National Longitudinal Survey of Children and Youth
Primary file Cycle 5, 2002-2003
Synthetic Files for Remote Access to the Master Files

Notice: The Synthetic Files (a.k.a. Dummy files) should not be used for purposes other than to develop and test computer programs that are to be submitted by remote access. The Synthetic Files contain modified data and must never be used to produce estimates for analysis.

1. Overview

The National Longitudinal Survey of Children and Youth (NLSCY) is a unique study of Canadians from birth to adulthood. The survey collects data to support longitudinal analysis on the prevalence of various biological, social and economic characteristics and risk factors among children and youths. A principal goal of the survey is to determine the reasons for poor outcomes and predictors for good outcomes. To date, five data collection cycles have been completed: NLSCY Cycle 1 (1994-1995) covering children aged 0 to 11, NLSCY Cycle 2 (1996-1997) covering children aged 0 to 13, NLSCY Cycle 3 (1998-1999) covering children 0 to 15 years of age, NLSCY Cycle 4 (2000-2001) covering children 0 to 17 years of age, and NLSCY Cycle 5 (2000-2001) covering children 0 to 19 years of age.

The survey and research programs were developed to support evidence-based policy using a human development view of the early decades of life. They seek to answer such fundamental questions as “Can good early child outcomes of development predict later success?” and “If so, are we under-investing in children?”

Usually, researchers can use the Public Use Microdata Files (PUMFs)[1]. However, some research projects require access to more variables found only on the Master Files. While individuals can obtain additional information through special "custom" tabulations, this process does not facilitate inferential statistical analysis which is the primary objective of most researchers. To facilitate access to the master files, a Remote Data Access (RDA) procedure has been in place for researchers to obtain outputs from the true and complete dataset. To expedite the process for users, the synthetic file has been created so researchers can test and fully develop their programs before submitting them through remote access.

To obtain remote access privileges, researchers must obtain approval from Statistics Canada. Requests must be submitted and must provide information about their research in accordance with the procedures outlined within: Researchers transmit programs electronically to Statistics Canada via Internet e-mail. The programs are then be moved into the Department's internal, secure environment. Next, the code is processed on a PC, the results vetted for confidentiality,

[1] See Section 3, Other Products

and shipped back to the client.

It should be noted that the onus is with the user to submit retrieval programs that are correct and tested. Statistics Canada reviews results only for confidentiality concerns and makes no assessment whatsoever as to whether or not the submitted program has worked properly. Initially, there should be some discussion between the researcher and Statistics Canada to ensure that a copy of the software used in the submitted program is available. Statistics Canada supports SAS and SPSS. The programs are run on microcomputers.

Procedures

ProceduresThe procedures are as follows:

1. Before beginning the remote access, researchers are required to contact Statistics Canada outlining the objectives for their research. Initial contact is made to **nlscy@statcan.ca**.
2. Upon approval of their access request, researchers are given access to the synthetic files on a CD and provided with the user guide for each survey cycle requested.
3. Researchers produce and test programs (SAS or SPSS) using the synthetic file. At this stage, clients can assess the feasibility of their requests and test their programs.
4. The client sends an e-mail message to Statistics Canada (**nlscy@statcan.ca**) which contains his/her program.
5. Statistics Canada runs the program, produces the results, vets them for confidentiality, and, if needed, suppresses results that do not meet the confidentiality criteria. Any frequency table, tabular output or other result based on less than five observations are deleted from the output.
6. If a large proportion of the output requires data suppression, researchers will be notified that their program needs modification in order to comply with confidentiality restrictions.
7. The results (ASCII), including the log for the program and a notification of data suppression (if necessary) are sent back to the researcher attached to an email message.
8. Statistics Canada does not fix any errors in the program. We will send back the log of the program back to the researcher.

Guiding principles

Guiding principles 1. The minimum cost is set at $80 per submission. A typical submission includes no more than 20 tables (or cross tabulations) and/or 20 other procedures (i.e., regression). In order to reduce costs, users are encouraged to test their programs carefully before sending them to Statistics Canada for submission.

2. Before releasing the results, Statistics Canada must vet the data for confidentiality. Any outputs with a cell count less than five will be suppressed. Clients can anticipate the likelihood of suppression by examining the results of their analysis on the synthetic file. Clients are encouraged to make the necessary changes to the programs (i.e., collapsing,

grouped categories, capping variables) such that their output will not result in cell counts of less than five.

3. The client is responsible for the logical correctness of his or her work. Statistics Canada will not make any assessment whatsoever as to whether or not the program has worked properly. If the submission results in an error due to the programming, the log will be sent to the client.
4. Additional support to the researcher to help correct or modify the program may be available on a cost recovery basis.
5. Generally, Statistics Canada will respond to a remote access request within a reasonable time frame (usually 2 to 5 working days) after receipt of the client's program. This does not include delays caused by unsuccessful submissions.
6. The client is responsible for the archival of all programs and output files. Statistics Canada will only retain these electronic files for a limited time.

2. Specification of Creation of the Synthetic File

The Synthetic file represents a subset of the original file having only one record for every five records found on the Master File. Researchers can easily estimate the sample counts from the master file by supposing that the true number of sampled units would be five times those found on the synthetic file. Weights were adjusted to account for this sub-sampling of records so no adjustment is required on the researcher's part when trying to produce estimates. It should be noted however, that the process of creating the synthetic file will distort the appropriate estimated representation of responses to certain variables.

In creating the synthetic file, only one in five records from the Master File is used. This allows a greater portability of the file and will reduce the risk of disclosure by protecting the identity of the respondents whose information was used as the primary building block for the file. Each record of the synthetic file is made up of both live data and of artificial data. Live data are data provided by a respondent, while artificial data are computer-generated plausible values. One of the objectives in creating synthetic records is to identify records with similar pathways through the questionnaire, so that when random swapping is done between these similar records, the resulting synthetic record looks more like a reasonable and likely response. The following are the steps for the first phase of the process used to create the individual records of the synthetic file.

- ♦ A core set of real data is identified as the pivotal information from which all other information will be swapped. This information is best described as the essence of a respondent and is used as the primary building block for each record of the file.
  - ♦ Three classes were identified within this core - the child, the Person Most Knowledgeable (PMK) and the spouse of the PMK.

- The remaining information is grouped into cohesive blocks of variables. The criterion for grouping variables in a block is determined by their fundamental interrelationship (usually corresponds to sections of the questionnaire).

- From the core variables, respondents are grouped according to their similarities to form domains for swapping data.
  - Each block is associated with one of the three classes of variables in the core; that is, the "Child class", the "PMK class" or the "Spouse of the PMK class".
    - 14 or 15 blocks are associated to the Child
    - 5 blocks are associated with the PMK
    - 5 blocks are associated with the spouse of the PMK
  - A minimum number of respondents is required in each of these domains to perform the swapping. The minimum is 30.

- Blocks of data are then swapped among respondents of the same domain.
  - All information, other than the core data, is swapped and is done so only once. This ensures that the sample responses have the same frequency counts before and after swapping.
  - The swapping is done one block at a time. This means that consistent response patterns within the blocks are preserved, but not necessarily among swapped blocks.
  - Each record can only contribute one block of variables to any synthetic record.

The next phase in the process is to secure the identity of the respondents whose information was used as the primary building block for the file. Because live data are used, Statistics Canada is responsible for ensuring that no breach of confidentiality ensues from the release of the synthetic file. Ultimately, no original record from the Master File is left intact or recognizable. Such an assessment was performed for the release of the Public Use Microdata File (PUMF), resulting in the suppression of many survey variables, or their values being capped or otherwise modified to eliminate the chance of identifying a child or his or her family. The following steps describe the process used to reduce the risk of disclosure.

- Once the swapping is completed, the same suppressions as the PUMF are applied to the variables in the core.

- Where data were removed completely or regrouped, artificial data were generated to replace the information. The values are generated to be plausible given the information that was not suppressed.
  - In certain cases, plausible values were generated to also maintain the same proportion of response found on the master file.

The last phase in creating the synthetic file is the sub-sampling and re-weighting of the records on the file. Given the breadth of information collected in the NLSCY and the large sample size, a sub-sample makes it easier for researchers with limited computer resources to handle the file. Moreover, the possibility of linking synthetic files over the cycles for the NLSCY (i.e., a longitudinal synthetic file) is only feasible if we have a sustained strategy of risk of disclosure. Linking information from one cycle to the next increases the risk of disclosure, since the linking of information reveals even more information about respondents. Although extraordinary steps have been taken thus far to reduce the risk of disclosure, a one in five sample of the swapped records would reduce the risk of disclosure from matching records from all cycles of the survey.

The re-weighting of the synthetic file represents an extraordinary effort to try to balance the effect of the original disproportional sampling scheme on the swapped information. Although some weighted estimates will look very similar to those produced from the master file, other variables will not be as fortunate. By design, we have attempted to preserve some realistic aspect of the file as it relates to the true master file. However, users should remember that the synthetic file is not intended to produce estimates of the population despite the fact that at times the estimates may look close to the published information.

3. Other Products

To allow broad access to the data from the NLSCY, Public Use Microdata Files (PUMFs) have been produced for the primary component of Cycles 1, 2 and 3 and for the self-completed portion of the questionnaires for Cycles 2 and 3. These files go through rigorous procedures to reduce the risk of disclosure of information that could identify any of the respondents. Consequently, many survey variables may have been suppressed, capped or otherwise modified to eliminate the chance of identifying a respondent or their family. The remaining variables, except for the survey weights, have values identical to those found on the master file. For reasons of securing the confidentiality of the NLSCY respondents, the PUMFs cannot be linked between cycles. Survey weights were modified to reduce the risk of linking records from one cycle to the next and from linking information from the self-complete PUMF to the primary PUMF. These modifications represent reasonable sampling adjustments and will have a small impact on the estimates produced using either PUMF files, but will not introduce errors to the estimates.

Another venue for researchers to gain access to the NLSCY data for analysis is through the Research Data Centres (RDCs). The RDCs were created as a result of a joint task force assembled by the Social Sciences and Humanities Research Council (SSHRC) and Statistics Canada (more information is available at their WEB site at **www.sshrc.ca**). The RDCs are located throughout the country, so researchers will not have to travel to Ottawa to access Statistics Canada data. At the same time, the centres are operated in accordance with all the confidentiality rules required under the *Statistics Act*. The Research Data Centres will meet, in a single location, both the need to facilitate access to data for crucial social research and the need to protect the confidentiality and security of Canadians' information. Researchers wishing access to the confidential microdata in the Research Data Centres have to submit proposals to a

review committee operating under the auspices of SSHRC and Statistics Canada. The synthetic file could also be a valuable tool for people in the research data centres who travel a great distance to get to the RDCs and need to prepare their programs in advance.

For more information or any questions about the synthetic data files or their use should be directed to:

Client Services
Special Surveys Division
Tel: (613) 951-7355 OR 1-888-297-7355
Fax: (613) 951-3012
Internet e-mail: **ssd@statcan.ca** or **nlscy@statcan.ca**