Diversity and Sociocultural Statistics

General Social Survey

Cycle 31 : Families
Public Use Microdata File Documentation and User's Guide

Numéro au catalogue : 45250001
Volume numéro : 2019001

April 2020
Aussi disponible en français

Statistics Canada Statistique Canada

Canada

How to obtain more information

Specific inquiries about this product and related statistics or services should be directed to: Social and Aboriginal Statistics Division, Statistics Canada, Ottawa, Ontario, K1A 0T6 (E-mail: statcan.dssclientservices dssserviceaclientele.statcan@canada.ca).

For information about this product or the wide range of services and data available from Statistics Canada, visit our website at www.statcan.gc.ca or contact us by e-mail at STATCAN.infostats-infostats.STATCAN@canada.ca or by telephone from 8:30 a.m. to 4:30 p.m. Monday to Friday:

Statistics Canada National Contact Centre
Toll-free telephone (Canada and the United States):

Inquiries line	1-800-263-1136
National telecommunications device for the hearing impaired	1-800-363-7629
Fax line	1-877-287-4369

Local or international calls:

Inquiries line	1-514-283-8300
Fax line	1-613-951-0581

Depository services program

Inquiries line	1-800-635-7943
Fax line	1-800-565-7757

Accessing and ordering information

This product, Catalogue no. 45250001 Issue no. 2018001 is also available as a standard printed publication.

The printed version of this publication can be ordered by:

• Telephone (Canada and United States)	1-800-267-6677
• Fax (Canada and United States)	1-877-287-4369
• E-mail	STATCAN.infostats-infostats.STATCAN@canada.ca
• Mail	Statistics Canada R.H. Coats Bldg., 6th Floor 150 Tunney's Pasture Driveway Ottawa, Ontario K1A 0T6

• In person from authorized agents and bookstores.

When notifying us of a change in your address, please provide both old and new addresses.

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, the Agency has developed standards of service which its employees observe in serving its clients. To obtain a copy of these service standards, please contact Statistics Canada toll free at 1-800-263-1136. The service standards are also published on www.statcan.ca under "About us" > "Providing services to Canadians."

Statistics Canada
Diversity and Sociocultural Statistique

General Social Survey
Cycle 31: Families

Public Use Microdata File Documentation and User's Guide

By Pascale Beaupré

Published by authority of the Minister responsible for Statistics Canada

April 2020

Catalogue no: 45250001 Issue no. 2019001

Frequency : Occasional

Ottawa

Cette publication est aussi disponible en français (numéro de catalogue 45250001volume numéro : 2019001)

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued cooperation and goodwill.

2017 GSS: Families

User Guide for the Public Use Microdata File (PUMF)

Table of Contents

1. Introduction

This guide was prepared for users of the Public Use Microdata File (PUMF) of the 2017 General Social Survey (GSS) on *the Family*. Its objectives are to provide context and background information, to familiarize users with the content of the survey, and to describe procedures and concepts related to data quality, estimation, collection, processing and methodology.

The 2017 GSS, conducted from February 2nd to November 30th 2017, is a sample survey with cross-sectional design. The target population includes all non-institutionalized persons 15 years of age and older, living in the 10 provinces of Canada. The survey uses a new frame, created in 2013, that combines telephone numbers (landline and cellular) with Statistics Canada's Address Register, and collects data via telephone. Data are subject to both sampling and non-sampling errors.

The information in the following sections should be used to ensure a clear understanding of the basic concepts that define the data provided in the GSS Cycle 31 PUMF, the underlying methodology of the survey and the key aspects of data quality. This information will provide a better understanding of the strengths and limitations of the data, and how they can be effectively used and analyzed. The information may be of particular importance when making comparisons with data from other surveys or sources of information, and in drawing conclusions regarding change over time, or differences between sub-groups of the target population.

2. Objectives of the General Social Survey

The GSS program, established in 1985, conducts telephone surveys across the ten provinces. The GSS is recognized for its regular collection of cross-sectional data that allows for trend analysis, and its capacity to test and develop new concepts that address current or emerging issues.

The two primary objectives of the General Social Survey are:

a) To gather data on social trends in order to monitor changes in the living conditions and well-being of Canadians over time; and
b) To provide information on specific social policy issues of current or emerging interest.

To meet these objectives, data collected by the GSS comprise two components: core content and classification variables. Core content is designed to measure changes in society related to living conditions and well-being, and to supply data to inform specific policy issues. Classification variables (such as, age, sex, education and income) help delineate population groups for use in the analysis of core data.

3. Content of the 2017 GSS

The central role family plays in people's lives is indisputable. Today's family, however, must navigate through changing conjugal, family, and work trajectories. While our understanding of families in Canada has deepened considerably over the past few decades, the future of families remains a matter of interest as we see that families are becoming increasingly diverse. How many families are there in Canada? What are their characteristics and socio-economic conditions? What do families at different stages of life look like? How common are step or single-parent families? The GSS on families will provide answers to these and many other questions.

3.1 Concepts

The survey collected a large amount of data for each selected respondent as well as some information about each member of the respondent's household. The documentation for the PUMF includes an annotated list of all variables included in the files as well as the entire questionnaire. Section 3.2 of this document gives a summary of the questionnaire content. Here is a brief outline:

Entry component (respondent's date of birth)
Family origins
Leaving the parental home
Conjugal history
Intentions and reasons to form a union
Respondent's children
Fertility intentions
Maternity/parental leave
Organization and decision making within the household
Arrangements and financial support after a separation/divorce
Labour market new and education
Health and subjective well-being
Characteristics of respondent's dwelling
Characteristics of respondent of spouse/partner

3.2 Survey content

Entry component
The purpose of this section is to introduce the survey and select a respondent. A household roster is created, which assembles key demographic information on each member of the household, including age, sex, marital status and relationships to other household members. Given the theme of Cycle 31 of the GSS, relationships between each household member is established (full matrix).

Respondent's date of birth
In order to confirm that all respondents who will be answering GSS Cycle 31 are 15 years or over, a question is asked on their date of birth to confirm the age of the respondent. In addition to determining if certain questions are to be asked later in the survey, age (or date of birth) allowing validating of responses where ages are involved.

Family origins
This section examines the family origins of the respondent. It allows for detailed information on family structure experienced by respondents during their childhood to be collected. The module also includes questions about sociodemographic characteristics of the respondent's parents. The questions can also shed light on the relationship between family of origin and current family functioning.

Leaving the parental home
This section focuses on the launching of young adults from their family of origin. It takes a detailed look at the dates and reasons for leaving home as well as the circumstances under which some adults return to the parental home. This data have provided important information on new trends in family life produced by delayed leaving home among particular cohorts during times of economic recession.

Conjugal history
This section gathers core demographic information about Canadian society with respect to union (marriage, common-law unions, separation, and divorce). It determines the current legal marital status of the respondent and then collects a detailed conjugal history, including dates which determine the duration of marriages, separations and divorces, and the age at which these events occurred in the life of respondents. This section covers up to four marriages of the respondent. Some questions are also asked about the respondent's spouse or partner, such as their marital status prior to the union and their date of birth. This section also establishes whether or not children were born into each union. These data allow for rich historical analyses which are not possible using other sources.

This section also confirms the marital status of the respondent and allows for any necessary corrections.

Intentions and reasons to form a union
Why do Canadians decide to live in common-law, rather than marrying? This section captures the intentions with respect relationship. Questions are asked to respondents who are currently in a relationship as well as those who do not have a spouse or partner.

Respondent's children
This section involves another core demographic area, identifying all children of the respondent: birth or biological children, step-children, and adopted children. Questions on grandchildren are also included. This survey is in fact an important source of information on blended families in Canada. We are interested in children who are presently living both inside and outside the household as well as those who are deceased. Detailed information is collected on each child, including the child's living arrangements, time spent with each parent and, where applicable, the date when the child left home. Other questions include an in-depth look at childcare arrangements. Detailed questions ask about types of childcare, use, costs and satisfaction with childcare.

These questions have high relevance for public policy and program development, particularly for child care strategies. Information is not collected on foster children or children who were stillborn.

Fertility intentions
This module examines people's intentions to give birth to or father children, or to have more children. Other questions include what contraceptive methods they use. The data provide a clearer picture of fertility trends in Canada and their impact on natural population growth.

Maternity/parental leave
This section takes a closer look at transitions to parenthood. Detailed questions covering the period between 2012 and 2017 include the topics of leaving and returning to work, such as the use of maternity, paternity and parental leave involved when re-entering the work force. This section has high relevance for public policy and program development, particularly for parental leave programs.

Organization and decision making within the household
Living as a couple means dividing up certain tasks and making decisions in daily life. A number of surveys in other countries examine those issues. The data will help analysts and researchers understand and update information about the distribution of roles and the organization of everyday responsibilities within households. The section covers topics such as sharing housework, decision-making, managing finances, and sharing expenses.

Arrangements and financial support after a separation/divorce
A separation or divorce involves various changes and arrangements regarding child custody and spousal support. This section covers the period between 1997 and 2017. For break-ups of marriages and common-law relationships occurring during that time frame, questions determine whether arrangements were in place with respect to time spent with children, financial supports and major decision-making for children. This is followed by more in-depth questions on time spent with children by each parent.

The questions also look at the re-organization of family life for Canadians following couple dissolution. It includes in-depth questions about financial arrangements for children as well as for former spouses and common-law partners. The amount and frequency of payments made and received are detailed as well as compliance and satisfaction with arrangements.

Labour market activity and education
The modules on labour market activity and education of the respondent are used to better understand the labour market status of the couple.

Health and subjective well-being
This module covers the respondent's life satisfaction, self-rated general health, and mental health. All are important factors in assessing the well-being of Canadians.

Characteristics of respondent's dwelling
This section covers the respondent's housing characteristics with emphasis on the type of current dwelling, and kind of ownership.

Characteristics of respondent and spouse/partner
This section provides a variety of socio-demographic measures - many of which are repeated each year in the General Social Survey concerning respondents, their parents, and their spouses/partners in order to support the analysis of Canadian families and individuals. This cycle of the GSS includes place of birth, aboriginal identity and visible minority status, religion, language, and education.

Questions on income were removed from the questionnaire and these data are now obtained by merging records from the respondent's fiscal files and the GSS.

4. New content and changes in the 2011 content

The content for this survey has changed through time. Since its first iteration in 1990, it covers much of the same content, allowing for historical comparisons. There are repeated core questions (key demographic concepts), other questions more policy oriented (added or deleted depending on needs), and socio-demographic questions. For more detailed information for changes to content, please refer to the explanatory notes in the questionnaire.

4.1 Summary of changes

All new questions and revisions to some of the existing questions were tested by Statistic Canada's Questionnaire Design Resource Centre (QDRC). The questionnaire was developed based on the findings of qualitative tests (face to face interviews), and interviewer feedback. No pilot test was administered for the 2017 GSS.

1) Socio-demographic classification
In 2014, many survey specific socio-demographic questions were replaced by Statistics Canada's harmonized content questions (i.e., standardized modules for household survey variables, such as marital status, education, labour force, aboriginal identity, birth place, citizenship, self-rated health and religion). Harmonized content modules contain standard concepts, definitions, classification and wording for multiple collection modes. This new standardized content is for the most part very similar to the previous concepts used by the GSS, but in some cases required adjustments to the traditionally derived variables.

2) Income
In 2017, income questions were no longer asked. Income information was obtained instead through a linkage to tax data. Respondents were notified of the planned linkage before and during the survey. Those who objected to the linking had their objections recorded. For these individuals, no linkage to their tax data took place. Linking to tax data diminishes respondent burden and increases data quality both in terms of accuracy and in response rates. The 2017 GSS will include personal and family incomes. (See Section 7.3 for more details.)

3) Frame
The 2017 GSS on Families uses the redesigned GSS frame, which integrates data from sources of telephone numbers (landline and cellular) available to Statistics Canada and the Address Register (AR). This new frame includes "cell phone only" households, a growing population not covered by the previous Random Digit Dialling (RDD) frame. The sampling unit is also different and is defined as groups of telephone numbers. (See Section 5.3 for more details.)

4) Coding
The North American Industry Classification System (NAICS) 2012 and National Occupational Classification (NOC) 2016 were used for industry and occupation coding.

5) Processing
Most of the ongoing data processing steps are standard, including consistency edits and family edits. Two aspects of processing, however, are still relatively new and merit a more detailed description.

One is Common Tools, used across the Social, Health and Labour Statistics field at Statistics Canada. From the start of questionnaire development through processing and dissemination, these new common tools are designed to streamline questionnaire specifications and processing steps with the aim of improving efficiency, coherence and consistency across surveys. The majority of new procedures are invisible to users, except for those related to the data dictionary. All surveys processed using common tools have variable names of 8 characters or less and the following reserve codes:

6 Valid skip
7 Don't know
8 Refused
9 Not stated

6) Tax data linkage: Linkage with tax records, was successful and fiscal information was available for 84.9% of 2017 GSS respondents that did not object to the linkage, which corresponds to 83.1% of all respondents.

7) Weights
Starting in 2013, the use of a new sampling frame and a new definition of the sampling unit have led to a new weighting strategy for the GSS (see Section 8.1). Additionally, the bootstrap weighting strategy has been changed from mean bootstrap to **standard bootstrap weights** (see Appendix A for more information on how to use standard bootstrap weights).

4.2 Comparability of estimates

It is important to point out that any significant change in survey methodology (as outlined above) can affect the comparability of the data over time. It is impossible to determine with certainty whether, and to what extent, differences in a variable are attributable to an actual change in the population or to changes in the survey methodology. Consequently, at every stage of processing, verification and dissemination, considerable effort was made to produce data that are as precise in their level of detail, and to ensure that the published estimates are of good quality in keeping with Statistics Canada standards.

Like other GSS cycles, trend monitoring is an important component of the 2017 GSS on Families. Analysts can count on the same concepts and high level indicators of activities to make comparisons between Cycle 25 and earlier iterations.

5. Survey and sample design

Data for 2017 GSS on Families was collected from February 2 to November 30, 2017. Please see the following sections for descriptions of the target population, stratification, the frame, the sampling strategy, the sample size and sample allocation.

5.1 Target population

The target population for the 2017 GSS included all persons 15 years of age and older in Canada, excluding:

1. Residents of the Yukon, Northwest Territories, and Nunavut; and
2. Full-time residents of institutions.

5.2 Stratification

In order to carry out sampling, each of the ten provinces were divided into strata (i.e., geographic areas). Many of the Census Metropolitan Areas[1] (CMAs) were each considered separate strata. This was the case for St. John's, Halifax, Saint John, Montreal, Quebec City, Toronto, Ottawa, Hamilton, Winnipeg, Regina, Saskatoon, Calgary, Edmonton and Vancouver.

All CMAs not on this list are located in Quebec, Ontario and British Columbia, with the exception of Moncton. Three more strata were formed by grouping the remaining CMAs (except Moncton) in each of Quebec, Ontario and British Columbia. Finally, the non-CMA areas of each of the ten provinces were also grouped to form ten more strata, for a total of 27 strata. Moncton was added to the non-CMA stratum for New Brunswick.

5.3 Frame

The survey frame was created using two different components:

- Lists of telephone numbers in use (both landline and cellular) available to Statistics Canada from various sources (telephone companies, Census of population, etc.);
- The Address Register (AR): List of all dwellings within the ten provinces.

The Address Register (AR) was used to group together all telephone numbers associated with the same valid address. About 86% of available telephone numbers were linked to the AR. The records resulting from this linkage could possess more than one telephone number (grouped by the address). The other 14% of telephone numbers not linked to the AR were also included on the frame[2]. The combination of those two components resulted in the survey frame. The rationale for using all the telephone numbers (linked and not linked to the AR) was to ensure a good coverage of all households with telephone numbers.

When more than one telephone number was attached to a record, they were sorted by source and by type of telephone number (landline telephone numbers first and cellular telephone numbers last). The first telephone number was considered the best telephone number available to reach the household.

Please note that for the remaining sections of this document, the word "record" will refer to the grouping of telephone numbers that consists of our sampling unit on the survey frame.

[1] Based on 2011 Census geography

[2] About 9% of these telephone numbers were grouped using address information from administrative sources. Each of the remaining telephone numbers constitutes a single record on the frame.

5.4 Sampling strategy

Each record in the survey frame was assigned to a stratum within its province. A simple random sample without replacement of records was next performed in each stratum.

The frame for GSS was created using several linked sources, such as the Census of population, administrative data files and billing files. Compared to the previous random digit dialing frame, coverage was improved (though over coverage and under coverage may still exist). All respondents in the ten provinces were rostered and interviewed by telephone. Households without telephones were excluded from the survey population. Survey estimates were adjusted (weighted) to represent all persons in the target population, including those not covered by the survey frame.

For the 2017 GSS, 91.8% of the selected telephone numbers reached eligible households. To be eligible, a household had to include at least one person 15 years of age or older. During collection, households that did not meet the eligibility criteria were terminated after an initial set of questions.

A respondent was then randomly selected from each household to participate in a telephone interview.

5.5 Sample size and allocation

The target sample size (i.e. the desired number of respondents) for the 2017 GSS was 20,000 while the actual number of respondents was 20,602. For each province, minimum sample sizes were determined that would ensure certain estimates would have acceptable sampling variability at the stratum level. Once these stratum sample size targets had been met, the remaining sample was allocated to the strata in a way that balanced the need for precision of both national-level and stratum-level estimates.

6. Collection and response rate

6.1 Collection

Data for the 2017 GSS was collected via computer assisted telephone interviews (CATI). Respondents were interviewed in the official language of their choice. Proxy interviews were not permitted.

All interviewing took place using centralized telephone facilities in five of Statistics Canada's regional offices, with calls being made from approximately 9:00 a.m. to 9:30 p.m. Mondays to Fridays. Interviewing was also scheduled from 10:00 a.m. to 5:00 p.m. on Saturdays and 1:00 p.m. to 9:00 p.m. on Sundays. The five regional offices were: Halifax, Sherbrooke, Sturgeon Falls, Winnipeg and Edmonton. Interviewers were trained by Statistics Canada staff in telephone interviewing techniques using CATI, as well as in survey concepts and procedures. The majority of interviewers had experience interviewing for previous GSS cycles.

Interviewers were instructed to make all reasonable attempts to obtain a completed interview with the randomly selected member of the household. Those who at first refused to participate were re-contacted up to two more times to explain the importance of the survey and to encourage their participation. For cases in which the timing of the interviewer's call was inconvenient, an appointment was arranged to call back at a more convenient time. For cases in which there was no one home, numerous call backs were made.

Interviewer manuals are not included in this documentation package but can be made available by contacting Statistics Canada (see Section 10).

6.2 Response rate

The overall response rate for the 2017 GSS was 52.4%.

The 2017 sample was selected using the new GSS frame, which necessitated some adjustments in the methodology used to calculate the response rates. Addition of “cell phone only” households to the frame was essential since this population constitutes a constantly growing portion of the population and coverage had been steadily declining with the previous frame. While the addition of these households is necessary for coverage of the Canadian population, this population is harder to reach. Another factor that affects comparability of the response rate over time is the way in which status (in-scope, out-of-scope) is determined under the new design.

7. Processing

7.1 Data capture

Using CATI, responses to survey questions were entered directly into computers as the interview progressed. The CATI data capture program allowed a valid range of codes for each question, had built in edits, and automatically followed the flow of the questionnaire. The data output was transmitted electronically to Ottawa.

7.2 Coding

Several questions allowed for write-in responses. These responses were coded into existing categories (where a match was possible), grouped into new categories or left in “other-specify” (if a match with an existing category was not possible or the frequencies were too small to create a new category). Where possible (e.g., occupation, industry, language, education, country of birth, religion), coding followed standard classification systems used by the population census and Statistics Canada’s harmonized content program.

7.3 Edit and imputation

All survey records were subjected to computer edits throughout the course of the interview. The CATI system identified “out-of-range” values as they were entered. As a result, the interviewer could immediately resolve such problems with the respondent. If interviewers were unable to correctly resolve the detected errors, they could bypass the edit and forward the data to Head Office for resolution. Interviewer comments were reviewed and taken into account during Head Office editing.

Head Office edits performed the same checks as the CATI system, as well as more detailed edits. Records with missing or incorrect information were, in a small number of cases, completed, corrected deterministically, or imputed from other information on the questionnaire.

The flow editing carried out by head office followed a ‘top down’ strategy, in that whether or not a given question was considered ‘on path’ was based on the response codes to the previous questions.

If the response to a given question was missing and ‘on path’, two codes are considered:

- If the respondent was clearly identified as belonging to a sub-population for which the current question was not applicable, the current question was coded as ‘Valid Skip’, i.e., 6 (96 or 996, etc.).
- When the question was asked, and the respondent reported ‘don’t know’ or preferred not providing an answer to the question, the responses to the current question were retained, though ‘Don’t Know’ was recoded as 7 (97 or 997, etc.) and refusals were recoded as 8 (98 or 998, etc.).

However, if the response codes to the previous questions indicated that the current question was ‘off path’, in order words, the respondent skipped a question which should have been asked, it was coded as ‘Not

Stated,' i.e., 9 (99 or 999, etc.). This happens when there are inconsistencies (eg. when a respondent reports contradictory responses) or errors within the questionnaire.

Non-response was not permitted for questions required for weighting. In the very rare case where the sex or the age of the respondent was missing, values were imputed. The imputation was based on a detailed examination of the data and of other useful variables, such as the age and sex of other household members, age when family events have been experienced, and the interviewer's comments.

In 2017, personal income questions were not asked as part of the survey. Personal income information was obtained instead through a linkage to tax data for respondents who did not object to this linkage. Income information was obtained from the 2017 T1FF for 83.1% of the respondents. As in GSS 2016, the **family income** (i.e., linking directly to a variable on the T1FF that corresponds to the census family income) was used for GSS 2017. In total, a family income value was obtained for 82.6% of households. Missing information for all other respondents was imputed.

Finally, the timing that a respondent experienced a given event is a key piece of information. It is from the reported dates that one can recreate the individual's life trajectory. First, the month and year of a particular event is asked. If the respondent is unable or refuses to provide the year, he or she is asked the age at which the event occurred. To provide as much information as possible in the public data file, some derived variables were created to determine the respondent's age at a specific event. For some situations, these variables require imputation. Below are different reporting scenarios according to the decision made in regards to imputation.

Scenarios	Reported Event Month	Reported Event Year	Reported Age of Event	Value of Derived Age at Event
1	Valid	Valid	n/a	Exact age
2	Missing	2017	n/a	Halfway between the age in January and the age at time of survey
3	Missing	< 2017	n/a	If month of birth between January and June, half a year was added to respondent's age at time of event If month of birth between July and December, half a year was subtracted from respondent's age at time of event
4	Valid	Missing	Valid	Exact age
5	Missing	Missing	Valid	Scenario 2 or 3
6	Valid	Missing	Missing	Missing
7	Missing	Missing	Missing	Missing

While imputed values are available for the derived age variable, the individual month, year and age variables included on the PUMF are as they were reported during collection.

7.4 Creation of combined and derived variables

A number of variables on the file were derived from information collected in the questionnaires. In some cases, the derived variables are straightforward and involve collapsing categories. In other cases, two or more variables were combined to create a new variable. The data dictionary identifies which variables are derived and the nature of their derivation.

8. Estimation

When a probability sample is used, as is the case for the GSS, the principle behind estimation is that each person selected in the sample represents (in addition to himself or herself) several other persons not in the sample. For example, in a simple random sample of 2% from a population size of 1,000, each person in the sample represents 50 persons in the population (themselves plus 49 others). The number of persons represented by a given person in the sample is usually known as the weight or weighting factor of the sampled person.

There are two microdata files from which GSS Cycle 31 estimates can be made: The Analytical File and the PUMF, which will be published in 2020. The PUMF contains responses and associated information from 20,602 respondents.

One weighting factor was placed on the PUMF and is explained below:

WGHT_PER: This is the basic weighting factor for analysis at the person level, i.e. to calculate estimates of the number of persons (non-institutionalized and aged 15 or over) having one or several given characteristics. WGHT_PER should be used for all person-level estimates.

In addition to the estimation weights, bootstrap weights have been created for the purpose of design-based variance estimation.

8.1 Weighting of persons

As mentioned previously, the records on the survey frame are groups of telephone numbers. A simple random sample of those records was selected in each stratum. Therefore, each record within a stratum has an equal probability of selection.

This probability is equal to:

$$\frac{Number\ of\ records\ sampled\ in\ the\ stratum}{Number\ of\ records\ in\ the\ stratum\ from\ the\ survey\ frame}$$

1) Initial weight calculation

Certain households in the survey frame had a probability of being reached through more than one record. This was possible since groupings of telephone numbers were subject to error.

As mentioned previously, telephone numbers belonging to the same valid address were grouped together on the survey frame. However, for a few cases, the grouping of those telephone numbers might be erroneous (i.e. all the telephone numbers grouped together do not belong to the same household). In addition the remaining telephone numbers that could not be linked to addresses were also included in the frame. It is possible that some of those telephone numbers could reach households already covered by the telephone numbers linked to addresses.

As a result, a series of questions were added to the survey to establish the prevalence of these situations. Several adjustments were made to the initial probability of selection to account for the fact that such households had a higher probability of being selected (i.e. they could be contacted through more than one group of telephone numbers). Therefore, the initial weight is the inverse of this adjusted probability of selection. The resulting initial weight is a household weight.

2) Removal of out-of-scope records

Telephone numbers associated with businesses, institutions or other out-of-scope dwellings, as well as numbers not in service or any other non-working numbers are all examples of out of-scope telephone numbers for this survey. Records with all telephone numbers out-of-scope are simply removed from the process, leaving only in-scope records in the sample. These in-scope records keep the same initial weight as described in the previous step.

3) Three-stage non-response adjustment

Weights for responding telephone numbers were adjusted to represent non-responding telephone numbers.

Non-responding telephone numbers were grouped into three types: those with some auxiliary information available (in particular, a complete roster of household members), those with auxiliary information from various sources available to Statistics Canada and those with no auxiliary information.

This non-response adjustment was done in three stages. In the first stage, adjustments were made for complete non-response (i.e., households for which no auxiliary information was available). This was done independently within each stratum. In the second stage, adjustments were made for non-response with auxiliary information from sources available to Statistics Canada. These households had some auxiliary information which was used to model propensity to respond. In the third stage, adjustments were made for partial non-response. These households had some auxiliary information which was used to model propensity to respond. The last two adjustments were done independently within each wave. The combination of these three adjustments is referred to as Factor 1.

Non-responding telephone numbers were then dropped.

4) Person weight calculation

A person weight was calculated for the respondent by multiplying the household weight by the number of persons 15 years of age or older in the household.

This step produces a person weight, which can be calculated as:

Initial Household Weight x Factor 1 x Number of Eligible Household Members.

5) Adjustment of person weights to external totals

The person weights were adjusted several times using a raking ratio procedure. This procedure ensures that, based on the survey's total sample, estimates are produced that match certain external reference totals. Two sets of external references were used for this survey, both of them population totals: for stratum (geographic), and for age-sex groups by province.

It should be noted that persons living in households without telephone service (or telephone service not covered by the frame) are included in the external references even though such persons were not sampled.

5a) Stratum Adjustment

An adjustment was made to the person weights for records within each stratum (geographic) in order to make population estimates consistent with the corresponding projected population counts. This was done by multiplying the person weight for each record within the stratum by the following ratio:

$$\frac{Projected\ Population\ Count\ for\ the\ Stratum}{Sum\ of\ the\ Person\ Weights\ for\ the\ Stratum}$$

5b) Province - age - sex adjustment

The next weighting step adjusts the weights so they agree with projected province-age-sex population distributions. Projected population counts were obtained for males and females within the following sixteen age groups:

15-19	20-24	25-29	30-34
35-39	40-44	45-49	50-54
55-59	60-64	65-69	70-74
75-79	80-84	85-89	90 +

For each of the resulting classifications, the person weights for records within the classification were adjusted by multiplying by the following ratio:

$$\frac{Projected\ Province - Age - Sex\ Group\ Population\ Count}{Sum\ of\ the\ Province - Age - Sex\ Group\ Person\ Weights}$$

When sample sizes were small, adjacent age group data for the same province and sex were combined before this adjustment was made.

5c) Raking Ratio Adjustments

The weights of each respondent were adjusted several times using a raking ratio procedure. This procedure ensured that estimates produced for stratum and province-age-sex totals would agree with the external reference totals. This adjustment was made by repeating steps 5a) and 5b) of the weighting procedures until each repetition of the step made a minimal adjustment to the weights.

6) Final person weight

The weight produced at the end of step 5) is the final person weight (WGHT_PER) placed on the PUMF.

8.2 Weighting policy

Users are cautioned against releasing unweighted tables or performing any analysis based on unweighted survey results. As was discussed in Section 8.1, several weight adjustments were performed that depended on the province, stratum, age and sex of the respondent. Sampling rates as well as non-response rates varied significantly from province to province, and non-response rates varied with demographic characteristics. For example, non-respondents are often more likely to be males and more likely to be younger. In the responding sample, 3.7% were males between the ages of 15 and 24, while in the overall population, approximately 7.5% were males between 15 and 24. Therefore, it is clear that unweighted sample counts cannot be considered representative of the survey target population.

The total number of households in the survey's scope was estimated at 39,323**.** Among these households, 20,602 usable responses were obtained, which gives a response rate of 52.4%. The distribution of the non-response and response categories is given in the table below:

Source	Number	%
1. Household non-response	14,687	37.4
2. Refusal (person level)	1,525	3.9
3. Non-response (person level)	2,509	6.4
4. Response (CATI)	20,602	52.4

Total Households	**39,323**	**100.0**

In all, the number of non-response cases is estimated to 18,721 cases. Categories 2 and 3 show non-response occurring after the respondent was selected. The “Non-response’ category (3) includes cases where no response could be obtained because of language difficulties or other reasons.

8.3 Types of estimates

Two types of “simple” estimates are possible from the results of the General Social Survey. These are qualitative estimates (estimates of counts or proportions of people possessing certain qualities or characteristics) and quantitative estimates involving quantities of averages. More complex estimation and analyses are covered in Section 8.5.

8.3.1 Qualitative estimates

The target population for the GSS was non-institutionalized persons aged 15 and older, living in the ten provinces. Qualitative estimates are estimates of the number or proportion of this target population possessing certain characteristics. The number of people (7,019,338) who describe their state of health as excellent (SRH_Q110 = 1) is an example of this kind of estimate. These estimates are readily obtained by summing the person weights (WGHT_PER) for the records possessing the characteristic of interest. This estimate does not, however, adjust for non-response to the question in any way.

If we make the assumption that those who either refused to answer the question or who responded ‘Don't Know’ have the same distribution as those who responded, then an adjusted estimate can be made. To do this, the proportion of the target population with this characteristic is estimated by excluding respondents with a ‘Not Stated’ or ‘Don't Know’ answer to question SRH_Q110 and calculating the ratio of the total of the weights of those respondents who answered that their state of health was ‘Excellent' (SRH_Q110=1) to that of all respondents who answered the question (SRH_Q110=1, 2, 3, 4, or 5). This proportion is then multiplied by the size of the target population to produce the final estimate (it should be noted that this adjustment does not have to be done, but it can be if needed):

$$7{,}072{,}579 \;=\; 30{,}302{,}287 \;\text{x}\; \frac{7{,}019{,}338}{30{,}074{,}178}$$

30,302,287 is the estimated number of persons aged 15 and over in the population (target population). 30,074,178 is the sum of the weights of all respondents who answered question SRH_Q110 (i.e. SRH_Q110 = 1,2,3,4 or 5). When the proportion of responses that are ‘Don't Know’ or ‘Refused’ are high, the differences between the two estimates will be large.

8.3.2 Quantitative estimates

Some variables on the General Social Survey PUMF are quantitative in nature (e.g. age, number of weeks worked in the past 12 months). From these variables, it is possible to obtain estimates such as the average number of weeks worked in the past 12 months. These quantitative estimates are of the following ratio form:

$$\text{Estimate (average)} = \frac{X}{Y}$$

The numerator (X) is a quantitative estimate of the total for the variable of interest (for example, the number of weeks worked in the past 12 months) for a given sub-population (for example, males who worked in the past 12 months). In this example, X would be calculated by multiplying the person weight (WGHT_PER) by the variable of interest (NWE_Q110) when it is known, 1 <= NWE_Q110 <= 52, (i.e. not equal to ‘96’, ‘97’, ‘98’ or ‘99’), and summing this product over all records for males who worked i.e. SEX=1 and (1 <= NWE_Q110 <= 52), which yields 510,542,546.

The denominator (Y) is the qualitative estimate of the number of persons within that sub-population (males who worked in the past 12 months). In this example, Y would be calculated by summing the person weight (WGHT_PER) over all male respondents with 1 <= NWE_Q110 <= 52, yielding 11,411,714.

The two estimates X and Y are derived independently and then divided to provide the quantitative estimate. The average number of weeks is then calculated to be:

$$\frac{510{,}542{,}546}{11{,}411{,}714} = 44.7$$

8.4 Guidelines for analysis

As detailed in Section 5 of this document, the respondents from the GSS do not form a simple random sample of the target population. Instead, the survey had a complex design, with stratification, multiple stages of selection and unequal selection probabilities for respondents. Using data from such complex surveys presents analytical challenges because the survey design and selection probabilities affect the estimation and variance calculations that should be used.

The GSS used a stratified design, with significant differences in sampling fractions between strata. Thus, some areas were over-represented in the sample (relative to their populations) while some other areas were relatively under-represented; this means that the unweighted sample was not representative of the target population, even if there was no non-response. Non-response rates may vary by demographic group, making the unweighted sample even less representative.

The survey weights must be used when producing estimates or performing analyses in order to account as much as possible for the geographic over- and under-representation and for the under- or over-representation of age-sex groups in the unweighted file. While many analysis procedures found in statistical packages allow weights to be used, the meaning or definition of the weight in these procedures often differs from that which is appropriate in a sample survey framework. As such, while in many cases the estimates produced by the packages are correct, the variances that are calculated are almost meaningless.

For many analysis techniques (for example linear regression, logistic regression, estimation of rates and proportions and analysis of variance), a method exists which can make the variances calculated by the standard packages more meaningful. If the weights for the data, or for the subset of the data that is of interest, are rescaled so that the average weight is one (1), then the variances produced by the standard packages will be more reasonable; they still will not take into account the stratification and clustering of the sample's design, but they will take into account the unequal probabilities of selection. This rescaling can be accomplished by dividing each weight by the overall average weight before the analysis is conducted. Section 9 describes sampling variability and data reliability in more detail.

8.5 Estimating number of persons by using WGHT_PER

As previously mentioned, a basic person weight has been assigned to each sampled individual and, as described in Section 7.1, these weights have been adjusted to reflect the age and sex composition of the various provincial populations as estimated by Statistics Canada for each month covered by Cycle 31.

$$\sum_{i=1}^{n=20{,}602} WGHT_{PER} = 30{,}302{,}287^{1}$$

[1] Estimate of the number of persons aged 15 and over in the population

In general, when an estimate is based on the person as the unit of observation, WGHT_PER should be used. Examples of this are the average number of weeks worked by persons aged 25 to 29 years old, the

percentage of persons who worked at a job or business last week, and the number of people aged between 25 and 44 who are currently attending school, college, CEGEP or university.

The last example would be calculated as follows: WGHT_PER would be summed up for all records on the main file with 2<= AGEGR10 <= 3 and ESC1_01 = 1, giving an estimate of 887,782 persons aged 25 to 44 who are currently attending school, college, CEGEP or university.

9. Release guidelines and data reliability

It is important for users to become familiar with the contents of this section before publishing or otherwise releasing any estimates derived from the General Social Survey PUMF.

This section of the documentation provides guidelines to be followed by users. With the aid of these guidelines, users of the PUMF should be able to produce figures consistent with those produced by Statistics Canada and in conformance with the established guidelines for rounding and release. The guidelines include four broad sections: Minimum Sample Sizes for Estimates; Sampling Variability Policy; Sampling Variability Estimation; and Rounding Policy.

9.1 Minimum sample size for estimates

Users should determine the number of records on the PUMF which contribute to the calculation of a given estimate. This number should be at least 15 in the case of persons or households. When the number of contributors to the weighted estimate is less than 15, the weighted estimate should generally not be released regardless of the value of the coefficient of variation. If it is, it should be with great caution and the insufficient number of contributors associated with the estimate should be prominently noted.

9.2 Sampling variability guidelines

The estimates derived from this survey are based on a sample of persons. Somewhat different figures might have been obtained if a complete census had been taken using the same questionnaire, interviewers, supervisors, processing methods, etc. as those actually used. The difference between the estimates obtained from the sample and the results from a complete count taken under similar conditions is called the sampling error of the estimate.

Errors which are not related to sampling may occur at almost every phase of a survey operation. Interviewers may misunderstand instructions, respondents may make errors in answering questions, the answers may be incorrectly entered into the CATI system, and errors may be introduced in the processing and tabulation of the data. These are all examples of non-sampling errors.

Over a large number of observations, randomly occurring errors will have little effect on estimates derived from the survey. However, errors occurring systematically will contribute to biases in the survey estimates. Considerable time and effort was made to reduce non-sampling errors in the survey. Quality assurance measures were used at each step of the data collection and processing cycle to monitor the quality of the data. These measures included the use of highly skilled interviewers, extensive training of interviewers with respect to the survey procedures and questionnaire, observation of interviewers to detect problems of questionnaire design or misunderstanding of instructions, and coding and edit quality checks to verify the processing logic.

9.2.1 Non-sampling errors

A major source of non-sampling errors in surveys is the effect of non-response on the survey results. The extent of non-response varies from partial non-response (failure to answer one or a few questions) to total non-response. Total non-response occurred because either the interviewer was unable to contact the respondent, no member of the household was able to provide the information (perhaps due to a language problem), or the respondent refused to participate in the survey. Total non-response was handled by adjusting the weight of households who responded to the survey to compensate for those who did not respond.

In most cases, partial non-response to the survey occurred when the respondent did not understand or misinterpreted a question, refused to answer a question, or could not recall the requested information.

9.2.2 Sampling errors

Since it is an unavoidable fact that estimates from a sample survey are subject to sampling error, sound statistical practice calls for researchers to provide users with some indication of the magnitude of this sampling error.

Although the exact sampling error of the estimate, as defined above, cannot be measured from sample results alone, it is possible to estimate a statistical measure of sampling error, the standard error, from the sample data. Using the standard error, confidence intervals for estimates (ignoring the effects of non-sampling error) may be obtained under the assumption that the estimates are normally distributed about the true population value. The chances are about 68 out of 100 that the difference between a sample estimate and the true population value would be less than one standard error, about 95 out of 100 that the difference would be less than two standard errors, and virtually certain that the differences would be less than three standard errors.

Since the absolute size of the sampling error of an estimate is often less important than its relative size (relative to the estimate itself) the standard error is not always the best measure of sampling error. For example, a standard error of 10 for an estimate of 20 would generally be taken as indicating that the estimate is a poor one, while the same standard error for an estimate of 1,000 would generally indicate a good estimate. For this reason the size of the sampling error is often expressed relative to the size of the estimate, as the coefficient of variation (c.v.). The coefficient of variation of an estimate is obtained by dividing the standard error of the estimate by the estimate itself, and the resulting fraction is usually expressed as a percentage. In the above example, the first estimate has a c.v. of 50% (10/20), while the second has a c.v. of 1% (10/1,000).

The choice between using the standard error or the CV as a measure of sampling variability is one the user should make based on his/her specific analysis. Guidelines for publishing estimates using the CV are given in the next section. With enough observations, the user can proceed to calculating variances and coefficients of variation using the bootstrap weights provided with the data (see Section 9.2.3 for guidelines to follow when using coefficients of variation and Appendix A for more details on the appropriate software to use for bootstrap weights).

9.2.3 Guidelines for release of estimates

When considering releasing *and*/*or* publishing an estimate from the PUMF, users should consult the table below and follow the guideline that matches the coefficient of variation of the estimate.

Type of Estimate	Coefficient of Variation	Policy Statement
1.With Moderate Sampling Variability	0.0% to 16.5%	Estimates can be considered for general unrestricted release. No special notation is required.
2. With High Sampling Variability	16.6% to 33.3%	Estimates can be considered for general unrestricted release but should be accompanied by a warning cautioning users of the high sampling variability associated with the estimates.
3. With Very High Sampling Variability	33.4% or over	Estimates should generally not be released, but when they are it should be with great caution and the very high sampling variability associated with the estimate should be prominently noted.

9.3 Variance estimation using bootstrap weights

In order to determine the quality of the estimate and to calculate the CV, the standard deviation must be calculated. Confidence intervals also require the standard deviation of the estimate. The GSS uses a multi-stage survey design and calibration, which means that there is no simple formula that can be used to calculate variance estimates. Therefore, an approximate method was needed. The bootstrap method is used because the sample design and calibration needs to be taken into account when calculating variance estimates. With the use of available software to compute variances and the help of bootstrap weights (discussed in the next subsection), the method is fairly easy for users. For more information on how to use the standard bootstrap weight, refer to Appendix A.

This technique involves dividing the records on the microdata file into subgroups (or replicates) and determining the variation in the estimates from replicate to replicate. The replicates are formed by selecting, independently within each stratum, a simple random sample with replacement of n-1 of the n units in the sample. Note that since the selection is with replacement, a unit may be chosen more than once. A bootstrap weight based on the bootstrap sample is calculated for each sample unit in the stratum. This process (selecting simple random samples, recalculating weights for each stratum) is repeated B times, where B is large, yielding B different initial bootstrap weights. The GSS typically uses B=500, to produce 500 bootstrap weights.

These weights are then adjusted according to the same weighting process as the regular person weights: non-response adjustment, calibration and so on. The end result is 500 final bootstrap weights for each unit in the sample. The variation among the 500 possible estimates based on the 500 bootstrap weights is related to the variance of the estimator based on the regular weights and can be used to estimate it.

9.4 Rounding

In order for estimates produced from the GSS microdata files to correspond to those produced by Statistics Canada, users are urged to adhere to the following guidelines regarding the rounding of such estimates. It may be misleading to release unrounded estimates, as they imply greater precision than actually exists.

9.4.1 Rounding guidelines

1) Estimates of totals in the main body of a statistical table should be rounded to the nearest thousand using the normal rounding technique (see definition in Section 9.4.2).

2) Marginal sub-totals and totals in statistical tables are to be derived from their corresponding unrounded components and then are to be rounded themselves to the nearest thousand units using normal rounding.

3) Averages, proportions, rates and percentages are to be computed from unrounded components and then are to be rounded themselves to one decimal using normal rounding.

4) Sums and differences of aggregates and ratios are to be derived from corresponding unrounded components and then rounded to the nearest thousand units or the nearest one decimal using normal rounding.

5) In instances where, due to technical or other limitations, a different rounding technique is used, resulting in estimates different from Statistics Canada estimates, users are encouraged to note the reason for such differences in the released document.

9.4.2 Normal rounding

In normal rounding, if the first or only digit to be dropped is 0 to 4, the last digit to be retained is not changed. If the first or only digit to be dropped is 5 to 9, the last digit to be retained is raised by one. For example, the number 8499 rounded to thousands would be 8000 and the number 8500 rounded to thousands would be 9000.

10. Additional information

For additional information please contact:

Survey Manager
Pascale Beaupré
Diversity and Sociocultural Statistics
(613) 854-0938
pascale.beaupre@canada.ca

Data from the survey is available through published reports, special request tabulations, and the microdata file. The microdata file will be available from the Social and Aboriginal Statistics Division of Statistics Canada. Tabulations can be obtained at a cost that will reflect the resources required to produce the tabulation.

To receive a copy of the microdata file or to order special tabulations, please contact:

Client Services
Diversity and Sociocultural Statistics
Statistics Canada
statcan.dssclientservices-dssserviceaclientele.statcan@canada.ca

Appendix A – Tips for using GSS standard bootstrap weights

A survey weight variable with a corresponding set of 500 standard bootstrap weight[3] variables are provided with many GSS microdata files in order that a full design-based approach may be taken for doing analysis with the data.

A design-based approach to analysis first involves using the survey weight variable for obtaining weighted estimates of the quantities of interest. Then, additional information about the survey design is used in order to make estimates of the variances[4] (and covariances) of these estimated quantities. In the case of many GSS microdata files, this additional information is in the form of 500 survey bootstrap weight variables. The design-based estimates and variance estimates can then be used for making the inferences required in the analysis.

The form of a bootstrap variance estimate can be described briefly as follows:

Let $\hat{\beta}$ be the weighted estimate of quantity of interest, β, computed using the survey weight variable w, and let $\hat{\beta}^{(b)}$ be an estimate obtained in exactly the same manner, except for substituting the bth bootstrap weight variable $w^{(b)}$ for the survey weight variable w, b=1,2,…500. This yields the bootstrap estimates $\hat{\beta}^{(1)},\ldots,\hat{\beta}^{(500)}$ of β. Then the usual bootstrap estimate of the variance of $\hat{\beta}$ is

$$\hat{V}_B(\hat{\beta}) = \frac{1}{500}\sum_{b=1}^{500}\left(\hat{\beta}^{(b)} - \hat{\beta}\right)^2. \qquad (1)$$

If $\hat{\beta}$ is a vector instead of a single value, such as if $\hat{\beta}$ is the set of coefficients of a model, then the matrix of estimates of the variances and covariances of the elements of $\hat{\beta}$ is $\hat{V}_B(\hat{\beta}) = \frac{1}{500}\sum_{b=1}^{500}\left(\hat{\beta}^{(b)} - \hat{\beta}\right)\left(\hat{\beta}^{(b)} - \hat{\beta}\right)'$. (The value "500" in the formula is due to the fact that we have 500 different series of bootstrap weights. If the number of bootstrap samples should change from 500, then the values in formula (1) would need to change.)

Survey bootstrapping is just one replication approach that may be used in order to obtain design-based variance estimates with survey data. While several commercial software packages for design-based analysis offer replication approaches for variance estimation, they usually do not specify bootstrapping as one of these approaches. However, due to the similarity in the form of the variance estimate for the bootstrap and for the particular replication method called BRR, programs that can carry out variance estimation by this latter approach with user-supplied replication weights can be used to obtain bootstrap variance estimates[5]. In particular, in these software, the 500 bootstrap weights provided in the GSS microdata files need to be designated as 500 BRR weights.

In the sections below, instructions will be given for implementing bootstrap variance estimation with GSS microdata, using 3 different commercial software packages that can carry out some design-based analysis for BRR: Stata 9 or 10, SUDAAN and WesVar. In all GSS cycles where bootstrap weights are provided,

[3] Since 2013, GSS has been using standard bootstrap weights. Special attention should be given to formula (1) as it is different from the formula for the mean bootstrap weights.
[4] The variance that is estimated in a design-based approach is the variability in an estimate due to re-sampling by exactly the same design from the same finite population.
[5] For a more detailed description see Gagné, Keown and Roberts (2014).

the names given to these bootstrap variables in the user documentation are wtbs_001 to wtbs_500[6]. The name of the survey weight variable is usually wght_per.

Stata 12

Beginning with Version 9, the commercial software package Stata added some replication approaches for carrying out design-based variance estimation in its survey analysis commands. Moreover, Stata 12 release included approaches specifically designed to work with Statistics Canada bootstrap weights. One replication approach offered is the bootstrap approach, and it is this approach that would be specified when analyzing GSS data. In order to specify this approach, the following is recommended:

1. Before using any of the survey analysis commands, use a "svyset" statement to declare the data to be survey data, to designate the variables that contain information about the survey design and to specify the method for variance estimation. Settings made by "svyset" are saved with a dataset when (or if) a dataset is saved. The form of the svyset statement to be used with a GSS analysis dataset would have the following form:

svyset [pweight=wght_per], bsrweight(wtbs_001-wtbs_500) vce(bootstrap) dof(500) mse

Declaring **pweight=wght_per** tells Stata that the survey weight (which is often called the probability weight) is the variable wght_per.

The option **vce(bootstrap)** states that the variance estimation approach to use is bootstrap.

The option **bsrweight (wtbs_001-wtbs_500)** states that the names of the bootstrap weight variables are **wtbs_001, wtbs_002, …, wtbs_500.** This option can also be designated as **bsrweight (wtbs_*)** provided there are no variables other than the bootstrap weight variables whose names begin with "wtbs_".

The **dof(500)** option sets the degrees of freedom to the default. i.e. the number of bootstrap weights. The number of primary sampling units containing sample in the population being analyzed minus the number of strata containing sample could be a good approximation. However, this information is rarely known by researchers. Most of the time, using the number of bootstrap weights will have negligible impact on the results.

Finally, the **mse** option tells Stata to calculate the variance using squared differences between bootstrap estimates and the full-sample estimate of the quantities of interest, as shown in equation (1). If this option is not included, Stata uses squared differences between each bootstrap estimate and the mean of all the bootstrap estimates. Both approaches should yield approximately the same result.

2. There is an extensive list of survey analysis commands in Stata, which take a design-based approach in their computations. These commands, described in the Stata documentation, are implemented through the use of the "svy" prefix along with the names of other estimators. For example, **svy: mean** is the command for estimating population and subpopulation means and estimates of variability taking a design-based approach. When the **svyset** statement precedes all survey commands, the survey commands do not have to contain any information about the design-based approach to be taken. It should be noted that, even though most of the commands that allow the "svy" prefix are also the names of commands for non-survey data, what is estimated, what options are available and what can be done through post-estimation change when the "svy" prefix is added.

[6] Please note that in previous GSS cycles (Cycle 26 and earlier), the variables wtbs_001 to wtbs_500 were mean bootstrap weights. Beginning with cycle 27 of GSS (2013), the variables wtbs_001 to wtbs_500 are standard bootstrap weights.

SUDAAN

SUDAAN is a commercial software package developed by the Research Triangle Institute specifically for analysis of data from complex sample surveys and other observational and experimental studies involving cluster-correlated data. The SAS-callable version of the software is particularly useful to people familiar with SAS.

Specification of the variance estimation approach to be used by SUDAAN is done in the procedure statement for a particular procedure. Additional sample design statements provide further information required by the program. In particular, to carry out bootstrapping with GSS data, the following is required:

- specify **DESIGN=BRR** in the procedure statement
- include the following WEIGHT statement to identify the survey weight variable:
  WEIGHT wght_per;
- include the REPWGT statement to indicate the names of the bootstrap variables on your data file. In particular, for GSS microdata files, this REPWGT statement would have the form:

REPWGT wtbs_001-wtbs_500;

WesVar

WesVar is a software package produced by Westat which carries out various analyses of survey data using exclusively replication methods for variance estimation. One of the methods offered is BRR. Quoting heavily from Phillips (2004), in WesVar, the variance estimation method is specified when creating a new WesVar data file. The resulting file is then used to define workbooks where table and regression requests are carried out. To define a WesVar data file with bootstrap weights:

- Move the replicate weight variables (i.e., wtbs-001 to wtbs_500) to the *Replicates* box..
- Move the survey weight variable (i.e., wght_per) to the *Full sample* box.
- For the mean bootstrap, specify the *Method* as BRR.
- Move analysis variables to the *Variables* box, a unique identifier to the ID box (optional), and save the file.

References

Gagné, C., Roberts, G. and Keown, L.-A. (2014) “Weighted estimation and bootstrap variance estimation for analyzing survey data: How to implement in selected software”. The Research Data Centres Information and Technical Bulletin. (Winter) 6(1):5-70. Statistics Canada Catalogue no. 12-002-X. http://www.statcan.gc.ca/bsolc/olc-cel/olc-cel?catno=12-002- X20040027032&lang=eng