Income Statistics Division

62F0026MIE - 01003

Methodology of the Survey of Household Spending

Prepared by:
Sophie Arsenault
Johanne Tremblay

October 2001

Statistics Canada Statistique Canada

Canada

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 product and related statistics or services should be directed to: Client Services, Income Statistics Division, Statistics Canada, Ottawa, Ontario, K1A 0T6 ((613) 951-7355; (888) 297-7355; income@statcan.ca) 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 dialing 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**

Ordering/Subscription information

All prices exclude sales tax

Catalogue no.62F0026MIE-01003, is available on internet for free. Users can obtain single issues at: http://www.statcan.ca/cgi-bin/downpub/research.cgi.

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner and in the official language of their choice. 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 your nearest Statistics Canada Regional Reference Centre.

Statistics Canada
Income Statistics Division

Methodology of the Survey of Household Spending

Published by authority of the Minister responsible for Statistics Canada

October 2001

Catalogue no. 62F0026MIE - 01003

Frequency: Irregular

Ottawa

La version française de cette publication est disponible sur demande

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 co-operation and goodwill.

Abstract

This document provides a detailed description of the methodology of the Survey of Household Spending. Topics covered include: target population; sample design; data collection; data processing; weighting and estimation; estimation of sampling error; and data suppression and confidentiality.

TABLE OF CONTENTS

1. INTRODUCTION 9

2. TARGET POPULATION 10

3. SAMPLE DESIGN 10

3.1 Size and allocation of the SHS sample 10
3.2 LFS sample design (cluster selection) 12
3.3 Selection of the SHS sample 14
3.4 Specifics of the design for the territories 15

4. DATA COLLECTION 15

4.1 Data collection methodology 16
4.2 Interviews and follow-up procedures 17
4.3 Supervision and controls 17
4.4 Non-response to the SHS 18

5. DATA PROCESSING 18

5.1 Coding and data entry 18
5.2 Edit and imputation 18

6. WEIGHTING AND EST IMATION 20

6.1 Sampling weight 20
6.2 Non-response adjustment 20
6.3 Adjustment using auxiliary information 21
6.4 Adjustment for influential data 22
6.5 Estimates 23

7. ESTIMATION OF SAM PLING ERROR 23

7.1 Model for approximating the CV for domain estimates 24
7.2 Model for approximating the CV based on the microdata file 24

8. DATA SUPPRESSION AND CONFIDENTIALITY 24

8.1 Suppression of unreliable data in estimate tables 24
8.2 Confidentiality of microdata files 25

9. CHANGES IN THE SURVEY METHODOLOGY 25

BIBLIOGRAPHY 26

APPENDIX 1

Formula for computation of the variance of estimates by the jackknife method 27

APPENDIX 2

Formula for approximating the CV for a domain (a population subgroup) 28

1. INTRODUCTION

The Survey of Household Spending (SHS) is an annual survey that collects information from Canadian households about spending habits, dwelling characteristics and household equipment.

The SHS was first conducted in January 1998 to collect household spending data for 1997. It replaced the periodic Family Expenditure Survey (FAMEX), generally conducted every four years.[1] The annual survey was introduced to meet a need for more accurate and more frequent provincial data for the National Accounts. As a result, it uses a larger sample. Another major change brought in with the new survey was a complete redesign of the questionnaire. The amount of detail required on expenditures was reduced substantially, which decreased interview times by about 33%.

The SHS has many objectives, as it serves as a valuable data source for a number of Statistics Canada products and for many users outside the Agency. For example, SHS estimates are used in preparing National Accounts estimates of personal spending on goods and services at the national and provincial levels. Data provided by the survey are also used to update the basket used in computing the Consumer Price Index. They are used to set low-income cut-offs, which are in turn used by the Survey of Labour and Income Dynamics to determine the percentages of low-income individuals and families. In addition, SHS data are one of the five data sources used to develop the Social Policy Simulation Model, which analyzes the impact of various economic and social policies. Since the Household Facilities and Equipment Survey was discontinued, the SHS has become the source of data on dwelling characteristics and household equipment needed by Canada Mortgage and Housing Corporation and many outside users.

SHS data are collected through personal interviews conducted with a sample of households in Canada’s ten provinces and three territories. The households are contacted early in the year, between January and March, and are asked about their expenditures for the previous calendar year. The data are then edited and weighted. Output from the survey includes tables and microdata files needed by the various users.

This document provides a detailed description of the survey methodology: sample design, data collection and processing, production of estimates and other products, and dissemination rules. For a more general description of the methodology, consult the Users’ Guide for each of the survey years [1].

[1] The last Family Expenditure Survey covered the 1996 reference year. National data were also collected in 1992, 1986, 1982 and 1978. In some years, such as 1990 and 1984, the survey was conducted in major cities only.

2. TARGET POPULATION

The target population of the SHS consists of individuals living in Canadian private households, excluding official representatives of foreign countries living in Canada and their families, and residents of Indian reserves and crown lands. The “private households” restriction means that the target population also excludes residents of institutions, such as prisons, chronic-care hospitals and senior citizens’ residences; members of religious orders and other communal colonies; members of the Armed Forces living in military camps; and individuals living permanently in hotels and rooming houses.

The survey covers nearly 98% of the population of the ten provinces. In the Yukon, Northwest Territories and Nunavut, the coverage is 81%, 92% and 89% of the population, respectively (or 80%, 93% and 90% of households). Note that in these regions, individuals living in very small communities (generally consisting of fewer than 100 households) or in unorganized areas are excluded from the target population.

Until 1996, coverage of the territories by the periodic Family Expenditure Survey was limited to the cities of Whitehorse and Yellowknife. In the 1997 Survey of Household Spending, that coverage was extended to 78% for Yukon and 70% for the Northwest Territories and Nunavut. The current coverage was achieved with the 1998 survey. Starting with the 2000 SHS, the territories are surveyed only once every two years to reduce the response burden.

3. SAMPLE DESIGN

Expenditure data are collected from a sample of households selected according to a multi-stage, stratified sampling design. This design varies according to the level of urbanization, but generally consists of a two-stage sample for which the first level is an area sample, i.e., a sample of geographic areas called clusters. In the second stage, dwellings are selected from a list of all private dwellings in the selected clusters. All the selected dwellings that are inhabited by individuals from the target population constitute the survey sample.

To minimize operating costs, the SHS uses largely the same sample design as the Labour Force Survey (LFS). The dwellings in the SHS sample are selected from LFS sample clusters, but the two surveys use different dwelling samples. The main aspects of the LFS cluster sample design are described in section 3.2. A more detailed description is available in the LFS methodology publication [2]. The characteristics used in selecting SHS dwellings from LFS sample clusters are covered in section 3.3. The sample design used in the territories is different; its specifics are described in section 3.4. The section immediately below explains how the sample is allocated among the provinces and territories.

3.1 Size and allocation of the SHS sample

The size of the SHS sample may vary from year to year. When the survey was launched in 1997, the sample size was set at about 24,000 households, some

67% more than in the FAMEX sample. Since then, the sample size has fluctuated up or down slightly depending on budgetary pressures and whether the territories were included in the survey. The total sample size for each year is shown in Table 3.1.

Each year, the total sample is allocated among the provinces and territories (when the latter are included) so as to obtain estimates of similar reliability. More specifically, the sample allocation is based on the variability of income in each province: a larger proportion of the sample is allocated to provinces where the difference between the highest and lowest incomes is greater. Population size is also taken into account, though to a much lesser extent. For the territories and Prince Edward Island, which has a much smaller population than the other provinces, the sample size is predetermined to ensure that the sample does not contain an excessive proportion of the population. At present, the SHS samples about 4% of the population in each territory and 2% of the population in Prince Edward Island.

Response rates for previous surveys are used to adjust the sample size of each province and territory. Vacancy rates (the proportion of unoccupied dwellings) are used for the same purpose at the provincial and subprovincial levels. The rates are provided by the most recent LFS data for the January-to-March period corresponding to the SHS collection period.

Lastly, each provincial or territorial sample is distributed in direct proportion to the size of the population in the census metropolitan areas.[2] The proportion of the sample allocated outside the census metropolitan areas matches the LFS allocation [2], except in the Northwest Territories and Nunavut, where the LFS is not conducted.

Table 3.1 presents the SHS's provincial and territorial sample sizes for each year since 1997 in terms of the number of households (excluding selected dwellings that were unoccupied or out of scope). In the 1997 survey, a larger proportion of the sample was allocated to Newfoundland, Nova Scotia and New Brunswick because we expected the sample size to increase in subsequent years and we wanted to assign the increased sample right away to the provinces that had signed the goods and services tax harmonization agreement. It turned out later that the survey budget limited the sample size to a maximum of 24,000 households, and the sample allocation was adjusted in subsequent years.

[2] And for the cities of Charlottetown, Summerside, Whitehorse, Yellowknife and Iqaluit.

Table 3.1
Sample size (number of households) by province or territory

Provinces and Territories	Sample size (number of households)			
	SHS 1997	SHS 1998	SHS 1999	SHS 2000[3]
Canada	**23,842**	**20,236**	**23,518**	**20,877**
Newfoundland	1,997	1,343	1,937	1,794
P.E.I.	795	807	822	822
N.S.	2,424	1,573	2,199	2,040
N.B.	2,044	1,406	1,957	1,821
Quebec	3,122	2,848	2,710	2,516
Ontario	3,362	3,056	3,453	3,202
Manitoba	1,772	1,739	2,034	1,882
Saskatchewan	1,478	1,721	1,837	1,697
Alberta	2,743	2,186	2,519	2,336
B.C.	3,010	2,590	2,985	2,768
Total, provinces	**22,747**	**19,269**	**22,453**	**20,877**
Yukon	451	383	403	0
N.W.T.	644	383	414	0
Nunavut		201	248	0
Total, territories	**1,095**	**967**	**1,065**	**0**

3.2 LFS sample design (cluster selection)

The LFS sample design is based on data from the Census of Population and is redesigned after each decennial census to reflect changes in the population. The current design is based on 1991 Census data.

The principles underlying the LFS sample design are the same for every province. First, each province is divided into a number of geographic regions based on the intersections of economic regions (ERs) and Employment Insurance Economic Regions (EIERs). In particular, every census metropolitan area forms a geographic region since it is an EIER.

Each geographic region is then divided into types of areas, primarily urban areas, rural areas and remote areas. The sample design varies according to the type of area.

Urban areas

In some major cities with large numbers of apartment buildings, both an apartment list frame and an area frame are used. In other urban areas, only an area frame is employed.

[3] For 2000, the figures are approximations based on the vacancy rate and the in-scope rate in the previous survey. The exact number of households in the sample is not known until interviewers have visited the selected dwellings and eliminated those which are unoccupied or are occupied by out-of-scope individuals.

An area frame is a list of geographic zones making up each area. These zones are combined to form strata. There can be up to three levels of stratification. At the top levels, the aim is generally to form geographically compact and contiguous strata, whereas at the bottom level, the requirement is for final strata that are as homogeneous as possible with respect to certain socio-economic characteristics. In a few large cities,[4] separate strata are formed from enumeration areas with high average household incomes (about $100,000 or more).

To reduce collection costs, the households that make up the final stratum are not selected directly. Instead, the stratum is divided into clusters. In urban areas, the clusters may be combinations of block-faces, enumeration areas (EAs) or parts of EAs. Then clusters are selected (usually six, sometimes a multiple of six) in each stratum with a probability proportional to cluster size. For example, if one cluster is twice as large as another is, the former will be twice as likely to be selected as the latter.

The apartment list frame is a list of apartments prepared using information from the Canada Mortgage and Housing Corporation. This frame provides better representation of apartment residents and minimizes the effect of cluster growth due to construction of new apartment buildings. In some cities,[5] apartment strata are divided into two categories: low-income (where the average household income is under $20,000) and regular. For each stratum in the frame, apartment buildings are selected for the first-stage sample with a probability proportional to the number of apartments in the building.

In low-density urban areas, which are highly dispersed towns, a different sample design is used. Sampling is done in three stages: first, towns are selected within the strata; then, clusters (block-faces) are chosen within the towns; and finally, dwellings are selected within the clusters.

Rural areas

Only an area frame is used in rural areas. Geographic strata are formed by combining two or three census divisions, which are then subdivided, where numbers permit, to form strata that are homogeneous with respect to certain socio-economic characteristics. In the first stage of sampling, enumeration areas are selected within each final stratum with a probability proportional to the number of households in the EA.

In low-density rural areas, a variation on the sample design is used. Two or three primary sampling units consisting of a group of six EAs are selected in the first stage, and then a sample of dwellings is selected within each of the EAs of the selected primary sampling units.

[4] Montreal, Ottawa, Toronto, Hamilton, London, Winnipeg, Calgary and Vancouver.

[5] Montreal, Ottawa-Hull, Toronto, Winnipeg, Calgary, Edmonton and Vancouver.

Remote areas

The northern parts of the provinces (excluding the Maritimes) are, for the most part, sparsely populated. Samples for those areas are usually selected in two steps. First, a sample of EAs and of agglomerations known as places is selected. Three-stage sampling is also used in one remote area in Quebec.

Places with fewer than 10 households or 25 persons are excluded from the sample design, as are EAs with fewer than 25 households. Despite these exclusions, the design covers 90% of the population of remote areas in the provinces.

3.3 Selection of the SHS sample

Interviewers visit the clusters selected in the LFS sample design and make a list of all the private dwellings they contain. From that list, one sample is chosen for the LFS and a different one is selected for the SHS. Dwellings are selected by systematic sampling.

Since the SHS uses a much smaller sample than the LFS, dwellings are not selected in every LFS cluster. The LFS is a panel survey in which households remain in the sample for six months. The LFS sample was designed so that it could be divided into six representative subsamples to permit rotation of one sixth of the sample each month. That is why six clusters (or a multiple of six) are selected in each final stratum, one per rotation group. This method makes it easy to select a smaller sample for another survey since a subset of the rotation groups can be used. This is generally the approach used for LFS supplementary surveys. For the SHS, the number of rotation groups is determined at the stratum level according to the survey’s specific needs with regard to provincial and subprovincial allocation of the sample. In some instances, only part of a rotation group is needed. Where that is the case, households are randomly removed from the sample.

The dwelling sample is obtained following the cluster listing operation. Since the sampling rates are predetermined, there may be a difference between the expected and actual sample sizes if the number of dwellings on the list differs from the number used in developing the survey’s sample design. To keep collection costs in check (since cluster sizes tend to increase) and prevent significant disparities in interviewer workloads, two methods are used to control the sample size.

The problem is usually corrected by randomly removing some of the originally selected dwellings. This process of keeping the sample size at the desired level is known as sample stabilization. When the number of dwellings increases sharply in certain urban areas, cluster subsampling is used instead. There are three options, depending on how large the increase is and how similar the new dwellings are to others in the same stratum: form subclusters; create a new stratum; or subsample the dwellings in the cluster.

3.4 Specifics of the design for the territories

Since the SHS has to cover all of the territories, unlike FAMEX, which included only Whitehorse and Yellowknife, a new sample design was introduced for the territories in the 1998 survey.[6] This design was different because it had to reflect the fact that a large portion of the population is scattered among low-density communities. This characteristic has a major impact on the survey's collection costs. Hence, individuals living in unorganized areas, very small communities (generally fewer than 100 households) or inaccessible areas, are excluded from the survey's target population.

Despite this difference in coverage levels, which results in the exclusion from the SHS of about 19% of the Yukon's population, the method used in this territory is similar to the one used in the provinces since the LFS is also conducted in the Yukon.

In the Northwest Territories and Nunavut, a specific design was developed for the SHS since the LFS did not collect information there. That design is based on 1996 Census data, whereas the LFS design used for Yukon and the provinces is based on 1991 Census data.

The cities of Yellowknife and Iqaluit each form a separate stratum divided into clusters, with a sample of dwellings selected in each cluster. Other communities are combined into two or three strata on the basis of socio-demographic characteristics such as population size, proportion of Native People and average household income. Each community makes up a cluster, and a sample of two or three clusters is selected in each stratum. Then a sample of about 30 dwellings is chosen in each of the selected clusters.

SHS data for the 2000 reference year were not collected in the three territories, following the decision to survey them only every other year so as to ease the heavy response burden on their populations.

4. DATA COLLECTION

The SHS collects information about the entire budget of Canadian households on a voluntary basis. This information includes expenditures, income, and changes in assets and debts over the 12-month period from January 1 to December 31 of the reference year.

The SHS also gathers information about dwelling characteristics and the household equipment owned by households. This information reflects the situation on December 31 of the reference year.

[6] The 1997 survey was transitional, as the probability selection method for the territories had not yet been finalized. On the basis of an arbitrary choice of communities and the sample for Whitehorse and Yellowknife, we were able to produce estimates representing 78% of Yukon's population and 70% of the Northwest Territories and Nunavut combined.

4.1 Data collection methodology

Interviewers collect the data in face-to-face interviews with respondents. The interviews are conducted in the first three months of the year following the survey's reference year.

The SHS is a recall survey, which means that respondents have to remember the expenditures they made during the one-year reference period. To reduce the recall effort and help them provide more accurate information, respondents are encouraged to consult records relating to the reference period, such as mortgage statements, cheque registers, credit card account statements, and income tax returns. For items purchased at regular intervals, information is generally collected by asking respondents the quantity they bought, the frequency of the purchases, and the typical price, and then those figures are used to derive estimates of annual expenditures for the household. In particular, annual figures for food expenditures are usually derived from data collected over a short period (a week or a month). The SHS collects only total food expenditures. Detailed data are collected once every four years by means of an expenditure diary in the Food Expenditure Survey.

The SHS questionnaire collects information about the household, such as expenditures for housing, furniture, food, transportation and recreation. Some information must be supplied for individual household members, such as personal income, taxes and clothing expenditures. This information is often obtained by proxy.

Members of a household

To obtain expenditure data for a household, we must first accurately identify its members. The person or group of people who occupies a dwelling constitutes a household. For the SHS, the members of that household are defined as follows:

i) all persons living in the dwelling at the time of the interview who have no permanent residence elsewhere and are not members of another household;

ii) any persons who were members of the household during the reference period, or part of the reference period, even if they do not live there at the time of the interview.

In reporting a household's income and expenditure, it is important to include the income and expenditure of members who have left the household and those who joined the household during the reference year.[7] In the latter case, the data must reflect the portion of the year during which the individual was a member of the household.

[7] For persons who joined a household, it is necessary to determine whether they were previously living in a household that no longer exists. If so, the former household had no chance of being selected. The data are collected on a different questionnaire for the portion of the reference period preceding the change in households.

Another possibility is that a household existed for only part of a reference year. That is the case, for example, where two young adults living with their parents get married and form a new household during the reference period. The expenditures of such households cover only part of the reference year. Such households receive special treatment in the computation of certain estimates. This point is discussed at greater length in section 5.5.

4.2 Interviews and follow-up procedures

The interviews are conducted by Statistics Canada interviewers, many of whom also collect information for the LFS. The interviewers receive special training for the SHS.

A week before visiting, the interviewer sends the occupants of the selected dwellings a letter of introduction emphasizing the survey's importance and the confidentiality of the information collected. He or she then visits the household to conduct the interview. If the timing is inconvenient, the interviewer makes an appointment to return at a more convenient time. If there is no one home, many additional attempts are made to contact the household, for example, visiting at different times of day, or consulting reverse directories to find out the occupant's telephone number.

Because a wide range of information is needed, lengthy interviews may be required in some cases, and the interviewer may have to visit more than once to get all the information. On average, the interview takes about one hour and forty minutes. At the end of the interview, respondents may keep a summary of the expenditures they have reported for their own records.

If a person refuses to take part in the SHS, the Regional Office mails a letter to the dwelling stressing the importance of the survey and the household's cooperation. Next, the interviewer makes a second visit (or call). If the interviewer is unable to persuade the household to take part, he or she will prepare a non-interview report. Depending on the comments provided, the senior interviewer will decide whether to make further refusal conversion attempts.

4.3 Supervision and controls

All SHS interviewers report to senior interviewers, who are responsible for ensuring that the interviewers are familiar with the survey's concepts and methodology, for periodically monitoring their work, and for reviewing completed documents. Senior interviewers are supervised by program managers working at Statistics Canada's regional offices.

The interviewer carries out the initial edit, making sure that the information is complete in all sections of the questionnaire. The questionnaires are then checked by senior interviewers.

Since respondents' recall is a key component of the quality of SHS data, one of the controls involves measuring the difference between receipts (income and other money received by the household) and disbursements (expenditures plus net change in assets and liabilities) reported by the household. If the difference is more than 10% of the larger of receipts or disbursements, the interviewer or senior interviewer will contact the respondents again to obtain further information and attempt to identify errors or omissions.

4.4 Non-response to the SHS

Despite all the effort put into collecting the information, there are always some non-respondent households. For example, contact could not be made; unusual circumstances such as illness or death prevented the interview; or the household members refused to take part in the survey. For each survey year, the report on data quality contains detailed information about the non-response rates [3]. The collection non-response rates for the past few years are shown in Table 4.1. Also shown in the table are the final non-response rates, which take into account those households, which were excluded following data processing (see section 5.4).

Table 4.1
SHS non-response rate

Reference year	Collection non-response rate			Final non-response rate (at estimation)
	TOTAL	No contact	Refusal	
1997	20.7	5.8	15.0	24.4
1998	20.7	4.9	15.8	23.6
1999	23.6	5.9	17.7	26.8

5. DATA PROCESSING

The main steps in the processing of SHS data are response coding, data entry, editing, imputation of partial non-response, identification of usable data, and weighting. The latter will be covered in section 6.

5.1 Coding and data entry

Very few questions in the SHS require coding. Coding is done by the interviewer and checked by the senior interviewer. Then the questionnaires are put into batches of 20 and the data is keyed in at Statistics Canada’s regional offices. Data entry is checked by selecting a sample of questionnaires from each data entry operator for rekeying. If the number of errors for a questionnaire exceeds a certain threshold, the entire batch is sent back for rekeying. The size of the sample selected for editing depends on the past performance of the data entry operators.

5.2 Edit and imputation

The first step in the automated edit process is carried out after each questionnaire has been checked manually by the interviewer and the senior interviewer. There are a number of “must pass” rules that check for consistency between answers in the questionnaire. The edit also identifies unusual situations that might require correction. This part of the automated edit is done at Statistics Canada’s regional offices, so that respondents can be contacted if additional information is needed to resolve inconsistencies in their responses. Problems identified during this edit are dealt with by members of specially trained questionnaire resolution teams. Subsequently, the data are transmitted to Head Office for further editing and correction of invalid responses.

In cases of partial non-response (where the respondent has failed to answer only some of the questions), the missing data are imputed. The imputation method depends on whether the data are categorical or continuous. Categorical variables can take only specific values (such as yes-or-no questions and type-of-dwelling questions), whereas continuous variables can take any numerical value (such as income and expenditure).

Categorical data, which occur mostly in the dwelling characteristics and household equipment sections of the questionnaire, are imputed by a “hot deck” method. In this procedure, a donor household is selected at random from a group of respondents that have similar characteristics.

Income and expenditure data are imputed by the nearest-neighbour method. This technique involves forming groups of similar households or individuals based on certain criteria (e.g., province of residence). Within those groups, each household requiring imputation (recipient) is matched to a household that has a complete questionnaire (donor) and resembles the other most closely with respect to certain characteristics (e.g., income, number of children, number of adults). The donor’s data are imputed to the recipient as long as they satisfy the edit requirement for consistency with the data reported by the recipient.

The SHS collects information about various aspects of household budgets. Imputation is not done for the whole questionnaire but by sections that generally correspond to the questionnaire’s sections, i.e., by groups of interrelated questions. This maximizes the number of potential donors in the sense that a household which leaves only one question unanswered, for example, can serve as a donor for those sections which it answered in full. This approach implies that one household could receive data from more than one donor. That possibility is minimized by the fact that we look for the household that exhibits the greatest possible similarity with respect to certain characteristics, which are often the same from section to section. It is important to note that all the questions in one section are imputed by the same donor, which preserves the relationships between questions.

5.3 Identification of usable data

The data for certain households whose questionnaires are at least partially complete may be rejected during processing. There are two main reasons for rejection. First, when a large portion of the income or expenditure questions is left unanswered, the questionnaire is deemed incomplete and is not used. Second, questionnaires are considered unusable if, following processing (editing for consistency and imputation of missing data where necessary), the difference between receipts (income and other money received by the household) and disbursements (expenditure plus net change in assets and liabilities) reported by the household is greater than 20%.

Once identified, the usable data are weighted to produce estimates.

6. WEIGHTING AND ESTIMATION

Estimates are based on the premise that each household in the sample represents a certain number of households in the target population, as it was defined in section 2. Accordingly, each respondent household is assigned a survey weight, which indicates how many households in the population it represents. The survey weight is generally the product of three factors: the sampling weight, which incorporates data from the sample design; a non-response adjustment factor, which compensates for non-respondent households; and an adjustment factor that reflects characteristics from sources other than the survey. Also included in the calculation of the survey weight is an adjustment factor for influential data, though it affects very few households.

6.1 Sampling weight

A household's sampling weight is the inverse of its probability of being included in the sample. Since the SHS is a probability survey, every household in the target population has a known probability of being selected for the sample. For example, if a household's selection probability is 1 over 200, its weight will be 200.

For a given sample allocation, the sampling weight is determined by the sample design. The SHS uses the LFS sample design, which is self-weighting by stratum (i.e., the sampling weights set when the design is developed are equal within each stratum). If the sample design and sample allocation remained unchanged, the initial weights could be used. However, the stabilization and subsampling steps described in section 3.3 alter the initial selection probabilities. The LFS sampling weights are adjusted to reflect those changes.

Since the SHS sample is a subset of the six LFS rotation groups, the SHS sampling weights are determined by adjusting the LFS sampling weights in accordance with the number of rotation groups used. This factor may vary from stratum to stratum since the number of rotation groups selected in each stratum is based on the SHS's specific sample allocation requirements.

6.2 Non-response adjustment

In instances where the respondent has failed to answer only some of the questions, the missing data are imputed by the methods described in section 5.2. In cases of total non-response (i.e., where the household cannot be contacted, where household members refuse to respond, or where the data provided cannot be used), the weights are adjusted.

Adjustment of weights for non-response is based on the premise that responding households can be used to represent all households, both responding and non-responding. For the purposes of this adjustment, the sample is first divided into non-response classes defined so as to increase the chances that respondents will have characteristics similar to non-respondents.

The non-response classes correspond to different levels of urbanization in each province or territory except Quebec, Ontario and British Columbia; these three provinces are first divided into two or three subprovincial regions. The urbanization levels are generally as follows: the principal metropolitan area, urban areas with a population of 100,000 to 500,000, smaller urban areas, and rural or remote areas. In some regions or provinces, some levels may have to be combined because their samples are too small. In each territory, there are only two non-response classes: the principal city and the rest of the target population.

High-income household strata also form specific non-response classes in each province where they exist. Note that non-response areas do not overlap; when combined, they cover the entire target population.

For each non-response class, a non-response adjustment factor corresponding to the inverse of the class's weighted response rate is computed. Expressed another way, this factor is the ratio of the number of households sampled, multiplied by the sampling weights so that they represent the class's households, to the number of weighted respondent households. To ensure that the non-response adjustment factors are not excessive, some non-response classes are combined when the adjustment factor is greater than 2.

6.3 Adjustment using auxiliary information

In theory, estimates can be produced by multiplying the sampling weight by the non-response adjustment. However, the estimates can be improved with auxiliary data about the target population. If the auxiliary data are correlated with the principal characteristics measured by the survey, more reliable estimates can be produced. For example, a household's expenditures are correlated with its size. A poor sample allocation with respect to household size would have an impact on the expenditure estimates. On the basis of auxiliary data on the number of households by size, the weights can be adjusted to obtain the real distribution of the number of households.

In the SHS, various sources of auxiliary data are used to adjust the weights. First, postcensal estimates produced by Statistics Canada's Demography Division provide population counts by age group and sex for each province and territory. Those counts are population projections for a given period based on census data and information from administrative records, such as births, deaths, immigration and emigration. After the counts have been adjusted to reflect the SHS's target population at the end of the survey's reference year, the estimates for 18 different age-sex groups for each province[8] are used to adjust the weights.

[8] In the territories, only four groups are used: two age groups for each sex.

The 18-and-over and under-18 population counts for selected metropolitan areas[9] are also used.

Estimates of the number of households by size (one, two or three or more persons) for each province and territory and by selected household type are also used to adjust the sample's representativeness in those groups. For household type, we use specifically the number of households composed of lone-parent families and the number of households composed of parents with never-married children.

In order to remedy certain problems observed in the income distribution of survey respondents, the population counts for some income classes based on administrative sources are also used to adjust the SHS weights. The data are the numbers of individuals who earned wage and salary income as reported by employers. Six income classes are used. Their boundaries are based on the following percentiles of the distribution: 25, 50, 60, 75, 90 and 95.[10] Since the administrative data for the reference year are not available when weighting of the SHS is carried out, the class counts are projected on the basis of administrative data for the previous year and trends in the LFS distribution of individuals by salary.

Adjustment of the weights to reflect all of the above figures is carried out simultaneously using a variant of the generalized regression (GREG) estimator based on the weighting method proposed by Lemaître and Dufour [4]. This method allows concordance between the survey estimates and the estimates from auxiliary sources, while ensuring that, after adjustment, all members of the same household still have the same weight. The adjustment factor produced by the GREG estimator is then applied to the sampling weight and the non-response adjustment factor to generate the household's final weight.

6.4 Adjustment for influential data

Since expenditures have a highly asymmetric distribution, the samples used in expenditure surveys are prone to having extreme values. If a particular household has a combination of extreme values and a high weight, it may contribute disproportionately to the estimates. The presence of such influential data has a serious effect on estimates of totals and averages, chiefly at the provincial level and for subsets of the population.

To minimize the negative impact of these influential data on interprovincial comparisons and trend estimates, the weights of some households are adjusted to reduce their contribution to the estimates. To make such an adjustment, we use auxiliary information in the form of the distribution of individuals' incomes, based on their tax data. Since total expenditure is very closely tied to income, the correction will weaken the impact on estimates of total expenditures.

[9] St. John's, Halifax, Saint John, Quebec City, Montreal, Ottawa, Toronto, Winnipeg, Regina, Saskatoon, Calgary, Edmonton, Vancouver and Victoria.

[10] For some provinces, the 98th or 99th percentile is used instead of the 95th percentile.

The approach involves identifying the few individuals who make a major contribution (usually more than 1%) to the provincial estimates of total income. If necessary, we then adjust the weight of the individuals' households to ensure that the estimated number of individuals with that level of income does not exceed the number obtained from the distribution of individuals' tax data. Then the auxiliary information adjustment described earlier is applied again to ensure consistency.

It is important to note that the influential data adjustment, which focuses on extreme data, affects the weight of very few households (usually fewer than five for the whole sample).

6.5 Estimates

When the SHS estimates of average household expenditure are produced, households that existed for only part of the reference year, called part-year households, are excluded. (See section 4.1). Part-year households are composed entirely of persons who were members of other households for part of the survey year, as in the example of the two young adults living with their parents who get married and form a new household during the reference period. There are also households composed solely of persons who immigrated to Canada during the reference period. Part-year households make up a very small proportion of the household sample (less than 4%).

On the other hand, when estimates of the total expenditures of the Canadian population or a subpopulation are produced, the entire sample of households is used.

7. ESTIMATION OF SAMPLING ERROR

After the estimates have been computed, their reliability must be measured; in other words, the sampling error associated with each estimate must be estimated. The usual measure of sampling error is the standard error or the coefficient of variation (which is simply the standard error expressed as a percentage of the estimate). The standard error is the degree of variation observed in the estimates following the selection of one particular sample rather than another. Since the SHS is a probability survey, the standard error of its estimates can be estimated.

In the SHS, the jackknife method is used to estimate the standard error. This technique involves creating replicates of the sample based on SHS data. The same number of replicates is generated, as there are primary sampling units (PSUs), with one PSU being removed from the sample for each successive replicate. Each PSU belongs to a stratum, and when the PSU is removed, the sampling weights of the other PSUs in the stratum are adjusted accordingly. Then the final estimates are recomputed with the auxiliary data adjustments described in section 6.3 applied to the replicates. By repeating this operation for each PSU in the sample, we obtain as many estimates as there are PSUs. The variability of these estimates is used to estimate the standard error of the estimate for the entire sample. The mathematical formula is shown in Appendix 1.

It is important to note that in the SHS, estimates of the standard error or coefficient of variation ignore the fact that some data were imputed. As a result, the computed CVs may underestimate the actual values. For most of the survey's variables, the effect of imputation is minimal. The impact of imputed data for each expenditure variable is included in the data quality report for each survey year.

7.1 Model for approximating the CV for domain estimates

For operational reasons, CVs cannot be produced for every characteristic collected by the survey at every level of aggregation of possible interest to users (e.g., by income quintile, household type, level of urbanization, tenure, selected metropolitan areas). The approach suggested to SHS users is to compute an approximate CV using a relationship between the number of households in the sample reporting expenditures for a category and the CV at an aggregate level (generally the national level). That relationship, based on the CV's tendency to grow in direct proportion to the decline in the square root of the number of households reporting a particular expenditure, is illustrated in Appendix 2.

7.2 Model for approximating the CV based on the microdata file

Microdata file users can take another approach to approximating the CVs of estimates. This approximation is generally more effective than the one described above. The method, which is fairly simple to use, is described in greater detail in reference [5]. It can be used only in combination with the microdata file, since the data and weights for all households are needed to compute the approximation.

8. DATA SUPPRESSION AND CONFIDENTIALITY

Steps are taken to ensure that the SHS estimates are sufficiently reliable to be published and that the anonymity of respondent households is maintained.

8.1 Suppression of unreliable data in estimate tables

Since the coefficient of variation is an indicator of data reliability, ideally we would use it to determine whether the estimates should be published or not. Estimates whose estimated CV is greater than 33% are not reliable enough to be released.

However, because so many estimates are produced for the SHS, it is impossible to compute the CV for each one. A study based on FAMEX data showed that the CVs generally reach 33% when the number of households reporting an expenditure approaches 30. This rule is used to determine whether SHS estimates can be published or not. Since it is a rule of thumb, some estimates will be released even if their CVs are above 33%, while others will not be published even though their CVs are under 33%. An assessment of this rule's performance is included in the 1997 data quality report [3].

It should be noted that even if the estimates of average expenditures for a certain type of purchase are not disseminated because they were reported by fewer than 30 households, the data are reflected in the estimates for aggregate components.

8.2 Confidentiality of microdata files

Even though a public use microdata file is produced from SHS data, it is different from the one used by Statistics Canada for the release of estimates. The differences are largely due to a series of measures taken to protect the anonymity of the responding households.

9. CHANGES IN THE SURVEY METHODOLOGY

The introduction of the annual Survey of Household Spending has provided more frequent, more reliable estimates of expenditures, particularly at the provincial level since the total sample was increased and its allocation was revised. A new questionnaire, much less detailed than the one used in the Family Expenditure Survey, was developed for the 1997 SHS. Nevertheless, the survey's methodology has changed little from year to year, with the exception of the auxiliary data adjustment applied during weighting.

For the 1999 survey, population projections based on the 1996 Census took the place of the 1991 projections used in previous surveys. In addition, the weighting strategy was altered as part of a project to harmonize auxiliary data adjustments in Statistics Canada's income surveys. For the SHS, the major changes were the use of many more age-sex groups for population counts, and the introduction of counts for household types and for certain wage and salary classes. As a result of all these changes, a historical revision of estimates from the 1997 and 1998 SHS and the 1996 and 1992 FAMEX is being carried out to ensure comparability in trend analyses.

BIBLIOGRAPHY

[1] *Users' Guide for the Survey of Household Spending,* Statistics Canada (available for each survey year), Catalogue No. 62F0026MIE

[2] *Methodology of the Canadian Labour Force Survey*, Catalogue Number 71-526-XPB

[3] *1997 Survey of Household Spending – Data Quality Indicators*, Household Survey Methods Division (HSMD), Internal document, Statistics Canada (also available for each survey year), Catalogue No. 62F0026MIE

[4] Lemaître and Dufour (1987), *An Integrated Method for Weighting Persons and Families, Survey Methodology*, Vol.13, no 2, pp.211-220, Statistics Canada

[5] Beaumont, J.-F. (2000), *Variance Estimation For A Public Use Microdata File From A Complex Survey*, HSMD Working Paper, HSMD-2000-002F/A, Statistics Canada

APPENDIX 1

Formula for computation of the variance of estimates by the jackknife method

In the jackknife method of estimating the variance of estimates, the variability of the estimates is measured with the following formula:

$$Var(\hat{Y}) = \sum_{h=1}^{H} \frac{n(h)-1}{n(h)} \sum_{i=1}^{n(h)} (\hat{Y} - \hat{Y}_{(hi)})^2$$

where

n(h)	is the number of PSUs in stratum h
$\hat{Y}_{(hi)}$	is the estimate of Y when PSU i is removed from stratum h.

The standard error is the square root of the variance.

APPENDIX 2

Formula for approximating the CV for a domain (a population subgroup)

If CV (Y) represents the CV for the estimate of the household average of a particular characteristic for the whole population, then we can compute an approximate CV for the estimate of that characteristic for a domain (which can be taken as a subgroup of the population, such as a household type, an income quintile or a level of urbanization) using the following equation:

$$CV\ (Y_d) = CV\ (Y) \times \sqrt{\frac{nP}{n_d P_d}}$$

where

n : *number of households in the sample*
P : *estimate of the proportion of households reporting a value > 0 for this characteristic in the population*
n_d : *number of households in the sample for domain d*
P_d : *estimate of the proportion of households reporting a value > 0 for this characteristic in domain d*

The CV, size *n* and proportion *P* for the national level are generally used to calculate the approximations for the various domains. Where we wish to compute an approximate CV for a metropolitan area, we can use provincial values since the domain is entirely within a single province and since provincial CVs will be published for the SHS.