Universidade de São Paulo
Faculdade de Saúde Pública

VOLUME 34
NÚMERO 6
DEZEMBRO 2000
p. 646-53

Revista de Saúde Pública

Journal of Public Health

Comparação de *softwares* para análise de dados de levantamentos complexos

Comparison of software programs for data analysis of complex surveys

Maria Helena de Sousa[a*] e Nilza Nunes da Silva[b]

[a]Centro de Pesquisas das Doenças Materno-infantis de Campinas (Cemicamp). Campinas, SP, Brasil.
[b]Departamento de Epidemiologia da Faculdade de Saúde Pública da Universidade de São Paulo. São Paulo, SP, Brasil

SOUSA[a*] Maria Helena de e Nilza Nunes da Silva[b] *Comparação de softwares para análise de dados de levantamentos complexos* Rev. Saúde Pública, 34 (6): 646-53, 2000 www.fsp.usp.br/rsp

Rev Saúde Pública 2000;34(6):646-53
www.fsp.usp.br/rsp

Comparação de *softwares* para análise de dados de levantamentos complexos

Comparison of software programs for data analysis of complex surveys

Maria Helena de Sousa[a*] e Nilza Nunes da Silva[b]

[a]Centro de Pesquisas das Doenças Materno-infantis de Campinas (Cemicamp). Campinas, SP, Brasil.
[b]Departamento de Epidemiologia da Faculdade de Saúde Pública da Universidade de São Paulo. São Paulo, SP, Brasil

Descritores
"Software", utilização#. Avaliação#. Interpretação estatística de dados#. Levantamentos epidemiológicos. Amostragem. Aplicação de informática médica. Eficiência. – Levantamentos amostrais complexos.

Resumo

Objetivo
Comparar "softwares" específicos para análise de dados de levantamentos amostrais complexos, em relação às características: facilidade de aplicação, eficiência computacional e exatidão dos resultados.
Métodos
Utilizaram-se dados secundários da Pesquisa Nacional sobre Demografia e Saúde, de 1996, cuja população-alvo foram as mulheres de 15 a 49 anos de idade, pertencentes a uma subamostra probabilística selecionada em dois estágios, estratificada, com probabilidade proporcional ao tamanho no primeiro estágio. Foram selecionadas da subamostra as regiões Norte e Centro-oeste do País. Os parâmetros analisados foram: a média, para a variável idade, e a proporção, para cinco outras variáveis qualitativas, utilizando os "softwares" Epi Info, Stata e WesVarPC.
Resultados
Os programas apresentam duas opções em comum para importação de arquivos: o dBASE e arquivos tipo texto. O número de passos anteriores à execução das análises foram 21, 11 e 9, respectivamente para o Epi Info, Stata e WesVarPC. A eficiência computacional foi alta em todos eles, inferior a três segundos. Os erros padrão estimados utilizando-se o Epi Info e o Stata foram os mesmos, com aproximação até a terceira casa decimal; os do WesVarPC foram, em geral, superiores.
Conclusões
O Epi Info é o mais limitado em termos das análises disponíveis, porém ele é simples de usar e gratuito. O Stata e o WesVarPC são bem mais completos nos recursos de análises, porém há a desvantagem do custo. A escolha do programa dependerá principalmente das necessidades específicas do usuário.

Keywords
Software, utilization#. Evaluation#. Data interpretation, statistical#. Medical informatics application. Sampling studies. Health surveys. Eficiency. – Complex sample surveys.

Abstract

Objective
To compare specific software programs for data analysis of complex surveys regarding the following characteristics: ease of application, computer efficiency and accuracy of the results.
Methods
Secondary data from the Pesquisa Nacional sobre Demografia e Saúde (National survey on demography and health) (1996) with a target population of women aged

Correspondência para/*Correspondence to*:
Maria Helena de Sousa
Cidade Universitária Zeferino Vaz, s/n - CP 6181
13081-970 Campinas, SP, Brasil
E-mail: mhsousa@usp.br

*Aluna de pós-graduação em epidemiologia da Faculdade de Saúde Pública da USP.
Edição subvencionada pela Fapesp (Processo nº 00/01601-8).
Recebido em 24/8/1999. Reapresentado em 24/8/2000. Aprovado em 11/9/2000.

15 to 49 years old were used. This was a probabilistic subsampling drawn in two stages, then stratified, with the probability proportional to size in the first stage. The northern and mid-western regions of the country were selected for the study. The parameters of interest were mean for the age variable, and the proportion for five other qualitative variables. The software programs used were Epi Info, Stata and WesVarPC.

Results

The programs have two common options for the files import: the dBASE and text type files. The number of steps previous to the execution of the analyses were twenty-one for Epi Info, eleven for Stata and nine for WesVarPC. Efficiency was high for all them, that is, less that three seconds. The standard errors estimated using Epi Info and Stata were the same, with approximation up to the third decimal; those for WesVarPC were generally higher.

Conclusions

Epi Info is the most limited software program regarding the analyses currently performed; however it is easy to use and free. Stata and WesVarPC are far more complete, however the disadvantage is their cost. The choice of the software program will depend mainly on the user's specific needs.

INTRODUÇÃO

Os "softwares" estatísticos convencionais geralmente não consideram as características dos levantamentos amostrais que os tornam complexos, ou seja, a estratificação, a existência de conglomerados, as probabilidades desiguais de seleção das unidades amostrais e os ajustes para não-resposta e pós-estratificação.[3]

Existem dois métodos gerais para estimação de variância em delineamentos complexos. São eles: linearização por série de Taylor e replicação repetida. O primeiro método produz estimativas das estatísticas de primeira ordem, baseadas na variância dos termos lineares da expansão de Taylor para as respectivas estatísticas. O segundo, de replicação, é uma família de aproximações que leva em consideração subamostras repetidas e recalcula a estimativa ponderada para cada réplica, obtendo-se então a variância baseada nos desvios dessas estimativas em função da amostra total.[4,12]

Nos últimos anos, diversos programas computacionais vêm sendo desenvolvidos, em ambiente de computador pessoal, para a análise de dados de levantamentos complexos.[4,12] Os que usam expansão por série de Taylor para estimação de variâncias são o SUDAAN, do "Research Triangle Institute";[14] o PCCARP, da "Iowa State University" e "Statistics Canada";[8] o Stata, do "College Station", Texas;[13] o CSAMPLE/Epi Info, do "Centers for Disease Control" e "World Health Organization".[6] Os que utilizam a técnica de replicação para estimação de variâncias incluem o WesVarPC, do "Westat Inc."[2] e o VPLX, do "U.S. Bureau of the Census".[7]

Em um trabalho recente, Cohen[5] avaliou três "softwares" específicos para análise de dados de levantamentos complexos em relação a algumas características, como o tempo de execução das análises e a exatidão dos resultados, entre outras. O Stata, o SUDAAN e o WesVarPC foram avaliados e concluiu-se que é difícil designar um ou outro como o melhor, sugerindo aos usuários que considerem as atrações e limitações dos diversos programas em função de suas necessidades próprias.

Segundo Landis et al,[11] muitos levantamentos de larga escala, como os conduzidos pelos censos, pelo "National Center for Health Statistics" (NCHS) nos EUA e pelo "World Fertility Survey" (WFS), coordenado pelo International Statistical Institute na Holanda, possuem planos amostrais de múltiplos estágios envolvendo estratificação e a existência de conglomerados, bem como necessitam de técnicas de estimação que incluam pós-estratificação e ajustes para não-resposta.[10]

Exemplos de levantamentos complexos no Brasil são a Pesquisa Nacional por Amostra de Domicílios (PNAD),[1] a Pesquisa Nacional sobre Saúde Materno-infantil e Planejamento Familiar (PNSMIPF),[15] de 1986, e a Pesquisa Nacional sobre Demografia e Saúde (PNDS),[16] de 1996. A PNAD teve início em 1967 e possui periodicidade anual, exceto nos anos do censo demográfico. A PNSMIPF e a PNDS fazem parte das pesquisas "Demographic and Health Survey", iniciadas em 1988 em diversos países em desenvolvimento da América Latina, África e Ásia, e foram subamostras das PNAD de 1984 (atualizada em 1985) e de 1995, respectivamente. A PNDS é a pesquisa mais recente, em nível nacional, a levantar dados na área da saúde

materno-infantil e, portanto, constitui uma rica fonte de informações demográficas, especialmente para pesquisadores da área de saúde pública.

Com a crescente demanda de pesquisadores na área da saúde interessados em analisar dados de levantamentos amostrais, que em geral são complexos (Carlson[4]), o presente trabalho objetiva avaliar alguns "softwares" específicos, comparando-os em relação a características como, por exemplo, facilidade de aplicação, eficiência computacional e exatidão dos resultados. Destaca-se a importância em divulgar as opções tecnológicas atuais, que atendem às características específicas dos planos de amostragem no processo de estimação de erros padrão e intervalos de confiança.

MÉTODOS

Delineamento da PNDS-96, variáveis selecionadas e parâmetros de interesse

A PNDS-96, que fez parte da terceira fase das pesquisas "Demographic and Health Surveys", foi implementada através de uma subamostra de domicílios, sorteada da amostra principal de setores censitários da PNAD, 1995.[1] Tratou-se de uma subamostra probabilística selecionada em dois estágios, estratificada, com probabilidade proporcional ao tamanho no primeiro estágio, sendo a unidade amostral o setor censitário, definido pelo IBGE como uma área geográfica delimitada por perímetro identificável e utilizada nos censos demográficos; a unidade amostral de segundo estágio foi o domicílio, definido como o local de moradia estruturalmente separado e independente, constituído por um ou mais cômodos. Os domínios geográficos dessa pesquisa foram as sete regiões contempladas na PNAD, ou seja, Rio de Janeiro, São Paulo, Sul, Centro-leste, Nordeste, área urbana da Região Norte e Centro-oeste. Uma primeira estratificação geográfica foi a dos 26 Estados da federação mais o Distrito Federal, e a amostra foi autoponderada dentro desses estratos. A estratificação final foi obtida pela combinação de cada dois setores censitários, para fins de estimação dos dados. A população-alvo da pesquisa PNDS-96 foi a de mulheres de 15 a 49 anos de idade, residentes em domicílios particulares no País, excetuando-se a área rural da Região Norte.

Para o presente trabalho foram selecionadas da amostra as Regiões Norte (área urbana) e Centro-oeste do País, ou seja, duas das sete áreas acima referidas. A primeira área apresentou um tamanho final de amostra igual a 1.340 mulheres, composto de 30 estratos finais, e a segunda de 1.406 mulheres, composto de 46 estratos finais combinados.

As variáveis selecionadas para a análise foram:

- Idade da mulher na entrevista, em anos completos (V012);
- Escolaridade da mulher, obtida pela junção de duas variáveis: nível educacional mais alto que ela obteve (V106) e ano de educação mais alto que foi alcançado (V107); com estas duas variáveis foi definida a escolaridade (Escol) e, em seguida, dicotomizada em escolaridade até a oitava série e estritamente maior que esta série;
- Estado marital atual da respondente (V501), sendo dicotomizada em com e sem companheiro;
- Trabalho (V714), refere-se a se a mulher está atualmente trabalhando, com categorias sim e não;
- Tem ouvido sobre a Aids (V751), refere-se a se a respondente tem ouvido falar da Aids, com categorias sim e não;
- Tem ouvido sobre os dois usos do "condom" (V764), refere-se a se a respondente tem ouvido falar de "condom" para uso contraceptivo e para uso para prevenir doenças sexualmente transmissíveis (DST); foi dicotomizada em sim (para ambos os usos) e não (para não tem ouvido falar ou tem ouvido falar para apenas um dos usos).

Dessas seis variáveis analisadas, a escolaridade e a variável indicadora de trabalho apresentaram ausência de informação para algumas mulheres (unidades de análise) e, portanto, estas foram excluídas da análise específica para essas variáveis.

Os parâmetros considerados de interesse foram a média, para a idade da mulher, e a proporção, para as demais variáveis: proporção de mulheres com escolaridade maior que oitava série, com companheiro, proporção de mulheres que estão trabalhando, que têm ouvido falar sobre a Aids e que têm ouvido falar sobre os dois usos do "condom". O estimador da proporção (ou média) é o estimador razão, que é o mesmo para qualquer um dos três programas.[10]

O arquivo de dados original, com 12.612 mulheres, foi estruturado utilizando-se o "software" ISSA – "Integrated System for Survey Analysis" – para a seleção dessas variáveis de interesse, bem como das variáveis que caracterizam o delineamento, quais sejam, estrato (V022), conglomerado (V001), peso (V005) e domínio geográfico (V023). Este arquivo, contendo 14 variáveis, foi gravado em SPSS, que é uma das três opções possíveis (as outras são SAS e arquivo do próprio ISSA). Em seguida, foram obtidos dois arquivos: um para a Região Norte (V023=6) e outro para a Região Centro-oeste (V023=7). Na seqüência, estes arquivos foram exportados para o Dbase III e, a seguir, para o Epi Info (com extensão "REC"), para o Stata (com extensão "dta") e para o WesVarPC (com extensão "VAR"). Para

este último foi necessário criar os pesos de replicação ("replicate weigths") para a análise.

Avaliação dos "softwares"

Os três "softwares" analisados datam de 1997: Epi Info, versão 6.04b, que apresenta um módulo denominado CSAMPLE, para análise de dados de levantamentos complexos; Stata, versão 5.0, que incorpora comandos específicos para este tipo de análise; e WesVarPC, versão 2.12, que foi especialmente desenvolvido para análises de tais levantamentos.

O Epi Info é um programa que opera em ambiente DOS. O tipo básico de delineamento que pode ser analisado é a amostragem por conglomerados, estratificada, em múltiplos estágios. A estimação da variância é feita por expansão de Taylor.[6] O Stata é disponível em ambiente Windows e a estimação da variância também é feita por aproximação de série de Taylor.[13] O WesVarPC, por sua vez, opera em Windows e o delineamento básico que ele atende é a amostragem estratificada em múltiplos estágios. A estimação da variância é feita através da técnica de replicação.[2]

A comparação dos três programas foi feita por meio da avaliação da "facilidade de aplicação", medida pelos recursos disponíveis de importação e exportação de arquivos de dados e pelo número de passos necessários até o início da execução das análises; da "eficiência computacional" medida pelo tempo de execução das análises; e da "exatidão dos resultados" avaliada pela aproximação decimal das estimativas obtidas.

O "hardware" utilizado foi um Pentium II, 400 Mhz, com disco rígido de 4.300 MB (4,3 GB) e 64 MB de memória RAM.

RESULTADOS

Facilidade de aplicação

A Tabela 1 apresenta os recursos de importação e exportação de arquivos, existentes em cada um dos programas. O Epi Info e o WesVarPc possuem poucas alternativas de importação, três e quatro, respectivamente. O Stata, através do "Stat/transfer", possui diversas opções de importação/exportação, que são as mesmas. Observa-se que uma opção em comum, para importação de arquivos, é o dBASE. Arquivo do tipo texto também é uma opção comum, porém, no Stata deve-se importar o arquivo usando o comando "infile" dentro do programa.

Em relação ao número de passos necessários antes do início da execução das análises, estão relacionados, a seguir, as linhas de comando (exceto para o Epi Info) ou os campos a serem preenchidos (exceto para o Stata).

O Epi Info, que é dirigido por menu, não permite programação, ou seja, as análises são obtidas somente após o preenchimento dos campos nas duas telas do CSAMPLE, a primeira para a escolha do banco de dados e a segunda para a especificação das variáveis e opções de saída "output". Os campos preenchidos na segunda tela foram:

- Main: v012 (ou escol, v501, v714, v751, v764)
- Strata: v022
- PSU: v001
- Weight: v005
- Output Options: File => n.epi (ou co.epi)

Em seguida, teclou-se conjuntamente Alt + m (Means) para a variável v012 e Alt + t (Tables) para as demais variáveis principais: escol, v501, v714, v751 e v764.

Foram gravados dois arquivos de saída, um para cada região analisada. Após o preenchimento das variáveis do delineamento, apenas a variável principal "Main" necessitou ser alterada. Em seguida, teclou-se Alt+t conjuntamente e, por último, a letra "a" para a opção "append", de atualização do arquivo de saída. Portanto, foram totalizados, separadamente para cada região, 21 passos para a execução das análises, embutindo-se aqui as teclas Alt+m ou Alt+t e a letra "a" acima referida.

O Stata, diferente do Epi Info, não possui menu de orientação para preenchimento de campos na tela. Portanto, deve-se digitar as linhas de comando uma a uma, ou preparar um arquivo em lote com extensão ".do", para posterior execução dessas mesmas linhas de comando, que foram as seguintes:

- use c:\diret\norte (ou coeste)
- svyset strata v022
- svyset psu v001
- svyset pweight v005
- log using c:\diret\n.sta (ou c:\diret\co.sta)
- svymean v012
- svymean escol
- svymean v501
- svymean v714
- svymean v751
- svymean v764

O Stata não reconhece letras maiúsculas nas linhas de comando, ou seja, deve-se digitar tudo em letra minúscula. As três linhas de caracterização do plano de amostragem, que se iniciam com "svyset", podem ser escritas uma única vez, desde que o arquivo seja

Tabela 1 - Recursos para importação e exportação de banco de dados nos três programas.

"Software"	Importação	Exportação
Epi Info	arquivo texto (ASCII); Lotus e dBASE	SYSTAT; SAS; arquivo texto (ASCII); Lotus; SPSS-X; Epistat; dBASE; BASIC; SPSS-PC; Statpac; MULTLR; Egret e xBASE
"Stat/transfer"	Lotus; dBASE; Excel; SAS trasport files; S-PLUS; SPSS export files; Stata; SYSTAT; Quattro pro; Paradox; Symphony; Foxbase; Clipper; Alpha Four; Crunch e Gauss File	Lotus; dBASE; Excel; SAS trasport files; S-PLUS; SPSS export files; Stata; SYSTAT; Quattro pro; Paradox; Symphony; Foxbase; Clipper; Alpha Four; Crunch e Gauss File
Stata*	arquivo texto (ASCII)	-
WesVarPC	SAS; SPSS for Windows; dBASE; arquivo texto (ASCII)	Não possui opção para exportação

*Diretamente no programa, utilizando o comando "infile"
Fontes: Epi Info v.6.04b, Stat/transfer v.3.53; Stata v.5.0; WesVarPC v.2.12

gravado novamente. Para cada uma das duas regiões, foram necessários 11 passos para a execução das análises, que poderiam ser diminuídos para oito se as linhas de caracterização do plano já estivessem incorporadas ao banco de dados.

Utilizou-se o WesVarPC, com o método de replicação de Jackknife. Dos três programas, este é o único que tem as duas alternativas de execução: permite o preenchimento dos campos no menu e também a execução desses mesmos comandos gravados em um arquivo com extensão ".wvq". As especificações foram as seguintes:

- No menu Tables => New => *c:\diret\norte* (ou *c:\diret\coeste*)
- Compute *m012*=mean*(v012)*
- Compute *mescol*=mean*(escol)*
- Compute *m501*=mean*(v501)*
- Compute *m714*=mean*(v714)*
- Compute *m751*=mean*(v751)*
- Compute *m764*=mean*(v764)*
- Close
- Run

O que está em itálico foi digitado e, para o restante, como "Compute" e "mean", foi necessário apenas apertar o botão do "mouse". Foram gravados dois arquivos: "norte.wvq" e "coeste.wvq", com o intuito de armazenar essas linhas de comando. Foram necessários nove passos antes da execução das análises.

Eficiência computacional

Para a análise dos dados no Epi Info, anotou-se o tempo total para executar as seis análises, uma a uma, referentes às variáveis de interesse. O tempo foi computado após o preenchimento da segunda tela do CSAMPLE e acumulado até a execução das seis análises desejadas, separadamente para cada região.

Quando os dados foram analisados utilizando-se o Stata, o tempo de execução foi obtido após a digitação da seguinte linha de comando: "do c:\diret\norte" ou "do c:\diret\coeste", que contém os comandos anteriormente descritos.

Para a análise utilizando-se o WesVarPC, anotou-se o tempo de execução após a abertura do programa gravado com extensão ".wvq" (norte ou coeste) e, em seguida, teclou-se "run".

Todos os "softwares" foram muito eficientes, com tempo de execução inferior a três segundos, no Epi Info, e inferior a um segundo no Stata e WesVarPC (Tabela 2).

Tabela 2 - Tempo de execução das análises em cada um dos três "softwares", para dados das Regiões Norte e Centro-oeste.

"Software"/ Região	Tempo (seg)
Epi Info- CSAMPLE	
Região Norte	<3"
Região Centro-oeste	<3"
Stata	
Região Norte	<1"
Região Centro-oeste	<1"
Wesvarpc	
Região Norte	<1"
Região Centro-oeste	<1"

Exatidão dos resultados

As Tabelas 3 e 4 apresentam os resultados dos parâmetros estimados para as Regiões Norte e Centro-oeste, respectivamente. Para a variável idade estimou-se a média e para as demais estimou-se a proporção de unidades elementares que pertenceram a determinada categoria. Os resultados para o erro padrão e o efeito do desenho (deff) estão apresentados nas linhas seguintes ao parâmetro estimado.

Optou-se por apresentar os resultados com três casas decimais, pois o CSAMPLE do Epi Info possui apenas este padrão de saída ("output").

Tabela 3 - Estimativas dos parâmetros de interesse para a Região Norte* segundo o "software" utilizado.

Parâmetro estimado	"Software"		
	Epi Info	Stata	WesVarPC
Média de idade (anos)	28,514	28,514	28,514
EP média	0,231	0,231	0,2**58**
Deff	-	**0,765**	**0,956**
Proporção ESCOL>8ª série (%)	31,306	31,306	31,306
EP proporção	2,421	2,421	2,**602**
Deff	3,6**51**	3,6**48**	**4,218**
Proporção v501:com companh. (%)	54,706	54,706	54,706
EP proporção	1,845	1,845	1,8**12**
Deff	1,8**40**	1,8**39**	1,**775**
Proporção v714: trabalha (%)**	49,923	49,923	49,923
EP proporção	1,989	1,989	**2,182**
Deff	2,11**4**	2,11**3**	2,**543**
Proporção v751: tem ouvido sobre a AIDS (%)	99,656	99,656	99,656
EP proporção	0,151	0,151	0,1**49**
Deff	0,88**8**	0,88**7**	0,8**71**
Proporção v764: tem ouvido sobre os dois usos do condom (%)	85,062	85,062	85,062
EP proporção	1,791	1,791	1,**920**
Deff	3,38**3**	3,38**1**	3,**886**

*(n=1.340)
**Faltou informação de 4 mulheres
EP = erro padrão
Deff = efeito do desenho

Tabela 4 - Estimativas dos parâmetros de interesse para a Região Centro-oeste* segundo o "software" utilizado.

Parâmetro estimado	"Software"		
	Epi Info	Stata	WesVarPC
Média de idade (anos)	29,856	29,856	29,856
EP média	0,231	0,231	**0,312**
Deff	-	**0,784**	**1,438**
Proporção ESCOL>8ª série (%)**	29,646	29,646	29,646
EP proporção	2,210	2,210	2,**847**
Deff	3,2**91**	3,2**89**	**5,461**
Proporção v501:com companh. (%)	62,714	62,714	62,714
EP proporção	1,927	1,927	1,9**57**
Deff	2,23**2**	2,23**0**	2,**302**
Proporção v714: trabalha (%)***	50,844	50,844	50,844
EP proporção	2,054	2,054	2,0**90**
Deff	2,36**4**	2,36**2**	2,**449**
Proporção v751: tem ouvido sobre a AIDS (%)	99,949	99,949	99,949
EP proporção	0,051	0,051	0,051
Deff	0,71**6**	0,71**5**	0,7**21**
Proporção v764: tem ouvido sobre os dois usos do condom (%)	78,398	78,398	78,398
EP proporção	1,278	1,278	1,**344**
Deff	1,35**7**	1,35**6**	1,**500**

*(n=1.406)
**Faltou informação de uma mulher
***Faltou informação de cinco mulheres

Os erros padrão estimados utilizando-se o Epi Info e o Stata (cujo método de estimação é por série de Taylor) foram os mesmos, com aproximação até a terceira casa decimal. O efeito do desenho (deff) apresentou variação na terceira ou mesmo na segunda casa decimal (em destaque). O CSAMPLE do Epi Info não produz o cálculo do deff quando o parâmetro é a média. Para o erro padrão obtido utilizando-se o WesVarPC, exceto para a variável "tem ouvido falar sobre a Aids", na Região Centro-oeste houve diferença pelo menos na segunda casa decimal. Em geral, as estimativas de erros padrão e deff´s obtidas pelo WesVarPC foram superiores às dos demais programas (Tabelas 3 e 4).

DISCUSSÃO

A principal limitação do presente trabalho é o tamanho dos bancos de dados, que foram inferiores a 1.500 mulheres (registros) em cada uma das duas regiões da PNDS-96. Ademais, apenas três "softwares" foram aqui considerados, dentre os diversos existentes.[4,12] Em relação ao Stata e WesVarPC, replicou-se aqui o trabalho de Cohen.[5] Utilizou-se, porém, bases de dados menores (cada uma das regiões analisadas correspondeu a menos de 5% da amostra do trabalho de Cohen), e o método de seleção também foi distinto.

Em decorrência do avanço que vem ocorrendo em informática, convém salientar que o "hardware" utilizado foi de melhor qualidade quando comparado ao utilizado por Cohen.[5] A memória RAM e a velocidade foram, respectivamente, 4 e 5 vezes superiores às do referido trabalho.

Para o tópico de avaliação da facilidade de aplicação dos programas, todos eles apresentam opções de importação e exportação de arquivos, que permitem, com isso, uma fácil e rápida conversão dos mesmos. Destaca-se a opção via arquivo dBASE, que pode ser importado em qualquer um dos três analisados. Há também a alternativa de utilização de uma única ferramenta, por exemplo o DBMS/Copy, que é específico para manipulação e conversão de arquivos. Em relação à programação, o Stata e o WesVarPC possuem recursos suficientes para armazenamento dos comandos, que inclusive podem ser utilizados em outras análises com bancos de dados semelhantes (por exemplo, estudos longitudinais periódicos). O usuário deve ter um conhecimento básico *a priori*, para utilizar o Stata. Tal conhecimento não é essencial aos novos usuários do WesVarPC, devido à existência de menu de tela com instruções sucessivas, que o usuário pode armazenar, ou não, antes de sair do programa. O Epi Info, por sua vez, não possui recurso de programação, ou seja, as instruções são feitas pelo menu e a cada nova variável o processo deve ser parcialmente repetido alterando as variáveis principais "Main". Este programa, inclusive, foi o que necessitou do maior número de passos antes da execução das análises.

Os parâmetros de interesse foram simples, ou seja, proporção e média, e os três programas analisados apresentaram alta eficiência, medida pelo tempo de execução. As estimativas pontuais foram iguais nos três, o que já era esperado, pois os estimadores são os mesmos. Para as estimativas de variância, o maior resultado, em geral obtido utilizando-se o WesVarPC, deve-se ao método utilizado (replicação), que reduz os graus de liberdade quando as unidades primárias de amostragem são separadas em réplicas. Portanto, em relação à exatidão dos resultados, medida em termos da aproximação decimal das estimativas obtidas, conclui-se que o método de replicação de JackKnife do WesVarPC apresenta estimativas de erro padrão, em geral superestimadas em relação às estimativas obtidas nos outros dois "softwares". Entretanto, tais valores superestimados não tiveram alto impacto no efeito do desenho (deff).

Segundo Kalton,[9] os métodos de estimação de variância como o de Taylor (no Epi Info e no Stata) e o de replicação (no WesVarPC) apresentam resultados semelhantes, como os aqui obtidos, sendo o primeiro mais indicado quando os estimadores são mais simples, como por exemplo proporção e média, e o segundo nos casos de estimadores de qualquer complexidade, como por exemplo coeficientes de regressão.

A pequena diferença observada nos efeitos do desenho quando se utilizou o Epi Info e o Stata, pode ser atribuída a erro de arredondamento, pois os erros padrão foram os mesmos.

Quanto à disponibilidade dos "softwares" analisados,* o Epi Info é gratuito, podendo ser obtido pela Internet; o Stata e o WesVarPC são vendidos, sendo que este último passou recentemente a fazer parte do SPSS, mas apresenta uma versão anterior de acesso gratuito pela Internet, utilizada no presente trabalho, versão 2.12 de 1997.

Em relação à documentação dos programas, o Epi Info e o WesVarPC apresentam manuais de fácil leitura, com informações sobre o módulo de análise de dados de levantamentos complexos, sendo que o primeiro é o único que possui versão do manual em português. O Stata possui incluído em seu extenso manual as informações específicas sobre análise de "surveys".

Assim como no trabalho de Cohen,[5] recomenda-se ao pesquisador que necessita do "software" específico para a análise de seus dados, que avalie de antemão quais são as análises disponíveis em cada um deles, para que possa decidir entre um ou outro programa, avaliando a relação custo/benefício dos mesmos. O Epi Info é, dos três, o mais limitado em termos das análises disponíveis, ou seja, é útil para estimação de parâmetros simples, como a média e a proporção. Em contrapartida, é bem conhecido na área de saúde pública, é gratuito e simples de ser usado, especialmente para análise de estudos epidemiológicos. O Stata e o WesVarPC possuem como principal vantagem os recursos de análises, bem mais completos, com diversas técnicas de análise multivariada. A desvantagem principal é o custo.

Quanto ao tamanho que os três "softwares" ocupam no disco rígido, o Epi Info 6.04b completo ocupa cerca de 8,5 MB, o Stata 5.0 aproximadamente 3,5 MB e o WesVarPC 2.12 1,4 MB. Todos eles podem ser considerados pequenos, baseando-se nos tamanhos de "winchester" atualmente disponíveis no mercado. Sobre a atualização dos "softwares",

*Epi Info: disponível no "site" http://www.cdc.gov/epo/epi/downepi6.htm; Stata, vendido ao custo de US$395; WesVarPC, vendido por US$600. Uma versão anterior está disponível no "site" http://www.westat.com/wesvar/wesvar.html#download.

que é uma característica importante, o Epi Info 6.04b continua sendo a última versão disponível, de 1997, mas o Stata já possui a nova versão 6, de 1999, e o WesVarPC está atualmente incorporado ao programa SPSS.

Os achados do presente trabalho merecem um paralelo com os de Cohen.[5] A facilidade de aplicação para o Stata é equiparável nos dois trabalhos, desde que excluídas as análises bivariadas que não foram feitas no presente trabalho. O mesmo vale para o WesVarPC, apesar de terem sido somados os comandos em blocos. Em relação à eficiência computacional, apesar do tamanho dos arquivos terem sido menores que os de Cohen, o tempo dispendido foi extremamente menor, sugerindo que "hardwares" mais modernos, com mais memória e maior velocidade, reduzem significativamente o tempo de execução.

Como sugestão ao usuário, se a demanda for para análises simples de seus dados, e para poucas variáveis, indica-se o Epi Info-CSAMPLE, que além de ser gratuito, é bem simples de ser utilizado. Se o número de variáveis e de análises for grande, e a demanda for para análises mais complexas, por exemplo modelagem, indica-se o Stata ou o WesVarPC, que são programáveis. Em síntese, a escolha dependerá principalmente das necessidades de análises e do volume das mesmas.

Sugere-se como seqüência ao presente trabalho, que dados de outras pesquisas de base populacional sejam analisados, de preferência com tamanhos de amostra maiores, para avaliação desses e mesmo de outros "softwares" específicos para análises de inquéritos.

AGRADECIMENTOS

A Maria José Martins Duarte Osis, pesquisadora do Centro de Pesquisas das Doenças Materno-infantis de Campinas (Cemicamp), pela revisão crítica final do trabalho. Ao revisor da *Revista de Saúde Pública* pelas valiosas sugestões.

REFERÊNCIAS

1. Albieri S, Bianchini ZM. *Uma revisão dos principais aspectos dos planos amostrais das pesquisas domiciliares realizadas pelo IBGE*. Rio de Janeiro:IBGE; 1998. (Série Textos para Discussão).

2. Brick JM, Broene P, James P, Severynse J. *A user's guide to WesVarPC*. Rockville, MD: Westat Inc; 1997.

3. Brogan DJ. Pitfalls of using standard statistical software packages for sample survey data. In: Armitage P, Colton T, editors. *Encyclopedia of biostatistics*, [serial on line] 1998. Avaliable from: <URL: http:// www.fas.harvard.edu/~stats/survey-soft/ donna_brogan.html> [2000 Mar 14]

4. Carlson BL. Software for statistical analysis of sample survey data. In: Armitage P, Colton T, editors. *Encyclopedia of biostatistics*, [serial on line] 1998. Avaliable from: <URL: http:// www.fas.harvard.edu/ ~stats/survey-soft/blc_eob.html> [2000 Mar 14]

5. Cohen SB. An evaluation of alternative PC-based software packages developed for the analysis of complex survey data. *Am Stat* 1997;51:285-92.

6. Dean AG, Dean JA, Coulombier D, Brendel KA, Smith DC, Burton AH, et al. *Epi Info, version 6:a word processing, database, and statistics program for epidemiology on microcomputers*. Atlanta, Georgia: Centers for Disease Control and Prevention; 1994.

7. Fay RE. VPLX: variance estimates for complex samples. *Proc Surv Res Meth Sec.* ASA 1990;266-71.

8. Fuller WA, Kennedy W, Schell D, Sullivan G, Park HJ. *PC CARP.* Ames, Iowa: Statistical Laboratory, Iowa State University; 1989.

9. Kalton G. *Introduction to survey sampling*. Beverly Hills: Sage Publications; 1983. (Series: Quantitative Applications in the Social Sciences, 35).

10. Kish L. *Survey sampling*. New York: John Wiley & Sons; 1965.

11. Landis JR, Lepkowski JM, Eklund SA, Stehouwer SA. A statistical methodology for analyzing data from a complex survey: the first National Health and Nutrition Examination Survey. *Vital Health Stat* 1982;92(2):1-52.

12. Lepkowski J, Bowles J. Sampling error software for personal computers. *Surv Stat* 1996;35:10-7.

13. StataCorp. *Stata statistical software: release 5.0.* College Station, TX: Stata Corporation; 1997.

14. Shah BV, Barnwell BG, Bieler GS. *SUDAAN users´s manual, version 6.4.* 2nd ed. Research Triangle Park, NC: Research Triangle Institute; 1996.

15. Sociedade Civil Bem-Estar Familiar no Brasil. *Pesquisa nacional sobre saúde materno-infantil e planejamento familiar, Brasil - 1986*. Rio de Janeiro; 1987.

16. Sociedade Civil Bem-Estar Familiar no Brasil. *Pesquisa nacional sobre demografia e saúde - 1996*. Rio de Janeiro; 1997.