

DUMMY FILE CREATION FOR THE REMOTE ACCESS PROGRAM OF THE NATIONAL POPULATION HEALTH SURVEY

Harold Mantel and Sylvain Nadon¹

ABSTRACT

Statistics Canada's National Population Health Survey (NPHS) is a longitudinal and cross-sectional survey that collects information on health and determinants of health for the Canadian population. As one of its data products the NPHS produces public-use microdata files (PUMFs) for use by researchers in universities and other agencies. These PUMFs are essentially copies of the master data files, but with some of the variables suppressed, capped or collapsed in order to protect the confidentiality of survey respondents. Because of these data losses, the PUMFs do not always satisfy the data needs of researchers. For this reason the NPHS Remote Access Program was set up. Statistics Canada produces NPHS dummy files that mimic the confidential master microdata files but that contain artificial records and are therefore not confidential. Researchers can now write their own analysis programs, testing them on the dummy files before submitting them to Statistics Canada. Members of the NPHS team run the programs using the master files as input and the output is then vetted for confidentiality before being returned to the researchers. Although the records are artificial some efforts were made to ensure that simple tabulations from the dummy files would yield reasonable results. For each dummy file a set of dummy bootstrap weights was also produced, allowing analysts to write their own bootstrap variance estimation programs. In this paper the methodology for creation of the dummy files is described.

KEY WORDS: Artificial data; Confidentiality; Public-use microdata file; Variance estimation.

RÉSUMÉ

L'Enquête nationale sur la santé de la population (ENSP), que mène Statistique Canada, est conçue pour recueillir des données longitudinales et transversales sur la santé et les déterminants de la santé de la population canadienne. À partir des données de l'ENSP, des fichiers de microdonnées à grande diffusion (FMGD) sont produits. Ces FMGD sont destinés principalement aux chercheurs universitaires ou à d'autres agences. Les FMGD sont essentiellement des copies des fichiers maîtres, toutefois certaines variables ont été supprimées, tronquées ou regroupées afin d'assurer la confidentialité des répondants de l'enquête. Puisque le contenu des FMGD est limité, ceux-ci ne satisfont pas toujours aux besoins en données des chercheurs. À cet égard, on a mis en place le programme de téléaccès de l'ENSP. Statistique Canada fournit aux utilisateurs autorisés des fichiers fictifs de l'ENSP, c'est-à-dire des fichiers qui ont la même structure que les fichiers maîtres, mais qui contiennent des données fictives. Puisque ces fichiers contiennent des enregistrements artificiels, ils ne sont donc pas confidentiels. Les chercheurs peuvent ainsi écrire leurs propres programmes d'analyse, les mettre à l'essai en se servant des fichiers fictifs, puis les transmettre à Statistique Canada. Les membres de l'équipe de l'ENSP exécutent les programmes sur les fichiers maîtres, vérifient les données de sortie pour s'assurer du respect des normes de confidentialité, puis les envoient aux chercheurs. Bien que les enregistrements soient artificiels, des efforts ont été réalisés afin d'assurer que des tabulations simples effectuées à partir des fichiers fictifs donnent des résultats raisonnables. De plus, pour chaque fichier fictif, un ensemble de poids bootstrap fictifs a également été produit, permettant aux analystes d'écrire leur propre programme d'estimation de la variance selon la méthode du bootstrap. Ce document décrit la méthodologie utilisée pour la création des fichiers fictifs.

MOTS-CLÉS : Données fictives; confidentialité; fichiers de microdonnées à grande diffusion; estimation de la variance.

1. INTRODUCTION

Statistics Canada's National Population Health Survey (NPHS) is a longitudinal and cross-sectional survey of the population living in private households in the ten

provinces. The survey collects information on health and related variables. Topics covered include, among others, health status, chronic conditions, utilization of health services, drug use, mental health, stress, social support, smoking and alcohol, physical activities, and sociodemographic and socioeconomic characteristics.

¹ Harold Mantel and Sylvain Nadon, Household Survey Methods Division, R.H. Coats - 16th floor, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6, manthar@statcan.ca

The design of the survey, which is intended to produce both longitudinal and cross-sectional estimates, is described in Tambay and Catlin (1995). It is based on the design of the Canadian Labour Force Survey, which is a stratified multistage household survey. The NPHS is complemented by two related surveys: (1) a survey of the North which is similar to the NPHS but which covers the Yukon, Nunavut and the Northwest Territories, and (2) an institutional survey which covers long term residents of hospitals and residential care facilities.

Data collection for the NPHS began in 1994/95. There were approximately 17,000 respondents who now form the longitudinal panel. Members of the panel will be contacted every second year for up to 20 years. The survey also has the capacity for provinces to purchase extra content or extra sample to improve provincial level estimates. Four provinces bought small supplementary samples for the first cycle of the survey. For the second cycle of the survey, for which data collection took place in 1996/97, Ontario, Manitoba and Alberta bought very large supplementary samples to allow estimation for Health Regions, *i.e.* for subprovincial areas. There were no supplementary samples purchased for the third cycle of the survey; however, there were two top-up samples to account for panel attrition and for new entrants to the population. The top-up samples are used only for cross-sectional analysis. Respondents from the cycle 3 top-up samples will also be followed into cycle 4 for cross-sectional purposes. It is expected that after cycle 4 the NPHS will no longer be used for cross-sectional estimates since a new cross-sectional survey, called the Canadian Community Health Survey with content similar to that of the NPHS, is being set up to provide cross-sectional estimates at the local health area level every second year.

The NPHS uses two different questionnaires: a health questionnaire which is asked for one selected person in each sample household, and a shorter, general questionnaire which is asked of all members of the household. Members of the longitudinal panel are among the selected individuals asked the health questionnaire questions; however, top-up and supplementary samples also have one selected individual per household. Thus the set of individuals asked the health questionnaire in any particular cycle of the survey may be larger than the longitudinal panel. There are therefore three different types of possible analyses: longitudinal analyses based on data from the longitudinal panel over two or more cycles of the survey, health cross-sectional analyses based on the selected individuals for one cycle of the survey,

and general cross-sectional analyses based on the selected individuals and their cohabitants for one cycle of the survey.

The main products of the NPHS are microdata files for analysis. The intended users of these files include analysts from Health Canada and the provincial health ministries, from universities, hospitals and research institutes, as well as from Statistics Canada. Three different types of microdata files are produced: master, share and public-use. The master microdata files contain information on all survey variables for all survey respondents. Information, such as name and address, that directly identifies the respondents is not on the master microdata files; however, postal codes and various derived geographic codes as well as exact age are present. As part of the interview, respondents are asked if they would allow their data to be shared with Health Canada and the provincial health ministries. About 95% of respondents agree to such sharing, and the share microdata files contain all of the master file variables for these respondents. Finally, the public-use microdata files (PUMFs) contain information on all respondents, but many of the variables are suppressed or aggregated to protect the confidentiality of respondents. For each of these three types we may have longitudinal, health and general microdata files for a total of nine different possible microdata files.

Analysts at Statistics Canada have access to the master microdata files. Other analysts who are deemed employees of Statistics Canada may also have access to the master microdata files, but this must be at a secure location (at the time of writing such secure locations are available only at Statistics Canada's Regional Offices). Analysts at Health Canada or at one of the provincial ministries of health have access to the share microdata files. Others have direct access only to the PUMFs.

2. NPHS REMOTE ACCESS PROGRAM

For reasons of confidentiality, the NPHS public-use microdata files (PUMFs) do not contain many of the original variables that are available on the master microdata files. Many of the PUMF variables are collapsed versions or derived summaries of the original data. For example, age and geographic variables are generally available on the PUMFs only at higher levels of aggregation. In addition, a longitudinal PUMF has not yet been released for cycle 2 of the survey, since such a PUMF could likely be used to link the already-released cross-sectional PUMFs from cycles 1 and 2, which could compromise

confidentiality. For these reasons, the PUMFs are not always adequate for analysis purposes.

To overcome these shortcomings of the PUMFs, a remote access program was set up to allow analysts to run analyses using the master microdata file, but in such a way that the data remain confidential. In order to run analyses using the master microdata through the remote access program, analysts must write their own analysis programs and send them to Statistics Canada where they are run with the master microdata file as input. Output is then vetted for confidentiality before being forwarded to the analyst. This vetting consists of checking that no confidential data are produced and that regressions or cell estimates in tables meet or exceed a minimum sample size cutoff.

The remote access program is available free of cost to those who purchase the PUMFs or have access to them through the Data Liberation Initiative (DLI, an agreement through which all of Statistics Canada's electronic data products are made available to a group of universities for use by their faculty and students for a flat annual fee). It is the analysts' responsibility to ensure that their analysis programs run properly. To this end Statistics Canada provides them with dummy files that they can use for development and testing of their programs. The dummy files have the same format as the master microdata files, but contain artificial data and only 5% or 10% as many records. The primary objective in the creation of the dummy files was that the artificial data be coherent, *i.e.*, that they be consistent with the skip patterns in the questionnaire. This was important since realistic data are needed to test analysis programs. A secondary objective was to preserve, at least approximately, the marginal distributions of variables and the relationships between closely related variables from the master microdata files. This was particularly important in the case of the longitudinal dummy file since no longitudinal PUMF is yet available.

In addition to remote access for analysis of the NPHS data, remote access can also be used to obtain bootstrap estimates of variance. For each dummy file, 500 sets of bootstrap weights were also created. This allows analysts to write their own variance estimation programs to be run at Statistics Canada using the master microdata files and their associated bootstrap weights. Examples of bootstrap variance estimation programs, written in SAS and SPSS, are also provided.

The Health Statistics Division of Statistics Canada currently has one person handling three or four remote access requests per day and spending about a third of

her time on these requests. It is expected that usage will increase as analysts become accustomed to the program and as they start to use the bootstrap weights for variance estimation. The remote access facility, in which analysts write their own programs for accessing the master microdata file, is much more efficient than custom tabulations and variance estimations.

3. CREATION OF THE NPHS DUMMY DATA FILES

As noted in Section 2, the creation of the dummy files had two objectives: (1) that the artificial data be coherent, *i.e.*, that they be consistent with the skip patterns in the questionnaires, and (2) to preserve, at least approximately, the marginal distributions of variables and the relationships between closely related variables from the master microdata files. Records on the dummy files were created by taking blocks of variables from different randomly selected donor records from the master microdata files. Thus each dummy record would contain data from several donor records and no single donor would be identifiable from the dummy files. In order to help meet the two objectives, there were two steps in the specification of the procedure: (1) the definition of donor classes, and (2) the definition of blocks of variables to be imputed together.

In the first step, records on the master file were divided into donor classes. One of the objectives was to form classes of records with similar pathways through the questionnaires, so that when random data swapping was applied within classes the resulting artificial records would be internally coherent. Since many of the skips in the questionnaire were based on age and sex, and a few on province, the classes were generally based on age/sex categories within province with a minimum class size to ensure that there would be no identifiable data in the dummy records. For the longitudinal file the classes were also based on the longitudinal response pattern (an indicator of response status for different cycles of the survey) to avoid mixing of data from full respondents and partial respondents. (A panel member who responds to the general questionnaire but not to the health questionnaire is called a partial respondent to that cycle of the survey.) For each class the class sample size, *i.e.*, the number of dummy records to be created from records in that class, was determined to be approximately 5% (for the general file) or 10% (for the health and longitudinal files) of the number of records from that class on the master file.

For the longitudinal dummy file, the definition of the classes was slightly different. Since the classes for the longitudinal files were based on the age at both the 1994 and 1996 as well as sex, the number of classes became very large. In order to maintain a minimum number (20) of master file records in each class, geography was not used in the formation of these classes. However, this created a problem since Alberta and Manitoba both had special additional content in 1994 and Alberta had special content in 1996. To get around this problem a dummy file was first created that did not include this special content, and the special content was then added to the dummy file records from Alberta and Manitoba at a second step. The problem of small classes did not arise in this second step since the number of age-based skips for the special content was quite small.

Certain questions on the NPHS questionnaires are asked only in the case of non-proxy interviews. Since one of the objectives is to form classes of records with similar pathways through the questionnaires, the variables indicating proxy/nonproxy status for an interview ought logically to be accounted for in the formation of classes. However, ignoring these variables is not completely unreasonable from an analytic perspective since a section in the questionnaire might also be skipped because of refusals. With this consideration in mind, and with a view to keeping the number of classes reasonably small, the proxy/non-proxy status was ignored in the formation of classes.

The second step was formation of the blocks of variables. The variables were first grouped into fifty basic blocks of closely related variables corresponding to sections of the questionnaire, and the blocks then consisted of collections of these basic blocks. Where the questions asked in one section depend on answers to questions asked in a previous section, the corresponding basic blocks of variables should be put into the same block, to maintain internal coherence. The exceptions to this rule are the variables used in the formation of the classes, whose effects on the pathway through the questionnaires were already accounted for. A guiding principle was that blocks should be analytically meaningful while also being small enough that they were safe from the point of view of confidentiality. Variables that in combination could identify individuals should be in different blocks.

Once the classes, class sample sizes, and blocks were determined, the dummy file was created. As mentioned above, records on the dummy file were created by taking blocks of variables from different

randomly selected donor records from the master file. In order to implement this, the master file was first randomly reordered within classes. Then for each class the required number of dummy records was formed by taking data for each block of variables sequentially from the randomly ordered records for that class. Once all of the records within a class were used the procedure would return to the beginning of the list. The class sample sizes were carefully chosen so that this procedure would not use the same donor record more than once in the creation of any particular dummy record.

After having created the dummy file, the weights were recalibrated by first ratio-adjusting the weights within each class to the total class weights from the master file, and then calibrating these adjusted weights using the same poststrata and control totals as had been used for creation of the master file. This recalibration was done to help preserve the marginal distribution of variables on the dummy files. Later, when ASCII versions of the dummy files were created the weights were divided by 20 (for the general file) or 10 (for the health and longitudinal files) to ensure that the weights were not too large for the space allotted for them. Thus estimates of totals from the dummy files will be about 1/10 or 1/20 of the corresponding estimates from the master files. Finally, to help ensure confidentiality, a number of administrative variables, which would be of no interest from an analytical viewpoint and which might possibly be dangerous in the sense of allowing identification of survey respondents, were suppressed (set to blank or 9) on the dummy files. Such variables included, for example, dates of data collection and an indicator of the person within the household who answered the survey questions.

4. DUMMY BOOTSTRAP WEIGHTS

4.1 Bootstrap Variance Estimation for the NPHS

Variance estimation for the NPHS is based on the bootstrap resampling method. The bootstrap method used is described in Rao, Wu and Yue (1992) and in Yung (1997). The method is appropriate for stratified multistage sampling. A more detailed description of the bootstrap variance estimation method in the context of the NPHS is given by Yeo, Mantel and Liu (1999). For each bootstrap sample $(n_h - 1)$ clusters are selected with replacement from the n_h sample clusters in stratum h . The bootstrap weights are then obtained by adjusting the sampling weights to account for the bootstrap resampling. A total of B independent bootstrap resamples are obtained using the same

procedure. The bootstrap variance estimator for $\hat{\theta}$ is then given by

$$\frac{1}{B} \sum_b (\hat{\theta}_{(b)}^* - \hat{\theta}_{(.)}^*)^2$$

where $\hat{\theta}_{(.)}^* = (1/B) \sum_b \hat{\theta}_{(b)}^*$. The estimator $\hat{\theta}$ may be as simple as an estimate of a total or proportion, or as complex as the estimator of a logistic regression parameter or a quantile of a distribution.

For cycle 2 of the NPHS, the number of bootstrap resamples, B , for the master microdata files was chosen as 500. Empirical investigation has suggested that this is large enough to produce stable variance estimates (in the sense that the estimate is close to what would be obtained from $B=1000$ or $B=2000$). For estimates of simple totals or proportions a smaller number of bootstrap resamples may be acceptable. The practical difficulty of loading 500 sets of weights and evaluating an estimator 500 times can be problematic, particularly with the very large cross-sectional files for cycle 2 (due to large provincial sample buy-ins) and for complex estimators; however, with the rapidly increasing size and speed of computers this problem is quickly disappearing.

For the NPHS, as for most other surveys, the sampling weights are adjusted to account for non-response and then calibrated to known demographic totals. Logically these adjustments to the weights ought to be considered part of the estimation process. However, it is practically convenient to consider them separately, as part of the weighting process, since they are identical for all estimates derived from the survey. The survey weights provided with the NPHS microdata files incorporate these adjustments. The files of bootstrap weights are similarly adjusted, with the caveat that the nonresponse adjustment currently takes place before the bootstrap resampling. Thus estimates of variance derived from these sets of bootstrap weights do not account for a component of the variance which is due to nonresponse. Work is currently underway to extend the bootstrap procedures to include the nonresponse adjustments as well.

4.2 Variance Estimation by Remote Access

For each dummy data file that was created, a corresponding file containing 500 sets of dummy bootstrap weights was also created. The purpose of these dummy bootstrap weight files is to allow analysts to write and test their own bootstrap variance estimation programs, which are then submitted to Statistics Canada by email and run with the master

microdata files and the corresponding master bootstrap weight files as input.

In addition to the dummy bootstrap weight files, variance estimation programs in SAS and SPSS, which can be easily adapted for particular analyses, are also provided. These programs contain macros to produce variance estimates for estimators of totals, ratios, differences of ratios, linear and logistic regressions, and generalized linear models. Macros for other types of statistics, such as quantiles, may be developed in the future.

The remote access variance capability is an important improvement in the accessibility and utility of NPHS data. Up to this point, variance estimates for analysts using the PUMFs were available only as special requests or through the use of approximate CV lookup tables. Variance estimates derived from the bootstrap weights are much more precise than those based on the CV lookup tables. In addition, CV lookup tables can be used only for certain types of estimator such as totals, proportions, ratios, and differences between estimators of these types. On the other hand, bootstrap variance estimation by remote access avoids the costs associated with custom variance requests and is likely to have a quicker turnaround time as well. Use of remote access for variance estimation requires more work on the part of the analyst; however, the SAS and SPSS examples provided are helpful and after an initial learning period they become relatively easy to use.

ACKNOWLEDGEMENT

We are grateful to Ellen Chan, a co-op student from Simon Fraser University, for preliminary work on creation of the dummy files. We also thank Bryan Lafrance, Denise Hall and Douglas Yeo for many useful comments on an earlier version of the paper.

REFERENCES

- Rao, J.N.K., Wu, C.F.J., and Yue, K. (1992). Some Recent Work on Resampling Methods for Complex Surveys, *Survey Methodology*, 18, 209-217.
- Tambay, J.-L., and Catlin, G. (1995). Sample Design of the National Population Health Survey, *Health Reports*, 7, 29-38.
- Yeo, D., Mantel, H., and Liu, T.P. (1999). Bootstrap Variance Estimation for the National Population

Health Survey, 1999 *Proceedings of the Survey Research Methods Section*, American Statistical Association, to appear.

Yung, W. (1997). Variance Estimation for Public Use Microdata Files, *Proceedings of Symposium 97, "New Directions in Surveys and Censuses"*, 91-95, Statistics Canada.