Catalogue no. 12M0010GPE

1995 General Social Survey, Cycle 10: The Family

Public Use Microdata File Documentation and User's guide

Data in many forms . . .

Statistics Canada disseminates data in a variety of forms. In addition to publications, both standard and special tabulations are offered. Data are available on the Internet, compact disc, diskette, computer printouts, microfiche and microfilm, and magnetic tape. Maps and other geographic reference materials are available for some types of data. Direct online access to aggregated information is possible through CANSIM, Statistics Canada's machine-readable database and retrieval system.

How to obtain more information

Inquiries about this publication and related statistics or services should be directed to: General Social Survey, Housing, Family and Social Statistics Division, Statistics Canada, Ottawa, Ontario, K1A 0T6 (telephone (613) 951-4995) or to the Statistics Canada Regional Reference Centre in:

Halifax	(902) 426-5331	Regina	(306) 780-5405
Montréal	(514) 283-5725	Edmonton	(403) 495-3027
Ottawa	(613) 951-8116	Calgary	(403) 292-6717
Toronto	(416) 973-6586	Vancouver	(604) 666-3691
Winnipeg	(204) 983-4020		

You can also visit our World Wide Web site: http://www.statcan.ca

Toll-free access is provided for all users who reside outside the local dialling area of any of the Regional Reference Centres.

National enquiries line	**1 800 263-1136**
National telecommunications device for the hearing impaired	**1 800 363-7629**
Order-only line (Canada and United States)	**1 800 267-6677**

How to order publications

Statistics Canada publications may be purchased from local authorized agents and other community bookstores, the Statistics Canada Regional Reference Centre, or from:

Statistics Canada
Operations and Integration Division
Circulation Management
120 Parkdale Avenue
Ottawa, Ontario
K1A 0T6

Telephone: (613) 951-7277
Fax: (613) 951-1584
Toronto (credit card only): (416) 973-8018
Internet: order@statcan.ca

Standards of service to the public

To maintain quality service to the public, Statistics Canada follows established standards covering statistical products and services, delivery of statistical information, cost-recovered services and services to respondents. To obtain a copy of these service standards, please contact your nearest Statistics Canada Regional Reference Centre.

Statistics Canada
Housing, Family and Social Statistics Division

1995 General Social Survey, Cycle 10: The Family

Public Use Microdata File Documentation and User's Guide

Published by authority of the Minister responsible for Statistics Canada

January 1997

Price: Canada: $ 50 per issue
United States: US$50 per issue
Other countries: US$50 per Issue

Catalogue no. 12M0010GPE

Ottawa

La version française de cette publication est disponible sur demande (no. 12M0010GPF au catalogue).

Note of appreciation

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

Table of Contents

1. Introduction

This package is designed to enable interested users to access and manipulate the microdata file for the tenth cycle of the General Social Survey (GSS), conducted from January through December, 1995. It contains information on the objectives, methodology and estimation procedures, as well as guidelines for releasing estimates based on the survey.

Appendix A contains the Approximate Variance Tables. The GSS Cycle 10 questionnaires are reproduced in Appendix B. A topical index of variables, data dictionary and record layout for each of the three files (main, child and union) are available in Appendices C to K. Finally, Appendices L and M present the occupation and industry coding respectively.

Excluding the appendices, this package is available in machine readable form.

2. Objectives of the General Social Survey

Increased pressure during the past decade, to operate more efficient government funded programmes, has led to a related increase in the information needed for policy formulation, programme development and evaluation. Many of these needs could not be filled through existing data sources or vehicles because of the range or periodicity of the information required. The two primary objectives of the GSS aim at closing these gaps. These objectives are: to gather data on social trends in order to monitor temporal changes in the living conditions and well-being of Canadians; and to provide immediate information on specific social policy issues of current or emerging interest. The GSS is a continuing program with a single survey cycle each year.

To meet the stated objectives, the data collected by the GSS are made up of three components: Classification, Core and Focus.

Classification content consists of variables which provide the means of delineating population groups and for use in the analysis of Core and Focus data. Examples of classification variables are age, sex, marital status, language, place of birth, and income.

Core content is designed to obtain information which monitors social trends or measures changes in society related to living conditions or well-being. Cycle 10 marks the first repeat of the family core content area originally covered in Cycle 5. Cycle 10 collected data on family and marital history (marriage and common-law relationships), joint custody arrangements, child leaving, family origins, fertility intentions, values and attitudes towards certain areas of family life, and work interruptions.

Focus content is aimed at the second survey objective of GSS. This component obtains information on specific policy issues which are of particular interest to certain federal departments or other user groups. In general, focus content, is not expected to be repeated on a periodic basis. Focus content for Cycle 10 covered the effects of environmental tobacco smoke, and wartime service, which targeted persons aged 55 and over. Focus content was sponsored by the Health Canada and Veteran Affairs Canada respectively.

3. Content and Special Features of Cycle 10

As in the 1993 and 1994 GSSs, data for Cycle 10 were collected using Computer Assisted Telephone Interviewing (CATI). With CATI, the survey questions appeared on a computer monitor. The interviewer asked the respondent the questions, and entered the responses into the computer as the interview progressed. Built-in edits and fewer processing steps resulted in better quality data. CATI methodology also eliminated the need for paper and pencil questionnaires. As a result, the forms in Appendix B were produced as reference documents only. In Cycle 10, the CATI system provided the interviewer with two main "components" which can be imagined to represent two paper questionnaires.

QUESTIONNAIRE	AGE GROUP	TITLE
GSS 10-1	All age groups	Survey Control Form
GSS 10-2	Age 15 and over	Family Questionnaire

A GSS 10-1 control form is completed for each telephone number generated in the sample. When a private household is contacted, all household members are enumerated and basic demographic information (eg. age and sex) is collected for each. The person who is to answer the questionnaire is then randomly selected by the computer. Proxy interviews are not accepted.

The creation of the household matrix is new to Cycle 10. Prior to Cycle 10, the GSS control form gathered information of the relationship of each household member to the selected respondent or to a reference person. In Cycle 10, the selected respondent provides the marital status for each household member and the relationship of each household member to the other household members. This will allow the 1995 GSS to construct a detailed structure of the respondent's household.

The GSS 10-2 questionnaire has the following sections:

Introduction Relationship to Household Member
A. Family Origins
B. Brothers and Sisters
C. Values and Attitudes (Part I)
D. Children
F. Fertility Intentions
H. Marriages
J. Common-law Partnerships
K. Values and Attitudes (Part II)
L. Paid and Unpaid Work
M Work Interruptions
N. Environmental Tobacco Smoke
R. Other Classification (including wartime service questions)
S. Contacts for follow-up

Section A is designed primarily to collect information on the respondent's family when he/she was growing up. This also includes demographic information on the respondent's mother and father (or parental substitutes). As well, this section collects data on the respondent's departure from the parental household if applicable.

Section B focuses on the respondent's brothers and sisters. Details on the type of sibling (i.e. step-, half-, adopted and birth) the respondent grew up with and the number of siblings the respondent did not grow up with are collected. Section C contains a number of questions on gender roles and the importance of relationship, marriage, children and employment.

Section D collects demographic information on all of the children the respondent has ever given birth to/fathered and/or raised (foster children are excluded). Questions are also asked concerning child custody, living arrangements and child leaving the parental home.

Section F examines the fertility intentions of respondents aged 15 to 49. Male respondents aged 50 or older are also interviewed for this section if their spouse or partner is aged 15 to 49. Section H contains a range of questions on the marriage history of the respondent. Section J repeats these questions for persons who have been involved in common-law partnerships.

Section K collects information on the respondent’s perception of the current relationship, the subjects and frequency of arguments and the sufficiency of various reasons for splitting up a marriage or relationship. Section L focuses on the respondent's participation in paid and unpaid work, as well as questions on their educational background. Section M covers work interruptions during the respondent's employment history (up to 4 interruptions).

Section N pertains to the respondent's exposure to environmental tobacco smoke. Lastly, Section R covers the respondent’s sociodemographic characteristics such as language, religion, state of health, and income, as well as some information on their spouse's or partner’s activity, where applicable. Focus content for wartime service is also covered, where applicable.

CATI differs from the collection method used for Cycles 1 to 7 of the GSS. Specifically, random

selection of the respondent is now performed by the computer rather than by the interviewer using a preprinted selection label. In addition, CATI asks the respondent to provide information on the relationship of each household member to the selected respondent, whereas in previous cycles, only the relationship with a designated person in each economic family was requested. As well, although shown in the questionnaire in appendix, the skips are built into CATI and do not appear on the screen. Other differences involve items which appear on the forms but do not appear on the CATI version. For example, interviewer check items are visible on the questionnaire but exist only as internal edits in the CATI system. Similarly, skip patterns are visible on the questionnaire but exist internally in the CATI system. Additionally, a few questions, such as date of birth, are asked in a different manner using CATI (eg. instead of asking date of birth, CATI asks three separate questions - year of birth, month of birth and day of birth).

4. Survey and Sample Design

Data for Cycle 10 of the GSS were collected monthly from January 1995 to December 1995. The sample was evenly distributed over the 12 months to evenly represent the seasonal variation in the information gathered. Most of the sample was selected using the Elimination of Non-Working Banks technique of Random Digit Dialling (RDD). An additional sample of 1,250 respondents sponsored by the province of Quebec was added in May and spread equally over the remaining months. A description of these methods is provided in Section 4.3. The target population is discussed in Section 4.1. Stratification used in the survey design is outlined in Section 4.2, and Section 4.4 discusses sample size.

4.1 Target Population

The target population for the GSS was all persons 15 years of age and over in Canada, excluding:

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

In the survey, all respondents were contacted by telephone, mainly by employing Random Digit Dialling (RDD), a telephone sampling method. Households without telephones were therefore excluded; however, persons living in such households represent less than 2% of the target population. Survey estimates have been adjusted (i.e. weighted) to account for persons without telephones. The tacit assumption is that, given the small number of people without telephones, their characteristics are not different enough from those of the rest of the target population to have an impact on the estimates. Since no one without a telephone is in the sample, this assumption cannot be verified using GSS data. It has, however, been verified with data from the Labour Force Survey.

4.2 Stratification

In order to carry out sampling, each of the ten provinces was divided into strata or geographic areas. Generally, for each province, one stratum represented the Census Metropolitan Areas (CMAs) of the province and another represented the non-CMA areas.

There were two exceptions to this general rule:

- Prince Edward Island has no CMAs and so did not have a CMA stratum
- Montreal and Toronto were each separate strata.

4.3 Elimination of Non-working Banks RDD Design

The Elimination of Non-Working Banks (ENWB) sampling technique is a method of Random Digit Dialling in which an attempt is made to identify all working banks[1] for an area (i.e., to identify all banks with at least one household). Thus, all telephone numbers within non-working banks are eliminated from the sampling frame.

For each province, lists of telephone numbers in use were purchased from the telephone companies and lists of working banks were extracted. Each bank was assigned to a stratum within its province.

A special situation existed in Ontario and Quebec because some small areas are serviced by independent telephone companies rather than by Bell Canada. The area code prefixes for these areas were identified by matching the Bell file with a file of all area codes and prefixes. Area code prefixes from Ontario and Quebec and not on the Bell file were identified. All banks within these area code prefixes were generated and added to the sampling frame. Use of the Waksberg method[2] (an alternative RDD method) was not possible for these areas since it requires that an accurate population estimate be available for the survey area. Such an estimate was not available for the parts of Ontario and Quebec not covered by Bell.

A random sample of telephone numbers was generated in each survey month for each stratum (from the working banks).

The entire sample of telephone numbers was generated before the first day of interviewing. Therefore, a prediction of the percentage of numbers dialled that would reach a household had to be made (this is known as the "hit rate"). Hit rates from Cycle 9 were used to estimate the hit rates for the Cycle 10 RDD sample.

For Cycle 10 of the GSS, 49.0% of the numbers dialled reached households. An attempt was made to conduct a GSS interview with one randomly selected person from each household.

[1] A bank of telephone numbers is a set of 100 numbers with the same first eight digits (i.e. the same Area Code-Prefix-Bank ID). Thus 613-951-9180 and 613-951-9192 are in the same bank, but 613-951-9280 is in a different bank.

[2] Waksberg, J. 'Sampling methods for Random Digit Dialling,' *Journal of the American Statistical Association*, 73, (1978):40-46.

4.4 Sample Size

The sample consisted of 10,749 people. A GSS 10-1 was completed for each telephone number generated in the sample and the main questionnaire (GSS 10-2) was then completed for the selected person.

5. Collection

Data collection for the GSS was conducted by Computer Assisted Telephone Interviewing (CATI) methods and involved two possible questionnaires. Respondents were interviewed in the official language of their choice. The French and English versions of the main questionnaire were identical with the exception of question R25 "What language did you first speak in childhood?" Respondents were not asked if they still understood the language in which they were being interviewed. The questionnaires, the procedures and the CATI system were field tested in August, 1994 in Winnipeg and Montreal. Data collection began in January 1995 and continued through the second week of December 1995. The main sample was evenly distributed over the 12 months. All interviewing took place using centralized telephone facilities in four of Statistics Canada's regional offices with calls being made from approximately 09:00 until 21:00, Monday to Saturday inclusive. The four regional offices were: Halifax, Montreal, Winnipeg and Vancouver. Interviewers were trained by Statistics Canada staff in telephone interviewing techniques using CATI, survey concepts and procedures in a four day classroom training session. The majority of interviewers had computer and telephone interviewing experience.

It would be too lengthy to include all the survey manuals as part of this documentation package. However, more information can be obtained from the survey manager (see Chapter 10). Shown below is a list of the manuals used in the survey:

- Introduction to Computer-assisted Telephone Interviewing (CATI)
- Content Manual
- Computer-assisted Telephone Interviewing (CATI) Interviewer's Manual
- Home Study Program
- Training Guide
- Regional Office Procedures Manual

6. Processing

The following is an overview of the processing steps for Cycle 10 of the GSS.

6.1 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 and built-in edits, and automatically followed the flow of the questionnaire. The data were transmitted to Ottawa electronically.

6.2 Edit and Imputation

All survey records were subjected to computer edits throughout the course of the interview. With CATI, built-in edits identified invalid or inconsistent information as the interview progressed. As a result, such problems could be immediately resolved with the respondent.

The system principally edited the main questionnaire for possible flow errors, out of range values and missing values. Edits on the 10-1 were limited to a few edits for the respondent's age and sex. The CATI system implemented such edits throughout the course of the interview. If the interviewer was unable to correctly resolve the detected errors, it was possible for the interviewer to bypass the edit and forward the data to head office for resolution.

Head office edits performed the same checks as the CATI system as well as more detailed edits. Records with missing or incorrect information were assigned non-response codes and in a small number of cases corrected from other information from the respondent's questionnaire. In most cases when editing, if data were inconsistent with responses that came earlier, the earlier information was considered to be correct. For example, if a screening question introduced two or more mutually exclusive "branches" (or "paths") in the questionnaire and data existed for more than one branch, it was the response to the screening question that was deemed correct, and only data in the branch corresponding to this response was retained.

Due to the nature of the survey, imputation was not appropriate for most items and thus “not stated” codes were usually assigned for missing data. In some cases, the answer was not known but could be obtained deterministically by either the questions which followed or from information from other areas of the survey.

There are three reasons that can explain the absence of a response for a question: the question may have been skipped because of a previous response; the question may have been skipped because of a previous refusal; or the respondent may have refused to answer the question. In the first case, the question is considered ‘not applicable’ and is given a code of 7, 97, 997 or 9997. In the second case the applicability of the question is not known since the question that determines the applicability was refused, so the question is ‘not asked, applicability unknown’ and is given a code of 6, 96, 996 or 9996. In the third case the question is ‘refused’ and is given a code of 9, 99, 999 or 9999.

Non-response was not permitted for those items required for weighting. Values were imputed in the rare cases where the number of residential telephone lines was missing. The imputation was based on a detailed examination of the data and the consideration of any useful data such as the ages and sexes of other household members, and the interviewer's comments. The procedure

used to select the respondent ensured that there was always a value for age. When not provided by the respondent, DVTEL (number of residential phone lines) was assigned a value of one (1).

6.3 Coding

Several questions allowing write-in responses had the write-in information coded into either new unique categories, or to a listed category if the write-in information duplicated a listed category. Where possible (e.g., occupation, industry, language, education, country of birth and religion), the coding followed the standard classification systems as used in the Census of Population.

6.4 Creation of Combined and Derived Variables

A number of variables on the file have been derived by using items found on the GSS 10-1 and 10-2 Questionnaires. Derived variable names generally start with DV and are followed by characters referring to the question number or subject. In some cases, the derived variables are straightforward and involve collapsing of categories. In other cases, several variables have been combined to create a new variable. The data dictionary provides comments indicating the origin of these variables.

6.5 Amount of Detail on Microdata File

In order to guard against disclosure, the amount of detail included on this file is less than is available on the master file retained by Statistics Canada. Variables with extreme values have been capped and information for some variables have been aggregated into broader classes (e.g., occupation, religion, country of birth).

The measures taken to cap, group or collapse data have been indicated in the data dictionary. Variables with a very limited number of observations such as the reason the spouse or partner is not living in the household (question H4) and the age when the respondent left the household instead of the child leaving (question D90M and D90Y) were omitted. Since only 2 respondents reported a fifth common-law union, the generated set of variables pertaining to that union on the main file was dropped. However the information can be found in the current common-law union and as a record of the union file.

7. Estimation

When a probability sample is used, as was the case for the GSS, the principle behind estimation is that each person selected in the sample 'represents' (in addition to himself/herself) several other persons not in the sample. For example, in a simple random sample of 2% of the population, each person in the sample represents 50 persons in the population. 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.

For analysis of GSS Cycle 10 information the weighting factor WGHTFNL was placed on the microdata file. As described above, this factor represents the number of persons in the population that the record represents. It refers to the number of times a particular record should contribute to a population estimate. For example, to estimate the number of adults who have ever smoked (N2), the value of WGHTFNL is summed over all records with this characteristic.

The process of deriving the weighting factor, WGHTFNL, is described in Section 7.1.

7.1 Weighting

Where possible, each survey month was weighted independently. This was done in an attempt to ensure that each survey month contributed equally to estimates. If monthly sample sizes were not large enough, two or more survey months were combined in certain steps of the weighting.

1) Basic Weight Calculation

Each household (responding and non-responding) in the sample was assigned a weight equal to the inverse of its probability of selection. This weight was calculated independently for each stratum-month group as follows:

$$\frac{\text{Number of possible telephone numbers within the stratum-month group}}{\text{Number of sampled telephone numbers within the stratum-month group}}$$

(The total number of possible telephone numbers for a stratum is equal to the number of working banks for a stratum times 100).

2) Non-Response Adjustment

Weights for responding households were adjusted to represent non-responding households. This was done independently within each stratum-month group. Records were adjusted by the following factor:

$$\frac{\text{Total of the basic household weights of all sampled households in each stratum-month group}}{\text{Total of the basic household weights of responding households in each stratum-month group}} = \frac{\text{Number of sampled households within the stratum-month group}}{\text{Number of responding households within the stratum-month group}}$$

Non-responding households were then dropped.

3) Multiple Telephone Adjustment

Weights for households in the sample with more than one residential telephone number (i.e. not used for business purposes only) were adjusted downwards to account for the fact that such households had a higher probability of being selected. The weight for each household was divided by the number of residential telephone numbers that serviced the household.

4) Person Weight Calculation

A person weight was then calculated for each respondent to the survey by multiplying the household weight by the number of persons in the household who were eligible to be selected for the survey (i.e. the number of persons 15 years of age or older).

5) Adjustment of Person Weight to External Totals

The person weights were adjusted several times using a raking ratio procedure. This procedure ensured that, based on the survey's total sample, estimates produced of the sizes of strata or of province-age-sex groups would match external references. The two sets of groupings used for these adjustments were stratum-month and province-age-sex. The age groupings used were:

15-19, 20-24, 25-29, 30-34, 35-39, 40-44,
45-49, 50-54, 55-59, 60-64, 65-69, 70+.

Sample sizes were too small to allow the province-age-sex adjustments to be made at the survey month level. Also due to small sample sizes, there were cases where two or more adjacent age groups in the same province-sex group or two adjacent months in the same stratum were collapsed before the adjustments were made.

The reference totals for the stratum-month adjustments were one twelfth of the population projections for each month. The reference totals for the province-age-sex adjustments were the average of the population projections for each month. At each stage in the adjustment process the weights were adjusted by the factor:

$$\frac{\text{reference total for group}}{\text{sum of person weights for group}}$$

The groupings used for the adjustments alternated between province-age-sex and stratum-month until the weights converged.

It should be noted that persons 15 years and over living in households without telephone service are included in the reference totals even though they were not sampled.

7.2 Weighting Policy

Users are cautioned against releasing unweighted tables or performing any analysis based on unweighted survey results. As was discussed in Section 7.1, there were several weight adjustments performed independently to the records of each province. Sampling rates as well as non-response rates varied significantly from province to province.

Contact was made or attempted with 13,251 households during the survey. Of these, 1,266 (9.6%) were non-responding households. The non-responding households included 790 household refusals, 228 households that could not be reached during the survey period, and 248 cases where a response could not be obtained due to language difficulties, illness, or other problems. An interview was attempted with a adult randomly selected from the eligible household members of the 11,985 responding households. Usable responses were obtained from 10,749 respondents. The difference consists of 572 person-level refusals, 382 persons that could not be reached during the survey period, and 282 cases where the interview could not be completed due to language difficulties, illness, or other problems. A response rate of 80.7% was obtained when it is assumed that all of the households for which there was no response were "in scope" (i.e., had at least one eligible member).

It is known that non-respondents are more likely to be males and more likely to be younger. In the responding sample, 3.3% were males between the ages of 15 and 19, while in the overall population, approximately 4.4% are males between 15 and 19. Therefore, it is clear that the sample counts cannot be considered to be representative of the survey target population unless appropriate weights are applied.

7.3 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 characteristics) and quantitative estimates involving quantities or averages. More complex estimation and analyses are covered in Section 7.4.

7.3.1 Qualitative Estimates

It should be kept in mind that the target population for the GSS was non-institutionalized persons 15 years of age or over, living in the ten provinces. Qualitative estimates are estimates of the number or proportion of this target population possessing certain characteristics. The number of adults who lived with both of their birth parents when they were born (A2) is an example of this kind of estimate. The simplest estimate to produce is that of the number of people in the target population who would have responded that the lived with both of their birth parents when they were born if a census had been conducted. This estimate is simply the sum of the weights for those respondents with a value of ‘1' for question A2: 22,337,501. This estimate does not however adjust for nonresponse 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 ignoring the respondents who did not provide an answer to A2 and calculating the ratio of the total of the weights of those respondents who answered ‘Yes-both birth parents’ to the total of the weights of those who answered ‘Yes-both birth parents’ or Yes-both adoptive parents’ or ‘No’. This proportion is then multiplied by the size of the target population to produce the final estimate:

$$22{,}351{,}647 = 23{,}267{,}125 \times \frac{22{,}337{,}501}{23{,}252{,}400}$$

The difference in this example is small because only .2% of respondents didn’t answer this question. When the proportion of responses that are ‘don’t know’ or ‘refused’ is higher the differences between the two estimates will be larger.

7.3.2 Quantitative Estimates

Some variables on the 1995 General Social Survey microdata file are quantitative in nature (e.g. number of hours per week a person usually works - L14). From these variables, it is possible to obtain such estimates as the average number of hours per week a person usually works. These estimates are of the following ratio form:

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

The numerator (X) is a quantitative estimate of the total of the variable of interest (say, number of hours per week person usually works) for a given sub-population (say, persons who were employed in the past 12 months). X would be calculated by multiplying the person weight WGHTFNL by the variable of interest when it is known, i.e. not equal to '996', '997', '998' or '999', and summing this product over all records which are in the subpopulation. The denominator (Y) is the qualitative estimate of the number of participants within that subpopulation (those who worked in the past 12 months and for whom the weekly hours of work was known). Y would be calculated by summing the person weight, WGHTFNL, over all records for persons who reported working in the past 12 months and who were not unable or unwilling to give the number of hours per week they usually worked. The two estimates X and Y are derived independently and then divided to provide the quantitative estimate. The average weekly number of hours persons who worked in the past 12 months worked per week when they were working is then estimated to be:

$$39.06 = \frac{609{,}736{,}314}{15{,}608{,}846}$$

7.4 Guidelines for Analysis

As is detailed in Chapter 4 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 and multiple stages of selection, and unequal probabilities of selection of respondents. Using data from such surveys presents problems to analysts because the survey design and the selection probabilities affect the estimation and variance calculation procedures that should be used.

The GSS used a stratified design with significant differences in sampling fractions between strata. Thus, some areas are over-represented in the sample (relative to their populations) while some other areas are relatively under-represented. This means that the unweighted sample is not representative of the target population.

The survey weights must be used when producing estimates or performing analyses in order to account for this over- and under-representation. 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, with the result that 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 on the data, or any subset of the data, 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.

For example, if an analysis of respondents who have ever been married is desired, then the following steps are required:

- Select respondents from the file who have ever been married (DVMS = 1, 2, 3, 4, or 5) (8,102 respondents)

- Calculate the Average Weight equal to the average of WGHTFNL for these records (The total of WGHTFNL for these 8,102 respondents is 17,596,117, so the average weight is 2171.82 = 17,596,117 / 8,102)

- For each of these respondents calculate a "working" weight equal to WGHTFNL / 2,121.82

- Perform the analysis for these respondents using the "working" weight. The calculation of truly meaningful variance estimates requires detailed knowledge of the design of the survey. Such detail cannot be given in this microdata file because of confidentiality.

Variances that take the sample design into account can be calculated for many statistics by Statistics Canada on a cost recovery basis.

7.5 Methods of Estimation and Interpretation of Estimates

Person Weight: WGHTFNL

The basic sampling weight assigned to each sampled individual has been adjusted to reflect the age and sex composition of the various provincial populations as projected by the Labour Force Survey for each month of 1995.

$$\sum_{10,749}^{3} \text{WGHTFNL} = 23,267,125$$

= an estimate of the number of persons 15 years of age and older in the population.

Examples and Interpretation:

(i) 17% (4.0 million) of adult Canadians have ever been a partner in a common law partnership that was not followed by marriage (including those currently in a common law relationship) (DVMS=1 or (H9=1 and DVH1H2=5) or DVJ2=1 or J4=1).

(ii) 72% (16.7 million) of adult Canadians have at least one living brother (DVB8CAP is between 1 and 5).

(iii) 17% (1.5 million) of adult Canadians (49 years or less or males 50 or more with partners less than 50) who have biological children (9.1 million) intend to or are about to have more children) (DVF78CAP is between 1 and 5 and DVD12CAP is between 1 and 10).

(iv) 22% (0.5 million) of adult Canadians currently in common law partnerships (2.1 million, DVMS=1) report that their partner has children from a previous relationship that they did not raise (DVJ10CAP is between 1 and 5).

7.6 Using the Household Weight

A household weight could be calculated for this file. It is simply the ratio:

$$\text{WGHTHLD} = \frac{\text{WGHTFLN}}{\text{DVELLCAP}}$$

[3] The number of responding households (with one randomly chosen respondent per household).

$$\sum_{10,749} \text{WGHTHLD} = 11,697,242$$

= an estimate of the number of households that include someone in the target population

The household weight should be used with variables that represent characteristics of the household rather than of the respondent. Examples of such variables on the microdata file are: DVELLCAP, DVTEL, HHSIZCAP and MULTIGEN. For instance, to estimate the number of households of size 5 that include someone in the target population you would sum the household weights for those cases with HHLDSIZE=5:

6.1% (711,297) of Canadian households have five members.

7.7 Using the Child and Union Files

There are three data files included in this package of microdata for GSS-10, the main file, the child file, and the union file. More detail on the structure of these files is given in Section 9.

The main file consists of one record for each respondent and includes both variables that represent characteristics of the respondent and of the household. This file can be used directly, with the choice of weight depending on whether a person or household characteristic is being tabulated.

The child file consists of one record for each child of each respondent. Therefore there are no child file records for some respondents, while for others there are more than one child file record. There is a maximum of ten file records for biological children of the respondent; if the respondent reported more than ten children, the ten records are for the nine oldest children and for the youngest child. Similarly, a maximum of 5 step and 3 adopted children per respondent are included on this file. Because any particular child could have been reported to the survey by any parent of any type (birth, adopted, or step) who considered that they raised the child, we do not know with certainty the probability of each child being included in the child file and so we cannot calculate a sampling weight for the children on the child file. This means that the child file should be used to derive characteristics of the respondent which are then tabulated using the person weight WGHTFNL.

The union file consists of one record for each union. The situation where the partners lived common law and then married is represented by one record on this file. The potential maximum number of records per respondent on this file is 9 i.e. 4 marriages, 4 common-law unions and the most recent union. No sampling weight has been attached to these records and so the information on these records should be used to derive characteristics of the respondent to be tabulated using the person weight WGHTFNL.

8. 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 microdata file.

This section of the documentation provides guidelines to be followed by users.With the aid of these guidelines, users of the microdata should be able to produce figures consistent with those produced by Statistics Canada and that respect the established guidelines for rounding and release.The guidelines can be broken into four broad sections: Minimum Sample Sizes for Estimates, SamplingVariability Policy, Sampling Variability Estimation and Rounding Policy.

8.1 Minimum Sample Size For Estimates

Users should determine the number of records on the microdata file which contribute to the calculation of a given estimate. When the number of contributors to the weighted estimate is less then 15 the weighted estimate should not be released regardless of the value of the Approximate Coefficient of Variation.

8.2 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 that 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 on the questionnaire 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 implemented 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, procedures to ensure that data capture errors were minimized and coding and edit quality checks to verify the processing logic.

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 just one or some questions) to total non-response. Total non-response occurred because the interviewer was either unable to contact the respondent, a language problem prevented the interview from taking place, 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.

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 it is virtually certain that the differences would be less than three standard errors.

Because of the large range in size of estimates that can be produced from a survey, the standard error is usually expressed relative to the estimate to which it pertains. The resulting measure, known as the coefficient of variation of an estimate, is obtained by dividing the standard error of the estimate by the estimate itself and is expressed as a percentage of the estimate. Before releasing and/or publishing any estimates from the microdata file, users should determine whether the estimate is releasable based on the guidelines shown below.

Type of Estimate	Coefficient of Variation	Policy Statement
1. Unqualified	**0.0 to 16.5%**	**Estimates can be considered for general unrestricted release.**
2. Qualified	**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. Not for	**33.4% or over**	**Estimates should not be released Release in any form under any circumstances. In such statistical tables, such estimates should be deleted.**

Note: The sampling variability policy should be applied to rounded estimates.

8.3 Estimates of Variance

Variance estimation is described separately for qualitative and quantitative estimates.

8.3.1 Sampling Variability for Qualitative Estimates

Derivation of sampling variabilities for each of the estimates which could be generated from the survey would be an extremely costly procedure, and for most users, an unnecessary one. Consequently, approximate measures of sampling variability, in the form of tables, have been developed for use and are included in Appendix A "Approximate Sampling Variability Tables".

Variance tables for estimates using WGHTFNL are provided at the Canada, province and Canada less Quebec level, as well as for the Atlantic and Prairie Regions.

It should be noted that all coefficients of variation in these tables are approximate and therefore unofficial. Estimates of actual variance for specific variables may be purchased from Statistics Canada. Use of actual variance estimates may allow users to release otherwise unreleasable estimates, i.e. estimates with coefficients of variation in the "Not for Release" range (see the policy regarding the release of the survey estimates on preceding pages).

The Approximate Variance tables have been produced using the coefficient of variation formula based on a simple random sample and the straightforward expansion estimator. Since estimates for the General Social Survey were based on a complex sample design and the complicated raking ratio estimator alluded to earlier, a factor called the Design Effect was introduced into the variance formula.The Design Effect for an estimate is the actual variance for the estimate (taking into account the design and estimator that were used) divided by the variance that would result if the estimate had been derived from a simple random sample and a simple expansion estimator. The Design Effect used to produce the Approximate Variance Tables has been determined by first calculating Design Effects for a wide range of characteristics and then choosing among these a conservative value which will not give a false impression of high precision. These Design Effects are specified in the table below.

GENERAL SOCIAL SURVEY Cycle 10 DESIGN EFFECTS		
Geographic Area	*15+ Population*	*Households*
Canada	1.50	1.34
Newfoundland	1.14	1.04
Nova Scotia	1.18	1.07
P.E.I.	1.16	1.04
New Brunswick	1.14	1.05
Atlantic Region	1.21	1.07
Quebec	1.30	1.15
Ontario	1.17	1.05
Manitoba	1.16	1.05
Saskatchewan	1.20	1.07
Alberta	1.17	1.04
Prairie Region	1.24	1.10
British Columbia	1.19	1.06

8.3.2 Sampling Variability For Quantitative Estimates

Approximate variances for quantitative variables cannot be as conveniently summarized. As a general rule, however, the coefficient of variation of a quantitative total will be larger than the coefficient of variation of the corresponding qualitative estimate (i.e. the number of persons contributing to the quantitative estimate). If the corresponding qualitative estimate is not releasable, then the quantitative total will in general not be releasable.

8.4 Rounding

In order that estimates produced from the General Social Survey microdata file correspond to those produced by Statistics Canada, users are urged to adhere to the following guidelines regarding the rounding of such estimates. It is improper to release unrounded estimates, as they imply greater precision than actually exists.

8.4.1 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 8.4.2).

(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 in which, due to technical or other limitations, a different rounding technique is used, which results in estimates being released which differ from the corresponding estimates produced by Statistics Canada, users are encouraged to note the reason for such differences in the released document.

8.4.2 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 8 and the number 8500 rounded to thousands would be 9.

9. Structure Of File

In view of the nature of the data, the microdata file consists of three subfiles described below.

The **Main File** consists of one record per respondent and provides the data for most sections of the questionnaire. It has therefore 10,749 records. Each record contains 541 variables. An exhaustive list of these variables can be found in the data dictionary.

The **Child file** consists of all children ever raised by the respondent (excluding foster children). Each respondent generated a variable number of records. The maximum number of biological, step and adopted children reported are respectively 10, 5 and 3. In situations where the respondent had more than 10 biological children, the 9 oldest and the youngest were put on the file. The same logic was used for step and adopted children. The file has 19,542 records. Each record contains 82 variables.

The **Union file** consists of all unions reported by the respondent. Respondents who never reported a union are not included in this file. In cases where the union started as a common-law union followed by marriage, only one record was created. Each respondent generated from 0 to 7 records. The maximum number of marriages is 3 while up to 5 common-law unions were reported by a respondent. The file has 10,938 records. Each record contains 22 variables.

There is little duplication across the three files. The variable RECID can be used for linking the files. An exhaustive list of variables in each file and frequencies can be found in the data dictionaries provided as appendices.

10. Additional Information

Additional information about this survey can be obtained from the individuals listed below. Data from the survey are available through published reports, special request tabulations, and this microdata file. The microdata file is available from the Housing, Family and Social Statistics Division of Statistics Canada at a cost of $1500.00. Tabulations can be obtained at a cost that will reflect the resources required to produce the tabulation.

Sample Selection Procedures, Weighting and Estimation
Dave Paton
Household Survey Methods Division
(613) 951-1467

or

Paul Matthews
Household Survey Methods Division
(613) 951-1480

Subject Matter, Data Collection and Data Processing
Ghislaine Villeneuve
Housing, Family and Social Statistics Division
(613) 951-4995

APPENDIX A

Approximate Variance Tables

APPROXIMATE VARIANCE TABLES

By using the Approximate Variance Tables and the following rules, users should be able to determine approximate coefficients of variation for aggregates (totals), percentages, ratios, differences between totals and differences between ratios.

As noted in 8.2, estimates having a coefficient of variation (cv) of more than 33.3% are not releasable. In addition, as mentioned in 8.1, each estimate should be derived from at least 15 respondents in order to be released, regardless of the approximate coefficient of variation.

Rule 1: Estimates of Numbers Possessing a Characteristic (Aggregates)

The coefficient of variation (cv) depends only on the size of the estimated aggregate itself. In the Approximate Variance Table, locate the estimated aggregate in the left-most column of the table (headed "Numerator of Percentage") and follow the asterisks across to the first figure encountered. This figure is the estimated coefficient of variation.

Example 1:

A user estimates that in Canada 384,557 females aged 15 years and over describe their state of health as poor compared to other people their age (question R30=5). How does the user determine the approximate coefficient of variation for this estimate?

Refer to the approximate variance table for Canada level estimates. The estimated aggregate does not appear in the left-most column (the “Numerator of Percentage” column), so it is necessary to use the closest figure, namely 400,000. The coefficient of variation for an estimated aggregate is found by referring to the first non-asterisk entry for that row, in this case 8.9%. This cv falls within the range of cv's for “Unqualified” estimates (i.e. 0.0% -16.5%, see section 8.2) allowing the estimate to be released without restriction.

Rule 2: Estimates of Percentages or Proportions Possessing a Characteristic

The coefficient of variation of an estimated percentage or proportion depends on both the size of the percentage or proportion and the size of the total upon which the percentage is based. Estimated percentages or proportions are relatively more reliable than the corresponding estimates of the numerators of the percentages, particularly if the percentages are 50 percent or more. (Note that in the tables the cv's decline in value reading from left to right).

When the percentage or proportion is based upon the total population of the geographic area covered by the table, the cv of the percentage or proportion is the same as the cv of the numerator of the percentage. In this case, Rule 1 can be used.

When the percentage or proportion is based upon a subset of the total population (e.g., those in a particular age-sex group), reference should be made to the percentage (across the top of the table) and to the numerator of the percentage or proportion (down the left side of the table). The intersection of the appropriate row and column gives the coefficient of variation.

Example 2:

A user estimates that in Canada 3.25% of females aged 15 years and over describe their state of health as poor compared to others their age (question R30). This is the expression of the estimate obtained in Example 1 as a percentage of all females aged 15 years and over in Canada. How does the user determine the approximate coefficient of variation for this estimate?

Refer to the approximate variance table for Canada level estimates. Because the estimate is a percentage which is based on a subset of the population covered by the table, it is necessary to use both the percentage (3.25%) and the numerator portion of the percentage (384,557) to determine the approximate coefficient of variation. Since the numerator does not appear in the left-most column (the 'Numerator of Percentage' column), it is necessary to use the figure closest to it, namely 400,000. Similarly, the percentage estimate does not appear among the column headings, so it is necessary to use the figure closest to it, namely 2.0%. The figure at the intersection of the row and column selected, namely 8.9%, is the coefficient of variation. This cv falls within the range of cv's for 'Unqualified' estimates (i.e. 0.0% - 16.5) allowing the estimate to be released without restriction.

Rule 3: <u>Ratios</u>

In the case where the numerator is a subset of the denominator, the ratio should be converted to a percentage and Rule 2 applied.In the case where the numerator is not a subset of the denominator, the coefficient of variation of the ratio of two estimates is approximately equal to the square root of the sum of squares of each coefficient of variation considered separately. That is, the standard deviation of a ratio

$$R = X / Y$$

is

$$sd(R) = (R) * (cv(X)^2 + cv(Y)^2)^{½}$$

The coefficient of variation of R is approximately:

$$cv(R) = sd(R) / R$$
$$= (cv(X)^2 + cv(Y)^2)^{½}$$

This formula will tend to overstate the error if X and Y are positively correlated and understate the error if X and Y are negatively correlated.

Example 3:

A user estimates that in Canada among females aged 15 years and over, 384,557 describe their state of health as poor compared to other people their age (question R30) and 3,088,330 describe their state of health as excellent as compared to others their age. The user is interested in the ratio of females describing their health as excellent versus those describing their health as poor. How does the user determine the approximate coefficient of variation for this ratio estimate?

The numerator of the ratio estimate is 3,088,330 (X). Using Rule 1 (refer to Example 1), the coefficient of variation for this estimate is determined to be 3.0% cv(X). The denominator of the ratio estimate is 384,557 (Y). Again using Rule 1, the coefficient of variation is determined to be 8.9% cv(Y). Using Rule 3, the coefficient of variation of the ratio estimate is

$$cv = (0.030^2 + 0.089^2)^{1/2}$$

$$= 0.0939$$

Therefore at the Canada level, the ratio of females who describe their health as excellent versus females who describe their health as poor is 3,088,330/384,557 or 8.03 to 1. The coefficient of variation of this estimate is 9.39%, and so the estimate can be released without restriction.

Rule 4: <u>Differences Between Totals or Percentages</u>

The standard deviation of a difference between two estimates is approximately equal to the square root of the sum of squares of each standard deviation considered separately. That is, the standard deviation of a difference:

$$d = X - Y$$

is

$$sd(d) = ((X * cv(X))^2 + (Y * cv(Y))^2)^{1/2}$$

The coefficient of variation of d is approximately:

$$cv(d) = sd(d) / d$$

This formula is accurate for the difference between separate and uncorrelated characteristics but is only approximate otherwise.

Example 4:

A user estimates that in Canada, among those 15 years and over, 3.25% (X) of females describe their state of health as poor compared to others their age and 2.83% (Y - an estimated 323,785) of males describe their state of health as poor compared to other people their age. The user is

interested in the difference between these two estimates. How does the user determine the approximate coefficient of variation for the estimate of the difference?

From Rule 2 (refer to example 2), the coefficient of variation for the female estimate is 8.9%. The coefficient of variation for the male estimate is 10.3%.

The difference between the estimates is 0.42%. Using Rule 4, standard deviation of the difference between the estimates is

$$sd = ((0.0325 \times 0.089)^2 + (0.0283 \times 0.103)^2)^{½}$$
$$= 0.0041$$

and the coefficient of variation is

$$cv = \frac{0.0041}{0.0042}$$
$$= 0.9762$$

Therefore the coefficient of the difference between the estimates is 97.62% and the estimate should not be released.

Rule 5: Difference of Ratios

In this case, Rules 3 and 4 are combined. The cv's of the two ratios are first determined using Rule 3, and the cv of their difference is found using Rule 4.

Confidence Limits

Although coefficients of variation are widely used, a more intuitively meaningful measure of sampling error is the confidence interval of an estimate. A confidence interval constitutes a statement on the level of confidence that the true value for the population lies within a specified range of values. For example a 95% confidence interval can be described as follows:

If sampling of the population is repeated indefinitely, each sample leading to a new confidence interval for an estimate, then in 95% of the samples the interval will cover the true population value.

Using the standard error of an estimate, confidence intervals for estimates may be obtained under the assumption that under repeated sampling of the population, the various estimates obtained for a population characteristic are normally distributed about the true population value. Under this assumption, 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 about 99 out of 100 that the

differences would be less than three standard errors. These different degrees of confidence are referred to as the confidence levels.

Confidence intervals for an estimate are generally expressed as two numbers: one below the estimate and one above the estimate, that is, more explicitly, as $(\hat{Y}-k, \hat{Y}+k)$ where k is determined from the level of confidence desired and the sampling error of the estimate.

Confidence intervals for an estimate can be calculated directly from the Sampling Variability Tables by first determining from the appropriate table the coefficient of variation of the estimate, and then using the following formula to convert to a confidence interval CI:

$$CI_Y = \{\hat{Y} - (t)(\hat{Y})(\alpha\hat{Y}), \hat{Y} + (t)(\hat{Y})(\alpha\hat{Y})\}$$

where $\alpha\hat{Y}$ is the determined coefficient of variation of $\hat{Y}$ and

- t = 1 if a 68% confidence interval is desired
- t = 1.6 if a 90% confidence interval is desired
- t = 2 if a 95% confidence interval is desired
- t = 3 if a 99% confidence interval is desired

Example 5(a):

An estimated 708,342 persons described their state of health as poor (question R30) as compared to other people their age. This estimate has an approximate coefficient of variation of 6.4% (obtained from the 750,000 row, left-most column, of the Canada approximate variance table). The 95% confidence interval for this estimate is thus:

$$\begin{aligned} CI &= \{708{,}342 - (2)(708{,}342)(0.064), 708{,}342 + (2)(708{,}342)(0.064)\} \\ &= \{708{,}342 - 90{,}668, 708{,}342 + 90{,}668\} \\ &= \{617{,}674, 799{,}010\} \end{aligned}$$

Example 5(b):

An estimated 3.25% of females aged 15 years and over in Canada described their state of health as poor when compared to other people their age or .0325 expressed as a proportion. From Example 2 this estimate has an approximate coefficient of variation of 8.9%. A 95% confidence interval for this estimate (expressed as a proportion) is

$$\begin{aligned} CI &= \{.0325 - (2 \times .0325 \times 0.0896, .0325 + (2 \times .0325 \times .0.089)\} \\ &= \{0.0267, 0.0383\} \end{aligned}$$

With 95% confidence it can be said that between 2.67% and 3.83% of females aged 15 years and over in Canada, describe their state of health as poor, compared to other people their age.

Note: Release guidelines which apply to the estimate also apply to the confidence interval. For example, if the estimate is not releasable, then the confidence interval is not releasable either.

T-test

Standard errors may also be used to perform hypothesis testing, a procedure for distinguishing between population parameters using sample estimates. The sample estimates can be numbers, averages, percentages, ratios, etc. Tests may be performed at various levels of significance, where a level of significance is the probability of concluding that the characteristics are different when, in fact, they are identical.

Let X_1 and X_2 be sample estimates for 2 characteristics of interest. Let the standard error of the difference $X_1 - X_2$ be σ_d

$$\textit{If } t = \frac{X_1 - X_2}{\sigma_d} \quad \textit{is between -2 and 2,}$$

then no conclusion about the difference between the characteristics is justified at the 5% level of significance. If however, this ratio is smaller than -2 or larger than +2, the observed difference is significant at the 5% level (Note:at the 1% level, values of -3 and +3 should be used, etc.).

Example 6:

A user wishes to test at the 5% level of significance the hypothesis that at the Canada level there is no difference between percentage estimates of males and females who describe their state of health as poor, as compared to other people their age. From Example 4 the estimate of the standard deviation of the difference between the estimates is 0.0041.

$$\text{Hence } t = \frac{0.0325 - 0.0283}{0.0041} = 1.02$$

Since t = 1.02 is less than 2, there is no evidence to reject the hypothesis at the 5% significance level.

APPENDIX B

1995 General Social Survey Questionnaires

APPENDIX C

Topical Index to Variables for Main File

APPENDIX D

Data Dictionary for Main File

APPENDIX E

Record Layout for Main File

APPENDIX F

Index to Variables for Child File

APPENDIX G

Data Dictionary for Child File

APPENDIX H

Record Layout for Child File

APPENDIX I

Index to Variables for Union File

APPENDIX J

Data Dictionary for Union File

APPENDIX K

Record Layout for Union File

APPENDIX L

Occupation Coding

This appendix shows the collapsing of the Standard Occupational Codes (SOC) into the 34-category grouping.

34 Category Grouping	**Standard Occupational Codes (SOC)**
01: MANAGERS/ADMINISTRATORS	1111 to 1119 1130 to 1137 1141 to 1149
02: MANAGEMENT/ADMIN/RELATED	1171 to 1179
03: LIFE SCIENCES/MATH/COMPUTERS	2111 to 2119 2131 to 2139 2181 to 2189
04: ARCHITECTS/ENGINEERS/RELATED	2141 to 2147 2151 to 2159 2160 to 2169
05: SOCIAL SCIENCE/ RELIGION/ETC.	2311 to 2319 2331 to 2339 2341 to 2349 2350 to 2359 2391 to 2399 2511 to 2519
06: TEACHING/RELATED	2711 to 2719 2731 to 2739 2791 to 2799
07: HEALTH OCCUPATIONS/RELATED	3111 to 3119 3130 to 3139 3151 to 3158 3161 to 3169
08: ARTISTIC/LITERARY/RECREATIONAL	3311 to 3319 3330 to 3339 3351 to 3359 3360 3370 to 3379
09: STENOGRAPHIC/TYPING	4110 to 4113
10: BOOKKEEPING/ACCOUNT-RECORDING	4130 to 4139
11: EDP OPERATORS/MATERIAL RECORD	4140 to 4143 4150 to 4159
12: RECREATION/INFO./MAIL/MESSAGE	4170 to 4179
13: LIBRARY/FILE/OTHER CLERICAL	4160 to 4169 4190 to 4199
14: SALES COMMODITIES	5130 to 5135 5141 to 5149
15: SALES/SERVICES	5170 to 5179 5190 to 5199
16: PROTECTIVE SERVICES	6111 to 6119

17: FOOD/BEVERAGE/ACCOMMODATION	6120 to 6129 6130 to 6139
18: PERSONAL/APPAREL/FURNISHINGS	6141 to 6149 6160 to 6169
19: OTHER SERVICE OCCUPATIONS	6190 to 6199
20: FARM OCCUPATIONS	7113 to 7119 7180 to 7185 7195 to 7199
21: PRIMARY OCCUPATIONS	7311 to 7319 7510 to 7519 7710 to 7719
22: FOOD/BEVERAGE PROCESSING ETC.	8210 to 8217 8221 to 8229
23: PROCESSING OCCUP. (EXCEPT FOOD)	8110 to 8119 8130 to 8137 8141 to 8149 8150 to 8159 8160 to 8167 8171 to 8179 8230 to 8239 8250 to 8259 8260 to 8267 8271 to 8279 8290 to 8299
24: MACHINING/RELATED OCCUPATIONS	8310 to 8319 8330 to 8339 8350 to 8359 8370 to 8379 8390 to 8399 8510 to 8515 8523 to 8529
25: ELECTRICAL/ELECTRONICS/RELATED	8530 to 8539
26: TEXTILES/FURS/LEATHERS	8550 to 8557 8561 to 8569
27: WOOD PRODUCTS/RUBBER/PLASTICS	8540 to 8549 8570 to 8579 8590 to 8599
28: REPAIRMEN (EXCEPT ELECTRICAL)	8580 to 8589
29: EXCAVATING/PAVING/WIRE COMM.	8710 to 8719 8730 to 8739
30: OTHER CONSTRUCTION TRADES	8780 to 8787 8791 to 8799

31: TRANSPORT OPERATING OCCUPATIONS	9110 to 9119 9130 to 9139 9151 to 9159 9170 to 9179 9190 to 9199
32: MATERIAL HANDLING	9310 to 9319
33: OTHER CRAFTS AND EQUIPMENT	9510 to 9519 9530 to 9539 9550 to 9559 9590 to 9599
34: OTHER OCCUPATIONS (N.E.S.)	9910 9916 9918 9919
97: NOT APPLICABLE	9997
98: DO NOT KNOW	9998
99: NOT STATED	9999

APPENDIX M

Industry Coding

This appendix shows the collapsing of the Standard Industry Codes (SIC) into the 18-category grouping.

18 Category Grouping	**Standard Industry Codes (SIC)**
01: TRADITIONAL PRIMARY SECTOR	011 to 017 021 to 023 031 to 033
02: NON-TRADITIONAL PRIMARY SECTOR	041 to 051 061 to 063 071, 081, 082, 091, 092
03: NATURAL RESOURCES INTENSIVE MANUFACTURING INDUSTRIES	101 to 109 111 to 114 121, 122 161 to 163 169, 252, 254, 259, 279, 283 295 to 297 301 351, 352, 354, 355, 358, 359, 361 369, 397
04: LABOUR INTENSIVE MANUFACTURING INDUSTRIES	171 182, 183, 191, 192, 193, 199 243 to 245 249, 256, 261, 264, 269 302 to 304 309, 328, 392, 399
05: SCALE-BASED MANUFACTURING INDUSTRIES	151, 152, 159, 251 271 to 273 281, 282, 284, 291, 292, 294, 299, 305 323 to 327 329 356, 357 371 to 373
06: PRODUCT DIFFERENTIATED MANUFACTURING INDUSTRIES	306 to 309 311, 312, 319 331 to 333 336, 338, 339, 377, 393
07: SCIENCE-BASED MANUFACTURING INDUSTRIES	321, 334, 335, 337 374 to 376 379, 391

08: CONSTRUCTION	401, 402, 411, 412, 421 422 to 427,429, 441, 449
09: DISTRIBUTIVE SERVICES: TRANSPORTATION AND STORAGE	451 to 459 461, 471, 479
10: DISTRIBUTIVE SERVICES: COMMUNICATION AND PUBLIC UTILITIES	481 to 484 491 to 493 499
11: DISTRIBUTIVE SERVICES: WHOLESALE TRADES	501, 511 521 to 524 531, 532 541 to 543 551, 552 561 to 563 571 to 574 579 591 to 599
12: CONSUMER SERVICES: RETAIL TRADES	601 to 603 611 to 615 621 to 623 631 to 635 639, 641 651 to 659 691, 692
13: CONSUMER SERVICES: PERSONAL, RECREATIONAL AND MISCELLANEOUS	911 to 914 921, 922 961 to 966 969 971 to 974 979, 985, 986 991 to 996 001 (RECODED VALUE; ORIGINAL VALUE 999)
14: BUSINESS SERVICES: FINANCE, INSURANCE AND REAL ESTATE	701 to 705 709, 711, 712, 721, 722, 729 731 to 733 741 to 743 749, 751, 759, 761
15: SERVICES TO BUSINESS MANAGEMENT	771 to 777 779 982 to 984
16: COMMUNITY SERVICES: EDUCATION AND RELATED	851 to 855 859 981

17: COMMUNITY SERVICES: HEALTH AND WELFARE	861 to 869
18: PUBLIC ADMINISTRATION	811 to 817 822, 823 825 to 827 832 835 to 837 841
97: NOT APPLICABLE	997
98: DO NOT KNOW	998
99: NOT STATED	999