desvantagens, que os procedimentos são fechados, não permitindo verificar como operam, não possuindo ambiente gráfico e que suas saídas são difícies de serem utilizadas.

Oito aplicativos, sendo quatro livres, com capacidade de analisar dados de amostragem complexa foram analisados por Brogan (2005) comparando custo do aplicativo, métodos de estimação de variância, opções de análise, interface e vantagens/desvantagens. Análises foram realizadas com dados reais para cinco aplicativos dentre os oito (STATA, SAS, SUDAAN, WesVar e Epi-Info), obtendo-se resultados equivalentes para todos, quando os mesmos métodos de linearização por série de Taylor ou reamostragem foram usados.

STATA, SUDAAN e SAS foram usados por Chantala (2006) para análise de regressão múltipla obtendo-se os mesmos resultados para os coeficientes de regressão e seus erros padrões.

3.4 ANÁLISE MULTINÍVEL

A metodologia da análise multinível será abordada nesta seção, como uma metodologia alternativa para analisar dados provindos de plano amostral complexo, considerando-se diferentes características da amostra complexa como níveis distintos da análise. Inicia-se a seção com a definição de modelo multinível e na seqüência são apresentados aplicativos e algumas considerações para uso desta metodologia nos dados provindos de amostragem complexa.

Segundo Vieira (2001) e Pessoa e Silva (1998) há duas abordagens principais para tratar a estrutura dos dados de pesquisas amostrais complexas. Uma delas é a abordagem descrita na seção anterior, denominada abordagem de amostragem complexa (análise agregada ou marginal segundo alguns autores). A outra é a abordagem de análise multinível (ou análise desagregada) que será descrita nesta seção e na qual se incorpora mais explicitamente a estrutura dos dados no procedimento de análise. Na abordagem multinível, efeitos de conglomeração são vistos como parte integral da estrutura da população, que deve ser adequadamente modelada e pode contribuir para melhorar a compreensão das relações entre as variáveis (Pessoa e Silva, 1998).

Em investigações epidemiológicas é muito comum o estudo de indivíduos agrupados em níveis ou hierarquias. Esta estrutura pode estar presente de forma intrínseca, como por exemplo, alunos dentro de escolas, médicos ou pacientes dentro de hospitais, indivíduos dentro de famílias, ou a estrutura pode ser criada pelo tipo de delineamento da pesquisa, como nos estudos longitudinais, os quais geram grupos de observações no mesmo indivíduo (Zanini, 2007; Neuhaus e Segal, 1993), ou ainda em estudos que utilizam conglomerados e estratos no plano amostral.

A análise multinível possibilita examinar grupos (ou amostras de grupos) e indivíduos (ou amostras de indivíduos) dentro desses grupos, simultaneamente, considerando a variável resposta medida no nível individual e as variáveis explicativas, que podem ser medidas no nível dos indivíduos ou dos grupos aos quais pertencem (Snijders e Bosker, 1999).

O modelo multinível pode ser visto como um sistema hierárquico de equações de regressão, possibilitando a estimação dos efeitos intragrupo (efeitos individuais) e dos efeitos entregrupos (efeitos contextuais) segundo Goldstein (2003). Também é possível modelar a estrutura de variância em cada um dos níveis.

Um modelo multinível difere de um modelo tradicional por ser formado por dois componentes: um fixo e outro aleatório (Snijders e Boskers, 1999). A parte fixa indica a magnitude das associações entre as variáveis, enquanto que a parte aleatória mostra as diferenças dos grupos e as variâncias nos diferentes níveis (Zanini, 2007; Merlo, 2003).

Modelo multinível para desfecho contínuo

A seguir será apresentado o modelo multinível com dois níveis para desfecho contínuo por ser este o utilizado no presente trabalho, utilizando a notação comum introduzida em Zanini (2007), Goldstein (2003) e Snijders e Boskers (1999).

Nos modelos com dois níveis ocorrem n_j unidades do nível 1 para cada unidade $j(j = 1,2,...J)$ do nível 2. Os modelos para o nível 1 são desenvolvidos separadamente em cada unidade do nível 2, considerando a possibilidade de variação dos interceptos e das inclinações:

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + u_{0j} + e_{0ij} \qquad \text{com } i = 1,2,...,n_j \text{ e } j = 1,2,...J$$

em que:

Y_{ij} : desfecho do *i*-ésima unidade do nível 1, agrupada na *j*-ésima unidade do nível 2;

X_{ij} : variável preditora medida na *i*-ésima unidade do nível 1, agrupada na *j*-ésima unidade do nível 2;

β_0 : intercepto geral do modelo;

β_1 : coeficiente de inclinação, associado à variável preditora X;

u_{0j} : efeito aleatório do nível 2;

e_{0ij} : o efeito aleatório do nível 1;

Os resíduos u_{0j} e e_{0ij} são supostamente independentes e normalmente distribuídos, com média zero e variâncias σ^2_{u0} e σ^2_{e0}, respectivamente. A variância residual, ou seja, a variância condicionada a X é dada por:

$$Var(Y_{ij} \mid X_{ij}) = \text{var}(u_{oj}) + \text{var}(e_{oij}) = \sigma^2_{u0} + \sigma^2_{e0}$$

A covariância entre dois indivíduos (i_1, i_2) no mesmo grupo j é:

$$Cov(Y_{i1j}, Y_{i2j} \mid X_{i1j}, X_{i2j}) = \text{var}(u_{oj}) = \sigma^2_{u0}$$

O coeficiente de correlação intraclasse ou intragrupo, o qual expressa a fração da variabilidade total que pode ser atribuída ao nível 2, indicando o grau de agrupamento da população, correspondendo à correlação entre os valores de dois indivíduos de um grupo controlado para a variável X, é dado por:

$$\rho_I(Y_{ij} \mid X_{ij}) = \frac{\sigma^2_{u0}}{\sigma^2_{u0} + \sigma^2_{e0}}$$

Estende-se o modelo para:

$$Y_{ij} = \beta_{0j} + \beta_{1j} X_{1ij} + e_{0ij}$$

Então, no nível 2, serão dois modelos:

$$\beta_{0j} = \beta_0 + u_{0j}$$
$$\beta_{1j} = \beta_1 + u_{1j}$$

em que:

β_{0j} : intercepto para a *j*-ésima unidade do nível 2;

β_{1j} : coeficiente de inclinação, associado à variável X da i-ésima unidade do nível 1, agrupada na *j*-ésima unidade do nível 2;

β_0 : valor esperado dos interceptos no nível 2;

β_1 : valor esperado das inclinações no nível 2;

u_{0j} : efeito aleatório da *j*-ésima unidade do nível 2 no intercepto β_{0j} ;

u_{1j} : efeito aleatório da *j*-ésima unidade do nível 2 na inclinação β_{1j} .

Substituindo-se as equações do nível 2 tem-se:

$$Y_{ij} = \beta_0 + \beta_1 X_{1ij} + (u_{0j} + u_{1j} X_{1ij} + e_{0ij})$$

Esses modelos são denominados modelos de coeficientes aleatórios, uma vez que pressupõe que cada grupo tem um intercepto (β_{0j}) e uma inclinação (β_{1j}) diferentes, que variam através dos grupos, apresentando uma estrutura de variância-covariância complexa como:

$$Var\begin{bmatrix} u_{0j} \\ u_{ij} \end{bmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_{u0} & \sigma_{u0u1} \\ \sigma_{u1u0} & \sigma^2_{u1} \end{pmatrix}\right] \text{ e } Var(e_{0ij}) \sim N(0, \sigma^2_{e0})$$

A variância é modelada como uma função de uma variável preditora.

Os principais métodos de estimação pressupondo normalidade são: IGLS (*Iterative Generalized Least Squares*) equivalente ao ML (*Maximum Likelihood*) e o RIGLS (*Restricted/Reweighted Iterative Generalized Least Squares*) equivalente ao REML (*Residual Maximum Likelihood*).

Interceptos e inclinações são estimados para unidades dentro do nível 2 e os resíduos são obtidos de forma diferente da tradicional. Denominam-se resíduos reduzidos (*shrunken*), com uma menor ou maior extensão em torno do relacionamento médio global, e podem ser estimados multiplicando-se os resíduos brutos ($\tilde{y}_j$) por um fator de redução:

$$\hat{u}_j = \frac{n_j \sigma_{u0}^2}{n_j \sigma_{u0}^2 + \sigma_{eo}^2} \tilde{y}_j$$

Observa-se que o fator de redução depende do número de grupos do nível 2 e das estimativas de variância dos dois níveis.

Estes resíduos reduzidos ordenados podem ser representados em um gráfico, onde as barras em torno de cada resíduo representam o intervalo de confiança de 95% (Moraes, 2007).

O aplicativo MLwiN v.2.02 é específico para análise de modelos multinível. O SAS v.9.1.3 disponibiliza o procedimento PROC MIXED para análise de modelos multinível. Estes dois aplicativos computacionais foram utilizados neste estudo. Versões mais recentes de alguns outros aplicativos estão incorporando análise multinível, por exemplo, o STATA v.10.

No anexo C apresentam-se os comandos para modelagem multinível com 2 níveis, com desfecho contínuo, para o SAS e o MLwiN. Detalhes dos comandos são

disponibilizados por: Pretto (2003) e Singer (1998) no caso do SAS Institute (2004) e, Rasbash et al. (2005), Ferrão (2003) e Barros (2001) no caso do MLwiN.

Abordagem multinível considerando amostragem complexa

Modelos multinível são freqüentemente usados para analisar dados de planos amostrais complexos, permitindo estudar o efeito das variáveis ao nível de conglomerados sobre a variável dependente (desfecho) do nível individual (Asparouhov e Muthen, 2006). Cada estágio de amostragem corresponde a um nível na modelagem multinível, sendo que a UFA (Unidade Final de Amostragem) é a do nível 1 e os conglomerados de cada estágio de amostragem constituem os demais níveis (Rabe-Hesketh e Skrondal, 2006).

Asparouhov e Muthen (2006), Rabe-Hesketh e Skrondal (2006) e Zhang (2005) comentam que se os pesos amostrais são ignorados, no caso de usar modelagem multinível padrão, o parâmetro estimado pode estar incorreto. Porém, diferente da estratificação e conglomeração, a probabilidade de seleção desigual, indicada pelos pesos não é fácil de ser incorporada na análise (Asparouhov e Muthen, 2006).

A aplicação de modelos multinível em dados de amostragem complexa é bastante recente (Chantala et al., 2006). Os autores afirmam que aplicativos estatísticos incorporaram essa metodologia possibilitando também o uso de pesos amostrais, porém, deve-se ressaltar que os pesos para análise multinível devem ser construídos diferentemente daqueles usados na abordagem da amostragem complexa.

Chantala et al. (2006) apontam que os institutos que distribuem dados secundários e que os programadores de aplicativos estatísticos freqüentemente deixam a tarefa da construção e de definição dos pesos amostrais para o usuário. Complementam que estes métodos de definição podem ser diferentes para os diferentes aplicativos.

Os métodos de construção de pesos para análise multinível mais comuns em aplicativos são os definidos por e Asparouhov (2004) e Pfeffermann et al. (1998). Os aplicativos STATA, LISREL, MLwiN utilizam os métodos definidos por Pfeffermann et al. (1998), que usam os pesos dos dois estágios de amostragem para construir os pesos amostrais para o nível 1 e o nível 2. Por outro lado, o aplicativo MPLUS usa o método desenvolvido por Asparouhov (2004) que combina os pesos dos dois estágios de amostragem para construir um único peso amostral.

Aplicação de modelos multinível em dados de amostragem complexa com intervenção

Estudos com intervenções na área da saúde freqüentemente são implementados ao nível de grupo (Ukoumunne et al., 1999), como por exemplo, cidades, escolas, hospitais e clínicas. Nestes estudos, há a aleatorização de grupos, a intervenção é realizada ao nível de grupo e o desfecho é avaliado a nível individual ou ao nível de grupos. Indivíduos no mesmo grupo tendem a ser mais similares entre si do que indivíduos entre grupos, e isto deve ser levado em consideração na análise dos dados (Donner e Klar, 1994). As fórmulas tradicionais para determinar o tamanho da amostra tendem a subestimar o número de indivíduos necessários, pois

estas fórmulas consideram a variação dentro dos grupos, mas não entre os grupos (Ukoumunne et al., 1999). Da mesma forma, os métodos tradicionais de tratamento estatístico dos dados não são apropriados para estudos com conglomerados, pois assumem que os indivíduos são independentes. Assim, os erros padrões do efeito da intervenção tendem a ser subestimados, resultando em intervalos de confiança mais estreitos e valores p (significância) menores.

Segundo Donner e Klar (1994), a análise de desfecho contínuo de um delineamento randomizado em grupo pode ser realizada usando modelos mistos, com o objetivo de verificar o efeito do tratamento. Covariáveis para tratamento, estratos e medida *baseline* podem ser incluídas no modelo misto.

Veugelers e Fitzgerald (2005) utilizaram modelos multinível para avaliar o efeito de um programa de prevenção de obesidade infantil em escolas. O estudo realizou-se em 2003, incluindo 5.200 estudantes de cinco séries em 282 escolas. Os dados foram analisados por regressão logística multinível, considerando a escola como fator contextual (nível 2), para variáveis de desfecho binárias no nível 1, e também foram utilizados modelos de regressão linear multinível para desfechos contínuos do nível 1.

4. OBJETIVOS

4.1 Objetivo Geral

Este trabalho tem por objetivo descrever e comparar métodos para tratamento de dados provindos de planos amostrais complexos, freqüentemente utilizados em estudos epidemiológicos, através de duas abordagens: amostragem complexa e modelo multinível.

4.2 Objetivos Específicos

- Descrever métodos de estimação para amostragem complexa.
- Obter e comparar estimativas de média, proporção e seus respectivos erros padrões na análise de dados em diferentes estudos, considerando amostragem complexa em relação à análise de dados supondo AAS.
- Obter e comparar estimativas para parâmetros do modelo de regressão Poisson e seus respectivos erros padrões, considerando amostragem complexa em relação à análise de dados supondo AAS.
- Obter e comparar resultados de análise de regressão linear utilizando abordagem de amostragem complexa e abordagem de modelos multinível.
- Descrever como se pode incorporar o delineamento complexo nas análises estatísticas de estudos epidemiológicos para obter estimativas corretas.
- Apresentar e descrever recursos computacionais, atualmente disponíveis, para análise de dados provenientes de amostragem complexa disponibilizados nos aplicativos.

5. REFERÊNCIAS BIBLIOGRÁFICAS

(1) An A, Watts D. New SAS Procedures for analysis of sample survey data. Proceedings of the SAS users group international 23. 22-25/03/1998. Nashville. Tennessee.

(2) Archer KJ, Lemeshow S. Goodness-of-fit test for a logistic regression model fitted using survey sample data. The Stata Journal 2006;6(1):97-105.

(3) Asparouhov T. Weighting for Unequal Probability of Selection in Multilevel Modeling. Mplus Web Notes: No 8. Em: http://www.statmodel.com/download/webnotes/MplusNote81.pdf, acesso em 27/03/2007.

(4) Asparouhov T, Muthen B. Multilevel Modeling of Complex Survey Data. Proceedings of the Joint Statistical Meeting in Seattle. 08/2006. Seattle.

(5) Barata RB, Moraes JC, Antonio PRA, Dominguez M. Inquérito de cobertura vacinal: avaliação empírica da técnica de amostragem por conglomerados proposta pela Organização Mundial da Saúde. Revista Panamericana de Salud Publica 2005;17(3):184-90.

(6) Barros AJD. Modelos multinível: primeiros passos. Departamento de Medicina Social. Faculdade de Medicina. Universidade Federal de Pelotas. 2001.

(7) Barros AJD, Bertoldi AD. Desigualdades na utilização e no acesso a serviços odontológicos: uma avaliação a nível nacional. Ciência e Saúde Coletiva 2002;7(4):709-17.

(8) Berglund PA, Arbor A. Analysis of complex sample survey data using the SURVEYMEANS and SURVEYREG procedures and macro coding. Proceedings of the SAS users group international 27. 14-17/04/2002. Orlando. Florida.

(9) Bolfarine H, Bussab WO. Elementos de Amostragem. São Paulo: Edgar Blücher; 2005.

(10) Brogan DJ. Pitfalls of using standard statistical software packages for sample surveys data. Em: http://www.fas.harvard.edu/~stats/survey-soft/donna_brogan.html, acesso em 19/09/2003.

(11) Brogan D. Sampling error estimation for survey. Household sample surveys in developing an transition countries. New York: United Nations Publication; 2005. p. 447-90.

(12) Bueno MB, Marchioni DML, Fisberg RM. Evolução nutricional de crianças atendidas em creches públicas no município de São Paulo, Brasil. Revista Panamericana de Salud Publica 2003;14(3):165-70.

(13) Cassell DL, Rousey A. Complex sampling designs meet the faming turkey of glory. Proceedings of the SAS users group international 28. 30/03-02/04/2003. Seattle, Washington.

(14) Carlson BL. Software for statistical analysis of sample survey data. Em: http://www.fas.harvard.edu/~stats/survey-soft/blc_eob.html, acesso em 25/09/2003.

(15) Chambers RL, Skinner CJ. Analysis of Survey Data. Chichester: John Wiley; 2003.

(16) Chantala K. Guidelines for analysing add health data. Em: http://www.cpc.unc.edu/projects/addhealth/files/wt_guidelines.pdf, acesso em 17/09/2007.

(17) Chantala K, Blanchette D, Suchindran CM. Software to compute sampling weights for multilevel analysis. Em: http://www.cpc.unc.edu/restools/data_analysis/ml_sampling_weights/Compute%20Weights%20for%20Multilevel%20Analysis.pdf, acesso em 17/01/2008.

(18) Chantala K, Tabor J. Strategies to perform a design-based analysis using the add health data. Em: http://www.cpc.unc.edu/projects/addhealth/files/weight1.pdf, acesso em 20/09/2003.

(19) Ciol MA, Hoffman JM, Dudgeon BJ, Shumway-Cook A, Yokston KM, Chan L. Understanding the Use of Weights in the Analysis of Data From Multistage Surveys. Archives of Physical Medicine Rehabilitation 2006;87:299-303.

(20) Cochran W. Sampling Techniques. New York: John Wiley; 1977.

(21) Cohen SB. An evaluation of alternative PC-based software packages developed for the analysis of complex survey data. The American Statistician 1997; 51(3):285-92.

(22) Cordeiro R. Efeito do desenho em amostragem de conglomerado para estimar a distribuição de ocupações entre trabalhadores. Revista de Saúde Pública 2001; 35(1):10-5.

(23) Donner A, Klar N. Cluster randomization trials in epidemiology: theory and application. Journal of Statistical Planning and Inference 1994;42(1-2): 37-56.

(24) Egede LE. Lifestyle modification to improve blood pressure control in individuals with diabetes. Diabetes Care 2003; 26(3):602-7.

(25) Egede LE. Diabetes, Major Depression and Functional Disability Among U.S. Adults. Diabetes Care 2004; 27(2): 421-8.

(26) Epi-Info, Center for Disease Control and Prevention. Version 6.0.3

(27) Ferrão ME, Beltrão KI, Santos DP. Modelo de regressão multinível: aplicação ao estudo do impacto da política de não-repetência no desempenho escolar dos alunos da 4ª série. Pesquisa e Planejamento Economico 2002; 32(3).

(28) Figueiredo CC. Análise de Regressão Incorporando o Esquema Amostral. Dissertação de mestrado em Estatística. Universidade de São Paulo; 2004.

(29) Giatti L, Barreto SM. Situação do indivíduo no mercado de trabalho e iniqüidade em saúde no Brasil. Revista de Saúde Pública 2006;40(1):99-106.

(30) Goldstein H. Multilevel statistical models. 3 ed. Edward Arnold; 2003.

(31) Gossett JJ, Simpson P, Parker JG, Simon WL. How complex can complex survey analysis be with SAS. Proceedings of the SAS users group international 27. 14-17/04/2002. Orlando. Florida.

(32) Gossett JJ, Jo C, Simpson P. U. S. Health and Nutrition: SAS survey procedures and NHANES. Proceedings of the SAS users group international 31. 26-29/03/2006. San Francisco. California.

(33) Guillén M, Juncá S, Rué M, Aragay JM. Efecto del diseño muestral en el análisis de encuestas de deseño complejo. Aplicación a la encuesta de salud de Catalunya. Gaceta Sanitaria 2000;14(5):399-402.

(34) Heeringa SG, Liu J. Complex sample design effects and inference for mental health survey data. International Journal of Methods in Psychiatric Research 1997;7(1):56-65.

(35) Hope AD, Shannon ED. A comparison of two procedures to fit multi-level data: PROC GLM verus PROC MIXED. Proceedings of the SAS users group international 30. 10-13/04/2005. Philadelphia. Pennsylvania.

(36) Hernández B, Haene J, Barquera S, Monterrubio E, Rivera J, Shamah T, et al. Factores asociados con la actividad física en mujeres mexicana en edad reproductiva. Revista Panamericana de Salud Publica 2003;14(4):235-45.

(37) Hosmer DW, Lemeshow S. Applied Logistic Regression. 2 ed. New York: John Wiley & Sons; 2000. p. 211-22.

(38) Horton NJ, Fitzmaurice GM. Regression analysis of multiple source and multiple informant data from complex survey samples. Statistics in Medicine 2004;23:2911-33.

(39) Kish L. Survey Sampling. New York: John Wiley; 1965.

(40) Kneipp SM, Yarandi HN. Complex sampling designs and statistical issues in secondary analysis. Western Journal of Nursing Research 2002;24(5):552-66.

(41) Korn EL, Graubard BI. Epidemiologic studies utilizing surveys: accounting for the sampling design. American Journal of Public Health 1991;81(9):1166-73.

(42) Korn EL, Graubard BI. Analysis of large health surveys: accounting for the sampling design. Royal Statistical Society 1995;158(2):263-95.

(43) Korn EL, Graubard BI. Examples of differing weighted and unweighted estimates from a sample survey. The American Statistician 1995;49(3):291-5.

(44) Kreuter F, Valliant R. A survey on survey statistics: what is done and can be done in Stata. The Stata Journal 2007;7(1):1-21.

(45) Lee R. Strengths and Limitations of Using SUDAAN, Stata and WesVarPC for Computing Variances from NCES Data Sets. National Center for Education Statistics 2000. Em: http://nces.edu.gov/pubs2000/200003.pdf, acesso em 18/01/2007.

(46) Lehtonen R, Pahkinen E. Practical Methods for Design and Analysis of Complex Survey. 2 ed. Chichester. England: John Wiley & Sons Ltd; 2004.

(47) Lehtonen R, Djerf K, Härkänen T, Laiho J. Design-based and model-based methods in analysing complex health survey data: a case study. Proceedings of Statistics Canada Symposium. 2002

(48) Leite PGPG, Silva DBN. Análise da situação ocupacional de crianças e adolescentes nas regiões sudeste e nordeste do Brasil utilizando informações da PNAD 1999. Proceedings XIV Encontro Nacional de Estudos Populacionais. 4-8/11/2002. Ouro Preto. Minas Gerais.

(49) Lemeshow S, Letenneur L, Dartigues JF, Lafont S, Orgogozo JM, Commenges D. Illustration of analysis taking into account complex survey considerations: the association between wine consumtion and dementia in the PAQUID study. American Journal of Epidemiology 1998;148(3):298-306.

(50) Lemeshow S, Cook ED. Pratical considerations in the analysis of complex sample survey data. Revue d'epidémiologie et de santé publique 1999;47:479-87.

(51) Lumley T. Analysis of complex survey samples. Journal of Statistical Software 2004;9(8):1-19.

(52) Lumley T. The survey package. Manual survey R Em: http://cran-r.c3sl.ufpr.br/, acesso em 18/02/2008.

(53) Moraes AB. Baixo peso de nascidos vivos no Rio Grande do Sul, Brasil: uma análise estatística multinível. Tese de doutorado em Epidemiologia. Faculdade de Medicina. Universidade Federal do Rio Grande do Sul; 2007.

(54) MLwiN, version 2.02. Multilevel Models Project, c. 2000. London: Institute of Education, University of London.

(55) Mundstock EC. Amostragem I. Cadernos de Matemática e Estatística. Série B Instituto de Matemática. Universidade Federal do Rio Grande do Sul. 2005.

(56) Natarajan S, Lipsitz SR, Fitzmaurice GM, Moore CG, Gonin R. Variance estimation in complex survey sampling for generalized linear models. Applied Statistics 2008;57(1):75-87.

(57) Neuhaus JM, Segal MR. Design effects for binary regression models fitted to dependent data. Statistics in Medicine 1993;12:1259-68.

(58) Oliveira MMC. Presença e extensão dos atributos da atenção primária à saúde entre os serviços de atenção primária em Porto Alegre: uma análise agregada. Dissertação de mestrado em Epidemiologia. Faculdade de Medicina. Universidade Federal do Rio Grande do Sul; 2007.

(59) Pérez MC, Utra IB, León AA, Roche RG, Sagué KA, Rosa MC, et al. Estimationes usadas en diseños muestrales complejos: aplicaciones en la encuesta de salud cubana del año 2001. Revista Panamericana de Salud Publica 2004;15(3):176-84.

(60) Pessoa DGC, Silva PLN. Análise de dados amostrais complexos. São Paulo: ABE-Associação Brasileira de Estatística; 1998.

(61) Pfeffermann D. The use of sampling weights for survey data analysis. Statistical Methods in Medical Research 1996;5:239-61.

(62) Pfeffermann D, Skinner C, Goldstein H, Holmes DJ, Rasbash J. Weighting for unequal selection probabilities in multilevel models. Journal of the Royal Statistical Society 1998;60(Série B):23-40.

(63) Pretto K. Modelos multiníveis: caracterização e aplicação. Monografia Conclusão de Curso. Bacharelado em Estatística. Instituto de Matemática. Universidade Federal do Rio Grande do Sul. 2003.

(64) Rasbash J, Steele F, Browne W, Prosser B. A User's Guide to MLwinN. Centre for Multilevel Modelling. University of Bristol; 2005.

(65) Rabe-Hesketh S, Skrondal A. Multilevel modelling of complex survey data. Journal of the Royal Statistical Society A 2006;169(4):805-27.

(66) Ray JJ. A comparison between cluster and "random" sampling. The Journal of Social Psychology 1983;121:155-6.

(67) Rodgers-Farmer A, Davis D. Analyzing complex survey data. Social Work Research 2001 Jan 1;25(3):185-92.

(68) SAS Institute, Inc. SAS statistical software, release 9.1. Cary, NC: SAS Institute, Inc, 2002-2003.

(69) Schaefer E, Potter F, Williams S, Diaz-Tena N, Reschovsky JD, Moore G. Comparison of selected statistical software packages for variance estimation in the CTS surveys. Em: http://hschange.com/CONTENT/575/575.pdf, acesso em 10/04/2006.

(70) Siller AB, Tompkins L. The Big Four: analyzing complex sample survey data using SAS, SPSS, STATA and SUDAAN. 2005. Proceedings of the SAS users group international 30. 10-13/04/2005. Philadelphia. Pennsylvania.

(71) Silva NN. Amostragem Probabilística. São Paulo: EDUSP; 1998.

(72) Silva PLN, Pessoa DGC, Lila MF. Análise estatística de dados da PNAD: incorporando a estrutura do plano amostral. Ciência e Saúde Coletiva 2002;7(4):659-70.

(73) Singer J. Using SAS PROC MIXED fo Fit Multilevel Models Hierarchical Models, and Individual Growth Models. Journal of Educational and Behavioral Statistics 1998;24(4):323-55.

(74) Skinner CJ, Holt D, Smith TMF. Analysis of complex survey. New York. John Wiley & Sons; 1989.

(75) Snijders T, Bosker R. An introduction to basic and advanced multilevel modeling. 1 ed. London: SAGE Publications; 1999.

(76) Sousa MH, Silva NN. Comparação de software para análise de dados de levantamentos complexos. Revista de Saúde Pública 2000;34(6):646-53.

(77) Sousa MH, Silva NN. Estimativas obtidas de um levantamento complexo. Revista de Saúde Pública 2003;37(5):662-70.

(78) StataCorp. STATA, release 9. Stata Corporation. 1985-2005.

(79) Sturgis P. Analysing complex survey data: clustering, stratification and weights. Social Research Update. Department of Sociology. University of Surrey. 2004.

(80) Szwarcwald CL, Damacena GN. Amostras complexas em inquéritos populacionais: planejamento e implicações na análise de dados. Em: http://www.ensp.fiocruz.br/biblioteca/home/exibedetalhesBiblioteca.cfm?ID=4502&tipo=B, acesso em 31/12/2007.

(81) Teixeira AMFB, Kanauth DR, Fachel JMG, Leal AF. Adolescentes e uso de preservativos: as escolhas dos jovens de três capitais brasileiras na iniciação e na última relação sexual. Cadernos de Saúde Pública 2006;22(7):1385-96.

(82) Tompkins L, Siller AB. Analyzing Complex Sample Survey Data: a new beginning. 2000. Proceedings of the SAS users group international 25. 9-12/04/2000. Indianapolis. Indiana.

(83) Thomas SL, Heck RH. Analysis of large-scale secondary data in higher education research. Research in Higher Education 2001;42(5):517-40.

(84) UCLA. Statistical computing seminars survey data analysis in Stata. Em: http://www.ats.ucla.edu/stat/stata/seminars/svy_stata_intro/default.htm, acesso em 27/06/2005.

(85) Ukoumunne OC, Gulliford MC, Chinn S, Sterne JAC, Burney PGJ, Donner A. Methods in health service research: evaluation of health interventions at area and organisation level. British Medical Journal 1999;319:376-9.

(86) Vasconcellos MTL, Portela MC. Índice de massa corporal e sua relação com variáveis nutricionais e sócio-econômicas: um exemplo de uso de regressão linear para um grupo de adultos brasileiros. Cadernos de Saúde Pública 2001;17(6):1425-36.

(87) Vasconcellos MTL, Silva PLN, Szwarcwald CL. Sampling design for the World Health Survey in Brazil. Cadernos de Saúde Pública 2005;21(Sup):89-99.

(88) Veugelers PJ, Fitzgerald AL. Effectiveness of school programs in preventing childhood obesity: a multilevel comparison. American Journal of Public Health 2005;95(3):432-5.

(89) Vieira MT. Um estudo comparativo das metodologias de modelagem de dados amostrais complexos - uma aplicação ao SAEB 99. Dissertação de mestrado em Ciência da Engenharia Elétrica. Pontifícia Universidade Católica do Rio de Janeiro; 2001.

(90) Vieira MT. Amostragem Repetida no Tempo: o uso de painéis. IBGE - Rio de Janeiro/RJ: 1ª Escola de Amostragem e Metodologia de Pesquisa; 2007.

(91) Wang ST, Yu ML, Lin LY. Consequences of analysing complex survey data using inappropriate analysis and software computing packages. Public Health 1997;111:259-62.

(92) Wang MQ. Research notes: analysis of data from complex survey designs. American Journal of Health Behavior 2001;25(1):72-4.

(93) Zanini RR. Modelos multiníveis aplicados ao estudo da mortalidade infantil no Rio Grande do Sul, Brasil, de 1994 a 2004. Tese de doutorado em Epidemiologia. Faculdade de Medicina. Universidade Federal do Rio Grande do Sul; 2007.

(94) Zhang F, Salvucci S, Cohen M. Multilevel Linear Regression Analysis of Complex Survey Data. Em: http://www.amstat.org/sections/srms/Proceedings/papers/2000_029.pdf, acesso em 10/04/2006.

6. ARTIGOS

6.1 ARTIGO 1

IMPACTO DO PLANO AMOSTRAL COMPLEXO NAS ESTIMATIVAS DE COEFICIENTES DE REGRESSÃO DE POISSON EM UM ESTUDO EPIDEMIOLÓGICO

EFFECT OF COMPLEX SAMPLING DESIGN ON ESTIMATES OF POISSON REGRESSION COEFFICIENTS IN AN EPIDEMIOLOGICAL STUDY

[1, 3] Iara Denise Endruweit Battisti

[1, 2] João Riboldi

[2] Elsa Mundstock

[1, 2] Jandyra Maria Guimarães Fachel

[1]Programa de Pós Graduação em Epidemiologia – Universidade Federal do Rio Grande do Sul, Porto Alegre/RS, Brasil

[2]Departamento de Estatística, Instituto de Matemática, Universidade Federal do Rio Grande do Sul, Porto Alegre/RS, Brasil

[3]Departamento de Física, Estatística e Matemática, Universidade Regional do Noroeste do Estado do Rio Grande do Sul, Santa Rosa/RS, Brasil

Endereço de correspondência do autor:
Iara Denise Endruweit Battisti
Av. Borges de Medeiros 550/701 Centro
Santa Rosa/RS CEP:98900-000
iara.battisti@unijui.edu.br

A ser enviado aos Cadernos de Saúde Pública

RESUMO

Introdução: Muitos estudos epidemiológicos utilizam amostragem complexa para coleta dos dados, podendo incluir estratos, conglomerados e probabilidades desiguais de seleção. Este estudo teve por objetivo avaliar o efeito da incorporação do plano amostral complexo em um estudo epidemiológico.

Método: Utilizaram-se os dados da busca ativa domiciliar dos participantes na Campanha Nacional de Detecção de Diabetes Mellitus – CNDDM de 2001, obtidos por amostragem estratificada por conglomerado em três estágios. Estimativas de média, proporção, coeficientes de regressão de Poisson e seus correspondentes erros padrões foram obtidos considerando amostragem aleatória simples, amostragem complexa e cada componente do plano amostral isoladamente.

Resultados: As estimativas pontuais de média e proporção são semelhantes comparando-se amostragem complexa e amostragem aleatória simples, porém observou-se grande diferença nos erros padrões. O mesmo foi observado nas estimativas de coeficientes de regressão de Poisson, mas com menor efeito do plano amostral.

Conclusão: Os resultados mostraram diferenças nos erros padrões da média, proporção e coeficientes de regressão de Poisson entre amostragem complexa e amostragem aleatória simples, indicando a necessidade de incorporação da complexidade do plano amostral na análise dos dados.

Palavras-chave: amostragem complexa; efeito do plano amostral; conglomerado.

ABSTRACT

Introduction: Many epidemiological studies use complex sampling designs and may include stratification, clustering or unequal selection probabilities. This study evaluated the effect of using a complex sampling design in an epidemiological study.

Method: Data were retrieved from a house-to-house survey of participants in the 2001 Brazilian Diabetes Detection Campaign (Campanha Nacional de Detecção de Diabetes Melitus - CNDDM), which were collected by stratified cluster sampling in three stages. Estimates of means, proportions, Poisson regression coefficients and their corresponding standard errors were calculated for simple random sampling, complex sampling, and individual analysis of each component of the sampling design.

Results: Mean and proportion point estimates were similar when complex sampling and simple random sampling were compared, but there was a great difference in standard errors. The same was found for estimates of Poisson regression coefficients that were less affected by sampling design.

Conclusion: Results showed differences in the standard errors of means, proportions and Poisson regression coefficients between complex and simple random sampling, which indicates that sample design complexity should be incorporated into data analysis.

Key words: Complex sample, design effect; cluster.

INTRODUÇÃO

Muitos estudos epidemiológicos utilizam amostragem complexa para coleta dos dados, podendo incluir estratos, conglomerados e probabilidades desiguais de seleção. A amostragem complexa tem a vantagem de diminuir os custos de levantamento dos dados[1], bem como, evitar a necessidade de uma listagem de todos os elementos que compõem a população[2]. No entanto, a análise dos dados deve incorporar a estrutura do plano de amostragem na obtenção das estimativas.

Têm-se dois desafios para análise de dados provindos de amostragem complexa[3,4]: (1) obter estimativas pontuais corretas, isto é, adequadas ao plano amostral e (2) calcular corretamente variâncias e erros padrões. As três características – estratos, conglomerados e probabilidades desiguais - possíveis de estarem presentes numa amostragem complexa têm diferentes efeitos no viés e na variância do estimador.

Os métodos de análise tradicionalmente utilizados por muitos pesquisadores pressupõem que os dados foram obtidos a partir de uma amostra aleatória simples (AAS). A maioria dos aplicativos de análise estatística somente oferecia recursos para análise de dados supondo AAS até recentemente. As suas novas versões incorporaram recentemente o tratamento de dados provenientes de amostras complexas.

Considerando que grandes inquéritos na área da saúde e bases de dados disponibilizados por institutos de pesquisa, como por exemplo, o IBGE, os quais envolvem dados obtidos com planos amostrais complexos e considerando a recente disponibilidade de rotinas computacionais que incorporam metodologias para tratamento adequado de dados obtidos por amostragem complexa, realizou-se este estudo que teve por objetivo avaliar o efeito da incorporação do plano amostral complexo em um estudo epidemiológico.

MÉTODOS

Para avaliar o impacto do plano amostral complexo utilizaram-se os dados da busca ativa domiciliar dos participantes na Campanha Nacional de Detecção de Diabetes Mellitus – CNDDM[5,6], Brasil, 2001, que teve, entre outros objetivos, a confirmação diagnóstica dos pacientes com rastreamento positivo. A CNDDM foi o primeiro rastreamento para detecção de casos suspeitos de Diabetes Mellitus no Brasil, com mais de 22 milhões de exames de glicemia capilar realizados.

Quadro 1 – Distribuição dos participantes e dos municípios por estrato na campanha e na busca ativa

Estrato	Número de participantes na campanha	Número de participantes selecionados na busca ativa (amostra)	Número de municípios na região	Número de municípios selecionados na busca ativa (amostra)	Peso amostral do indivíduo
N	1.162.303	298	417	3	3874,343
NE	6.276.269	1.378	1.688	14	4483,049
CO	1.522.295	293	426	3	5074,317
SE	9.440.561	2.012	1.640	21	4495,505
S	3.668.477	925	1.130	9	4076,052
Total	22.069.905	4.906	5.301	50	-

Na busca ativa, realizou-se uma amostragem estratificada com conglomerado em três estágios em cada estrato. Os estratos foram definidos pelas 5 grandes regiões geográficas do Brasil (Norte, Nordeste, Centro-Oeste, Sudeste e Sul). No primeiro

estágio, foram pesquisados 50 municípios. O número de municípios para compor a amostra em cada região definiu-se proporcionalmente ao número de indivíduos participantes (exames realizados) na campanha, conforme Quadro 1. Os municípios foram selecionados com probabilidade proporcional ao número de indivíduos participantes na campanha.

No segundo estágio, selecionou-se uma unidade básica de saúde (UBS) em cada município, com probabilidade proporcional ao número de participantes na campanha de cada UBS. No terceiro estágio, foram selecionados 100 participantes em cada UBS com igual probabilidade (Figura 1).

Figura 1 – Processo de amostragem

UA: unidade de amostragem; PS: processo de seleção; PPT: probabilidade proporcional ao tamanho; AAS: amostragem aleatória simples

A amostra final foi composta por 3 municípios da região norte, 14 da nordeste, 3 da centro-oeste, 21 da sudeste e 9 da região sul (Quadro 1). Os dados da pesquisa foram armazenados no aplicativo Epi-Info 6.03[7], assim como as variáveis que caracterizam o

delineamento, possibilitando a análise que incorpora o delineamento amostral complexo.

Os conglomerados do primeiro estágio de amostragem, neste estudo, são os municípios ou unidades primárias de amostragem (UPA). No segundo estágio de amostragem, dentro de cada município foi selecionada uma unidade básica de saúde ou unidade secundária de amostragem (USA). Assim, foram utilizadas somente as UPAs para estimação da variância no STATA 9.0[8].

O peso de cada indivíduo na amostra é composto pelo produto dos pesos das unidades amostrais em cada estágio. A probabilidade do k - ésimo indivíduo do estrato h pertencer à amostra é definida pelo produto das probabilidades de seleção em cada estágio:

$$\pi_{h1} = \frac{n_h . M_{hi}}{M_h} \ , \ \pi_{h2} = \frac{u_{hi} M_{hij}}{M_{hi}} \ , \ \pi_{h3} = \frac{100}{M_{hij}}$$

em que:

π_{h1} : probabilidade do i-ésimo município ser sorteado dentro da região h;

π_{h2} : probabilidade da j-ésima unidade básica de saúde ser sorteada dentro do município i da região h;

π_{h3} : probabilidade do k-ésimo indivíduo ser sorteado dentro da unidade básica de saúde j no município i na região h;

n_h : número de municípios selecionados na região (estrato) h;

M_{hi} : número de indivíduos no município i do estrato h;

M_h : número de indivíduos na região (estrato) h na população;

u_{hi}: número de unidades básicas de saúde selecionadas no município i do estrato h;

M_{hij}: número de indivíduos da unidade básica de saúde j no município i do estrato h.

Então:

$$\pi_{hk} = \pi_{h1}.\pi_{h2}.\pi_{h3}$$

$$\pi_{hk} = \frac{n_h.M_{hi}}{M_h}.\frac{u_{hi}M_{hij}}{M_{hi}}.\frac{100}{M_{hij}}$$

Simplificando os termos e como u_{hi} = 1, a probabilidade final é:

$$\pi_{hk} = \frac{100n_h}{M_h}$$

O peso final é o inverso da probabilidade de seleção de cada unidade amostral dentro do estrato h, sendo definido por:

$$w_{hk} = \frac{1}{\pi_{hk}}$$

Realizaram-se cinco estratégias de análises distintas para avaliar o efeito de cada componente do delineamento, assim definidas: AAS (considera os dados como se tivessem sido gerados por amostragem aleatória simples), AC (considera o plano amostral complexo, incorporando estratos, conglomerados e peso amostral), ACO

(considera apenas o conglomerado, excluindo estratos e pesos amostrais), AE (considera apenas o estrato, excluindo conglomerado e pesos amostrais), AP (considera apenas peso amostral, excluindo conglomerados e estratos).

Estimativas para média e proporção, erros padrões, intervalos de 95% de confiança para as variáveis da pesquisa foram obtidas para cada uma das cinco estratégias de análises.

Foi ajustado um modelo de regressão de Poisson com variância robusta para a variável glicemia em jejum (CDGLN) dicotomizada (não diabetes, provável ou muito provável diabetes) em função do sexo, da idade (40 a 49 anos, 50 anos ou mais) e do índice de massa corporal (menos que 25 Kg/m^2, 25 Kg/m^2 a 29,99 Kg/m^2, 30 Kg/m^2 ou mais) considerando AAS, AC e cada componente isoladamente (ACO, AE, AP). O método de estimação utilizado foi o de máxima pseudo-verossimilhança (MPV) com linearização por série de Taylor.

Utilizou-se a *deviance* para verificar o ajuste global do modelo na AAS. O teste de Wald ajustado[8] foi utilizado para avaliar a significância global dos demais modelos.

Para avaliar a diferença nos valores das estimativas de variabilidade usou-se o EPA (efeito do plano amostral ou DEFF – *design effect*) proposto por Kish (1965)[9]. O EPA é obtido pela razão entre a variância estimada considerando o plano efetivamente utilizado e a variância estimada considerando uma AAS.

As análises foram realizadas no aplicativo STATA 9.0[8], o qual possui rotinas que incorporam o delineamento amostral complexo. Os comandos utilizados para regressão de Poisson estão apresentados no Quadro 2.

Quadro 2 – Comandos utilizados para ajuste da regressão de Poisson no STATA 9.0

Plano amostral / componente	Comando
AAS[1]	*. xi: poisson cdgln1 sexo1 idadec1 i.imc_ptos3, robust*
Complexo[2]	*. svyset municipi [pweight=pesof], strata(regiao) vce(linearized)* *. xi: svy, vce(linearized): poisson cdgln1 sexo1 idadec1 i.imc_ptos3*
Conglomerado	*. svyset municipi, vce(linearized)* *. xi: svy, vce(linearized): poisson cdgln1 sexo1 idadec1 i.imc_ptos3*
Estrato	*. svyset _n, strata(regiao) vce(linearized)* *. xi: svy, vce(linearized): poisson cdgln1 sexo1 idadec1 i.imc_ptos3*
Peso amostral	*. svyset _n [pweight=pesof], vce(linearized)* *. xi: svy, vce(linearized): poisson cdgln1 sexo1 idadec1 i.imc_ptos3*

[1] "*irr*" foi adicionado no final do comando para obter a razão de prevalência

[2] "*, irr*" foi adicionado no final da segunda linha do comando para obter a razão de prevalência

RESULTADOS

As estimativas pontuais para proporção de sexo e de CDGLN, além das estimativas da média de idade, do IMC e da glicemia obtidas em cada uma das cinco estratégias de análise são apresentadas na Tabela 1, juntamente com seus respectivos erros padrões, intervalos de confiança e EPA.

Considerando a AC estima-se que 56,93% (IC95%: 54,38 - 59,48) dos indivíduos são do sexo feminino, que a idade média é de 59,47 anos (IC95%: 58,83 -

60,11), que o IMC médio é de 26,79 (IC95%: 26,58 – 26,99) e que 19,71% (IC95%: 16,89 - 22,52) dos indivíduos são classificados como apresentando diabetes “provável” ou “muito provável” diabetes.

Observa-se que não há praticamente diferença na estimativa pontual da proporção do sexo feminino na população entre AAS (56,91%) e AC (56,93%). Porém, a variabilidade associada à estimativa é mais afetada, obtendo-se um erro padrão na AAS de 0,71% e na AC de 1,27%, resultando num intervalo de confiança mais amplo na AC.

Os erros padrões da idade média, IMC médio, glicemia média e proporção de CDGLN são subestimados quando foi ignorado o plano complexo. O erro padrão da glicemia média sofreu o maior impacto aumentando de 1,0141 na AAS para 3,2638 na AC o que ocasionou o alto valor de EPA (10,34).

Quando o conglomerado foi considerado isoladamente o EPA variou de 3,22 a 9,75; quando o estrato foi considerando isoladamente, o EPA foi igual a 1,00 para todas as variáveis. Nos resultados da análise considerando o peso isoladamente, observa-se a diferença na estimativa pontual da média e proporção em relação à AAS, porém não houve diferença nos erros-padrões.

Na Tabela 2, apresentam-se os resultados da modelagem observando-se pequenas diferenças nas estimativas dos coeficientes de regressão de Poisson entre as análises considerando AAS e AC. Contudo, há diferenças acentuadas nas estimativas de variabilidade. Na análise da componente estrato isoladamente, os resultados são similares à AAS, sem diferença entre as estimativas pontuais e de variabilidade. No entanto há um aumento no erro-padrão dos coeficientes na análise da componente conglomerado e diferença nos coeficientes na análise da componente peso comparada à AAS.

O impacto do plano amostral complexo é maior para as estimativas de média e proporção do que para os coeficientes de regressão de Poisson. Os EPAs para média e proporção variam de 2,13 a 10,34 (Tabela 1) e os coeficientes de regressão de Poisson variam de 1,00 a 1,34 (Tabela 2).

Os intervalos de confiança para as razões de prevalência das variáveis em análise são mais estreitos quando se ignora o plano complexo (Tabela 3).

Percebe-se o impacto do plano amostral complexo no teste de comparação de médias da glicemia entre os sexos (Tabela 4). O teste é significativo (p = 0,034) considerando AAS e torna-se não significativo (p = 0,086) na AC e quando se avalia isoladamente a componente conglomerado (p = 0,090). Não se percebe impacto nas significâncias quando se avalia a componente estrato e a componente peso amostral isoladamente.

DISCUSSÃO

O objetivo principal deste estudo é abordar a necessidade da incorporação dos componentes (conglomerados, estratos e probabilidades desiguais de seleção) do plano amostral utilizado na pesquisa para se obter resultados corretos, demonstrando o impacto que cada componente tem na precisão e acurácia das estimativas. Assim, não serão discutidos os resultados epidemiológicos, já que o interesse está nos aspectos estatísticos da análise.

Observa-se nos resultados, que o impacto nas estimativas por ignorar o plano amostral efetivamente utilizado na pesquisa é maior nas estimativas de variabilidade do que nas estimativas pontuais. Neste estudo, as estimativas pontuais para média e

proporção são semelhantes, explicado pelo peso atribuído a cada elemento da amostra foi o mesmo dentro de cada estrato. Korn e Graubard[10] descrevem que as estimativas considerando a AAS são similares às obtidas ao considerar a ponderação, quando os pesos amostrais têm pouca variabilidade.

O plano amostral com probabilidade proporcional ao tamanho (PPT) produz frações desiguais de seleção em cada estágio, e o produto das frações parciais produz uma fração final igual associada a cada unidade de terceiro estágio, se a seleção no primeiro estágio de amostragem for PPT e no último estágio for AAS com tamanhos iguais[11], dentro de cada estrato, como é o caso deste estudo.

Os achados mostram que as estimativas pontuais somente diferiram quando o peso amostral foi considerado na análise, tanto na estimativa de média e proporção como na estimativa dos coeficientes de regressão. Isso é relatado na literatura (Pessoa e Silva[11], Brogan[12], Leite e Silva[13], Silva et al.[14]). Quando se tem grande variabilidade nos pesos, o impacto nas estimativas pontuais será considerável e afetará ainda mais as estimativas de variabilidade[15], portanto, a incorporação dos pesos é fortemente recomendada. Kreuter e Valliant[3] apontam que as estimativas obtidas quando se desconsideram os pesos amostrais não são válidas para toda a população.

A diferença nas estimativas de variabilidade é percebida nos erros padrões e nos intervalos de confiança. Quando se compara a AAS e AC verificaram-se alterações em todos os intervalos de confiança, observando-se amplitudes maiores na AC. Nas estimativas de média e proporção da AC os erros padrões foram todos maiores em relação a AAS e nas estimativas dos coeficientes de regressão de Poisson os EPAs foram menores. Esses achados são consistentes com estudos de Pérez et al.[16], Rodgers-Farmer e Davis[17], Guillén et al.[18], Heeringa e Liu[19].

Da mesma forma, quando incorporou-se somente o conglomerado na análise de dados, as estimativas pontuais foram iguais aos resultados da AAS, porém, os erros padrões foram maiores. Estes resultados são consistentes com o estudo de Sousa e Silva[20]. Os achados mostram que o conglomerado engloba a variabilidade, pois todos os EPAs para AC e para o plano por conglomerado são maiores que 1,00 (exceto para idade foi igual a 1,00 na AC), refletindo o aumento na variância dos estimadores que ocorre devido à homogeneidade dentro dos conglomerados.

A homogeneidade dentro dos conglomerados torna o plano amostral por conglomerados menos eficiente[21], aumentando o erro padrão das estimativas, quando comparado a AAS, evidenciando-se a necessidade de considerar o plano amostral complexo na análise dos dados, principalmente o efeito do conglomerado[20]. Desejam-se conglomerados heterogêneos dentro de si, porém na prática isso dificilmente ocorre, como constatou-se nesse estudo.

De forma diferente, quando se incorporou somente o estrato na análise de dados, obtiveram-se estimativas pontuais e de variabilidade muito semelhantes às obtidas com a análise considerando AAS. Isso se explica por não se obter estratos homogêneos já que a estratificação foi realizada por facilidade logística e não com o objetivo de se ter estratos mais homogêneos. No entanto, desconsiderar o efeito da estratificação pode ter pequeno impacto nas estimativas dos parâmetros, mas aumenta o erro padrão no caso dos estratos serem homogêneos em relação à variável de estudo. Desta forma, é importante a incorporação do estrato na análise.

Esse estudo encontrou maior diferença nos erros padrões associados à média e à proporção do que nos erros padrões associados aos coeficientes do modelo de regressão de Poisson, quando comparou-se AC com AAS. Esta mesma constatação foi apontada nos estudos de Lemeshow et al.[22] e Lemeshow e Cook[23], embora estes estudos tenham

utilizado regressão logística. Não foram encontrados estudos comparando AAS e AC utilizando regressão de Poisson.

Deve-se considerar uma limitação neste estudo, referente à falta de ajuste para não-resposta, no sentido de refletir exatamente a fração populacional na amostra. Neste estudo, o peso é um fator de expansão utilizado na análise dos dados que teve o objetivo de projetar o número de indivíduos populacionais que cada indivíduo da amostra representa, sendo definido para cada região (estrato).

Mostrou-se a importância da incorporação do plano efetivamente utilizado na pesquisa para se obter conclusões corretas. Apesar de algumas diferenças assumirem pequenas magnitudes quando se compara AAS e AC, parecendo indicar pouco impacto principalmente nas estimativas pontuais, cabe lembrar que este estudo baseou-se em apenas um banco de dados. O impacto poderá ser maior em outros estudos, conduzindo a resultados errôneos caso o plano amostral efetivamente utilizado na coleta de dados não seja incorporado na análise estatística.

Recomenda-se, portanto, que os pesquisadores incorporem o delineamento amostral efetivamente utilizado em suas pesquisas, pois atualmente há disponibilidade de rotinas em aplicativos que permitem incorporar as características do plano amostral complexo em diversos tipos de análise. A conclusão deste estudo é a de que se podem obter diferentes conclusões sobre erros padrões, a significância e a conseqüente inclusão ou não de variáveis no modelo conforme o delineamento amostral considerado na análise.

Colaboradores

I. D. E. Battisti contribuiu com análise estatística, interpretação e redação do artigo.

J. M. G. Fachel, J. Riboldi e E. Mundstock contribuíram na orientação, redação e revisão do artigo.

REFERÊNCIAS

1 Cordeiro R. Efeito do desenho em amostragem de conglomerado para estimar a distribuição de ocupações entre trabalhadores. Revista de Saúde Pública 2001; 35(1):10-5.

2 Silva NN. Amostragem Probabilística. São Paulo: EDUSP; 1998.

3 Kreuter F, Valliant R. A survey on survey statistics: what is done and can be done in Stata. The Stata Journal 2007;7(1):1-21.

4 Vieira MT. Amostragem Repetida no Tempo: o uso de painéis. IBGE - Rio de Janeiro/RJ: 1ª Escola de Amostragem e Metodologia de Pesquisa; 2007.

5 Nucci LB. A Campanha Nacional de Detecção do Diabetes Mellitus: cobertura e resultados glicêmicos. Tese de doutorado em Epidemiologia. Faculdade de Medicina. Univerisdade Federal do Rio Grande do Sul; 2003.

6 Brasil. Ministério da Saúde. Organização Pan Americana da Saúde. Avaliação do Plano de Reorganização da Atenção à Hipertensão Arterial e ao Diabetes Mellitus no Brasil. Brasília: Ministério da Saúde, 2004.

7 Epi-Info, Center for Disease Control and Prevention. Version 6.0.3

8 StataCorp. STATA, release 9. Stata Corporation. 1985-2005.

9 Kish L. Survey Sampling. New York: John Wiley; 1965.

10 Korn EL, Graubard BI. Epidemiologic studies utilizing surveys: accounting for the sampling design. American Journal of Public Health 1991;81(9):1166-73

11 Pessoa DGC, Silva PLN. Análise de dados amostrais complexos. São Paulo: ABE-Associação Brasileira de Estatística; 1998.

12 Brogan DJ. Pitfalls of using standard statistical software packages for sample surveys data. Em: http://www.fas.harvard.edu/~stats/survey-soft/donna_brogan.html, acesso em 19/09/2003.

13 Leite PGPG, Silva DBN. Análise da situação ocupacional de crianças e adolescentes nas regiões sudeste e nordeste do Brasil utilizando informações da PNAD 1999. 2002.

14 Silva PLN, Pessoa DGC, Lila MF. Análise estatística de dados da PNAD: incorporando a estrutura do plano amostral. Ciência e Saúde Coletiva 2002;7(4):659-70.

15 Ciol MA, Hoffman JM, Dudgeon BJ, Shumway-Cook A, Yokston KM, Chan L. Understanding the Use of Weights in the Analysis of Data From Multistage Surveys. Archives of Physical Medicine Rehabilitation 2006;87:299-303.

16 Pérez MC, Utra IB, León AA, Roche RG, Sagué KA, Rosa MC, et al. Estimationes usadas en diseños muestrales complejos: aplicaciones en la encuesta de salud cubana del año 2001. Revista Panamericana de Salud Publica 2004;15(3):176-84.

17 Rodgers-Farmer A, Davis D. Analyzing complex survey data. Social Work Research 2001 Jan 1;25(3):185-92.

18 Guillén M, Juncá S, Rué M, Aragay JM. Efecto del diseño muestral en el análisis de encuestas de deseño complejo. Aplicación a la encuesta de salud de Catalunya. Gaceta Sanitaria 2000;14(5):399-402.

19 Heeringa SG, Liu J. Complex sample design effects and inference for mental health survey data. International Journal of Methods in Psychiatric Research 1997;7(1):56-65.

20 Sousa MH, Silva NN. Estimativas obtidas de um levantamento complexo. Revista de Saúde Pública 2003;37(5):662-70.

21 Bolfarine H, Bussab WO. Elementos de Amostragem. São Paulo: Edgar Blücher; 2005.

22 Lemeshow S, Letenneur L, Dartigues JF, Lafont S, Orgogozo JM, Commenges D. Illustration of anlaysis taking into account complex survey considerations: the association between wine consumtion and dementia in the PAQUID study. American Journal of Epidemiology 1998;148(3):298-306.

23 Lemeshow S, Cook ED. Pratical considerations in the analysis of complex sample survey data. Revue d'epidémiologie et de santé publique 1999;47:479-87.

TABELAS

Tabela 1. Estimativas pontuais, erros padrões, intervalos de confiança e efeitos do plano amostral para as variáveis: sexo, idade, IMC, glicemia e CDGLN, considerando-se os diferentes componentes do plano amostral

Variável	Componente	Estimativa	Erro Padrão	IC 95%	EPA
Sexo (feminino)		Proporção			
	Simples	0,5691	0,0071	[0,5552;0,5830]	-
	Complexo	0,5693	0,0127	[0,5438;0,5948]	3,21
	Conglomerado	0,5691	0,0126	[0,5437;0,5946]	3,22
	Estrato	0,5691	0,0070	[0,5553;0,5830]	1,00
	Peso amostral	0,5693	0,0071	[0,5554;0,5832]	1,00
Idade		Média			
	Simples	59,46	0,1637	[59,1414;59,7832]	-
	Complexo	59,47	0,3177	[58,8349;60,1127]	3,77
	Conglomerado	59,46	0,3431	[58,7737;60,1509]	4,39
	Estrato	59,46	0,1629	[59,1430;59,7816]	1,00
	Peso amostral	59,47	0,1639	[59,1525;59,7951]	1,00
IMC		Média			
	Simples	26,80	0,0697	[26,6665;26,9401]	-
	Complexo	26,79	0,1013	[26,5836;26,9909]	2,13
	Conglomerado	26,80	0,1273	[26,5479;27,0589]	3,33
	Estrato	26,80	0,0692	[26,6676;26,9392]	1,00
	Peso amostral	26,79	0,0693	[26,6513;26,9232]	1,00
Glicemia		Média			
	Simples	168,32	1,0141	[166,3343;170,3106]	-
	Complexo	168,45	3,2638	[161,8912;175,0158]	10,34
	Conglomerado	168,32	3,1666	[161,9683;174,6766]	9,75
	Estrato	168,32	1,0099	[166,3426;170,3023]	1,00
	Peso amostral	168,45	1,0179	[166,4579;170,4490]	1,01
CDGLN (diabetes)		Proporção			
	Simples	0,1961	0,0057	[0,1850;0,2072]	-
	Complexo	0,1971	0,0140	[0,1689;0,2252]	6,07
	Conglomerado	0,1961	0,0140	[0,1689;0,2232]	5,70
	Estrato	0,1961	0,0057	[0,1850;0,2072]	1,00
	Peso amostral	0,1971	0,0057	[0,1859;0,2082]	1,01

Tabela 2. Contribuições relativas dos componentes do plano amostral nos coeficientes de regressão de Poisson para a variável resposta CDGLN

Componente Variável	Estimativa	Erro padrão	IC 95%	P	EPA
Simples					
Sexo[a]	-0,0854	0,0580	[-0,1990; 0,0282]	0,141	-
Idade[b]	0,1538	0,0719	[0,0129; 0,2947]	0,032	-
IMC[c] sobrepeso	0,1081	0,0684	[-0,0258; 0,2421]	0,114	-
obeso	0,3093	0,0791	[0,1544; 0,4643]	<0,01	-
Constante	-1,8173	0,0857	[-1,9853;-1,6493]	<0,01	-
Complexo					
Sexo[a]	-0,0913	0,0608	[-0,2135; 0,0309]	0,139	1,11
Idade[b]	0,1566	0,0719	[0,0121; 0,3012]	0,034	1,00
IMC[c] sobrepeso	0,1057	0,0787	[-0,0526; 0,2640]	0,186	1,34
obeso	0,3092	0,0861	[0,1361; 0,4823]	0,001	1,19
Constante	-1,8099	0,1135	[-2,0380; -1,5817]	<0,01	1,77
Conglomerado					
Sexo[a]	-0,0854	0,0605	[-0,2068; 0,0360]	0,164	1,09
Idade[b]	0,1538	0,0765	[0,0003; 0,3073]	0,050	1,13
IMC[c] sobrepeso	0,1081	0,0802	[-0,0529; 0,2691]	0,184	1,38
obeso	0,3093	0,0855	[0,1377; 0,4810]	0,001	1,17
Constante	-1,8173	0,1181	[-2,0543;-1,5802]	<0,01	1,90
Estrato					
Sexo[a]	-0,0852	0,0580	[-0,1991; 0,0282]	0,141	1,00
Idade[b]	0,1538	0,0719	[0,0129; 0,2947]	0,032	1,00
IMC[c] sobrepeso	0,1081	0,0683	[-0,0258; 0,2421]	0,114	1,00
obeso	0,3093	0,0791	[0,1543; 0,4643]	<0,01	1,00
Constante	-1,8173	0,0856	[-1,9851; -1,6494]	<0,01	1,00
Peso amostral					
Sexo[a]	-0,0913	0,0580	[-0,2050; 0,0223]	0,115	1,01
Idade[b]	0,1566	0,0720	[0,0155; 0,2978]	0,030	1,01
IMC[c] sobrepeso	0,1057	0,0684	[-0,0284; 0,2397]	0,122	1,01
obeso	0,3092	0,0791	[0,1542; 0,4642]	<0,01	1,01
Constante	-1,8099	0,0858	[-1,9781; -1,6416]	<0,01	1,01

[a] masculino como categoria de referência; [b] 40 a 49 anos de idade como categoria de referência;

[c] IMC (menos que 25 Kg/m² como categoria de referência)

Tabela 3. Razão de prevalência (RP) e intervalo de confiança (IC) segundo o plano amostral simples e complexo

Variável	Plano amostral			
	Simples		Complexo	
	RP	IC 95%	RP	IC 95%
Sexo[a]	0,9181	[0,8195;1,0286]	0,9127	[0,8077;1,0313]
Idade[b]	1,1662	[1,0130;1,3428]	1,1696	[1,0122;1,3514]
IMC[c] sobrepeso	1,1142	[0,9745;1,2739]	1,1115	[0,9487;1,3021]
obeso	1,3625	[1,1669;1,5908]	1,3623	[1,1458;1,6197]

[a] masculino como categoria de referência; [b] 40 a 49 anos de idade como categoria de referência;

[c] IMC (menos que 25 Kg/m² como categoria de referência)

Tabela 4. Contribuições relativas dos componentes do plano amostral associadas ao teste de comparação de médias da glicemia entre sexo e entre faixa etária

Componente	Média ± EP	Média ± EP	p	EPA
	Masculino	Feminino		
Simples	170,8460 ± 1,5249	166,5077 ± 1,3599	0,034	-
Complexo	171,0449 ± 3,1057	166,5887 ± 3,7598	**0,086**	1,54
Conglomerado	170,8460 ± 3,1769	166,5077 ± 3,5474	**0,090**	1,51
Estrato	170,8460 ± 1,5200	166,5077 ± 1,3584	0,034	1,00
Peso amostral	171,0449 ± 1,5309	166,5887 ± 1,3644	0,030	1,01
	40 a 49 anos	50 anos ou mais		
Simples	163,0613 ± 2,0812	169,8897 ± 1,1600	0,004	-
Complexo	163,2139 ± 3,7899	170,0113 ± 3,2991	0,007	1,03
Conglomerado	163,0613 ± 3,7403	169,2188 ± 3,2188	0,010	1,16
Estrato	163,0613 ± 2,0782	169,8897 ± 1,1559	0,004	1,00
Peso amostral	163,2139 ± 2,0895	170,0113 ± 1,1641	0,005	1,01

6.2 ARTIGO 2

ESTIMATIVAS OBTIDAS POR ABORDAGENS DA AMOSTRAGEM COMPLEXA E DE MODELOS MULTINÍVEL: APLICAÇÃO EM UM ESTUDO COM ESCOLARES DE ENSINO FUNDAMENTAL NO MUNICÍPIO DE IJUÍ/RS

ESTIMATES OF COMPLEX SAMPLE AND MULTILEVEL MODEL APPROACH: APPLICATIONS IN A STUDY WITH FIFTH GRADERS IN IJUÍ, BRAZIL.

[1, 3] Iara Denise Endruweit BATTISTI

[1, 2] João RIBOLDI

[2] Elsa MUNDTOCK

[1, 2] Jandyra Maria Guimarães FACHEL

[1]Programa de Pós Graduação em Epidemiologia – Universidade Federal do Rio Grande do Sul, Porto Alegre/RS, Brasil

[2]Departamento de Estatística, Instituto de Matemática, Universidade Federal do Rio Grande do Sul, Porto Alegre/RS, Brasil

[3]Departamento de Física, Estatística e Matemática, Universidade Regional do Noroeste do Estado do Rio Grande do Sul, Santa Rosa/RS, Brasil

Endereço de correspondência do autor:
Iara Denise Endruweit Battisti
Av. Borges de Medeiros 550/701 Centro
Santa Rosa/RSCEP:98900-000
iara.battisti@unijui.edu.br

A ser enviado a Revista Brasileira de Epidemiologia

RESUMO

Introdução: Em investigações epidemiológicas é comum o estudo de indivíduos agrupados em níveis ou hierarquias. Para tratar estas estruturas complexas existem duas principais abordagens: abordagem da amostragem complexa e abordagem de modelos multinível. Este estudo tem o propósito de avaliar as duas abordagens para análise de dados provindos de estruturas ou delineamentos complexos utilizando dados de um estudo do desempenho das crianças na avaliação de conhecimento, percepções e crenças sobre aleitamento materno realizado com escolares da quinta série do ensino fundamental, no município de Ijuí/RS.

Método: Trata-se de um estudo aleatorizado, com amostra estratificada por conglomerados, tendo como desfecho um escore de desempenho. Estimativas de média, proporção e seus erros padrões foram obtidos considerando amostragem aleatória simples e amostragem complexa. Foram ajustados modelos de regressão linear com e sem pesos amostrais para a abordagem da amostragem complexa e para a abordagem de modelos multinível.

Resultados: Observou-se diferença no valor da significância de um coeficiente do modelo final entre a análise ponderada e não ponderada na abordagem da amostragem complexa. Não houve diferença de significância dos coeficientes entre os modelos com ponderação e sem ponderação, considerando a abordagem da análise multinível. Há diferença dos erros padrões dos coeficientes da regressão entre as duas abordagens, sendo que os mesmos são maiores na amostragem complexa. Na abordagem da amostragem complexa ponderada, o coeficiente associado à variável sexo não foi significativo (p=0,068) enquanto para a análise multinível ponderada este coeficiente foi significativo (p=0,030).

Conclusão: Os resultados mostram diferenças nos erros padrões entre amostragem complexa e amostragem aleatória simples, indicando a necessidade de incorporar a complexidade do plano amostral na análise dos dados. A escolha de qual abordagem é a mais apropriada depende da questão de pesquisa, isto é, se o pesquisador deseja resultados, tanto ao nível de grupo como ao nível individual, opta-se pela abordagem multinível. Caso contrário, opta-se pela abordagem da amostragem complexa.

Palavras-chave: amostragem complexa; modelo multinível; conglomerado.

ABSTRACT

Introduction: Epidemiological studies often use samples of individuals grouped according to levels. Two approaches are used to deal with these complex structures: the complex sample approach, and the multilevel model approach. This study evaluated these two approaches to complex survey data analysis using the database of a study about the performance of children in the evaluation of knowledge, perceptions and beliefs about maternal breastfeeding conducted with fifth graders in Ijuí, Brazil.

Method: This random study had a stratified cluster sampling design, and its outcome was a performance score. Estimates of means, proportions and standard errors were calculated for a simple random sample and for a complex sample. Linear regression models with and without sample weights were adjusted for the complex sample and multilevel model approaches.

Results: There was a difference in the p-value of the final model coefficient between weighted and unweighted analysis in the complex sample approach. There was no difference in the significance of coefficients between the weighted and unweighted analyses in the multilevel model approach. The complex sample approach showed significantly greater standard errors of the regression coefficients than the multilevel model approach. The coefficient associated with the variable “gender” was not significant (p=0.068) in the weighted complex sample approach, but was significant (p=0.030) in the weighted multilevel model approach.

Conclusion: Differences in standard errors between complex and simple random sampling showed the need to incorporate the complexity of the sample level into data analysis. The choice of an adequate approach depends on research questions: when researchers look for results at both group and individual levels, the choice should be the multilevel model approach. Otherwise, the complex sample approach should be chosen.

Key words: Complex sample; multilevel model; cluster.

INTRODUÇÃO

Em investigações epidemiológicas é comum o estudo de indivíduos agrupados em níveis ou hierarquias. Esta estrutura pode estar presente de forma intrínseca (alunos dentro de escolas, médicos dentro de hospitais, indivíduos dentro de famílias) ou pelo tipo de delineamento utilizado na pesquisa (por exemplo, utilização de setores censitários como conglomerados e níveis sócio-econômicos como estratos). A estrutura presente nos dados, na forma intrínseca ou não, deve ser incorporada na análise dos dados, sob pena de se obter estimativas incorretas[1,2,3].

O plano amostral complexo implica na necessidade de adoção de alguns procedimentos na estimação de modelos estatísticos. Esses procedimentos exigem o uso de aplicativos com rotinas que possibilitem o uso dos pesos amostrais e a especificação correta do plano amostral na etapa de estimação dos modelos[4].

Existem duas abordagens principais para tratar a estrutura dos dados, disponibilizadas recentemente em aplicativos estatísticos: (1) abordagem da amostragem complexa, denominada análise agregada[1], na qual a estrutura de dados é um fator complicador que invalida o uso de procedimentos padrões de análise e (2) abordagem de modelos multinível, denominada análise desagregada[1], que incorpora mais explicitamente a estrutura da população no procedimento de análise[5].

Na modelagem considerando a abordagem da amostragem complexa todos os fatores são considerados em um único nível, sendo este o mesmo da variável resposta (desfecho). Incorpora-se na análise tanto estratos quanto conglomerados em um ou mais estágios de amostragem. Neste caso, procedimentos padrões de análise de dados que consideram amostragem aleatória simples (AAS) não são adequados, já que estes

pressupõem que as observações são independentes e, na presença de conglomerados, essa pressuposição falha, pois indivíduos dentro de conglomerados geralmente são mais semelhantes do que indivíduos entre conglomerados.

Vieira[6] encontrou grandes diferenças entre as estimativas dos erros padrões das estimativas dos parâmetros dos modelos ajustados considerando independência das observações utilizando procedimentos computacionais padrões, e aquelas obtidas quando considerou a estrutura do delineamento amostral utilizando procedimentos computacionais específicos para tratamento de dados de amostragem complexa.

Na análise multinível, os fatores são considerados em dois ou mais níveis e a variável resposta no nível individual (nível 1). Os conglomerados são considerados na análise como níveis e estratos são incorporados explicitamente no modelo. Procedimentos padrões de análise de dados são inadequados, pois estes desconsideram a estrutura complexa existente na população. A vantagem do modelo multinível é a possibilidade de estimação de efeitos intragrupo (efeitos individuais) e entre grupos (efeitos contextuais) e, ainda é possível modelar a estrutura da variância em cada um dos níveis[7].

Pessoa e Silva[1] e Lehtonen e Pahkinen[2] apresentam a análise multinível como uma alternativa de análise de dados provindos de amostragem complexa. Lehtonen e Pahkinen[2] realizaram estudos na área da educação para comparar resultados obtidos na modelagem considerando AAS com os obtidos nas abordagens da amostragem complexa e da análise multinível, observando diferenças tanto para efeitos dos coeficientes do modelo quanto para a significância destes coeficientes.

A utilização de probabilidade desigual de seleção de indivíduos num processo de amostragem deve ser incorporada tanto na abordagem da amostragem complexa[1,2] como na de modelos multinível[8], visando evitar a ocorrência de vícios na estimação dos

parâmetros do modelo[5,9]. O motivo principal é que procedimentos padrões de análise de dados consideram os dados como provindos de amostra aleatória simples (AAS), na qual todos os indivíduos têm a mesma probabilidade de serem sorteados. Pfeffermann et al.[10] apresentam uma forma de incorporar pesos no ajuste de modelos hierárquicos para compensar diferentes probabilidades de inclusão das unidades na amostra.

Assis[8] ajustou um modelo de regressão linear múltipla para dados obtidos por amostragem complexa na Pesquisa de Padrões de Vida (PPV) do IBGE. Os resultados indicam diferenças nas estimativas dos parâmetros quando se compara a análise multinível sem ponderação e análise multinível ponderada pelos pesos amostrais. Quando são comparadas as estimativas dos erros padrões dos coeficientes estimados para os dois métodos, as diferenças são ainda maiores.

É comum encontrar análises de dados de amostragem complexa realizadas inadequadamente, isto é, desconsiderando a complexidade da estrutura dos dados, evidenciando a necessidade de estudos que demonstrem procedimentos corretos para análise deste tipo de dados[11,12,13].

Desta forma, este estudo tem o propósito de avaliar as duas abordagens – amostragem complexa e modelos multinível - para análise de dados provindos de delineamento complexo, utilizando dados de um estudo com escolares da quinta série do ensino fundamental do município de Ijuí, Rio Grande do Sul (RS)[14].

MÉTODOS

Delineamento

Trata-se de um estudo aleatorizado, com amostra estratificada por conglomerados, tendo como desfecho o desempenho das crianças na avaliação de conhecimento, percepções e crenças sobre aleitamento materno, avaliado através de um escore (obtido através do número de acertos em questões objetivas do questionário). A população-alvo do estudo foi constituída de escolares de ambos os sexos, matriculados na quinta série das escolas estaduais, municipais e particulares do ensino fundamental do município de Ijuí/RS[14].

Para a seleção da amostra, utilizaram-se subgrupos populacionais - conglomerados – representados, primeiramente, pelas escolas e, posteriormente, pelas turmas de quinta série existentes nas escolas sorteadas no primeiro estágio. O esquema amostral está demonstrado na Figura 1. Levou-se em consideração a proporção de alunos que freqüentavam as escolas estaduais, municipais e particulares, bem como a sua distribuição geográfica na zona urbana e zona rural.

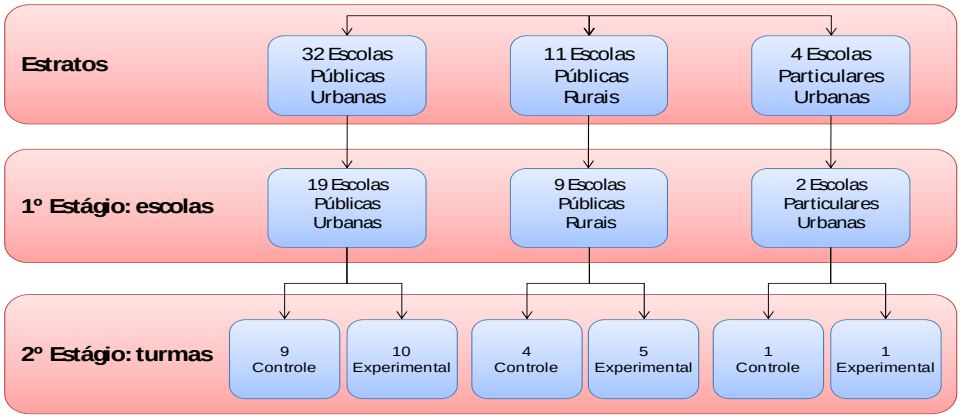


Figura 1- Esquema amostral do estudo

A seleção das escolas foi realizada através de amostragem aleatória simples (AAS) e a seleção das turmas dentro das escolas também foi por AAS, sendo que quando a escola tinha apenas uma turma, esta teve probabilidade igual a 1 de ser incluída na amostra.

Das 47 escolas existentes no município com turmas de quinta série (32 urbanas públicas, 11 rurais públicas e 4 urbanas particulares), 30 foram sorteadas: 19 públicas urbanas (10 no grupo experimental e 9 no grupo controle) ; 9 públicas rurais (5 no grupo experimental e 4 no grupo controle) e 2 urbanas particulares (1 experimental e 1 controle). Participaram do estudo 561 escolares, sendo 252 sujeitos no grupo controle e 309 no grupo experimental.

As escolas foram sorteadas para o grupo controle e experimental, considerando a proporcionalidade quanto ao tipo de escola (urbana pública, rural pública e urbana particular). Todos os alunos responderam um questionário, composto por 25 questões objetivas e 5 descritivas e, para cada pergunta (objetiva) respondida com a opção mais favorável ao aleitamento materno, o aluno recebia um ponto. Assim, o escore varia de 0 a 25 pontos. O grupo controle não foi exposto à intervenção e o grupo experimental sofreu a intervenção desenvolvida em três momentos na sala de aula ou dependências da escola. O uso de dinâmicas diferenciadas proporcionou o enfoque da intervenção sob três formas: vídeo, dramatização e cartilha[14]. Após três meses do início da intervenção, o mesmo questionário foi aplicado em todas as turmas, tanto no grupo controle como no experimental.

Estratégia de análise dos dados

A amostragem complexa é freqüentemente utilizada em pesquisas de grande porte. Um plano amostral complexo é definido pela utilização de estratos e conglomerados em um ou mais estágios. É comum selecionar indivíduos com probabilidades variáveis, como é o caso da amostragem com probabilidade proporcional ao tamanho (PPT). Desta forma, existem três elementos – conglomerados, estratos e pesos que podem estar presentes num plano amostral complexo.

Porém, muitas análises ainda são realizadas sem considerar esses elementos, podendo afetar as estimativas pontuais e o erros padrões e, conseqüentemente, produzir resultados incorretos[12,13]. A metodologia para tratamento de dados provindos de amostragem complexa é discutida há bastante tempo, porém, somente recentemente os aplicativos estatísticos estão incluindo recursos para tratamento destes dados[11].

A aplicação de modelos multinível em dados de amostragem complexa é bastante recente[15]. Modelos multinível permitem estudar o efeito das variáveis ao nível de conglomerados sobre a variável dependente (desfecho) do nível individual[16]. Cada estágio de amostragem corresponde a um nível na modelagem multinível, sendo que a UFA (unidade final de amostragem) corresponde ao nível 1 e os conglomerados de cada estágio de amostragem constituem os demais níveis[17]. Os estratos são definidos como variável independente do modelo multinível[1] e os pesos são incluídos na análise multinível de forma diferente que na abordagem da amostragem complexa[15].

Primeiramente, estimativas para proporção, erro padrão, intervalo de 95% de confiança e efeito do plano amostral foram obtidas para as variáveis de caracterização dos escolares, considerando-se que os dados provêm de amostragem aleatória simples (AAS) e de amostragem complexa.

Nas modelagens para as duas abordagens foram obtidas estimativas dos coeficientes de regressão, erros-padrões e nível de significância associados aos coeficientes, com o principal interesse de verificar se a variável desfecho escore-pós difere entre o grupo experimental e o controle. Utilizaram-se como variáveis independentes as características dos escolares (sexo e idade), escore-pré e tratamento.

Na análise, considerando amostragem complexa, as escolas foram definidas como conglomerados (UPA) e tipo de escola, como estratos. Dois modelos de regressão linear foram obtidos, um considerando pesos amostrais e outro desconsiderando pesos amostrais. O teste de Wald foi usado para verificar a influência de cada variável e a significância geral do modelo. O processo de inclusão das variáveis iniciou com a inclusão da variável escore-pré (medida *baseline*), seguida das variáveis sexo, idade e tratamento.

O método de estimação utilizado foi o de máxima pseudo-verossimilhança (MPV) com linearização por série de Taylor, o qual considera os pesos amostrais dos elementos da amostra[18]. O efeito do plano amostral complexo foi calculado por EPA $(\hat{\theta})=\frac{V_{VERD}(\hat{\theta})}{V_{AAS}(\hat{\theta})}$, onde $V_{verd}(\hat{\theta})$ é a variância do estimador $\hat{\theta}$ considerando o plano amostral complexo e $V_{AAS}(\hat{\theta})$ é a variância pressupondo que a amostra foi coletada a partir de uma AAS, com o mesmo número n de elementos selecionados. Com o EPA é possível avaliar o impacto no erro padrão da estimativa quando se desconsidera o plano amostral complexo.

O peso amostral do i-ésimo aluno foi definido por $w_i = \frac{1}{\pi_i} = \frac{1}{\pi_{1hk}}.\frac{1}{\pi_{2\,jhk}}$, em que π_{1hk} é a probabilidade da h-ésima escola ser sorteada no estrato k, $\pi_{2\,jhk}$ é a

probabilidade j-ésima turma ser sorteada na escola h e π_i é a probabilidade final do i-ésimo aluno ser sorteado.

Na análise de modelos multinível considerou-se a hierarquia dos dados, definindo escolas como nível 2 e alunos como nível 1. Também construíram-se modelos multinível com e sem pesos amostrais. Ajustaram-se os seguintes modelos: modelo não-condicional (vazio), modelo de interceptos aleatórios e modelo de intercepto e inclinações aleatórias. O teste de Wald foi utilizado para verificar a significância de cada variável no modelo e o teste de razão de verossimilhança para significância do modelo. Com os modelos de intercepto e inclinação aleatórios é possível modelar a variância do escore-pós em função de variáveis explanatórias.

O modelo multinível com dois níveis para desfecho contínuo é apresentado utilizando a notação comum introduzida por Goldstein[7], Snijders e Bosker[19] e Zanini[20].

Nos modelos com dois níveis ocorrem n_j unidades do nível 1 para cada unidade $j(j = 1,2,...J)$ do nível 2. Os modelos para o nível 1 são desenvolvidos separadamente em cada unidade do nível 2, considerando a possibilidade de variação dos interceptos e das inclinações.

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + u_{0j} + e_{0ij} \qquad \text{com } i = 1,2,...,n_j \text{ e } j = 1,2,...J$$

em que:

Y_{ij} : desfecho do i-ésima unidade do nível 1, agrupada na j-ésima unidade do nível 2;

X_{ij} : variável preditora medida na i-ésima unidade do nível 1, agrupada na j-ésima unidade do nível 2;

β_0 : intercepto geral do modelo;

β_1 : coeficiente de inclinação, associado à variável preditora X;

u_{0j} : efeito aleatório do nível 2;

e_{0ij} : o efeito aleatório do nível 1;

Os resíduos u_{0j} e e_{0ij} são supostamente independentes e normalmente distribuídos, com média zero e variâncias σ^2_{u0} σ^2_{e0}, respectivamente. A variância residual, ou seja, a variância condicionada a X é dada por:

$$Var(Y_{ij} \mid X_{ij}) = \text{var}(u_{oj}) + \text{var}(e_{oij}) = \sigma^2_{u0} + \sigma^2_{e0}$$

A covariância entre dois indivíduos (i_1, i_2) no mesmo grupo j é:

$$Cov(Y_{i1j}, Y_{i2j} \mid X_{i1j}, X_{i2j}) = \text{var}(u_{oj}) = \sigma^2_{u0}$$

O coeficiente de correlação intraclasse ou intragrupo, que corresponde à correlação entre os valores de dois indivíduos de um grupo, controlado para a variável X, é dado por:

$$\rho_I(Y_{ij} \mid X_{ij}) = \frac{\sigma^2_{u0}}{\sigma^2_{u0} + \sigma^2_{e0}}$$

Estende-se o modelo para:

$$Y_{ij} = \beta_{0j} + \beta_{1j} X_{1ij} + e_{0ij}$$

Então, no nível 2,

$$\beta_{0j} = \beta_0 + u_{0j}$$
$$\beta_{1j} = \beta_1 + u_{1j}$$

em que:

β_{0j}: intercepto para a j-ésima unidade do nível 2;

β_{1j}: coeficiente de inclinação, associado à variável X da i-ésima unidade do nível 1, agrupada na j-ésima unidade do nível 2;

β_0: valor esperado dos interceptos no nível 2;

β_1: valor esperado das inclinações no nível 2;

u_{0j}: efeito aleatório da j-ésima unidade do nível 2 no intercepto β_{0j};

u_{1j}: efeito aleatório da j-ésima unidade do nível 2 na inclinação β_{1j}.

Substituindo-se as equações do nível 2:

$$Y_{ij} = \beta_0 + \beta_1 X_{1ij} + (u_{0j} + u_{1j} X_{1ij} + e_{0ij})$$

Denominam-se modelos de coeficientes aleatórios, o que pressupõe que cada grupo tem um intercepto (β_{0j}) e uma inclinação (β_{1j}) diferentes, que variam através dos grupos, apresentando uma estrutura de variância-covariância complexa como:

$$Var\begin{bmatrix} u_{0j} \\ u_{ij} \end{bmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_{u0} & \sigma_{u0u1} \\ \sigma_{u1u0} & \sigma^2_{u1} \end{pmatrix}\right] \text{ e } Var(e_{0ij}) \sim N(0, \sigma^2_{e0})$$

A variância é modelada como uma função de uma variável preditora.

Os parâmetros foram estimados usando IGLS (*Iterative Generalized Least Squares)*. Os pesos amostrais foram definidos para cada nível, sendo o inverso da probabilidade de seleção das unidades de amostragem em cada nível, definidos como w_1 e w_2 para o nível 1 e para o nível 2, respectivamente.

Utilizaram-se os aplicativos STATA 9.0[21], SAS 9.1.3[22] e MLwiN 2.02[23], conforme a disponibilidade de técnicas para as análises propostas. Os comandos utilizados são apresentadas nos quadros 1 e 2.

Questões Éticas

O projeto (processo número 01-429) foi aprovado pelo Grupo de Pesquisa e Pós-graduação do Hospital de Clínicas de Porto Alegre e pela Comissão Científica e Comissão de Pesquisa e Ética em Saúde.

O projeto foi aprovado pela 36ª Coordenadoria de Educação do município de Ijuí, Secretaria Municipal de Educação de Ijuí e pela direção das escolas particulares deste município. As escolas selecionadas foram previamente visitadas e informadas sobre a pesquisa, ficando livres a participar ou não da pesquisa. O Termo de Consentimento Livre e Esclarecido foi enviado aos pais dos escolares antes do início da pesquisa.

RESULTADOS

Na Tabela 1 são apresentadas as estimativas da média e dos intervalos de 95% de confiança para o escore-pré, considerando que os dados foram obtidos através de AAS e por amostragem complexa. Também são apresentadas as estimativas de proporção e intervalos de 95% de confiança para a idade (9 a 11 anos; 12 a 17 anos), o sexo e o tipo de escola.

O escore médio antes da intervenção (escore-pré) é de 12,523 para os 561 escolares que participaram do estudo, considerando a amostragem complexa. Destes, 397 têm entre 9 e 11 anos, os demais entre 12 a 17 anos e 283 são do sexo feminino. Na escola pública urbana foram pesquisados 386 escolares, na escola pública rural 126 escolares e na escola particular urbana 49 escolares.

Os intervalos de 95% de confiança são mais amplos na amostragem complexa quando comparados à AAS. No caso do escore-pré, o intervalo de 95% de confiança é (12,257-12,820) para AAS e (12,073-12,973) para amostragem complexa.

A diferença de amplitude nos intervalos de 95% de confiança entre AAS e amostragem complexa pode ser detectada pelo EPA associado a cada uma das variáveis, os quais foram todos maiores que 1, variando de 1,64 a 14,70. No caso do escore-pré, o EPA é 2,31, indicando que o erro padrão da média do escore-pré é 2,31 vezes maior na AC em relação à AAS.

As variáveis apresentadas na Tabela 1 foram utilizadas na modelagem das abordagens da amostragem complexa e da análise multinível, sendo os resultados apresentados nas Tabelas 2 a 5.

Abordagem da amostragem complexa

Nas Tabelas 2 e 3 apresentam-se os resultados da modelagem por amostragem complexa sem e com ajustes para pesos amostrais, respectivamente. O modelo 1 (modelo nulo) foi ajustado somente com o intercepto, o modelo 2 (modelo somente com variáveis individuais) foi ajustado com o escore-pré, idade e sexo e, no modelo 3 (modelo com variáveis individuais e ao nível de escola), foi incluído o tratamento.

Na Tabela 2, observa-se para o modelo 1, sem ponderação, que o intercepto (escore-pós médio) é 15,562 e EPA de 9,50, indicando uma grande diferença do erro padrão associado ao intercepto entre AAS e AC, sendo maior na AC. No modelo 2, todos os EPAs, com exceção do da variável sexo, foram maiores que 1. E no modelo 3, os EPAs foram todos maiores que 1, variando de 1,10 a 2,15, demonstrando a necessidade da incorporação do plano amostral efetivamente utilizado na pesquisa. No modelo 3, foi incluída a variável tipo de escola, a qual não foi significativa e, portanto, retirada do modelo.

Na Tabela 3 são apresentados os modelos para a análise ponderada: os EPAs foram maiores que 1 para os 3 modelos, com exceção da variável sexo no modelo 2, demonstrando o mesmo comportamento da análise não ponderada. Porém no modelo 3, da análise ponderada, a variável sexo não foi significativa ao nível de 5%, com resultado diferente quando comparada à análise não ponderada.

Abordagem da análise multinível

Nas Tabelas 4 e 5 apresentam-se os resultados da análise multinível sem e com ajuste para pesos amostrais, respectivamente. Na análise multinível foi considerado

modelo com interceptos aleatórios. Neste caso, os coeficientes de regressão são os mesmos para todas as escolas, mas os interceptos podem variar. Também foram considerados modelos com variância complexa (intercepto e inclinação aleatórios), porém não foram significativamente superiores ao modelo com intercepto aleatório.

Na Tabela 4 encontram-se os resultados para os modelos multinível ajustados sem ponderação, observando-se que a média geral do escore-pós é 15,431. A variância residual no nível de alunos é 16,463, enquanto que a variância residual no nível de escola é 9,419, obtendo-se um coeficiente de correlação intra-conglomerado (CCI) de 0,3639. Com base no valor de CCI indica-se a necessidade de ajuste de um modelo multinível.

O modelo 2 foi ajustado somente com as variáveis do nível 1, obtendo-se todos coeficientes significativos e CCI de 0,4053. No modelo 3 foram incluídas as variáveis do nível 2, sendo o tratamento altamente significativo e o tipo de escola não significativo. O CCI no modelo 3 é 0,0238, observando-se que o tratamento contribuiu muito para a explicar a redução da variabilidade ao nível 2 (escolas).

Na Tabela 5 são apresentados os modelos multinível para a análise ponderada. O CCI para o modelo 1 é 0,3638, sendo a variância residual do nível de alunos e do nível de escola de 25,736 e 14,719, respectivamente. Nos modelos 2 e 3, todas as variáveis foram significativas ao nível de 5% de significância. No modelo 3, a variável tipo de escola não contribuiu para a significância do modelo, sendo retirada. Os CCIs para o modelo 2 e 3 são 0,4111 e 0,0277, respectivamente, obtendo-se uma redução significativa da variabilidade no nível 2 (escolas) quando o tratamento é incluído no modelo 3.

Abordagem da amostragem complexa versus abordagem de modelo multinível

Comparando-se as abordagens da amostragem complexa e modelagem multinível sem ponderação (Tabelas 2 e 4) e com ponderação (Tabelas 3 e 5) verificaram-se diferenças nas estimativas dos coeficientes para as variáveis nos modelos, sendo essas diferenças de pouco impacto. Há grande diferença dos erros padrões entre as duas abordagens, sendo que os erros padrões assumem valores maiores na amostragem complexa. No modelo 3, na análise sem ponderação, todas as variáveis foram significativas, tanto para a amostragem complexa como para a análise multinível, porém, o coeficiente associado à variável sexo foi significativo (p=0,025) para a amostragem complexa e altamente significativo (p=0,002) para a análise multinível. Na análise ponderada, o coeficiente associado à variável sexo não foi significativo (p=0,068) para a amostragem complexa, entretanto foi significativo (p=0,030) para a análise multinível. Todos os demais coeficientes foram significativos ao nível de 5% tanto para a amostragem complexa como para a análise multinível.

DISCUSSÃO

Na amostragem complexa (AC) geralmente estão presentes estratos, conglomerados e probabilidades desiguais de seleção. Este tipo de amostragem é muito utilizado em estudos epidemiológicos e de saúde pública[24]. O principal motivo da utilização da AC é de não estar disponível uma lista de indivíduos na população ou se está disponível, o custo da implementação de uma AAS é muito alto, quase sempre

inviabilizando o trabalho. No presente estudo, é incontestável a maior viabilidade em se pesquisar crianças em escolas ao invés de residências.

Apesar de existir recursos disponíveis, tanto metodológicos quanto computacionais, ainda há pouca utilização dos métodos e aplicativos para tratamento correto dos dados provindos de AC. Isto pode acontecer tanto por falta de conhecimento dos recursos computacionais considerando AC disponíveis quanto do erro que se pode cometer ao usar recursos computacionais considerando AAS.

O uso de aplicativos tradicionais (não incorporando AC) na análise pode levar a resultados incorretos na estimação de parâmetros, tanto da média e proporção e respectivos erros-padrões (EP), quanto de coeficientes de modelos e respectivos erros padrões[12,17,18,]. No presente estudo, isso é percebido pelos valores de EPA, todos maiores que 1 para estimativas de média e proporção, conduzindo a intervalos de confianças mais estreitos sob AAS do que realmente seriam sob AC. Nos coeficientes dos modelos, os EPAs foram todos maiores que 1, com uma exceção, tanto na análise ponderada como na análise não ponderada. Assim, confirma-se a necessidade de adotar alguma abordagem de análise que considera o delineamento amostral.

O agrupamento de dados no ensaio aleatorizado em grupo (ERG) tem muito em comum com o agrupamento dos dados observados na abordagem da amostragem complexa. Desta forma, métodos de análise de dados desenvolvidos para esta abordagem também podem ser aplicados para dados de ERG[25]. Intervenções em saúde são freqüentemente implementadas ao nível de grupo (hospital, clínica, escola, etc) ao invés de, ao nível individual. Portanto, deve-se estar atento à dependência existente entre indivíduos de um mesmo grupo, necessitando tanto de delineamento apropriado, quanto análise apropriada dos dados[26].

Pessoa e Silva[1], Lehtonen e Pahkinen[2] e Vieira[6] descrevem as duas metodologias

apresentadas como apropriadas para tratamento de dados provindos de AC. Estas metodologias, modelagem multinível e amostragem complexa, incorporam as características do delineamento complexo como estratos, conglomerados e probabilidades desiguais dos indivíduos, assim como a estrutura hierárquica da população para obter resultados corretos. No presente estudo, na abordagem da AC, os conglomerados foram definidos pelas escolas e os estratos pelos tipos de escolas. Cabe ressaltar que as turmas dentro das escolas não foram definidas como conglomerados, pois somente uma turma foi selecionada por escola, já que o aplicativo utilizado para análise não possibilita tal situação. Para contornar esse problema, poder-se-ia unir dois conglomerados do nível superior daquele que possui um único conglomerado. Neste trabalho, decidiu-se não unir duas escolas, pois o principal objetivo era comparar as abordagens. Na abordagem de modelos multinível, os conglomerados (escolas) foram definidos no nível 2 (superior) e os estratos foram incluídos como efeitos fixos.

Nas duas abordagens há a inconveniência de que os indivíduos no mesmo conglomerado tendem a ser mais similares que indivíduos entre conglomerados, o que poderá levar a desfechos similares[27]. Na análise multinível, avalia-se essa similaridade através do CCI (correlação intraclasse ou intra-conglomerados), o qual neste estudo foi superior a 0,3 tanto na análise multinível sem ponderação como na análise multinível com ponderação no modelo 1. A modelagem multinível é indicada sempre que CCI for maior que 0,05 por Thomas e Heck[28] e maior que 0,01 por Hope e Shannon[29].

Na análise sem ponderação, percebe-se que os coeficientes são semelhantes nos modelos finais (modelo 3) obtidos na abordagem da AC e na abordagem da análise multinível, porém há diferença nos erros-padrões associados aos coeficientes. Na análise ponderada, têm-se diferença entre coeficientes das duas abordagens, observando-se diferença maior ainda nos erros-padrões associados aos coeficientes.

Essas diferenças nos resultados entre as duas abordagens se dá pelo motivo da intervenção ter sido realizada no nível de escola (nível 2). Assim, a variável de intervenção (tratamento) engloba a variância desse nível na modelagem multinível e repercute de forma mais acentuada, nas variáveis do nível individual (nível 1), produzindo coeficientes diferentes e erros-padrões menores, modificando a significância. Com isso, a modelagem multinível apresenta vantagem em relação à abordagem de amostragem complexa que utiliza os dados de forma agregada.

Ainda que o presente estudo baseia-se somente em um conjunto de dados, os resultados são consistentes com outros estudos[2,8,24,29].

A escolha de qual abordagem é a mais apropriada depende da questão de pesquisa. Na análise dos dados do SAEB (Sistema Nacional de Avaliação da Educação Básica), Vieira[6] utilizou a abordagem da amostragem complexa e Ferrão et al.[30] utilizaram a abordagem de modelos multinível. A identificação para qual nível a inferência é desejada permitirá avaliar o método mais adequado de análise de dados[31]. Quando deseja-se uma modelagem no nível de grupo e no nível individual, tratando a homogeneidade intragrupo implicitamente no modelo, opta-se pela abordagem multinível. Entretanto, quando deseja-se modelagem em um único nível, opta-se pela abordagem da amostragem complexa, na qual a homogeneidade dentro dos grupos é considerada na análise, removendo o efeito de conglomerados[1,2,28].

Uma vantagem do modelo multinível é a possibilidade de modelar a variação entre conglomerados[24]. Modelou-se a variação entre conglomerados, ou seja, entre escolas (nível 2) pelo modelo multinível, que não pôde ser modelada na abordagem da amostragem complexa.

Assim, os conglomerados são tratados como níveis de hierarquia, as covariáveis são utilizadas para tratar estratos, tratamentos (intervenção) e medida *baseline* no modelo multinível[32].

Correa[5] indicou a necessidade de incorporar pesos na análise multinível, quando a probabilidade de seleção dos indivíduos é diferente, visando a melhorar a qualidade das estimativas dos parâmetros do modelo. Então, deve-se estar atento quando unidades em qualquer nível da hierarquia são selecionadas com probabilidades desiguais, para evitar a ocorrência de vícios na estimação dos parâmetros do modelo[5,8,17]. Pfeffermann et al.[10] apresentaram uma forma de incorporar pesos no ajuste de modelos hierárquicos para compensar diferentes probabilidades de inclusão das unidades na amostra. No presente trabalho, incorporou-se o peso de cada nível (escola e alunos) pela opção disponível no MLwiN. Cabe salientar, que o peso usado foi o peso bruto.

Chantala et al.[15] tem discutido o uso de pesos amostrais em modelos de dois níveis (modelos mistos). Alertam que a incorporação dos pesos na abordagem multinível é realizada de forma diferente que na abordagem da amostragem complexa.

Sendo assim, sugere-se comparações mais acuradas através da realização de estudos de simulação para comparar a abordagem de amostragem complexa e a abordagem de modelos multinível, considerando diferentes planos amostrais e diferentes tamanhos de amostra para avaliar o efeito dos aspectos da amostragem (conglomerado, estrato, probabilidade diferente de seleção) nos resultados da modelagem.

REFERÊNCIAS

1 Pessoa DGC, Silva PLN. Análise de dados amostrais complexos. São Paulo: ABE-Associação Brasileira de Estatística; 1998.

2 Lehtonen R, Pahkinen E. Practical Methods for Design and Analysis of Complex Survey. 2 ed. Chichester. England: John Wiley & Sons Ltd; 2004.

3 Vieira MT. Amostragem Repetida no Tempo: o uso de painéis. IBGE - Rio de Janeiro/RJ: 1ª Escola de Amostragem e Metodologia de Pesquisa; 2007.

4 Andrade AO. Aplicação do Modelo Logístico Multinomial no Estudo da Decisão do voto. Dissertação de mestrado em Estudos Populacionais e Pesquisas Sociais. Escola Nacional de Ciências Estatísticas; 2006.

5 Correa ST. Modelos lineares hierárquicos em pesquisas por amostragem – relacionando o Índice de Massa Corporal às variáveis da pesquisa sobre os padrões de vida/IBGE. Dissertação de mestrado em Estudos Populacionais e Pesquisas Sociais. Escola Nacional de Ciências Estatísticas; 2001.

6 Vieira MT. Um estudo comparativo das metodologias de modelagem de dados amostrais complexos - uma aplicação ao SAEB 99. Dissertação de mestrado em Ciência da Engenharia Elétrica. Pontifícia Universidade Católica do Rio de Janeiro; 2001.

7 Goldstein H. Multilevel statistical models. 3 ed. Edward Arnold; 2003.

8 Assis JM. Modelos multiníveis em pesquisas amostrais complexas – uma aplicação à valorização de aluguéis de imóveis residenciais segundo suas característica/atributos. Dissertação de mestrado em Estudos Populacionais e Pesquisas Sociais. Escola Nacional de Ciências Estatísticas; 2005.

9 Skinner CJ, Holt D, Smith TMF. Analysis of complex survey. New York. John Wiley & Sons; 1989.

10 Pfeffermann D, Skinner C, Goldstein H, Holmes DJ, Rasbash J. Weighting for unequal selection probabilities in multilevel models. Journal of the Royal Statistical Society 1998;60(Série B):23-40.

11 Sousa MH, Silva NN. Estimativas obtidas de um levantamento complexo. Revista de Saúde Pública 2003;37(5):662-70.

12 Kreuter F, Valliant R. A survey on survey statistics: what is done and can be done in Stata. The Stata Journal 2007;7(1):1-21.

13 Vieira MT. Amostragem Repetida no Tempo: o uso de painéis. IBGE - Rio de Janeiro/RJ: 1ª Escola de Amostragem e Metodologia de Pesquisa; 2007.

14 Bottaro SM. Avaliação de Estratégia de Promoção do Aleitamento Materno em Escolas do Ensino Fundamental. Tese de doutorado em Ciências Médicas Pediatria, Faculdade de Medicina. Universidade Federal do Rio Grande do Sul; 2006.

15 Chantala K, Blanchette D, Suchindran CM. Software to compute sampling weights for multilevel analysis. Em: http://www.cpc.unc.edu/restools/data_analysis/ml_sampling_weights/Compute%20Weights%20for%20Multilevel%20Analysis.pdf, acesso em 17/01/2008.

16 Asparouhov T, Muthen B. Multilevel Modeling of Complex Survey Data. Proceedings of the Joint Statistical Meeting in Seattle. 08/2006. Seattle.

17 Rabe-Hesketh S, Skrondal A. Multilevel modelling of complex survey data. Journal of the Royal Statistical Society A 2006;169(4):805-27.

18 Silva PLN, Pessoa DGC, Lila MF. Análise estatística de dados da PNAD: incorporando a estrutura do plano amostral. Ciência e Saúde Coletiva 2002;7(4):659-70.

19 Snijders T, Bosker R. An introduction to basic and advanced multilevel modeling. 1 ed. London: SAGE Publications; 1999.

20 Zanini RR. Modelos multiníveis aplicados ao estudo da mortalidade infantil no Rio Grande do Sul, Brasil, de 1994 a 2004. Tese de doutorado em Epidemiologia. Faculdade de Medicina. Universidade Federal do Rio Grande do Sul; 2007.

21 StataCorp. STATA, release 9. Stata Corporation. 1985-2005.

22 SAS Institute, Inc. SAS statistical software, release 9.1. Cary, NC: SAS Institute, Inc, 2002.

23 MLwiN, version 2.02. Multilevel Models Project, c. 2000. London: Institute of Education, University of London.

24 Utra IB, Pérez MC, Marqués LL. Influencia de la estructura de los datos en la selección de los métodos de análisis estadísticos. Revista Española de Salud Pública 2002;76:95-103.

25 Murray DM, Varnell SP, Blitstein JL. Design and analysis of group-randomized trials: a review of recent methodological developments. American Journal of Public Health 2004;94:423-32.

26 Ukoumunne OC, Gulliford MC, Chinn S, Sterne JAC, Burney PGJ, Donner A. Methods in health service research: evaluation of health interventions at area and organisation level. British Medical Journal 1999;319:376-9.

27 Barros AJD. Modelos multinível: primeiros passos. Departamento de Medicina Social. Faculdade de Medicina. Universidade Federal de Pelotas. 2001.

28 Thomas SL, Heck RH. Analysis of large-scale secondary data in higher education research. Research in Higher Education 2001;42(5):517-40.

29 Hope AD, Shannon ED. A comparison of two procedures to fit multi-level data: PROC GLM verus PROC MIXED. Proceedings of the SAS users group international 30. 10-13/04/2005. Philadelphia. Pennsylvania.

30 Ferrão ME, Beltrão KI, Santos DP. Modelo de regressão multinível: aplicação ao estudo do impacto da política de não-repetência no desempenho escolar dos alunos da 4ª série. Pesquisa e Planejamento Economico 2002; 32(3).

31 Campbell MK, Mollison J, Steen N, Grimshaw JM, Eccles M. Analysis of cluster randomized trials in primary care: a practical approach. Family Practice 2000;17(2):192-6.

32 Donner A, Klar N. Cluster randomization trials in epidemiology: theory and application. Journal of Statistical Planning an Inference 1994;42(1-2): 37-56.

QUADROS

Quadro 1 – Comandos utilizados para ajuste da regressão no STATA 9.0 considerando a abordagem da amostragem complexa

Modelo	Comando
Modelo 1 sem ponderação	*. svyset escola, strata(tipesc) vce(linearized)* *.svy, vce(linearized): regress escorta2*
Modelo 2 sem ponderação	*. svyset escola, strata(tipesc) vce(linearized)* *. svy, vce(linearized): regress escorta2 escorta1_cent idadecod sexo*
Modelo 3 sem ponderação	*. svyset escola, strata(tipesc) vce(linearized)* *. svy, vce(linearized): regress escorta2 escorta1_cent idadecod sexo trat*
Modelo 1 com ponderação	*.svyset escola [pweight=final_wt], strata(tipesc) vce(linearized)* *.svy, vce(linearized): regress escorta2*
Modelo 2 com ponderação	*.svyset escola [pweight=final_wt], strata(tipesc) vce(linearized)* *. svy, vce(linearized): regress escorta2 escorta1_cent idadecod sexo*
Modelo 3 com ponderação	*.svyset escola [pweight=final_wt], strata(tipesc) vce(linearized)* *. svy, vce(linearized): regress escorta2 escorta1_cent idadecod sexo trat*

Quadro 2 – Comandos utilizados para ajuste da regressão no SAS 9.1.3 e MLwiN 2.02 considerando a abordagem de modelos multiníveis

Modelo	Comando
Modelo 1 sem ponderação (SAS)	*proc mixed data=_PROJ_.banco_18_01_atual COVTEST;* *class ESCOLA;* *model ESCORTA2 = / solution;* *random intercept / subject = ESCOLA solution;* *run; quit;*
Modelo 2 sem ponderação (SAS)	*proc mixed data=_PROJ_.banco_18_01_atual COVTEST;* *class ESCOLA TRAT IDADECOD SEXO;* *model ESCORTA2 = ESCORTA1_CENT IDADECOD SEXO / solution;* *random intercept / subject = ESCOLA solution;* *run; quit;*
Modelo 3 sem ponderação (SAS)	*proc mixed data=_PROJ_.banco_18_01_atual COVTEST;* *class ESCOLA TRAT IDADECOD SEXO;* *model ESCORTA2 = ESCORTA1_CENT IDADECOD SEXO TRAT/ solution;* *random intercept / subject = ESCOLA solution;* *run; quit;*
Modelo 1 com ponderação[1] (MLwiN)	$\text{escorta2}_{ij} \sim \text{N}(XB,\ \Omega)$ $\text{escorta2}_{ij} = \beta_{0ij}\text{cons}$ $\beta_{0ij} = 15{,}381(0{,}597) + u_{0j} + e_{0ij}$
Modelo 2 com ponderação[1] (MLwiN)	$\text{escorta2}_{ij} \sim \text{N}(XB,\ \Omega)$ $\text{escorta2}_{ij} = \beta_{0ij}\text{cons} + 0{,}584(0{,}037)\text{escorta1\_cent}_{ij} +$ $1{,}205(0{,}385)\text{idadecod}_{ij} + 0{,}884(0{,}398)\text{sexo}_{ij}$ $\beta_{0ij} = 14{,}077(0{,}500) + u_{0j} + e_{0ij}$
Modelo 3 com ponderação[1] (MLwiN)	$\text{escorta2}_{ij} \sim \text{N}(XB,\ \Omega)$ $\text{escorta2}_{ij} = \beta_{0ij}\text{cons} + 0{,}578(0{,}038)\text{escorta1\_cent}_{ij} +$ $1{,}130(0{,}374)\text{idadecod}_{ij} + 0{,}875(0{,}402)\text{sexo}_{ij} +$ $5{,}823(0{,}372)\text{trat}_{j}$ $\beta_{0ij} = 11{,}071(0{,}480) + u_{0j} + e_{0ij}$

[1]Use raw weight: *level 2 = psu_wt* e *level 1 = fsu_wt*

TABELAS

Tabela 1. Estimativas de média, proporção, intervalos de 95% de confiança e efeito do plano amostral para as variáveis da pesquisa considerando amostragem complexa e amostragem aleatória simples (AAS)

Variáveis	Estimativa (IC95%)			
	n	AAS	AC	EPA
Idade[2]				
9 a 11 anos	397	70,766 (66,991-74,542)	72,823 (66,956-78,690)	2,31
12 a 17 anos	164	29,234 (25,458-33,009)	27,177 (21,310-33,044)	
Sexo[2]				
Feminino	283	50,446 (46,296-54,596)	51,393 (45,638-57,148)	1,76
Masculino	278	49,554 (45,404-53,704)	48,607 (42,852-54,362)	
Tipo de escola[2]				
Pública urbana	386	68,806 (64,960-72,651)	77,740 (68,028-87,452)	7,25
Pública rural	126	22,460 (18,996-25,924)	11,612 (8,052-15,172)	1,64
Particular urbana	49	8,734 (6,391-11,078)	10,648 (0,395-20,901)	14,70
Escore-pré[1]	561	12,538 (12,257-12,820)	12,523 (12,073-12,973)	2,31

[1]média; [2]proporção

Tabela 2. Estimativas, erro padrão e nível de significância dos coeficientes de regressão linear para o desfecho escore-pós na abordagem da amostragem complexa, sem ponderação

Variáveis	Estimativa (erro padrão)		
	Modelo 1	Modelo 2	Modelo 3
Intercepto	15,562 (0,664)	14,173 (0,568)	11,004 (0,478)
	p<0,0001;EPA=9,50	p<0,0001;EPA=2,57	p<0,0001;EPA=2,15
Escore-pré		0,634 (0,058)	0,564 (0,046)
		p<0,0001;EPA=1,24	p<0,0001;EPA=1,10
Idade		1,288 (0,645)	1,266 (0,365)
		p=0,056;EPA=2,49	p=0,002;EPA=1,23
Sexo		0,894 (0,349)	0,901 (0,380)
		p=0,016;EPA=0,82	p=0,025;EPA=1,63
Tratamento			5,781 (0,394)
			p<0,0001;EPA=1,81
p		<0,0001	<0,0001
		$F_{(3,25)}$	$F_{(4,24)}$

Masculino como referência; 12 a 17 anos como referência; tratamento (referência=grupo controle);
p corresponde ao teste de Wald ajustado
Modelo 1: modelo sem covariável
Modelo 2: modelo com variáveis a nível de indivíduos
Modelo 3: modelo com variáveis a nível de conglomerados

Tabela 3. Estimativas, erro padrão e nível de significância dos coeficientes de regressão linear para o desfecho escore-pós na abordagem da amostragem complexa, com ponderação

Variáveis	Estimativa (erro padrão)		
	Modelo 1	Modelo 2	Modelo 3
Intercepto	15,282 (0,755)	14,136 (0,596)	11,155 (0,526)
	p<0,0001;EPA=12,47	p<0,0001;EPA=2,79	p<0,0001;EPA=2,70
Escore-pré		0,643 (0,059)	0,585 (0,047)
		p<0,0001;EPA=1,33	p<0,0001;EPA=1,29
Idade		1,083 (0,716)	1,003 (0,394)
		p=0,142;EPA=3,01	p=0,017;EPA=1,43
Sexo		0,664 (0,341)	0,800 (0,421)
		p=0,062;EPA=0,78	p=0,068;EPA=2,10
Tratamento			5,899 (0,431)
			p<0,0001;EPA=2,21
p		<0,0001	<0,0001
		$F_{(3,25)}$	$F_{(4,24)}$

Masculino como referência; 12 a 17 anos como referência; tratamento (referência=grupo controle);
p corresponde ao teste de Wald ajustado
Modelo 1: modelo sem covariável
Modelo 2: modelo com variáveis a nível de indivíduos
Modelo 3: modelo com variáveis a nível de conglomerados

Tabela 4. Estimativas, erro padrão e nível de significância dos coeficientes de regressão linear para o desfecho escore-pós na abordagem da análise multinível, sem ponderação

Variáveis	Estimativa (erro padrão)		
	Modelo 1	Modelo 2	Modelo 3
Efeitos fixos			
Nível 1			
Intercepto	15,431 (0,588)	14,011(0,602)	10,982 (0,369)
	p<0,001	p<0,001	p<0,001
Escore-pré		0,577 (0,046)	0,564 (0,044)
		p<0,001	p<0,001
Idade		1,403 (0,348)	1,326 (0,331)
		p<0,001	p<0,001
Sexo		0,912 (0,309)	0,905 (0,299)
		p=0,003	p=0,002
Nível 2			
Tratamento			5,739 (0,356)
			p<0,001
Efeitos aleatórios			
Variância do nível 1 (σ^2_{e0})	16,463 (1,010)	11,874 (0,729)	11,831 (0,725)
	p<0,001	p<0,001	p<0,001
Variância do nível 2 (σ^2_{u1})	9,419 (2,679)	8,093 (2,266)	0,288 (0,240)
	p<0,001	p<0,001	p=0,230
CCI	0,3639	0,4053	0,0238
-2*loglikelihood*(IGLS)	3233,105	3057,602	2989,249

Masculino como referência; 12 a 17 anos como referência; tratamento (referência=grupo controle);
Modelo 1: modelo nulo
Modelo 2: interceptos aleatórios, com variáveis de nível 1
Modelo 3: interceptos aleatórios, com variáveis de nível 1 e nível 2

Tabela 5. Estimativas, erro padrão e nível de significância dos coeficientes de regressão linear para o desfecho escore-pós na abordagem da análise multinível, com ponderação

Variáveis	Estimativa (erro padrão)		
	Modelo 1	Modelo 2	Modelo 3
Efeitos fixos			
Nível 1			
Intercepto	15,381 (0,597)	14,077 (0,500)	11,071 (0,480)
	p<0,001	p<0,001	p<0,001
Escore-pré		0,584 (0,037)	0,578 (0,038)
		p<0,001	p<0,001
Idade		1,205 (0,385)	1,130 (0,374)
		p<0,001	p=0,003
Sexo		0,884 (0,398)	0,875 (0,402)
		p=0,026	p=0,030
Nível 2			
Tratamento			5,823 (0,372)
			p<0,001
Efeitos aleatórios			
Variância do nível 1 (σ^2_{e0})	25,736 (1,932)	18,270 (1,626)	18,160 (1,614)
	p<0,001	p<0,001	p<0,001
Variância do nível 2 (σ^2_{u1})	14,719 (1,756)	12,752 (1,798)	0,517 (0,432)
	p<0,001	p<0,001	p=0,231
CCI	0,3638	0,4111	0,0277
-2*loglikelihood*(IGLS)	3284,612	3103,430	3035,034

Masculino como referência; 12 a 17 anos como referência; tratamento (referência=grupo controle);
Modelo 1: modelo nulo
Modelo 2: interceptos aleatórios, com variáveis de nível 1
Modelo 3: interceptos aleatórios, com variáveis de nível 1 e nível 2

7. CONCLUSÕES E CONSIDERAÇÕES FINAIS

Nesta tese fez-se a descrição dos métodos para tratamento de dados provindos de planos amostrais complexos, freqüentemente utilizados em estudos epidemiológicos, através de duas abordagens: amostragem complexa e modelo multinível.

Para ilustrar os métodos, analisaram-se dois conjuntos de dados já utilizados em outras pesquisas, sendo que a teoria envolvida nesses dois casos pode ser estendida para outros casos, tanto na abordagem da amostragem complexa como na abordagem de modelos multinível.

De um modo geral, observa-se que as estimativas pontuais de média, proporção e coeficientes de regressão de Poisson são semelhantes entre amostragem complexa e amostragem aleatória simples, mas há diferenças nos erros padrões. As diferenças nos erros padrões da média e proporção são maiores do que dos coeficientes de regressão de Poisson entre amostragem complexa e amostragem aleatória simples.

Desta forma, os resultados analisados e discutidos apontam para a necessidade da incorporação da complexidade do plano amostral na análise dos dados pelos pesquisadores da área epidemiológica e de outras áreas sempre que o plano amostral considerado for complexo.

Também, mostrou-se que os resultados relativos à regressão linear diferem quanto aos erros padrões dos coeficientes entre a abordagem da amostragem complexa e a abordagem de modelos multinível, sendo maiores na amostragem complexa.

Pode haver diferenças entre a significância das variáveis entre as modelagens por amostragem complexa e por multinível, o que implicaria, na prática, modelos finais

diferentes. Também, pode haver diferença da significância de variáveis e, portanto, da inclusão de variáves nos modelos entre a análise ponderada e não ponderada, na abordagem da amostragem complexa.

Na análise multinível os dados analisados não mostraram diferença de significância dos coeficientes entre os modelos com ponderação e sem ponderação, mas, nos casos em que os pesos tiverem maior variabilidade, pode ocorrer diferenças entre modelos com ponderação e sem ponderação, mostrando a importância da incorporação dos pesos amostrais na análise.

A escolha de qual abordagem é a mais apropriada depende da questão de pesquisa: se o pesquisador deseja resultados tanto no nível de grupo como individual deve optar pela abordagem multinível; caso contrário, pela abordagem da amostragem complexa.

Como conclusão geral, os estudos apresentados neste trabalho evidenciam a necessidade de incorporar a complexidade do plano amostral na análise dos dados.

Sugere-se, para estudos futuros, a aplicação dos métodos de estimação por re-amostragem para obtenção dos estimadores na análise de dados incorporando planos amostrais complexos. E, ainda, um estudo por simulação comparando a abordagem da amostragem complexa com a abordagem de modelos multinível.

ANEXO A – COMANDOS PARA DIFERENTES TÉCNICAS DE ANÁLISE DE DADOS EM APLICATIVOS

Quadro A.1 – Comando para diferentes técnicas de análise de dados no STATA 9.0

Técnica de análise	**Comando**
Definir as variáveis do plano amostral	**svyset** var_conglomeado [**pweight**=var_peso], **strata**(var_estrato) **vce**(linearized)
Média, erro padrão e intervalo de 95% de confiança	svy, vce(linearized): **mean** var_name
Proporção, erro padrão e intervalo de 95% de confiança	svy, vce(linearized): **proportion** var_name
Regressão linear simples	svy, vce(linearized): **regress** var_resposta_name var_1_name var_2_name
Regressão logística	svy, vce(linearized): **logistic** var_resposta_name var_1_name var_2_name
Teste de ajuste do modelo, segundo Archer e Lemeshow (2006)	**svylogitgof**
Regressão Poisson	svy, vce(linearized): **poisson** var_resposta_name var_1_name var_2_name
Regressão Poisson, com variável dummy definida como i.	xi: svy, vce(linearized): **poisson** var_resposta_name var_1_name **i.**var_2_name

var_conglomerado = nome da variável que define conglomerado; var_estrato = nome da variável que define estrato; var_peso = nome da variável que define pesos; var_name = nome da variável de pesquisa; var_reposta_name = nome da variável resposta de pesquisa.

Quadro A.2 – Comando para diferentes técnicas de análise de dados no SAS 9.1.3

Técnica de análise	**Comando**
Média, erro padrão e intervalo de 95% de confiança	**proc surveymeans** data=banco_dados; **cluster** var_conglomerado; **strata** var_estrato; var var_name; **weight** var_peso; run; quit;
Proporção, erro padrão e intervalo de 95% de confiança	**proc surveymeans** data=banco_dados; **class** var_name; **cluster** var_conglomerado; *strata* var_estrato; var var_name; **weight** var_peso; run; quit;
Regressão linear simples	**proc surveyreg** data=banco_dados; **cluster** var_conglomerado; **model** var_reposta_name = var_1_name var_2_name/ solution covb deff; **strata** var_estrato; **weight** var_peso; run; quit;
Regressão logística	**proc surveylogistic** data=banco_dados; **cluster** var_conglomerado; **model** var_reposta_name = var_1_name var_2_name; **strata** var_estrato; **weight** var_peso; run; quit;
Regressão Poisson	Não disponível.

Quadro A.3 – Comando para diferentes técnicas de análise de dados no SPSS 15

Técnica de análise	**Comando**
Definir as variáveis do plano amostral	**CSPLAN** ANALYSIS /PLAN FILE='banco_dados' /PLANVARS **ANALYSISWEIGHT**=var_peso /SRSESTIMATOR TYPE=WOR /PRINT PLAN /DESIGN STAGELABEL= '1' **STRATA**= var_estrato **CLUSTER**= var_conglomerado /ESTIMATOR TYPE=WR.
Média, erro padrão e intervalo de 95% de confiança	**CSDESCRIPTIVES** /PLAN FILE = 'banco_dados' /SUMMARY VARIABLES =var_name /MEAN /STATISTICS SE CIN (95).
Proporção, erro padrão e intervalo de 95% de confiança	**CSTABULATE** /PLAN FILE = 'banco_dados' /TABLES VARIABLES = var_name /CELLS POPSIZE TABLEPCT /STATISTICS SE CIN(95).
Regressão linear simples	**CSGLM** var_resposta_name BY var_1_name WITH var_2_name /PLAN FILE = 'banco_dados' /MODEL var_resposta_name var_1_name
Regressão logística	**CSLOGISTIC** var_resposta(HIGH) BY var_1_name /PLAN FILE = 'banco_name' /MODEL var_1_name
Regressão Poisson	Não disponível

Quadro A.4 – Comando para diferentes técnicas de análise de dados no EPI – INFO 3.4.3

Técnica de análise	**Comando**
Média, erro padrão e intervalo de 95% de confiança	**MEANS** var_name **STRATAVAR**= var_estrato **WEIGHTVAR**=var_peso **PSUVAR**=var_conglomerado
Proporção	**FREQ** var_name **STRATAVAR** = var_estrato **WEIGHTVAR**=var_peso **PSUVAR**=var_conglomerado
Regressão linear simples	Não disponível
Regressão logística	Não disponível
Regressão Poisson	Não disponível

Quadro A.5 – Comando para diferentes técnicas de análise de dados no R

Técnica de análise	**Comando**
Definir as variáveis do plano amostral	plano <- **svydesign**(**id**=~conglomerado, **strata**=~estrato, **weights**=~peso, data=nome_matriz_dados, nest=TRUE)
Média, erro padrão e intervalo de 95% de confiança, EPA	**svymean** (~var_name, plano, deff=TRUE)
Proporção, erro padrão	**svymean** (~var_name, plano)*
Regressão linear simples	modelo <- **glm**(var_resposta_name~var_1_name+var_2_name, data=nome_matriz_dados) summary(modelo)
Regressão logística	**logitmodel** <- svyglm(I(var_resposta_name == 1) ~var_1_name + var_2_name, design = plano, family = quasibinomial()) summary(logitmodel)
Regressão Poisson	**poissonmodel** <- svyglm(I(var_resposta_name == 1) ~var_1_name + var_2_name, design = plano, family = quasipoisson()) summary(poissonmodel)

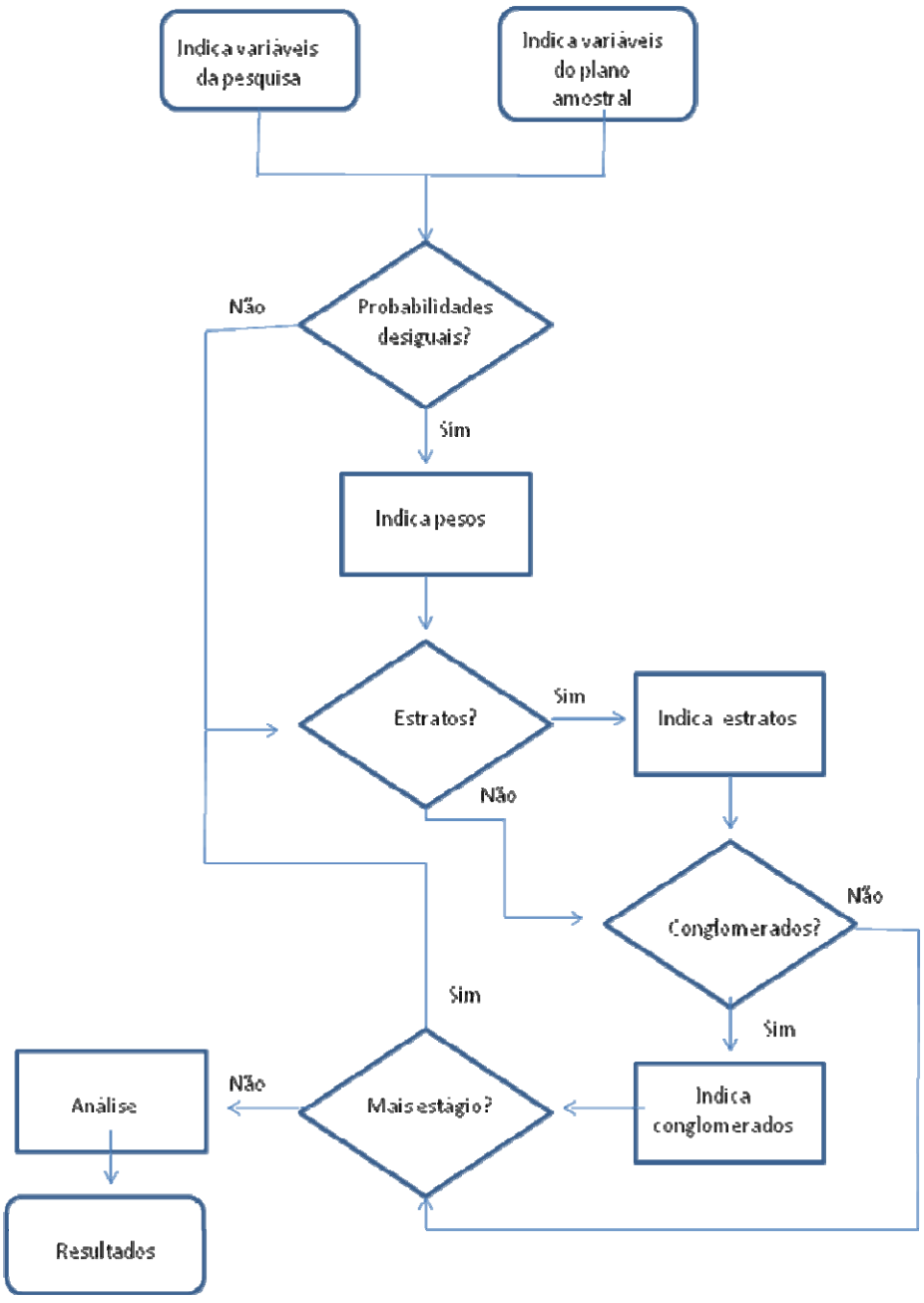
*definindo as categorias como 0 e 1

Para leitura dos dados: dados <- **read.table**("D:\\nome_arquivo.txt", header = TRUE, sep = "\t").

SUDAAN: as técnicas de análise estatística apresentadas nos aplicativos (Quadros A1 a A5) também estão disponíveis no SUDAAN, porém seus comandos não são apresentados aqui por falta de acesso ao referido programa.

WesVar: não estão sendo apresentados os comando para o WesVar, pois este aplicativo utiliza pesos pelo método de replicação, não sendo este o objeto deste estudo.

ANEXO B – Fluxograma para análise de amostragem complexa



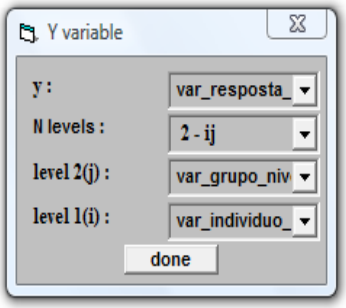
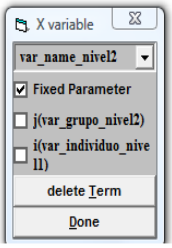
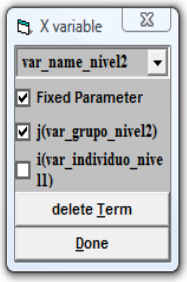
ANEXO C – COMANDOS PARA ANÁLISE MULTINÍVEL EM APLICATIVOS

Quadro C.1 – Comando para modelagem multinível com 2 níveis com desfecho contínuo no SAS 9.1.3

Modelo	**Comando**
Modelo nulo	**proc mixed** data=banco_dados covtest; **class** var_name; **model** var_resposta_name_nivel1 = / **solution**; **random intercept** / **subject** = var_grupo_nivel2 **solution**; run; quit;
Modelo com intercepto aleatório entre grupos do nível 2	**proc mixed** data=banco_dados covtest; **class** var_name; **model** var_resposta_name_nivel1 = var_name_nivel1 var_name_nivel2 / **solution**; **random intercept** / **subject** = var_grupo_nivel2 **solution**; run; quit;
Modelo com intercepto e inclinação aleatórios entre grupos do nível 2	**proc mixed** data=banco_dados covtest; **class** var_name; **model** var_resposta_name_nivel1 = var_name_nivel1 var_name_nivel2 / **solution**; **random intercept** / **subject** = var_grupo_nivel2 var_name_nivel1 **solution**; run; quit;

var_reposta_name_nivel1 = nome da variável resposta de pesquisa; var_name_nivel1 = nome da variável explicativa do nível 1; var_name_nivel 2 = nome da variável explicativa do nível 2; var_grupo_nível 2 = nome da variável que define o grupo;

Quadro C.2 – Comando para modelagem multinível com 2 níveis com desfecho contínuo no MLwiN 2.02

Modelo	**Comando**
Modelo nulo	$var\_resposta\_name\_nivel1_{ij} \sim N(XB, \Omega)$ $var\_resposta\_name\_nivel1_{ij} = \beta_{0ij}cons$ $\beta_{0ij} = 15{,}431(0{,}588) + u_{0j} + e_{0ij}$ $\begin{bmatrix} u_{0j} \end{bmatrix} \sim N(0, \Omega_u) : \Omega_u = \begin{bmatrix} 9{,}419(2{,}679) \end{bmatrix}$ $\begin{bmatrix} e_{0ij} \end{bmatrix} \sim N(0, \Omega_e) : \Omega_e = \begin{bmatrix} 16{,}463(1{,}010) \end{bmatrix}$ 
Modelo com intercepto aleatório entre grupos do nível 2	$var\_resposta\_name\_nivel1_{ij} \sim N(XB, \Omega)$ $var\_resposta\_name\_nivel1_{ij} = \beta_{0ij}cons + 0{,}591(0{,}046)var\_name\_nivel1_{ij} + 1{,}093(0{,}310)var\_name\_nivel2_{ij}$ $\beta_{0ij} = 14{,}886(0{,}562) + u_{0j} + e_{0ij}$ $\begin{bmatrix} u_{0j} \end{bmatrix} \sim N(0, \Omega_u) : \Omega_u = \begin{bmatrix} 8{,}045(2{,}259) \end{bmatrix}$ $\begin{bmatrix} e_{0ij} \end{bmatrix} \sim N(0, \Omega_e) : \Omega_e = \begin{bmatrix} 12{,}239(0{,}751) \end{bmatrix}$ 
Modelo com intercepto e inclinação aleatórios entre grupos do nível 2	$var\_resposta\_name\_nivel1_{ij} \sim N(XB, \Omega)$ $var\_resposta\_name\_nivel1_{ij} = \beta_{0ij}cons + 0{,}585(0{,}046)var\_name\_nivel1_{ij} + \beta_{2j}var\_name\_nivel2_{ij}$ $\beta_{0ij} = 14{,}893(0{,}522) + u_{0j} + e_{0ij}$ $\beta_{2j} = 1{,}105(0{,}372) + u_{2j}$ $\begin{bmatrix} u_{0j} \\ u_{2j} \end{bmatrix} \sim N(0, \Omega_u) : \Omega_u = \begin{bmatrix} 6{,}766(2{,}100) & \\ 1{,}001(1{,}048) & 1{,}218(1{,}031) \end{bmatrix}$ $\begin{bmatrix} e_{0ij} \end{bmatrix} \sim N(0, \Omega_e) : \Omega_e = \begin{bmatrix} 11{,}940(0{,}753) \end{bmatrix}$ 

var_reposta_name_nivel1 = nome da variável resposta de pesquisa; var_name_nivel1 = nome da variável explicativa do nível 1; var_name_nivel 2 = nome da variável explicativa do nível 2; var_grupo_nível 2 = nome da variável que define o grupo; var_individuo_nivel1 = nome da variável que define o indivíduo.

ANEXO D – INTRUMENTO DE COLETA DE DADOS DO ARTIGO 2

Opinião de alunos da 5ª série sobre a alimentação de bebês

Data:___/___/___

A. Dados gerais:

1. Número do questionário	NUMQ □□□□
2. Número do entrevistador	NUME □□

B. Dados da Escola:

1 . Nome da escola	
2. Tipo de escola [1] estadual urbana [2] estadual rural [3] municipal urbana [4] municipal rural [5] particular	TIPE □

C. Dados do aluno:

1. Nome do aluno:.....	
2. Idade [___] anos completos	IDADE □□
3. Sexo [1] masculino [2] feminino	SEXO □
4. Turma [ ][ ]	TURMA □□
5. Turno [1] manhã [2] tarde	TURNO □

D. Para cada questão, escolha apenas UMA resposta, a que melhor expresse a que sua opinião:

1. Na sua opinião, qual a principal função das mamas? [1] Para dar mamar para as crianças pequenas. [2] Embelezar o corpo das mulheres. [3] Para diferenciar o corpo das mulheres do corpo dos homens. [4] Não sei.	MAMA1 □
2. Marquinhos acabou de nascer e é o primeiro filho de Dona Ana e seu Paulo. Eles estão muito felizes e querem que Marquinhos receba o melhor leite. Qual leite você acha que seria o melhor para o Marquinhos? [1] Leite em pó. [2] Leite tirado da vaca. [3] Leite de saquinho ou caixinha. [4] Leite do peito da mãe. [5] Não sei.	ALIMF1 □
3. Imagine que você é o médico do Marquinhos, que acabou de nascer. Que conselho você daria à mãe dele? [1] Dê só o peito para o seu bebê. [2] Dê o peito e mamadeira para o seu bebê. [3] Dê só mamadeira para o seu bebê. [4] Tanto faz dar o peito ou a mamadeira. [5] Não sei.	MELB1 □
4. Marquinhos está com 15 dias de vida e continua mamando só no peito. Dona Teresa, sua	AVON1 □

avó, acha que o neto precisa receber também água e chazinho. Na sua opinião: [1] Dona Teresa está certa, porque todo o bebê precisa receber água e chazinho desde que nasce. [2] Marquinhos não precisa de água ou chá no primeiro mês, mas depois que completar 1 mês vai precisar tomar água e chazinho nos intervalos das mamadas no peito. [3] Marquinhos deve receber chazinho só se tiver cólicas. [4] Marquinhos não precisa de água ou chá, porque a criança que mama só no peito não precisa de outros líquidos até os 6 meses. [5] Não sei.	
5. Marquinhos agora tem 1 mês de idade, não chupa bico e só mama no peito. Dona Ana e seu Paulo levaram Marquinhos ao médico porque eles acham que Marquinhos chora muito. O médico disse que Marquinhos está bem e está crescendo bem. Se você fosse a mãe ou o pai de Marquinhos, o que você faria em primeiro lugar para ele chorar menos? [1] Daria bico. [2] Daria mais colo. [3] Daria mamadeira com outro leite. [4] Daria chazinho. [5] Daria umas palmadas. [6] Não sei.	MECO1
6. Marquinhos já fez 2 meses e continua mamando só no peito. Os pais estão preocupados porque Marquinhos não tem horário certo para mamar. Mama seguido (8 a 12 vezes por dia) e acorda à noite para mamar. O médico de Marquinhos constatou que ele está muito bem de saúde. Se você fosse o médico, o que diria para os pais de Marquinhos? [1] É normal bebê dessa idade não ter horário para mamar e mamar várias vezes ao dia. [2] É preciso dar de mamar em horários regulares (de 3 em 3 horas ou de 4 em 4 horas), para a criança se disciplinar. [3] É preciso dar água e chazinho para a criança mamar menos no peito. [4] Durante o dia, a criança pode mamar quanto quiser, mas à noite ela deve se acostumar a não mamar. [5] Não sei.	HORA1
7. Marquinhos vai completar 3 meses na semana que vem e ainda mama no peito. Ele está crescendo bem. Os pais levaram Marquinhos ao médico para saber quando ele deve começar a receber outros alimentos (suquinhos, frutinhas, sopinhas). O que você acha que o médico respondeu? [1] Marquinhos já deveria estar comendo outros alimentos. [2] Marquinhos deve começar a comer outros alimentos quando completar 3 meses. [3] Marquinhos deve começar a comer outros alimentos quando tiver 4 meses. [4] Marquinhos deve começar a comer outros alimentos perto dos 6 meses. [5] Marquinhos só deve começar a comer outros alimentos depois que completar 1 ano de idade. [6] Não sei.	ALIA1
8. Até quanto tempo você acha que Marquinho deve mamar no peito? [1] Por 3 meses. [2] Por 6 meses. [3] Por 1 ano. [4] Por 2 anos ou mais. [5] Não sei.	MESES1
9. Marquinhos cresceu, casou-se com Marcela e tiveram uma filha chamada Linda. Marquinhos decidiu com Marcela que a filha seria amamentada no peito. Na sua opinião, qual a melhor	LINDA1

maneira de Marquinhos ajudar para que a amamentação seja boa? [1] Dando mamadeira de vez em quando para a mãe descansar. [2] Dando força, carinho e atenção à Marcela, fazendo os serviços da casa, fazendo as compras da casa e trocando as fraldas do bebê. [3] Trabalhando para manter as despesas da casa. [4] Marquinhos não pode ajudar, pois só a Marcela pode dar o peito para Linda mamar. [5] Não sei.	
ESCORE 1:	ESCORE1 □□

E. Marque com um X as palavras que na sua opinião completam melhor a frase:

10. O bebê gosta mais ..... [1] do peito [2] da mamadeira. [3] Não sei.	GOSTO2 □
11. É mais fácil alimentar um bebê ..... [1] com mamadeira. [2] dando o peito. [3] Não sei	FACIL2 □
12. Amamentar uma criança..... [1] dói. [2] não dói. [3] Não sei.	DORA2 □
13. O leite do peito de algumas mulheres..... [1] é fraco e não sustenta o bebê. [2] não é fraco e sustenta o bebê. [3] Não sei.	FRACOL2 □
14. As crianças que mamam no peito pegam..... [1] menos doenças. [2] mais doenças. [3] Não sei.	DOEN2 □
15. Quando o bebê machuca o peito da mãe é porque..... [1] eles estão com fome e sugando muito forte. [2] ele está mamando de mau jeito. [3] é normal o bebê machucar o peito da mãe. [4] Não sei.	JEITO2 □
16. A mulher que tem mamas grandes.....que as mulheres com mamas pequenas. [1] produz mais leite. [2] produz a mesma quantidade de leite. [3] Não sei.	MAGR2 □
17. A mãe..... parar de dar o peito para o bebê quando ele começa a ter dentinhos. [1] não precisa.	DENT2 □

[2] precisa. [3] Não sei.	
18. O bebê que mama só no peito nos primeiros 6 meses..... [1] fica muito magrinho. [2] fica muito gordinho. [3] Não sei.	LEPO2 □
19. Na maioria das vezes .....oferecer bico para acalmar o bebê. [1] não é preciso. [2] é preciso. [3] Não sei.	BICO2 □
20. Bebê que mama só no peito..... entre as mamadas. [1] precisa tomar chá e água. [2] não precisa tomar chá e água. [3] Não sei.	ACHA2 □
21. Quando o bebê tem cólicas ou chora muito.....dar chazinho para ele. [1] é preciso. [2] não é preciso. [3] Não sei.	COLIC2 □
22. Para que uma mulher amamente o seu bebê com sucesso..... a participação do pai. [1] é muito importante [2] não é importante. [3] Não sei	PPAI2 □
23. O bebê quando nasce..... [1] precisa aprender a mamar porque ele não nasce sabendo. [2] não precisa aprender a mamar porque ele nasce sabendo. [3] Não sei.	NASC2 □
24. Se a mãe trabalha fora de casa.....continuar amamentando seu filho no peito. [1] não é possível. [2] é possível. [3] Não sei.	TRABA □
25. A amamentação..... do peito da mulher ficar caído. [1] é a principal causa. [2] não é a principal causa. [3] Não sei.	PEICA □
ESCORE 2:	ESCORE2 □□
ESCORE TOTAL:	ESCORE TOTAL □□

F. Escolha a resposta e marque com X, quando for pedido explique a sua resposta:

1. Você mamou no peito? [1] Sim [2] Não [3] Não sei	MAMOU □
2. Você já viu alguém amamentando? [1] Sim [2] Não [3] Não sei	ERAM □
3. Se você já viu alguém sendo amamentado, quem você já viu? (Marque com X e pode marcar mais de uma resposta) [ ] Meu irmão ou minha irmã. [ ] Bebê da família. [ ] Bebês de pessoas conhecidas, mas não da família. [ ] Desconhecidos, na rua. [ ] Personagem da televisão. [ ] Em livros e revistas. [ ] Outros Quais os outros? ----------	MUEMI □ BFAM □ BCON □ DRUA□ PERST □ LIVRE □ VOUTRO □
4. Nas suas brincadeiras da infância, as bonecas mamavam no peito? [ ] Sim. [ ] Não. [ ] Não sei.	BRIBO □
5. Você acha que é feio dar o peito para o bebê na frente de outras pessoas? [1] Sim [2] Não [3] Não sei	FEIO □
6. Se você (ou sua esposa) tivesse um filho hoje, como gostaria que ele fosse alimentado? [1] Só com leite em pó, na mamadeira. [2] Só com leite de vaca (direto da vaca ou de saquinho), na mamadeira. [3] Só com leite do peito. [4] Com leite do peito e outro leite na mamadeira. [5] Não tenho opinião ainda.	FIMA □

G. Complete as frases a seguir:

1. Diga três coisas boas da amamentação.
   1.....
   2.....
   3.....

2. Diga três coisas não boas da amamentação.
   1.....
   2.....
   3.....

3. Diga três coisas boas da mamadeira:

1.....

2.....

3.....

4. Diga três coisas não boas da mamadeira:

1.....

2.....

3.....

5. Descreva como o pai do bebê pode participar na fase de amamentação da criança.

.....

.....

.....

H. Informações socioeconômicas sobre a família:

1. Com quem você mora? (Marque com um X e pode marcar mais que uma resposta) [ ] mãe [ ] madrasta [ ] pai [ ] padrasto [ ] irmãos [ ] com outros. Quais são as outras pessoas?	MAE □ MADRAS □ PAI □ PADRAS □ IRMAOS □ OUTRO □
2. A sua mãe sabe ler ? [1] Sim. [2] Não. [3] Não sei.	MESLER □
3. A sua mãe sabe escrever ? [1] Sim. [2] Não. [3] Não sei.	MSESC □
4. A sua mãe estudou na escola até: [1] ensino fundamental (da primeira a oitava série) [2] ensino médio (segundo grau). [3] Ensino superior (faculdade). [4] Não sei.	MPSF □
5. Quantos anos completos a sua mãe estudou? [___] anos. [88] Não sei.	[____] ANOS MESC □□
6. O seu pai sabe ler? [1] Sim. [2] Não. [3] Não sei.	PESLER □

7. O seu pai sabe escrever? [1] Sim. [2] Não. [3] Não sei.	PSECR □
8. Seu pai estudou na escola até: [1] ensino fundamental (da primeira a oitava série) [2] ensino médio (segundo grau). [3] Ensino superior (faculdade). [4] Não sei.	PSESC □
9. Quantos anos completos o seu pai estudou? [____] anos. [88] Não sei.	[_____] ANOS PESC □□

ANEXO E – DECLARAÇÃO

DECLARAÇÃO

Pelo presente instrumento, eu, Iara Denise Endruweit Battisti, brasileira, casada, professora, residente e domiciliada em Santa Rosa/RS, à Av. Borges de Medeiros, nº 550/701, Centro, portador da identidade nº 7051186246, declaro que a confiabilidade dos dados abaixo nomeados, está garantida, pois nos bancos disponibilizados não é possível a identificação dos indivíduos envolvidos na pesquisa. Também, declaro que os referidos dados serão utilizados na execução do projeto *Análise de dados epidemiológicos incorporando esquemas amostrais complexos*, sendo este o projeto de doutorado no Programa de Pós-Graduação em Epidemiologia da Faculdade de Medicina da UFRGS.

Salienta-se que os projetos de onde provem os dados a serem utilizados no projeto de doutorado foram aprovados por Comitê de Ética em Pesquisa os quais estão nomeados a seguir:

Banco de dados:

I. Campanha Nacional de Detecção de Diabetes Mellitus. Pesquisadores: Bruce B. Duncan e Maria Inês Schmidt, PPG Epidemiologia / UFRGS

Projeto aprovado pelo Comitê de Ética em Pesquisa da UFRGS.

II. Avaliação de estratégia de aleitamento materno em escolas de ensino fundamental. Pesquisadores: Elsa Regina Justo Giugliani e Silvania Bottaro, PPG Ciências Médicas: Pediatria / UFRGS

Projeto aprovado pelo Grupo de Pesquisa e Pós-graduação do Hospital de Clínicas de Porto Alegre e pela Comissão Científica e Comissão de Pesquisa e Ética em Saúde.

Ciente do exposto acima, assino o presente instrumento.

Porto Alegre, 21 de agosto de 2006.

Iara Endruweit Battisti

Iara Denise Endruweit Battisti

De acordo:

Jandyra Maria Guimarães Fachel

HCPA - HOSPITAL DE CLÍNICAS DE PORTO ALEGRE
Grupo de Pesquisa e Pós-Graduação
COMISSÃO CIENTÍFICA E COMISSÃO DE PESQUISA E ÉTICA EM SAÚDE

RESOLUÇÃO

A Comissão Científica e a Comissão de Pesquisa e Ética em Saúde, que é reconhecida pela Comissão Nacional de Ética em Pesquisa (CONEP)/MS como Comitê de Ética em Pesquisa do HCPA e pelo Office For Human Research Protections (OHRP)/USDHHS, como Institucional Review Board (IRB0000921) analisaram o projeto:

Projeto: 02-170

Pesquisadores:
SOTERO SERRATE MENGUE
BRUCE B. DUNCAN
MARIA INES SCHIMIDT

Título: AVALIAÇÃO DAS CAMPANHAS NACIONAIS PARA DETECÇÃO DE DIABETES MELLITUS E HIPERTENSÃO (CNDDM / CNDHAS)

Este projeto foi Aprovado em seus aspectos éticos e metodológicos, de acordo com as Diretrizes e Normas Internacionais e Nacionais, especialmente as Resoluções 196/96 e complementares do Conselho Nacional de Saúde. Toda e qualquer alteração do Projeto deverá ser comunicada ao CEP/HCPA.

Porto Alegre, 05 de julho de 2002.

Profa. Themis Reverbel da Silveira
Coordenadora do GPPG e CEP-HCPA

HCPA - HOSPITAL DE CLÍNICAS DE PORTO ALEGRE
Grupo de Pesquisa e Pós-Graduação
COMISSÃO CIENTÍFICA E COMISSÃO DE PESQUISA E ÉTICA EM SAÚDE

RESOLUÇÃO

A Comissão Científica e a Comissão de Pesquisa e Ética em Saúde, que é reconhecida pela Comissão Nacional de Ética em Pesquisa (CONEP)/MS como Comitê de Ética em Pesquisa do HCPA e pelo Office For Human Research Protections (OHRP)/USDHHS, como Institucional Review Board (IRB0000921) analisaram o projeto:

Projeto: 01-429

Pesquisadores:
ELSA REGINA JUSTO GIUGLIANI
SILVANIA MORAES BOTTARO

Título: AVALIAÇÃO DE ESTRATÉGIA DE PROMOÇÃO DO ALEITAMENTO MATERNO EM ESCOLARES DE 5ª SÉRIES DO ENSINO FUNDAMENTAL

Este projeto foi Aprovado em seus aspectos éticos e metodológicos, inclusive quanto ao seu Termo de Consentimento Livre e Esclarecido, de acordo com as Diretrizes e Normas Internacionais e Nacionais, especialmente as Resoluções 196/96 e complementares do Conselho Nacional de Saúde. Os membros do CEP/HCPA não participaram do processo de avaliação dos projetos onde constam como pesquisadores. Toda e qualquer alteração do Projeto, assim como os eventos adversos graves, deverão ser comunicados imediatamente ao CEP/HCPA.

Porto Alegre, 16 de janeiro de 2002.

Profa. Themis Reverbel da Silveira
Coordenadora do GPPG e CEP-HCPA

ANEXO F – PROJETO DE PESQUISA

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
FACULDADE DE MEDICINA
PROGRAMA DE PÓS-GRADUAÇÃO EM EPIDEMIOLOGIA

Projeto de Pesquisa

Análise de dados epidemiológicos incorporando esquemas amostrais complexos

Iara Denise Endruweit Battisti
Orientadora: Jandyra Maria Guimarães Fachel
Co-orientador: João Riboldi
Colaboradora: Elsa Cristina de Mundstock

Porto Alegre, 29 de agosto de 2006.

SUMÁRIO

1 INTRODUÇÃO

1.1 Caracterização do problema

Muitos estudos epidemiológicos utilizam amostragem probabilística para coleta de dados, definidas de quatro formas diferentes: amostragem aleatória simples, amostragem estratificada, amostragem sistemática e amostragem por conglomerado. É muito comum, em grandes inquéritos para estudo da saúde de indivíduos, a aplicação de duas ou mais destas formas de amostragem ao mesmo tempo, e ainda, em dois ou mais estágios de seleção das unidades amostrais, definindo-se como amostragem complexa.

A maioria dos softwares estatísticos contempla técnicas de análise dados provindos de amostragem aleatória simples, sendo que estas não são apropriadas para os demais tipos de amostragem probabilística. Felizmente, atualmente têm-se rotinas disponíveis para análise de dados provindos de diferentes amostragens probabilísticas em softwares estatísticos, porém ainda subutilizados. Isto se deve a pouca disponibilidade de referências aplicadas à epidemiologia, assim como a recente incorporação destas técnicas nos softwares.

Sabe-se que a análise de dados desconsiderando a amostragem complexa resulta em estimativas incorretas, porém isso é ignorado em muitas pesquisas.

Desta forma, tem-se a necessidade de demonstrar o erro que se comete ao ignorar a complexidade do delineamento na análise de dados, assim como comparar essa metodologia com outras alternativas. Para isso serão utilizados dados de dois estudos já realizados.

1.2 Questão de pesquisa

- Qual é o impacto (magnitude) nas estimativas quando ignora-se o plano efetivamente utilizado na coleta de dados?
- Modelagem incorporando métodos de *complex survey* leva aos mesmos resultados de modelagem por multinível?

1.3 Justificativa

Várias pesquisas utilizam amostragem complexa para a coleta de dados, porém muitos pesquisadores ainda não utilizam tratamento estatístico adequado para esses dados.

Pouca disponibilidade de referência metodológica aplicada à epidemiologia, principalmente na definição de parâmetros da amostragem complexa.

Recursos computacionais recentemente disponibilizados nos pacotes estatísticos, por isso são desconhecidos por muitos.

Necessidade de comparação de resultados de *complex survey* e modelos multiníveis, sendo essas duas técnicas disponibilizadas recentemente nos pacotes estatísticos e de grande importância para resultados mais específicos.

2 OBJETIVOS

2.1 Geral

Comparar estimativas e respectivos erros padrão (precisão) para várias técnicas estatísticas incorporando métodos de *complex survey* com as estimativas

de amostra aleatória simples e método multinível, utilizando dois bancos de dados provenientes de amostragem complexa.

2.2 Específicos

- Obter e comparar estimativas considerando uma amostra padrão (AAS) e amostra complexa, utilizando análise simples (média, proporção, desvio-padrão), para os diferentes estudos.
- Obter e comparar estimativas considerando uma amostra padrão (AAS) e amostra complexa, utilizando modelagem (regressão logística, regressão linear), para os diferentes estudos.
- Obter e comparar resultados de análise utilizando metodologia de amostragem complexa e metodologia multinível.
- Investigar como se pode incorporar o delineamento complexo nas análises estatísticas de estudos epidemiológicos, para obter estimativas corretas.
- Implementar análise de resíduos ajustados em tabelas cruzadas no *software* R e outros procedimentos que necessários para análise de dados
- Comparar recursos atualmente disponíveis para análise de dados provenientes de amostragem complexa disponibilizados nos *softwares*.

3 REVISÃO DE LITERATURA

Grandes inquéritos populacionais são comuns em estudos da saúde da população de um país, de uma região ou de um município. Como geralmente envolvem grandes áreas geográficas e diferentes níveis populacionais, o delineamento amostral não é simples, isto é, envolve um delineamento complexo.

Como os inquéritos têm o propósito de inferir características da amostra para a população, a amostragem aleatória é a mais apropriada. Em um delineamento complexo tem-se a junção de pelo menos dois tipos de amostragens aleatórias entre as seguintes: amostragem aleatória simples ou sistemática,

amostragem estratificada, amostragem por conglomerado em uma ou várias etapas. Ainda, pode haver amostragem com probabilidades variáveis.

A coleta de dados por amostragem apropriada evita desperdício de tempo, de recursos materiais e financeiros, garantindo maior qualidade às medidas, uma vez que um número menor de entrevistas permite maior cuidado em sua realização (BARATA et al., 2005).

Dados provindos de grandes inquéritos são muito freqüentes em pesquisas epidemiológicas, disponíveis como dados primários ou como dados secundários. Nos Estados Unidos, organizações como o Centro Nacional para Estatísticas de Saúde (*National Center for Health Statistics*) e o Departamento de Censo Americano (*United States Bureau of the Census*) disponibilizam dados secundários sobre saúde da população (LEMESHOW e COOK, 1999). No Brasil, dados secundários de saúde são disponibilizados pelo IBGE e pelo DATASUS.

A vantagem desses grandes estudos epidemiológicos é que muitos fatores de riscos podem ser examinados, porém trazem uma dificuldade maior na identificação das técnicas mais adequadas para análise dos dados destes estudos (KORN e GRAUBARD, 1991).

AMOSTRAGEM

Amostragem é o processo pelo qual obtém-se uma ou mais amostras da população. Na amostragem seleciona-se uma parte de uma população para observá-la com a finalidade de estimar parâmetros populacionais. É importante que os diferentes procedimentos amostrais satisfaçam os seguintes critérios (MUNDSTOCK, 2005):

1) que sejam amostras representativas da população;

2) que forneçam estimativas precisas das características da população, podendo medir sua confiabilidade;

3) que tenham pequeno custo para selecionar a amostra.

O processo de obtenção de uma amostra está relacionado com (MUNDSTOCK, 2005):

- o tamanho da amostra;
- a maneira de selecionar a amostra;
- a definição dos métodos para obter os dados;
- a escolha dos tipos de dados para registrar.

Unidades amostrais

As unidades amostrais são as unidades observadas com a finalidade de fazer estimativas. As unidades podem ser pessoas, residências, hospitais, entre outras.

Amostragem probabilística

Na amostragem probabilística é possível calcular, com antecedência, a probabilidade de se obter cada uma das amostras possíveis, sendo que todas as unidades da população têm probabilidade maior que zero de entrar na amostra. É importante observar que a aleatoriedade não é uma característica de uma determinada amostra, mas sim do processo pelo qual foi obtida.

A seleção probabilística de amostras exclui fontes humanas de erro, tais como tendências conscientes ou inconscientes de selecionar unidades com valores maiores (ou menores) da variável de interesse.

Também na amostra probabilística é possível quantificar os erros de amostragem, ou discrepâncias entre as estimativas amostrais e os valores populacionais que seriam obtidos observando todas as unidades da população. O uso de amostragem probabilística permite que sejam feitas estimativas da magnitude média desses erros. Também, permite pré-fixar o tamanho de amostra, de maneira que a magnitude média dos erros de amostragem não ultrapasse um valor pré-determinado com uma probabilidade pré-determinada.

Os métodos probabilísticos permitem o controle da precisão das estimativas amostrais dentro de determinados limites fixados com antecedência (intervalos de confiança). Existem diferentes métodos de obtenção de uma amostra probabilística de uma população, os quais são descritos a seguir (MUNDSTOCK, 2005;MUNDSTOCK, 2004):

Amostragem aleatória simples (AAS)

Consiste na seleção de n unidades de uma população de tamanho N, de maneira que cada uma das amostras possíveis tenha a mesma probabilidade de ser selecionada.

As unidades da população são numeradas de 1 a N e depois são obtidos números aleatórios da tabela ou do computador. Uma amostra aleatória simples é selecionada extraindo-se uma unidade de cada vez. As unidades correspondentes aos n números sorteados constituem a amostra.

Amostragem estratificada

A população é dividida em subpopulações mutuamente exclusivas chamadas de estratos. A amostragem estratificada consiste em selecionar amostras em cada estrato e combinar estas amostras numa única amostra para estimar parâmetros da população.

Tem como vantagem o aumento da precisão das estimativas, possibilidade de obtenção de informações a nível de estrato e facilidade na coleta de dados, por razões físicas ou administrativas.

Amostragem por conglomerados

A população é dividida em M grupos ou conglomerados que servem como unidades primárias de amostragem (UPA), de maneira que cada unidade da população é associada com um e somente um conglomerado. Cada conglomerado é formado por N_i unidades, chamadas unidades secundárias. Das M unidades primárias (conglomerados) na população é selecionada uma amostra de tamanho

m pelo método aleatório simples ou sistemático. A amostragem por conglomerados pode ser realizada em etapa única ou em mais etapas, como segue:

Amostragem em etapa única: todas as unidades do conglomerado selecionado são incluídas na amostra.

Amostragem em duas etapas (bietápica): nos conglomerados selecionados são extraídas amostras de n_i unidades secundárias.

Amostragem multietápica: o processo pode ser estendido a várias etapas de amostragem sendo que não é necessário ter o mesmo método de amostragem em todos os níveis (UCLA, 2005)

Amostragem sistemática

Quando uma listagem de indivíduos da população está disponível, pode-se selecionar, aleatoriamente, uma unidade amostral entre as k primeiras unidades populacionais e, a partir daí, selecionar as restantes a intervalos fixo em cada k unidade.

Amostragem com probabilidades variáveis

Em alguns procedimentos amostrais algumas unidades da população são “mais importantes” por terem uma contribuição maior no valor do parâmetro, neste caso estabelece-se probabilidades desiguais de seleção às diferentes unidades da população. Assim, quando as unidades variam em tamanho e a variável em estudo está correlacionada com tamanho, as probabilidades de seleção podem ser estabelecidas em proporção ao tamanho da unidade. Neste caso, o procedimento amostral é definido como amostragem com probabilidade proporcional ao tamanho (PPT).

A vantagem em utilizar amostragem PPT é obter uma amostra mais representativa da população e assim aumentar a precisão dos estimadores quando comparados à Amostragem Aleatória Simples (AAS).

Tipos de seleção das unidades amostrais

A amostragem é dita sem reposição, quando um elemento selecionado em uma extração é excluído da população para as extrações subseqüentes. Neste esquema, todos os elementos da mesma amostra devem ser diferentes.

A amostragem é dita com reposição, quando os N elementos da população permanecem em todas as extrações, isto é, uma unidade selecionada em uma extração é reposta e pode ser extraída novamente. Assim, um elemento pode-se repetir na mesma amostra.

TÉCNICAS PARA ESTIMAÇÃO DE PARÂMETROS DE AMOSTRAGEM COMPLEXA

Para estimação de médias utiliza-se estimadores de razão e para estimar variâncias utilizam-se diferentes métodos aproximados, que não são expressões algébricas exatas: linearização por série de Taylor e técnica de replicação (COHEN, 1997); (HEERINGA e LIU, 1997), os quais produzem resultados semelhantes (RODGERS-FARMER e DAVIS, 2001). Mais tarde, o método bootstrap, foi incluído entre os métodos de reamostragem (HEERINGA e LIU, 1997). Ainda mais sofisticadas, são as técnicas para estimação de coeficientes de regressão, como o método de estimação de equações generalizadas (PÉREZ et al., 2004).

A estimação de variância pelo método de Taylor ou técnicas de replicação sofre uma importante limitação, pois as técnicas são aproximadas e não dão conta de todos os possíveis planos de amostragem e estimadores desejados. Segundo Wolter *apud* (SOUSA e SILVA, 2003) qualquer uma das técnicas resultará estimativas aproximadas de variância.

(RAO e WU, 1988)) desenvolveram métodos que tem as características dos métodos jackknife e replicação repetida balanceada (BRR), que usam uma fórmula única de variância para todas estatísticas não lineares (por exemplo, razão, regressão e coeficiente de correlação) necessitando de mais “esforço computacional”, porém permitem extensão para delineamentos mais complexos com amostragem sem reposição. Ao contrário, a linearização por série de Taylor utiliza uma fórmula separada para cada estatística não-linear, necessitando maiores recursos de programação”.

A correção para população finita (FPC – *finite population correction*), definida como $((N-n)/(N-1))^{½}$ é usada quando a fração de amostragem é grande, neste caso é aconselhável incluí-la no cálculo do erro padrão da estimativa. Se o FPC for 1, o impacto é pequeno, podendo ser ignorado na análise (UCLA, 2005).

ESTIMATIVAS OBTIDAS EM AMOSTRAGEM COMPLEXA

Já se sabe da teoria de amostragem, apresentada por (COCHRAN, 1965), que quando um delineamento complexo é usado para coleta de dados, este deve ser considerado na análise dos dados, para que as estimativas pontuais e seus respectivos erros-padrão sejam corretos (UCLA, 2005). Porém, ainda é comum encontrar resultados de análise onde o delineamento complexo foi ignorado (LEMESHOW e COOK, 1999). Isso ocorre mais freqüentemente em países em desenvolvimento, onde ainda são obtidas estimativas em softwares estatísticos que não possuem módulos específicos para amostragem complexa (SOUSA e SILVA, 2003). Alguns dos motivos disso acontecer é o desconhecimento do impacto nas estimativas quando se ignora o delineamento complexo e também a falta de disponibilidade das técnicas adequadas nos softwares estatísticos até bem pouco tempo atrás.

Assim, estudos vêm sendo realizados para demonstrar esses impactos e a comparação de softwares específicos para tratamento de dados de delineamentos complexos, porém percebe-se pouca abordagem na saúde.

No entanto, já se encontram estudos que incorporam o delineamento amostral complexo e que mostram o cálculo da fração global de amostragem e o valor do efeito do delineamento como o estudo de (BUENO et al., 2003) e que consideram na análise o delineamento amostral como o estudo de (BARROS e BERTOLDI, 2002), (HERNÁNDEZ et al., 2003), (EGEDE, 2003), (EGEDE, 2003) e (EGEDE, 2004).

Amostras são escolhidas freqüentemente para ter conglomerados geográficos para reduzir custos administrativos (CARLSON, 2003), permitindo um custo menor por indivíduo amostrado comparadas a uma amostra aleatória simples,

porém têm a desvantagem de análise estatística mais complexa e geralmente incrementos nas variâncias dos estimadores (CORDEIRO, 2001).

É importante que os conglomerados sejam heterogêneos. Espera-se que exista maior heterogeneidade entre os indivíduos, resultando em maior variância nas estimativas obtidas. Por esse motivo, aumenta-se o tamanho da amostra e o efeito do conglomerado deve ser considerado no cálculo dos intervalos de confiança (BARATA et al., 2005). Desconsiderar o efeito do conglomerado na análise de dados fornecerá pouco impacto nas estimativas dos parâmetros, porém resultará em subestimação da variabilidade, isto é, erros-padrão subestimados e intervalos de confiança menores (HORTON e FITZMAURICE, 2004).

Segundo (CORDEIRO, 2001), para um determinado delineamento amostral e um determinado tamanho de conglomerado, o efeito do delineamento é função do grau de homogeneidade intraconglomerado para as classes amostradas. Complementarmente, quando o efeito de desenho é igual ou próximo a 1, pode-se estimar a variância do estimador como AAS.

O grau de viés na variância estimada a partir de dados provindos de amostragem por conglomerados é uma função do coeficiente de correlação intraconglomerado (ICC), sendo que quanto maior o ICC, maior será o viés (THOMAS e HECK, 2001). Quando o ICC for maior que 0,05 há necessidade de ajuste para efeito do delineamento do conglomerado (THOMAS e HECK, 2001). Pesos amostrais corrigem para superamostragem, porém não para similaridade dentro dos conglomerados, o estudo de (THOMAS e HECK, 2001) apresenta quatro formas de considerar o efeito dos conglomerados.

Outra característica muito comum numa amostra complexa é a estratificação, por proporcionar facilidade logística na fase de coleta de dados, no entanto desconsiderar o efeito da estratificação tem pequeno impacto nas estimativas dos parâmetros, mas superestima a variabilidade, isto é, os erros-padrão serão superestimados e intervalos de confiança maiores (HORTON e FITZMAURICE, 2004).

A estratificação tem o objetivo de reduzir a variância das estimativas amostrais, isto é, a variância das estimativas amostrais é reduzida ao ponto que a variabilidade entre unidades do mesmo estrato é menor que sua variância em toda a população (HORTON e FITZMAURICE, 2004). Cada estrato é independente dos demais, desta forma, estratos poderão ter pesos diferentes (UCLA, 2005).

Também, é comum, em amostragem complexa a probabilidade maior para determinado subgrupo populacional para assegurar representatividade na amostra final, tornando possíveis análises separadas para grupos diferenciados. Neste caso, os indivíduos de grupos diferentes da amostra não tiveram a mesma probabilidade de seleção, sendo necessário ajustes através de pesos e também ajustes para não resposta ou outros fatores, como pós estratificação (CARLSON, 2003), (KORN e GRAUBARD, 1991) e estratificação (TRAVASSOS et al., 2002).

(HERNÁNDEZ et al., 2003) denominam o peso como um fator de expansão, sendo que o valor do peso indica o número de indivíduos na população que cada observação na amostra representa (BROGAN, 2003), (KORN e GRAUBARD, 1995).

Não considerar o peso amostral resultará em estimativas dos parâmetros populacionais viesados e em subestimação da variabilidade, isto é, erros padrão subestimados e intervalos de confiança mais estreitos (HORTON e FITZMAURICE, 2004) (GUILLÉN et al., 2000).

Existem vários tipos de pesos que podem estar associados a uma pesquisa, sendo o mais comum o peso amostral, que é o inverso da probabilidade do indivíduo ser selecionado na amostra, calculado como N/n, em que N é o tamanho da população e n é o tamanho da amostra. No caso de delineamento em dois estágios, o peso é calculado como $f_1.f_2$, respectivamente o peso no primeiro e no segundo estágio do delineamento. Em muitos planos amostrais, a soma dos pesos será igual ao tamanho da população (UCLA, 2005).

Os autores (KNEIPP e YARANDI, 2002) consideram duas categorias de ponderação: (a) ponderação, representando a pessoa ou nível familiar e (b) ponderação da variância estimada. A primeira ajusta a estimativa para refletir o número total ou proporção de elementos na população baseada no delineamento e tamanho amostral, sendo possível executar nos softwares SPSS e STATA. A segunda ajusta para o aumento da variância em função do delineamento amostral, sendo possível somente no STATA. Estas duas categorias de ajuste para dados provenientes de delineamento complexo refletem melhor o desfecho na população de interesse e são definidos em função do tipo de delineamento.

No estudo realizado por (THOMAS e HECK, 2001), dois tipos de pesos são abordados: brutos e relativos. Ainda, comenta-se que alguns softwares usam o

tamanho da amostra como a soma dos pesos para o tamanho da população, sendo que isso não afeta a estimativa pontual da média, porém afeta o erro padrão, tornando-o menor do que realmente seria. Desta forma, os resultados dos testes de hipóteses podem ser significativos quando na realidade não seriam (THOMAS e HECK, 2001). Os autores recomendam o uso de peso relativo.

A importância de se considerar os pesos amostrais é maior quando os mesmos têm uma correlação inversa com a variável analisada, já que nesse caso a estimação não ponderada será maior que o valor real. Quando os pesos amostrais têm pouca variabilidade as estimativas pontuais considerando AAS são similares ao considerar ponderação. Nas análises considerando-se AAS também se subestima as medidas de variabilidade, sendo que suas magnitudes depende da magnitude do coeficiente de correlação das variáveis analisadas dentro do conglomerado. Quando a correlação é grande, a variabilidade é subestimada. A análise ponderada é uma boa estratégia se o coeficiente de correlação é pequeno (29). As ponderações podem aumentar ou diminuir o efeito do delineamento das estimativas, dependendo da correlação dos valores dos pesos com o desvio-padrão da(s) variável(s) usada na estimação da estatística (HEERINGA e LIU, 1997).

Para (KORN e GRAUBARD, 1995), a escolha entre análise ponderada e não ponderada depende de cada estudo, da escolha entre robustez e eficiência, isto é, viés ou variância.

O impacto do delineamento deve sempre ser considerado na análise, porém é difícil identificar qual o aspecto do delineamento é mais importante (KORN e GRAUBARD, 1991). Os autores sugerem que se a diferença entre o número de UPA e o número de estratos é maior ou igual a 20, então se considera o efeito do conglomerado. Caso contrário métodos de análise mais avançados devem ser usados. O peso amostral deve ser considerado se a ineficiência for menor que 10%. Ainda sugerem sempre a documentação das variáveis relacionadas ao delineamento.

As estimativas pontuais dos parâmetros são influenciadas pela ocorrência de pesos amostrais distintos, enquanto as estimativas de variância (precisão) dos estimadores dos parâmetros do modelo são influenciadas também pelos efeitos de estratificação e conglomerados (BROGAN, 2003), (LEITE e SILVA, 2002). O cálculo da variância das estimativas desempenha papel essencial na realização da inferência analítica permitindo a avaliação da precisão das estimativas, bem como

de intervalos de confiança e a formulação de testes de hipóteses sobre os parâmetros dos modelos (LEITE e SILVA, 2002).

O autor (BROGAN, 2003) complementa, que a magnitude do viés em estimativas não ponderadas dependerá da variabilidade dos dados e da variabilidade dos pesos. Também, a relação entre o valor dos pesos e a variável de análise é outro fator que contribui no viés da estimativa baseada na análise não ponderada, como por exemplo, se o grupo for superamostrado e neste caso a prevalência do desfecho é grande, então a estimativa do desfecho será viesada (superestimada). Ainda, estimativas pontuais viesadas e erros-padrão subestimados são obtidos em análises não ponderadas devido a conglomerados e variabilidade dos pesos (BROGAN, 2003). Sendo que o grau de subestimação depende da grandeza do ICC da variável em análise, isto é, quanto maior o ICC, maior é a subestimação da variabilidade, quando se ignora o delineamento complexo na análise de dados (BROGAN, 2003).

A seguir seguem resultados encontrados em estudos com abordagem metodológica na análise de dados provenientes de amostragem complexa.

No estudo de (LEMESHOW e COOK, 1999) os dados da NHANES III (*National Health and Nutrition Examination Survey III*, EUA, 1988-1994) e do PAQUID (Personnes Agées Quid Study, França, 1988) foram analisados considerando duas estratégias de análise: amostra aleatória complexa (estratificado em múltiplos estágios) e amostra aleatória simples. Neste estudo, obteve-se diferença entre as duas estratégias de análise maiores para estimativas pontuais e erros padrão de médias do que para as estimativas de coeficientes de regressão e razão de chances. Concluem que as diferenças encontradas nos resultados entre as duas estratégias de análise e ainda a disponibilidade de softwares, como o STATA e SUDAAN, demonstram a necessidade do uso das técnicas apropriadas para a análise de dados provindos de delineamento amostral complexo.

Detalhando os resultados para NHANES III, percebe-se diferença das estimativas pontuais da média entre as duas estratégias de análise. Também, que o erro-padrão é consideravelmente maior considerando a amostra complexa. Neste caso específico, não houve sobreposição dos intervalos de confiança para média. Houve pouca diferença nas estimativas pontuais dos coeficientes de regressão linear entre as duas estratégias de análise, porém os erros-padrão foram duas vezes maiores quando o delineamento complexo foi considerado. Já, para a

regressão quadrática houve grande diferença nas estimativas pontuais e nos erros-padrão entre as duas estratégias de análise.

Detalhando os resultados para PAQUID, percebe-se que, incluindo a ponderação, a razão de chances é similar para as duas estratégias de análise, porém o efeito de proteção obtido através da tabela cruzada é 18% maior na análise considerando o delineamento complexo e 12% maior utilizando a regressão logística. Para este caso específico, as estimativas pontuais para as médias são pouco menores e os erros-padrão são maiores considerando o delineamento complexo.

A incorporação de pesos pode causar grandes diferenças nos resultados da análise entre as duas estratégias (LEMESHOW e COOK, 1999); (LEMESHOW et al., 1998). (LEMESHOW e COOK, 1999) concluíram a partir do estudo que variâncias de estimativas lineares, como para média, foram subestimadas quando o delineamento complexo foi desconsiderado. Porém, os efeitos de delineamento para coeficientes de regressão não foram tão altos quanto para as médias. Isto pode ser justificado pela expressão do efeito do delineamento = $1+(n-1)\rho_x\rho_y$, que depende do produto do coeficiente de correlação intraconglomerado da variável dependente ρ_y e independente ρ_x. Já para o caso de estimativas lineares, como a média, somente depende do coeficiente de correlação intraconglomerado de uma variável ρ. Também, é possível que os coeficientes de correlação intraconglomerado não tenham a mesma direção, no caso de dados de conglomerados serem heterogêneos para uma variável e homogêneo para outra, resultando em efeito de delineamento menor que 1.

Também, (LEMESHOW e COOK, 1999) comentam que a incorporação de pesos pode afetar muito os coeficientes de regressão estimados. Estas situações incluem especificação incorreta (*misspecification*) do modelo, omissão de uma variável do modelo que tem forte interação com as variáveis independentes e é altamente correlacionada com os pesos e quando a razão de amostragem é muito diferente dentro dos níveis da variável resposta. Em contraste, sempre que a razão de amostragem dentro das categorias da variável resposta difere moderadamente, a ponderação ou não-ponderação fornecerá resultados diferentes. Variâncias da média foram subestimadas considerando amostragem simples, porém coeficientes de regressão e erros-padrão não foram tão diferentes entre os dois delineamentos (LEMESHOW e COOK, 1999).

Anteriormente, (LEMESHOW et al., 1998) analisaram os dados da PAQUID considerando o delineamento complexo da amostragem, já que originalmente a análise foi realizada considerando uma AAS. Os autores utilizaram o STATA, definindo a variável de estratificação, de conglomeração e de ponderação. O valor de ponderação associado a cada indivíduos foi definido em duas etapas, já que o processo de amostragem englobou estratificação e pós-estratificação. Na primeira etapa, a ponderação devido a estratos e conglomerados, o peso foi definido como o inverso da probabilidade de seleção de cada indivíduo e na segunda etapa, a ponderação devido a pós-estratificação foi definida pela razão entre a proporção populacional e proporção amostral nas categorias estratificadas posteriormente. O resultado desta segunda etapa foi utilizado para corrigir o peso da primeira etapa, obtendo-se assim, o peso final associado a cada indivíduo.

Complementa, (LEMESHOW et al., 1998) que a diferença de variância entre as estratégias pode ser devido as ponderações ou pelo procedimento de inferência ou ainda pelo coeficiente de correlação intraconglomerado, sendo que geralmente a variância é maior no caso da amostragem complexa (LEMESHOW et al., 1998).

Dois exemplos são apresentados por (WANG et al., 1997), considerando amostragem estratificada em dois estágios. Os resultados demonstraram que os erros padrão das estimativas de média desconsiderando o delineamento complexo (estratificação, conglomerado e peso) foram menores que os correspondentes para estimativas ajustadas para o delineamento. Na estimativa de proporção também houve diferença e seus erros padrão foram menores que os correspondentes considerando a ponderação. Concluem que desconsiderar ponderação, quando a probabilidade não for a mesma para todos elementos da amostra, na estimação de média e proporção resulta em estimativas viesadas.

O impacto do plano por conglomerado e o efeito de ponderação foi avaliado por (SOUSA e SILVA, 2003) utilizando dados da PNDS96 (Pesquisa Nacional sobre Demografia e Saúde) através do Epi Info 6.04b (CSAMPLE). No banco de dados, variáveis definem conglomerados, estratos e peso global, este último obtido pelo produto do peso devido ao plano de amostragem e o peso devido à ausência de resposta e, após padronizado, com o objetivo de obter o total ponderado da amostra igual ao total não ponderado. A padronização foi realizada pela multiplicação de cada peso não padronizado por k, sendo que k corresponde ao tamanho da amostra dividido pela soma dos produtos de cada peso não

padronizado pelo tamanho da amostra no estrato, conglomerado UPA e unidade secundária de análise (USA) correspondente.

Neste estudo, quatro estratégias de análise foram consideradas: (1) conglomerado, sem ponderação, (2) ponderação devido ao plano de amostragem, (3) ponderação devido a ausência de resposta e (4) ponderação global. Para todas as estratégias foram obtidas estimativas de prevalência, erros-padrão, intervalo de confiança, efeito do desenho e vícios das estimativas. Concluíram, para esse estudo, que a ponderação não aumentou a precisão das estimativas, além da já incluída pelo plano por conglomerado. Também, sugerem que o cálculo dos efeitos do desenho e sua publicação devem tornar-se prática usual nas pesquisas.

As autoras descrevem que quando se utiliza a amostragem com probabilidade proporcional ao tamanho estimado, o tamanho final da amostra não é fixo, constituindo uma variável aleatória, neste caso, pode-se utilizar o estimador razão. Este estimador é viciado, porém se o coeficiente de variação do tamanho final da amostra for pequeno (CV<20%), considera-se o vício desprezível. (Kish, em (SOUSA e SILVA, 2003). Vícios superiores a 0,20 têm um impacto direto na inferência que se faz utilizando os intervalos de confiança, isto é, corresponde a se ter intervalos que na realidade possuem menos de 95% de confiança.

(GUILLÉN et al., 2000) ilustram como incorporar o desenho amostral complexo na análise de dados para obter estimativas corretas de média, proporção, erro-padrão e regressão logística, utilizando dados reais provenientes de um delineamento estratificado em dois estágios no STATA. Concluem que ignorar o delineamento amostral resulta em estimativas viesadas dos parâmetros. Sendo que a análise ponderada produz estimativas pontuais não viesadas, porém os erros-padrão são muito menores.

As particularidades de métodos de estimação de parâmetros simples como a média, o total e o percentual e seus respectivos erros padrão e também modelos de regressão logística, para dados provenientes de amostragem complexa, considerando AAS, incorporando pesos e incorporando o delineamento complexo foram avaliadas por (PÉREZ et al., 2004). Obtiveram prevalências estimadas similares, para as três estratégias de análise, podendo ser explicado pela pequena variabilidade das ponderações amostrais, porém a precisão diferiu, sendo menores para AAS e bem menores na análise ponderada, comparando esses dois com o delineamento complexo.

Dados do NHIS (National Center Health Interview Survey 1994) foram analisador por (RODGERS-FARMER e DAVIS, 2001) para verificar estimativas pontuais viesadas, erros-padrão subestimados quando se utiliza software tradicional para analisar dados sob amostragem complexa. Ajustaram modelo de regressão múltipla em cinco estratégias: (1) ignorando peso amostral e delineamento amostral complexo, (2) considerando pesos amostrais e ignorando delineamento complexo, (3) considerando pesos normalizados e ignorando delineamento complexo (pesos normalizados são transformações lineares nos pesos originais tal que o peso relativo das observações é mantido), (4) usando peso amostral e ignorando delineamento complexo (assumindo um único estrato com reposição) e por último, (5) considerando peso amostral e delineamento complexo (com reposição). Os três primeiros casos foram executados no SPSS e os dois últimos no SUDAAN.

Comparando as duas primeiras estratégias, percebe-se diferença entre os coeficientes não padronizados e erros padrão, segundo (RODGERS-FARMER e DAVIS, 2001) os erros padrão obtidos foram menores considerando pesos, pois os softwares tradicionais subestimam as variâncias das estatísticas ponderadas. Na terceira análise, percebe-se uma suave redução no erro da variância estimada. A quarta análise apresenta erros-padrão alterados, nível de significância alterado e coeficientes não padronizados iguais, comparando com a estratégia 3, com efeito de delineamento para todos os coeficientes maiores que 1, mostrando que os erros-padrão foram subestimados quando se considerou AAS. Ainda, os autores (RODGERS-FARMER e DAVIS, 2001) afirmam que desconsiderar o delineamento quando o efeito do delineamento é maior que 2 resulta em uma significativa subestimação do erro-padrão. Na estratégia 5, comparado a estratégia 4, os coeficientes não alteram, mas os erros padrão são alterados e conseqüentemente o nível de significância.

Estimadores ponderados tem a desvantagem de maior variabilidade que estimadores não-ponderados ((KORN e GRAUBARD, 1995). No estudo, os autores analisaram resultados de regressão linear simples, diferença entre proporção, regressão logística e diferença entre médias obtidos para os dados da *National Maternal and Infant Health Survey – 1998* no software SUDAAN. Os resultados obtidos mostraram estimativas diferentes para análise ponderada e não ponderada.

(KNEIPP e YARANDI, 2002) realizaram estudo sobre questões inerentes a delineamento amostral de grandes pesquisas nacionais, explicando a estimação da variância, quando há ponderação. Também, compararam análise com ponderação apenas para amostra, realizada no SPSS e a ponderação para amostra e variância, realizada no STATA. Analisando dados do *Medical Expenditure Panel Survey,* obtiveram intervalo de confiança mais largo incorporando ponderação na análise realizada no STATA. Ao contrário da análise realizada no SPSS, neste caso o intervalo de confiança é mais estreito, devido a redução da variabilidade. Também, observaram que o teste *t* de *Student* para comparar médias entre duas amostras independentes e a regressão linear foram menos afetados que o teste de qui-quadrado quando a ponderação não é considerada.

Os resultados do estudo de (HEERINGA e LIU, 1997) realizado com quatro bases de dados referente a saúde mental mostrou que o efeito do delineamento é maior que 1 para quase todas as estimativas de prevalência e que o efeito do delineamento varia entre os estudos. Também, usaram modelos de regressão logística para verificar o efeito do delineamento com três diferentes estratégias: não fazendo ajustamento nos dados, ajustando os dados com ponderações e ajustando os dados com ponderações e para delineamento (estratificação e conglomerado). Os resultados mostraram que ajustando os dados com pesos, as estimativas pontuais se alteram, sendo iguais considerando ajuste somente com pesos e com pesos junto com delineamento. O erro-padrão é maior no caso de ajustamento para pesos e delineamento. A significância dos coeficientes do modelo pode ser alterada com a inclusão do efeito do delineamento amostral no cálculo do testes estatísticos, como é mostrada para os dados do estudo.

Dados da *Behavioral Risk Factor Surveillance System* do CDC (*Centers for Disease Prevention and Control*) foram analisados por (BROGAN, 2003), desconsiderando fator de correção para população finita na estimativa de variância. Análise foi realizada no SUDAAN, considerando o delineamento complexo e no SAS considerando quatro diferentes aproximações para o peso: (1) análise não ponderada, isto é, apontando 1 para todos os valores da variável peso; (2) análise ponderada; (3) análise ponderada com peso normalizado, multiplicando cada peso pela amostra e dividindo pelo tamanho da população; e, (4) análise ponderada com peso normalizado pelo estrato.

Os resultados deste estudo mostram que a prevalência é superestimada em 10% utilizando o SAS sem ponderação comparado ao SUDAAN, sendo também superestimada, mas em menor grandeza nos três tipos de ponderação no SAS. O erro-padrão é superestimado utilizando o SUDAAN em 35% comparando com a análise não ponderada do SAS e quase a mesma diferença para as 3 análises ponderadas do SAS. Também, que a prevalência e seu erro-padrão estimado foram iguais para o peso final e peso normalizado. E, não houve diferença do peso normalizado na análise por estrato, porém houve diferença na estimativa geral (todos estratos conjuntamente).

Ainda, os resultados do teste qui-quadrado mostram que a análise não ponderada do SAS resulta em valor de qui-quadrado maior, gerando valor p mais significativo, comparada a análise do SUDAAN, considerando o geral. Já, na análise separada por estrato não houve um padrão de variação, sendo alguns valores de p maiores e outros menores. Utilizando a análise ponderada do SAS (peso bruto), obteve-se valores muito grandes da estatística de qui-quadrado, pois o tamanho da amostra é considerado igual a soma de todos os pesos (isto é, igual ao tamanho da população). Considerando o peso normalizado, em alguns estratos é maior e em outros menores. Para o peso normalizado por estrato no SAS, os valores de qui-quadrado são duas vezes maiores em relação aos encontrados no SUDAAN.

Nem sempre estão disponíveis as variáveis do delineamento no banco de dados, impossibilitando o ajuste para delineamento, sendo que (VASCONCELLOS e PROTELA, 2001) encontrou como alternativa a seleção de uma subamostra autoponderada.

EFEITO DO PLANO AMOSTRAL (EPA)

Um dos métodos para avaliar o impacto da incorporação do delineamento amostral sobre a precisão das estimativas foi desenvolvido por Kish, denominado efeito do plano amostral (EPA), definido pela razão entre a variância estimada incorporando o plano amostral efetivamente utilizado e a variância estimada

supondo uma AAS (LEITE e SILVA, 2002) (HEERINGA e LIU, 1997). Esta medida é utilizada na fase de planejamento da pesquisa, pois valores elevados do EPA destacam a importância da consideração do plano amostral efetivamente utilizado ao estimar as variâncias associadas às estimativas dos parâmetros. EPA menor que 1, indica variância considerando AAS superestimada, EPA igual a 1 indica que não há diferença entre as estimativas de variância e EPA maior que 1 indica variância considerando AAS subestimada (PESSOA e SILVA, 1998).

O EPA ampliado (*meff-misspecification effect*) é mais indicado para medidas analíticas, avaliando a tendência de um estimador usual (consistente), calculado sob hipótese de IID (independente identicamente distribuídos), subestimar ou superestimar a variância verdadeira do estimado pontual (LEITE e SILVA, 2002).

Quanto maior o valor do EPA e do EPA ampliado, menor será a probabilidade real de cobertura do intervalo de confiança para o parâmetro de interesse, caso o intervalo seja calculado sem considerar o plano amostral da pesquisa (LEITE e SILVA, 2002).

O EPA é um importante indicador para o erro de amostragem, permitindo avaliar subestimativas ou até superestimativas dos erros padrão, utilizando-se as diferentes características do delineamento amostral e diferentes métodos de estimação (SOUSA e SILVA, 2003). Os resultados do estudo de (SOUSA e SILVA, 2003) demonstraram que os conglomerados influenciaram a precisão das estimativas, para duas das seis variáveis estudadas, com EPA superiores a 1,5, indicando a importância de considerar os conglomerados na análise. Apontam a possível existência de heterogeneidade intraconglomerados para outras variáveis, em que EPA foram inferiores a 1.

Muthen e Satorra *apud* (RODGERS-FARMER e DAVIS, 2001), colocam que é essencial considerar as características do delineamento complexo, se o efeito do delineamento for maior ou igual a 2. Ainda, quando o efeito do delineamento for maior que 1 e menor que 2, a quantidade que excede 1 é uma medida de ineficiência da amostragem complexa (LaVange et al, em (RODGERS-FARMER e DAVIS, 2001).

(BARATA et al., 2005) analisaram a representatividade da amostra e a precisão das estimativas obtidas com o uso da metodologia por conglomerados (30 por 7) proposta pela Organização Mundial da Saúde, utilizando dados do inquérito realizados em duas cidades de SP, no ano de 2000. Consideraram satisfatória a

precisão quando o EPA foi inferior a 2 e a amplitude dos intervalos de confiança foi inferior a 10%, sendo estes valores utilizados na determinação do tamanho da amostra.

Um efeito do desenho próximo a 1 significa que, para fins práticos, o grau de homogeneidade das medidas nos conglomerados pode ser desprezado e que a estimativa da proporção é equivalente à que seria obtida em uma AAS (CORDEIRO, 2001).

O efeito do delineamento pode ser usado para determinar o tamanho efetivo da amostra, dividindo o tamanho amostral nominal pelo EPA. O tamanho efetivo da amostra fornece o número de elementos que produziriam com equivalente precisão considerando amostra IID (independente, identicamente distribuído) (CARLSON, 2003).

Dessa forma, o efeito do delineamento é importante para corrigir procedimentos padrão de análise estatística, por exemplo, no estágio de delineamento é incorporado no cálculo do tamanho da amostra (NEUHAUS e SEGAL, 1993).

SOFTWARES PARA ANÁLISE DE DADOS PROVINDOS DE AMOSTRAGEM COMPLEXA

Os softwares padrão estão incorporando a análise de dados provindos de amostragem complexa, como os seguintes:

- STATA (*Stata Corporation*, www.stata.com)
- SAS (*SAS Institute Inc*., www.sas.com)
- SPSS (www.spss.com)
- EPI INFO (OMS www.cdc.gov/epiinfo)
- R (www.r-project.org)

O STATA e o SAS estão incorporando rotinas para tratamento de dados de amostragem complexa há algum tempo, porém o SPSS recentemente. O EPI INFO e o R têm a vantagem de serem livres, sendo que o R tem a vantagem adicional de ter o código aberto, permitindo verificação e implementação de rotinas.

Também, software específico para análise de dados provindos de amostragem complexa está disponível, como:

- SUDAAN (*Research Triangle Institute*, www.rti.org)

Este, como o STATA, SAS e SPSS precisam ser adquiridos.

Pacotes estatísticos padrão geralmente não consideram quatro características comuns em pesquisas com delineamento transversal: probabilidade desigual de seleção dos indivíduos, conglomerados, estratificação e ajustamento para não resposta ou outro fator (BROGAN, 2003).

Estudos com dados obtidos com amostragem complexa não podem ser analisados com softwares tradicionais, pois produzem erros-padrão subestimados, intervalos de confiança não apropriados e testes de significância enganosos. A especificação do delineamento complexo na análise de dados é fortemente recomendada (RODGERS-FARMER e DAVIS, 2001).

Ainda, pacotes estatísticos padrão permitem o uso de pesos na análise, permitindo obter estimativas pontuais corretas, porém, não para as estimativas de variância (BROGAN, 2003) (WANG et al., 1997).

Em alguns softwares, como o SUDAAN, é necessário informar se a amostragem foi realizada com ou sem reposição, neste último caso, deve-se usar o FPC (UCLA, 2005).

STATA, SUDAAN e WESVarPC foram comparados por (COHEN, 1997), possibilitando ao usuário considerar as limitações e atrações para seu caso particular de amostragem complexa. Análise estatística de dados provindos de amostragem complexa considerando-os como provindos de uma AAS geralmente subestima a variância, resultando em intervalo de confiança mais estreito e teste de hipótese menos conservatório, isto é, maior probabilidade de cometer o erro tipo I.

A facilidade de analisar dados considerando a amostragem complexa, utilizando o STATA foi demonstrada em (LEMESHOW e COOK, 1999) e (GUILLÉN

et al., 2000) também concluem pela facilidade de incorporar o plano amostral nas estimativas obtidas pelos programas estatísticos, sendo que isto deve encorajar os pesquisadores a adotar essa estratégia de análise.

A diferença entre análise executada em um software tradicional e um software que permite incorporar o delineamento amostral utilizando um exemplo no SUDAAN foi demonstrada por (WANG, 2001). Ainda, demonstra que os procedimentos disponíveis nos softwares para análise de amostra complexa necessitam dos parâmetros que indicam estágios, estratos, conglomerados, probabilidade de seleção igual ou desigual, probabilidade proporcional ao tamanho e amostragem com ou sem reposição.

Softwares para analisar dados de levantamentos amostrais complexos foram comparados por (SOUSA e SILVA, 2000), avaliando facilidade de aplicação, eficiência computacional e exatidão dos resultados. Utilizaram dados da PNDS (Pesquisa Nacional sobre Demografia e Saúde, 1996), para analisar média e proporção utilizando os softwares CSAMPLE/Epi Info v. 6.04, Stata v. 5 e WesVarPC v 2.12 (incorporado ao SPSS). Concluíram que o Epi Info é mais limitado na disponibilidade de técnicas de análise, porém seu uso é simples e gratuito. O Stata e WesVarPC têm diversidade de técnicas de análise, porém tem custo de aquisição. Também, que a escolha pelo software depende das necessidades e volume de análise.

No Brasil, o primeiro estudo realizado no R foi de (FIGUEIREDO, 2004), que tem a vantagem de ter o código fonte aberto. Analisou dados obtidos de pesquisas amostrais complexas, considerando o delineamento complexo e considerando AAS, utilizando o SUDAAN e a biblioteca ADAC da linguagem R para modelos lineares de regressão normal e logística. Foram considerados aspectos teóricos e aplicações em dados reais.

Oito softwares com capacidade de analisar dados de amostragem complexa foram analisados por (BROGAN, 2005) de acordo com custo, métodos de estimação de variância, opções de análise, interface e vantagens/desvantagens. Destes, 4 são livres. Análises foram realizadas com dados de Burundi para 5 softwares (STATA, SAS, SUDAAN, WesVar e Epi-Info), obtendo-se resultados equivalentes para todos quando linearização por série de Taylor e replicação repetida balanceada foi usada.

As estimativas e erros padrão foram são não viesados na análise de um conjunto de dados da saúde no SUDAAN e STATA, quando incorpora-se características do delineamento amostral (CHANTALA e TABOR, 1999). Os autores discutem que os resultados são influenciados de diferentes modos: estimativas pontuais (média, parâmetros de regressão, proporções,...) são afetados somente pelos pesos e variâncias estimadas são afetadas pelo conglomerado, estratificação, peso e tipo de delineamento. Também, comentam a análise de subpopulações e que neste caso o uso somente dos dados da subpopulação resulta em estimativa pontual correta, porém o erro padrão poderá não ser correto, já que a estrutura do delineamento não está disponível. Os softwares para análise de delineamento complexo, disponibilizam análise de subpopulações.

Por fim, (PESSOA e SILVA, 1998) discutem que certos cuidados precisam ser tomados para utilização correta dos dados de pesquisas amostrais como as que o IBGE realiza. Estes dados provem de população finita envolvendo probabilidades distintas de seleção, estratificação e conglomeração das unidades, ajustes para compensar não-resposta e outros ajustes. Os autores afirmam que pacotes tradicionais de análise ignoram estes aspectos, podendo resultar em estimativas incorretas de parâmetros e variâncias das estimativas. A publicação aborda as metodologias de tratamento de dados provenientes de amostragem complexa e também um capítulo sobre análise desagregada (modelagem multinível), sendo esta última abordada na seqüência.

ANÁLISE MULTINÍVEL

As metodologias adequadas para a análise de dados amostrais complexos podem ser agrupadas em duas abordagens, como comentado por (VIEIRA, 2001): abordagem agregada, baseando-se na incorporação de pesos e efeitos do plano amostral no ajuste dos modelos estatísticos e abordagem desagregada, incorporando efeitos devidos à amostragem complexa, utilizando modelos lineares hierárquicos ou multinível. No estudo, o autor abordou aspectos teóricos das técnicas de estimação pontual de parâmetros de modelos de regressão e respectivas variâncias. Também, discute o pacote SUDAAN na abordagem de efeito do plano amostral, intervalos de confiança e testes de hipóteses. A aplicação foi realizada com dados do SAEB 1999 (Sistema Nacional de Avaliação da Educação Básica).

Segundo (THOMAS e HECK, 2001), há duas aproximações que podem ser usadas na análise de dados obtidos de amostragem complexa: *designed-based* e *model-based*. Na análise *designed-based*, isto é, análise por nível único, é realizada ajustamento para probabilidade desigual de seleção e observações não independentes (*cluster*) e na análise *model-based*, isto é, regressão multinível, os efeitos do delineamento amostral são incorporados no modelo analítico. Sendo que, *model-based* é eficaz para tratamento de conglomerados, porém ainda necessita ajuste para probabilidade desigual de seleção.

No estudo de (TRAVASSOS et al., 2002), a análise foi realizada em dois níveis: individual e familiar. Porém, nenhuma característica das famílias foi capaz de explicar a variabilidade observada na variável desfecho entre famílias, desta forma foi possível substituir o modelo hierárquico por modelo de regressão logística tradicional, ajustado para o EPA.

4 METODOLOGIA

Para realização do estudo serão utilizados dois bancos de dados, detalhados a seguir.

4.1 Banco de dados

Banco de dados I: Campanha de Diabetes – Brasil

Campanha Nacional de Detecção de Diabetes Mellitus, é um programa de rastreamento nacional direcionado a usuários do Sistema Único de Saúde com 40 ou mais anos de idade (Nucci, 2003). A pesquisa por amostragem probabilística da população, com busca ativa dos pacientes fez parte desta pesquisa.

- Delineamento amostral: estratificado em dois estágios
- Variáveis de amostragem: estratos, conglomerados
- Variáveis de pesquisa: sexo, idade, peso, altura, IMC, glicemia e diabete (sim, não)

- Técnicas para análise univariada e multivariada: proporção e respectivo intervalo de 95% de confiança; regressão logística de diabete em função de sexo, idade e peso. Comparar estimativas considerando amostra aleatória simples e considerando amostra complexa. Ainda verificar o efeito do conglomerado e do estrato.

Banco de dados II: Avaliação de estratégia de aleitamento materno em escolas de ensino fundamental

- Delineamento amostral: estratificado em dois estágios, aleatorização em *cluster*
- Variáveis de amostragem: estratos, conglomerados, peso
- Variáveis de pesquisa: sexo, idade, tipo de escola, procedência da escola, escore no pré-teste, escore no re-teste
- Técnicas para análise univariada e multivariada: proporção, média e respectivos intervalos de 95% de confiança; teste para comparação entre duas médias; teste para comparação entre duas proporções; regressão múltipla; regressão logística. Comparar modelagem considerando amostra aleatória simples com modelagem considerando amostra complexa e modelagem multinível.

Na análise dos dois bancos de dados serão utilizados os software estatísticos STATA, SAS, SUDAAN e R, conforme a disponibilidade das técnicas de análise propostas.

Demais detalhes sobre metodologia seguem no esboço dos artigos.

4.2 Questões éticas

Os três bancos de dados tiveram seus projetos aprovados por Comitê de Ética em Pesquisa, como segue o detalhamento abaixo. Ainda, a confiabilidade dos dados está garantida, pois nos bancos não é possível identificar os indivíduos do estudo. A declaração de confiabilidade consta em anexo.

Banco de dados I

O projeto foi aprovado pelo Comitê de Ética em Pesquisa da UFRGS e termo de consentimento informado foi obtido de cada participante da pesquisa.

Banco de dados II

O projeto (processo número 01-429) foi aprovado pelo Grupo de Pesquisa e Pós-graduação do Hospital de Clínicas de Porto Alegre e pela Comissão Científica e Comissão de Pesquisa e Ética em Saúde.

O projeto foi aprovado também pela 36ª Coordenadoria de Educação do município de Ijuí, Secretaria Municipal de Educação de Ijuí e pela direção das escolas privadas deste município. As escolas selecionadas foram previamente visitadas e informadas sobre a pesquisa, ficando livre a participar ou não da pesquisa.

O Termo de Consentimento Livre e Esclarecido foi enviado aos pais dos escolares antes do início da pesquisa. A declaração de aceite para realização do estudo encontra-se arquivada. A forma de aplicação do termo encontra-se no Manual do Entrevistador.

5 ASPECTOS OPERACIONAIS

5.1 Recursos necessários

- *Softwares* com documentação completa: SUDAAN v. última, STATA v. última, SAS v. última, R v. última.
- *Hardware*: um microcomputador, com processador 2 GHz, 512 MB, 40 GB.
- Bibliografia com metodologia de amostragem complexa e modelagem multinível.

5.2 Cronograma

Atividade	2º Semestre de 2006 a 1º Semestre de 2007			
	Set-Out	Nov-Dez	Jan-Fev	Mar-Abr
Estudo das metodologias de *complex survey* e multinível				
Estudo das rotinas nos softwares e implementação rotinas				
Análise dos dados através dos softwares				
Interpretação dos resultados				
Apresentação boneco de tese (março)				
Apresentação tese (abril)				

6 ESBOÇO DOS ARTIGOS

6.1 Artigo 1

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
FACULDALDE DE MEDICINA
PROGRAMA DE PÓS-GRADUAÇÃO EM EPIDEMIOLOGIA

IMPACTO DO PLANO AMOSTRAL COMPLEXO NAS ESTIMATIVAS DE COEFICIENTES DE REGRESSÃO LOGÍSTICA EM UM ESTUDO EPIDEMIOLÓGICO

IARA DENISE ENDRUWEIT BATTISTI
JANDYRA MARIA GUIMARÃES FACHEL
ELSA CRISTINA DE MUNDSTOCK

PORTO ALEGRE, AGOSTO DE 2006

INTRODUÇÃO

Muitos estudos epidemiológicos utilizam amostragem complexa para coleta dos dados, como estratificação, conglomerados e probabilidade desiguais de seleção. A amostragem complexa tem a vantagem de diminuir os custos[1] e a não necessidade de uma listagem de todas os elementos que compõe a população[2]. No entanto, a análise dos dados deve considerar o plano de amostragem.

Quando o plano de amostragem não é considerado na análise de dados provenientes de amostragem complexa, podem-se obter estimativas incorretas. As estimativas pontuais de parâmetros da população são influenciadas por pesos distintos das observações e as estimativas de variância são influenciadas pela conglomeração, estratificação e pesos[3].

As metodologias tradicionais de inferência assumem que os dados foram obtidos a partir de uma amostra aleatória simples. A maioria dos softwares estatísticos, até há pouco tempo, somente abordavam essas metodologias. Em suas novas versões incorporaram o tratamento de dados provenientes de amostras complexas, como é o caso do SPSS e STATA. Também, existem softwares específicos para o tratamento de dados de amostras complexas como SUDAAN (SUrvey DAta ANalysis).

Considerando a grande demanda de pesquisas com planos amostrais complexos e a existências de softwares estatísticos que incorporam essas metodologias e, ainda que muitos pesquisadores de países em desenvolvimento ainda utilizam os estimadores de amostra aleatória simples, usando programas estatísticos que não possuem módulos específicos para amostragem com delineamentos complexos[4] realizou-se este estudo que teve por objetivo avaliar o impacto de um plano amostral complexo nas estimativas de um estudo epidemiológico.

MÉTODOS

Para avaliar o impacto do plano amostral complexo foram utilizados os dados da busca ativa domiciliar dos participantes na Campanha Nacional de Detecção de Diabetes Mellitus – CNDDM, Brasil, 2001, que teve, entre outros objetivos, a confirmação diagnóstica dos pacientes com rastreamento positivo.

A população alvo para o planejamento da campanha foi definida como os 31 milhões de brasileiros acima de 40 anos (41,4 milhões com mais de 40 anos x 0,75 usuários do SUS, aproximadamente 31 milhões)[5]. A CNDDM foi o primeiro rastreamento para detecção de casos suspeitos de Diabetes Mellitus no Brasil, com 22.069.905 exames de glicemia capilar realizados, sendo 3,5 milhões (16%) deles considerados suspeitos pelos critérios definidos[6].

Especificamente na busca ativa, foi realizada uma amostragem estratificada em dois estágios para a seleção de amostra probabilística de 50 municípios brasileiros sorteados por região, de acordo com a participação da população na campanha (exames realizados). Para o cálculo de seleção da amostra foram utilizados os dados da CNDDM, obtidos em 2001, referente aos 93% (5.186) municípios brasileiros, definindo-se a população base para a pesquisa, aproximadamente 22 milhões de brasileiros[6].

A amostra foi composta por 3 municípios da região norte, 14 do nordeste, 3 do centro-oeste, 21 do sudeste e 9 da região sul, proporcionalmente ao número de exames de cada região.

No segundo estágio, foi selecionada uma unidade básica de saúde (UBS) de cada município, proporcional ao número de participantes da campanha. Estimou-se que seriam necessárias 2.000 fichas por município para a busca ativa de 100 pacientes. Quando a UBS sorteada não possuía as 2.000 fichas, foi sorteada outra UBS no município.

Os dados foram armazenados no software EPI INFO, versão 6.03, da Organização Mundial de Saúde, assim como as variáveis que caracterizam o delineamento - o conglomerado e o estrato, relacionados em cada registro. Os softwares utilizados para a obtenção das estimativas foram o R e STATA v.9, sendo o primeiro de domínio público.

Neste estudo ajustou-se um modelo de regressão logística para a variável CDGLN que expressa os vários pontos de corte levando em conta estado de jejum do paciente (codificada como dicotômica 0 – não diabetes e 1 – provável ou muito provável diabetes) em função da idade, sexo e peso corporal.

Obtiveram-se estimativa para proporção, média, erro-padrão, intervalo de confiança e o efeito do delineamento amostral – EDA, considerando-se que os dados provem de amostra aleatória simples e amostra complexa, definida por conglomerados, estratos e pesos.

Também, foram obtidas estimativas para coeficiente de regressão logística, erro-padrão, intervalo de confiança, nível de significância e EDA, considerando-se que os dados provem de amostra aleatória simples e amostra complexa.

Para as estimativas simples e estimativas de coeficientes também foram verificadas o efeito de incorporar o conglomerado e a estrato.

A estimação da variância no STATA foi possível somente considerando o método do conglomerado primário, que neste caso desconsidera os demais estágios, pois no segundo estágio somente tem-se um conglomerado para cada conglomerado da etapa anterior.

As regiões foram definidas como estratos e os municípios como conglomerados no primeiro estágio, identificando a unidade primária de amostragem. Como o objetivo foi utilizar e avaliar o impacto dos três parâmetros que constituem a amostra complexa nos softwares, definiu-se peso para todas as unidades amostrais, representando um fator de expansão que caracteriza o número de indivíduos que cada elemento da amostra representa na população. Sabendo-se que a seleção das unidades amostrais considerou probabilidade proporcional ao tamanho – PPT, nos dois estágios, a probabilidade de o indivíduo pertencer à amostra (fração de amostragem) é definida por:

$$f = f_1 . f_2 . f_3$$

em que:

f_1: probabilidade de o k-ésimo município ser sorteado dentro da região.

f_2: probabilidade de a j-ésima unidade básica de saúde ser sorteada dentro do município.

f_3: probabilidade do i-ésimo indivíduo ser sorteado dentro da unidade básica de saúde J.

Desta forma, o peso final foi construído para cada unidade amostral:

$$w = \frac{1}{f}$$

RESULTADOS

TABELAS

Tabela 1. Estimativas, erro padrão, intervalo de confiança para as variáveis sexo, idade, IMC, segundo o plano amostral simples e complexo e o efeito do plano amostral - EPA

Variável	Plano amostral[1]	Estimativa	Erro Padrão	IC 95%	EPA
Sexo[a]					
Masculino					
	Simples				-
	Estrato				
	Conglomerado				
	Complexo				
Feminino					
	Simples				-
	Estrato				
	Conglomerado				
	Complexo				
Idade[b]					
	Simples				-
	Estrato				
	Conglomerado				
	Complexo				
IMC[b]					
	Simples				-
	Estrato				
	Conglomerado				
	Complexo				

[1] plano amostral simples considera amostra aleatória simples e plano amostral complexo considera amostra estratificada, conglomerado e peso;
[a] proporção; [b] média.

Tabela 2. Estimativas, erro padrão, intervalo de confiança, nível de significância e efeito do delineamento amostral dos coeficientes de regressão logística da variável CDGLN em relação ao sexo, idade e peso corporal segundo o plano amostral simples e complexo

Plano amostral[1] Variável	Estimativa	Erro Padrão	IC 95%	p	EDA
Simples					
Sexo[a]					
Idade[b]					
Peso[c]					
Constante					
Complexo					
Sexo[a]					
Idade[b]					
Peso[c]					
Constante					
Conglomerado					
Sexo[a]					
Idade[b]					
Peso[c]					
Constante					
Estrato					
Sexo[a]					
Idade[b]					
Peso[c]					
Constante					

[1] plano amostral simples considera amostra aleatória simples; plano amostral complexo considera amostra estratificada, conglomerado e peso; plano amostral conglomerado considera apenas o conglomerado, excluindo estratos e pesos;

[a] homem como categoria de referência; [b] 40 a 49 anos de idade como categoria de referência; [c] peso corporal para incremento de um kg

Tabela 3. *Odds ratio* e intervalo de confiança segundo o plano amostral simples e complexo

Variável	Plano amostral[1]			
	Simples		Complexo	
	OR	IC 95%	OR	IC 95%
Sexo[a]				
Idade[b]				
Peso[c]				

[1] plano amostral simples considera amostra aleatória simples e plano amostral complexo considera amostra estratificada, conglomerado e peso;

[a] homem como categoria de referência; [b] 40 a 49 anos de idade como categoria de referência; [c] peso corporal para incremento de um kg

REFERÊNCIAS

1. Cordeiro R. Efeito do desenho em amostragem de conglomerados para estimar a distribuição de ocupações entre trabalhadores. *Ver Saúde Pública* 2001; 35(1): 10-5.
2. Silva NN. Amostragem Probabilística: um curso introdutório. São Paulo: EDUSP; 1998.
3. Pessoa DGP, Silva PLN. Análise de Dados Amostrais Complexos. In: *13º Simpósio Nacional de Probabilidade e Estatística*; 1998 jul 27-31; Caxambu (MG). Caxambu: ABE; 1998.
4. Sousa MH, Silva NN. Estimativas obtidas de um levantamento complexo. *Ver Saúde Pública* 2003; 37(5): 662-70.
5. Brasil. Ministério da Saúde. Organização Pan Americana da Saúde. Avaliação do Plano de Reorganização da Atenção à Hipertensão Arterial e ao Diabetes Mellitus no Brasil. Brasília: Ministério da Saúde, 2004.
6. Nucci LB. A Campanha Nacional de Detecção do Diabetes Mellitus: cobertura e resultados glicêmicos. Tese de doutorado(Epidemiologia). Porto Alegre: UFRGS, 2003.

6.2 Artigo 2

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
FACULDALDE DE MEDICINA
PROGRAMA DE PÓS-GRADUAÇÃO EM EPIDEMIOLOGIA

ESTIMATIVAS OBTIDAS POR ABORDAGEM AGREGADA E ABORDAGEM DESAGREGADA: APLICAÇÃO EM UM ESTUDO COM ESCOLARES DE ENSINO FUNDAMENTAL NO MUNICÍPIO DE IJUÍ/RS

IARA DENISE ENDRUWEIT BATTISTI
JANDYRA MARIA GUIMARÃES FACHEL
JOÃO RIBOLDI
ELSA CRISTINA DE MUNDSTOCK

PORTO ALEGRE, AGOSTO DE 2006

INTRODUÇÃO

Muitos estudos epidemiológicos utilizam amostragem probabilística para coleta de dados, definidas de quatro formas diferentes: amostragem aleatória simples, amostragem estratificada, amostragem sistemática e amostragem por conglomerado. É muito comum, em grandes inquéritos para estudo da saúde de indivíduos, a aplicação de duas ou mais destas formas de amostragem ao mesmo tempo, e ainda, em dois ou mais estágios de seleção das unidades amostrais, definindo-se como amostragem complexa.

A maioria dos softwares estatísticos contempla técnicas de análise dados provindos de amostragem aleatória simples, sendo que estas não são apropriadas para os demais tipos de amostragem probabilística. Felizmente, atualmente têm-se rotinas disponíveis para análise de dados provindos de diferentes amostragens probabilísticas em softwares estatísticos, porém ainda subutilizados. Isto se deve a pouca disponibilidade de referências aplicadas à epidemiologia, assim como a recente incorporação destas técnicas nos softwares.

Sabe-se que a análise de dados desconsiderando a amostragem complexa resulta em estimativas incorretas, porém isso é ignorado em muitas pesquisas. Desta forma, este estudo tem o propósito de avaliar duas metodologias de análise para dados provindos de delineamento complexo, referentes a um estudo com escolares da quinta série do ensino fundamental, no município de Ijuí/RS.

MÉTODOS

Delineamento do estudo

Trata-se de um ensaio clínico randomizado em *cluster* (amostra estratificada em *cluster*) tendo como desfecho o desempenho das crianças na avaliação de conhecimento, percepções e crenças sobre aleitamento materno.

População-alvo

A população-alvo do estudo é constituída de escolares de ambos os sexos, matriculados na quinta série das escolas estaduais, municipais e particulares do ensino fundamental do município de Ijuí/RS.

Amostra

Seleção

Para a seleção da amostra, foram utilizados subgrupos populacionais - conglomerados - representados primeiramente pelas escolas e, posteriormente, pelas turmas de quinta série existentes nas escolas sorteadas no primeiro estágio. O esquema amostral está demonstrado na Figura 1.

Levou-se em consideração a proporção de alunos que freqüentavam as escolas estaduais, municipais e particulares, bem como a sua distribuição geográfica na zona urbana e zona rural.

A seleção das escolas foi realizada através de amostragem aleatória simples e a seleção das turmas dentro das escolas também foi por amostragem aleatória simples, sendo que quando a escola tinha apenas uma turma, esta teve probabilidade certa de ser incluída na amostra.

Foram considerados elegíveis para o estudo todos os alunos que freqüentavam a quinta série do ensino fundamental de Ijuí. Foram excluídos os alunos que freqüentavam a escola no período noturno, os de escolas em que o número de alunos matriculados na quinta série era inferior a dez e os de turmas com alunos de séries diferentes.

ESQUEMA AMOSTRAL

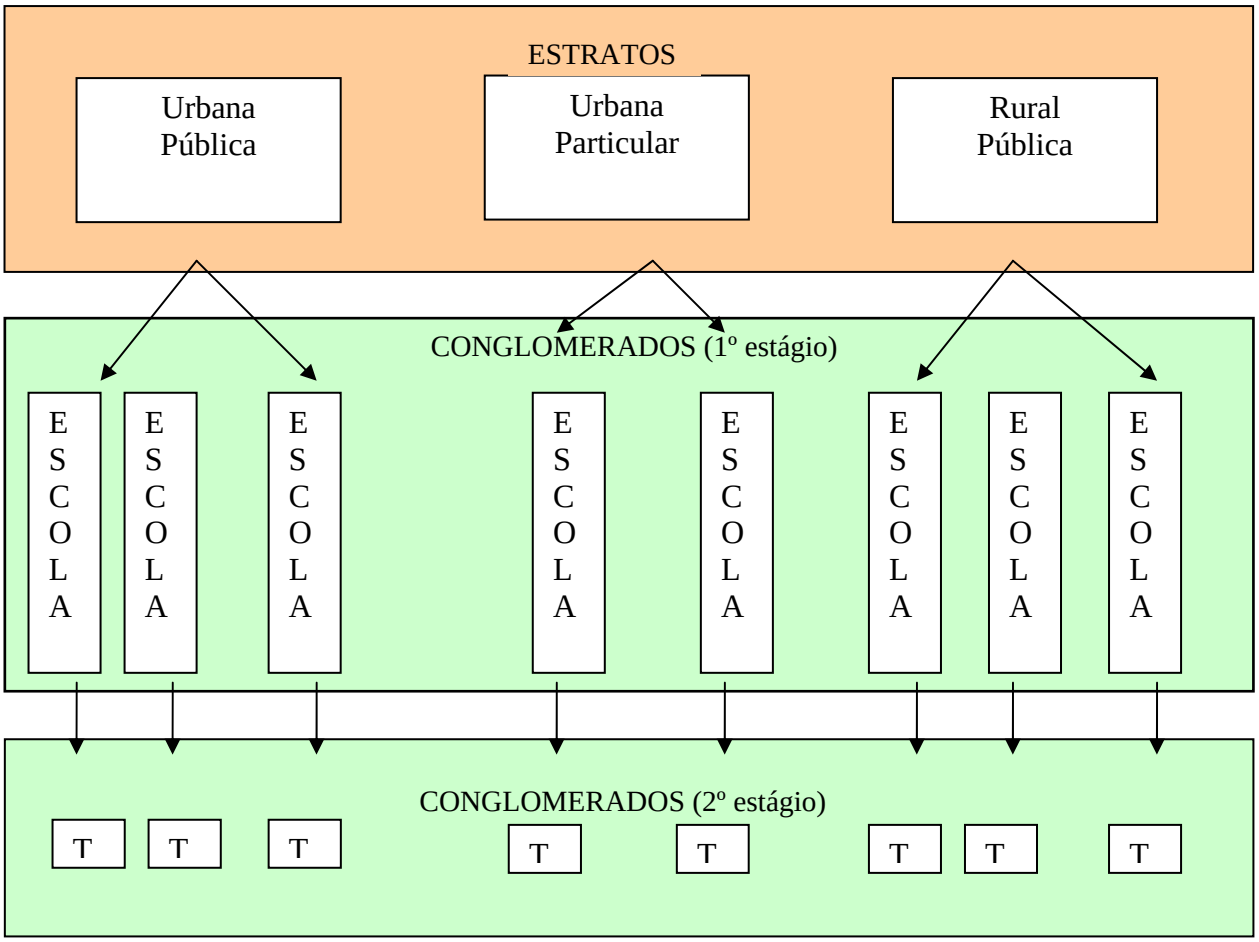


Figura 1- Seleção da amostra para o estudo.

Nota: T = Turmas de quinta-série

Alocação dos grupos

Primeiro estágio (escolas)

Das 47 escolas existentes no município com turmas de quinta série (32 urbanas públicas, 11 rurais públicas e 4 urbanas particulares), 30 foram sorteadas: 19 públicas urbanas; 9 públicas rurais e 2 urbanas particulares.

As escolas foram sorteadas para o grupo controle e experimental, considerando a proporcionalidade quanto ao tipo de escola (urbana publica, rural pública e urbana particular). O grupo controle não foi exposto a nenhum tipo de intervenção e o grupo experimental sofreu a intervenção detalhada mais adiante.

Segundo estágio (turmas)

Após as turmas foram sorteadas, pôde-se calcular o número de alunos selecionados para o estudo: foram 656 escolares, dos quais 564 participaram do estudo, 253 no grupo-controle e 311 no experimental. Quando comparados os 564 escolares que consituíram a amostra estudada com os 92 escolres não incluídos, observa-se pelo teste do qui-quadrado (p=0,901) que não houve diferença quanto à idade, sexo, tipo de escola, procedência do aluno e escolaridade dos pais.

Coleta de dados

Num primeiro momento, aplicou-se o questionário em sala de aula, em todos os alunos, tanto do grupo controle como do grupo experimental. Logo após o início da intervenção foi aplicado novo questionário (pós-teste) somente no grupo experimental. Após três meses do início da intervenção, o mesmo questionário aplicado logo após a intervenção foi novamente aplicado em todas as turmas, tanto do grupo controle como no experimental.

Variáveis

Variável desfecho: desempenho das crianças na avaliação de conhecimento, percepções e crenças sobre aleitamento materno. Para o presente estudo o desempenho foi avaliado através de um escore.

O questionário utilizado na pesquisa continha perguntas 25 questões objetivas e 5 descritivas. O escore das questões objetivas variou de zero a 25 pontos

Outras variáveis:

Tipo de escola (pública e particular), procedência (urbana e rural), turno (manhã e tarde), idade do escolar (em anos completos no dia da aplicação do questionário), sexo do escolar (feminino e masculino), escolaridade da mãe (até 8 anos, de 9 a 12 anos, mais que 12 anos) e escolaridade do pai (até 8 anos, de 9 a 12 anos, mais que 12 anos). Ainda, vivências prévias em aleitamento materno: se o escolar fora amamentado; se já havia visto alguém amamentando; se nas brincadeiras de infância as bonecas mamavam no peito.; se acha feio dar o peito na frente de outras pessoas; caso tivesse filho agora como alimentaria.

Análise estatística

Metodologia agregada, considera o plano amostral e metodologia desagregada considera a estrutura da população (PESSOA e SILVA, 1998).

Na análise agregada, isto é, análise por nível único, é realizada ajustamento para probabilidade desigual de seleção e observações não independentes (*cluster*) e na análise desagregada, isto é, regressão multinível, os efeitos do delineamento amostral são incorporados no modelo analítico (THOMAS e HECK, 2001) (VIEIRA, 2001). Sendo que, a análise desagregada é eficaz para tratamento de conglomerados, porém ainda necessita ajuste para probabilidade desigual de seleção (THOMAS e HECK, 2001).

Para comparar as metodologias agregada e desagregada de análise dos dados do estudo serão realizados os seguintes passos:

- obter estimativa para proporção, erro-padrão, intervalo de 95% de confiança e efeito do plano amostral para as variáveis de caracterização dos escolares, considerando-se que os dados provem de amostra aleatória simples e amostra complexa
- analisar variáveis (experiências prévias em aleitamento materno) entre sexo dos escolares com o teste de qui-quadrado, considerando-se que os dados provem de amostra aleatória simples e amostra complexa. Também, avaliar o efeito de ignorar-se somente o conglomerado e ignorar-se somente o peso.
- modelagem: estimativas, erro-padrão, intervalo de 95% de confiança e efeito do plano amostral dos coeficientes do modelo. Neste caso, verificar se a variável desfecho - desempenho (dicotomizada pelo escore obtido no questionário) difere entre o grupo experimental e controle, considerando como variáveis independentes as características dos escolares.

Software estatístico

Serão utilizados o STATA v.9, R v.2.1.1 ou SAS v.8, conforme a disponibilidade de técnicas para as análises propostas.

Reference List

(1) Barata RB, Moraes JC, Antonio PRA, Dominguez M. Inquérito de cobertura vacinal: avaliação empírica da técnica de amostragem por conglomerados proposta pela Organização Mundial da Saúde. Rev Panam Salud Publica 2005;17(3):184-90.

(2) Lemeshow S, Cook ED. Pratical considerations in the analysis of complex sample survey data. Rev Epidém et Santé Publ 1999;47:479-87.

(3) Korn EL, Graubard BI. Epidemiologic studies utilizing surveys: accounting for the sampling design. American Journal of Public Health 1991;81(9):1166-73.

(4) Mundstock EC. Amostragem I. 2005. Instituto de Matemática/UFRGS. Cadernos de Matemática e Estatística.
Ref Type: Serial (Book,Monograph)

(5) Mundstock EC. Amostragem II. 2004. Instituto de Matemática / UFRGS. Cadernos de Matemática e Estatística.
Ref Type: Serial (Book,Monograph)

(6) UCLA. Statistical computing seminars survey data analysis in Stata. Eletronic Citation 2005 [cited 2005 Jun 27];Available from: URL: http://www.ats.ucla.edu/stat/stata/seminars/svy_stata_intro/default.htm

(7) Cohen SB. An evaluation of alternative PC-based software packages developed for the analysis of complex survey data. The American Statistician 1997;51(3):285-92.

(8) Heeringa SG, Liu J. Complex sample design effects and inference for mental health survey data. International Journal of Methods in Psychiatric Research 1997;7(1):56-65.

(9) Rodgers-Farmer A, Davis D. Analyzing complex survey data. Social Work Research 2001 Jan 1;25(3):185-92.

(10) Pérez MC, Utra IB, León AA, Roche RG, Sagué KA, Rosa MC, et al. Estimationes usadas en diseños muestrales complejos: aplicaciones en la encuesta de salud cubana del año 2001. Rev Panam Salud Publica 2004;15(3):176-84.

(11) Sousa MH, Silva NN. Estimativas obtidas de um levantamento complexo. Rev Saúde Pública 2003;37(5):662-70.

(12) Rao JNK, Wu CFJ. Resampling inference with complex survey data. Journal of the American Statistical Association 1988;83:231-41.

(13) Cochran WG. Técnicas de Amostragem. Rio de Janeiro: Fundo de Cultura; 1965.

(14) Bueno MB, Marchioni DML, Fisberg RM. Evolução nutricional de crianças atendidas em creches públicas no município de São Paulo, Brasil. Rev Panam Salud Publica 2003;14(3):165-70.

(15) Barros AJD, Bertoldi AD. Desigualdades na utilização e no acesso a srviços odontológicos: uma avaliação a nível nacional. Ciência e Saúde coletiva 2002;7(4):709-17.

(16) Hernández B, Haene J, Barquera S, Monterrubio E, Rivera J, Shamah T, et al. Factores asociados con la actividad física en mujeres mexicana en edad reproductiva. Rev Panam Salud Publica 2003;14(4):235-45.

(17) Egede LE. Lifestyle modification to improve blood pressure control in individuals with diabetes. Diabetes Care 2003;26(3):602-7.

(18) Egede LE. Association between number of physician visits and influenza vaccination coverage among diabetic adults with access to care. Diabetes Care 2003;26(9):2562-7.

(19) Egede LE. Diabetes, Major Depression, and Functional Disability Among U.S. Adults. Diabetes Care 2004;27(2):421-8.

(20) Carlson BL. Software for statistical analysis of sample survey data. Eletronic Citation 2003 [cited 2003 Sep 25];Available from: URL: http://www.fas.harvard.edu/~stats/survey-soft/blc_eob.html

(21) Cordeiro R. Efeito do desenho em amostragem de conglomerado para estimar a distribuição de ocupações entre trabalhadores. Rev Saúde Pública 2001;35(1):10-5.

(22) Horton NJ, Fitzmaurice GM. Regression analysis of multiple source and multiple informant data from complex survey samples. Statistics in Medicine 2004;23:2911-33.

(23) Thomas SL, Heck RH. Analysis of large-scale secondary data in higher education research. Research in Higher Education 2001;42(5):517-40.

(24) Travassos C, Viacava F, Pinheiro R, Brito A. Utilização dos serviços de saúde no Brasil: gênero, características familiares e condição social. Rev Panam Salud Publica 2002;11(5/6):365-73.

(25) Brogan DJ. Pitfalls of using standard statistical software packages for sample surveys data. Eletronic Citation 2003 [cited 2003 Sep 19];Available from: URL: http://www.fas.harvard.edu/~stats/survey-soft/donna_brogan.html

(26) Korn EL, Graubard BI. Analysis of large health surveys: accounting for the sampling design. Royal Statistical Society 1995;158(2):263-95.

(27) Guillén M, Juncá S, Rué M, Aragay JM. Efecto del diseño muestral en el análisis de encuestas de deseño complejo. Aplicación a la encuesta de salud de Catalunya. Gac Sanit 2000;14(5):399-402.

(28) Kneipp SM, Yarandi HN. Complex sampling designs and statistical issues in secondary analysis. Western Journal of Nursing Research 2002;24(5):552-66.

(29) Leite PGPG, Silva DBN. Análise da situação ocupacional de crianças e adolescentes nas regiões sudeste e nordeste do Brasil utilizando informações da PNAD 1999. 2002.

(30) Lemeshow S, Letenneur L, Dartigues JF, Lafont S, Orgogozo JM, Commenges D. Illustration of anlysis taking into account complex survey considerations: the association between wine consumtion and dementia in the PAQUID study. American Journal of Epidemiology 1998;148(3):298-306.

(31) Wang ST, Yu ML, Lin LY. Consequences of analysing complex survey data using inappropriate analysis and software computing packages. Public Health 1997;111:259-62.

(32) Vasconcellos MTL, Protela MC. Índice de massa corporal e sua relaçào com variáveis nutricionais e sócio-econômicas: um exemplo de uso de regressão linear para um grupo de adultros brasileiros. Cadernos de Saúde Pública 2001;17(6):1425-36.

(33) Pessoa DGC, Silva PLN. Análise de dados amostrais complexos. São Paulo: ABE-Associação Brasileira de Estatística; 1998.

(34) Neuhaus JM, Segal MR. Design effects for binary regression models fitted to dependent data. Statistics in Medicine 1993;12:1259-68.

(35) Wang MQ. Research notes: analysis of data from complex survey designs. American Journal of Health Behavior 2001;25(1):72-4.

(36) Sousa MH, Silva NN. Comparação de software para análise de dados de levantamentos complexos. Rev Saúde Pública 2000;34(6):646-53.

(37) Figueiredo CC. Análise de Regressão Incorporando o Esquema Amostral [Dissertation] Universidade de São Paulo; 2004.

(38) Brogan D. Sampling error estimation for survey. Household sample surveys in developing an transition countries.New York: United Nations Publication; 2005. p. 447-90.

(39) Chantala K, Tabor J. Strategies to perform a design-based analysis using the add health data. Eletronic Citation 1999 [cited 2003 Sep 20];1-20. Available from: URL: http://www.cpc.unc.edu/projects/addhealth/files/weight1.pdf

(40) Vieira MT. Um estudo comparativo das metodologias de modelagem de dados amostrais complexos - uma aplicação ao SAEB 99 Pontifícia Universidade Católica do Rio de Janeiro; 2001.