

TABLE OF CONTENTS

1.	Introduction	5
	1.1 NPHS Supplements	5
	1.2 Which data file should you use?	5
2.	Background	5
3.	Objectives	7
4.	Survey Content	7
	4.1 Survey Content of NPHS Supplements	8
5.	Sample Design	9
	5.1 Sample Allocation	9
	5.2 The Rejective Approach	11
	5.3 Sample Selection	11
	5.4 Integration of NPHS with the National Longitudinal Survey of Children (NLSC)	12
	5.5 Sample Design in Quebec	13
	5.6 Sample design for the NPHS Supplement	14
6.	Data Collection	16
	6.1 Questionnaire Design and Data Collection Method	16
	6.2 Tests	16
	6.3 Interviewing	17
	6.4 Supervision and Control	17
	6.5 Non-Response to the NPHS	17
	6.6 Non-response follow-up	18
7.	Data Processing	19
	7.1 Data Capture	19
	7.2 Editing	19
	7.3 Coding	19
	7.4 Creation of Derived Variables	19
	7.5 Weighting	19
	7.6 Suppression of Confidential Information	20
8.	Data Quality	21
	8.1 Response Rates	21
	8.2 Response Rates to the NPHS Supplement	23
	8.2 Survey Errors	24
9.	Guidelines For Tabulation, Analysis And Release	26
	9.1 Rounding Guidelines	26
	9.2 Sample Weighting Guidelines for Tabulation	27
	9.2.1 Definitions of types of estimates: Categorical vs. Quantitative	27
	9.2.2 Tabulation of Categorical Estimates	28
	9.2.3 Tabulation of Quantitative Estimates	28
	9.3 Guidelines for Statistical Analysis	29
	9.4 Release Guidelines	30

10.	APPROXIMATE SAMPLING VARIABILITY TABLES	31
10.1	How to use the C.V. tables for Categorical Estimates	32
10.2	Examples of using the C.V. tables for Categorical Estimates	35
10.3	How to use the C.V. tables to obtain Confidence Limits	38
10.4	Example of using the C.V. tables to obtain confidence limits	38
10.5	How to use the C.V. tables to do a t-test	39
10.6	Example of using the C.V. tables to do a t-test	39
10.7	Exact Variances/ Coefficients of Variation	39
10.8	Release cut-off's for the NPHS Supplement	42
11.	Weighting	43
11.1	Weighting Procedure for the Provinces Outside of Quebec	43
11.1.1	LFS Basic Weights	43
11.1.3	Weight Adjustments to the Basic Weights	44
11.1.5	Further Weight Adjustments For Selected Members	47
11.2	Weighting Procedures for Quebec	51
11.2.1	ESS Weights	51
11.2.2	NPHS Basic Dwelling Weights	51
11.2.3	Further Weight Adjustments to the Basic Weights	52
11.2.5	Further Weight Adjustments for Selected Members	54
Appendix A: Questionnaires: Supplement and Main Survey		
Appendix B: Data Dictionary		
Appendix C: CV Tables		

1. Introduction

The National Population Health Survey (NPHS) is designed to collect information related to the health of the Canadian population. The first cycle of data collection began in 1994, and will continue every second year thereafter. The survey will collect not only cross-sectional information, but also data from a panel of individuals at two-year intervals.

The target population of the NPHS includes household residents in all provinces, with the principal exclusion of populations on Indian Reserves, Canadian Forces Bases and some remote areas in Quebec and Ontario. Separate surveys were conducted to cover the Yukon, the Northwest Territories and the Institutions (long term residents of hospitals and residential care facilities) and will be presented at a later stage.

The National Population Health Survey (NPHS) was conducted by Statistics Canada in 1994-1995. This manual has been produced to facilitate the manipulation of the microdata file of the survey results.

1.1 NPHS Supplements

In addition to the content of the main NPHS, Health Canada sponsored a set of supplementary questions on six subjects. The intent of this supplementary content was to collect more detailed information on certain areas which were of interest, but were not covered in detail in the main survey. The topics covered by the NPHS supplements were: nutrition, safety and accident prevention, smoking, breast-feeding, alcohol use during pregnancy, sexually transmitted diseases and use of health care services.

1.2 Which data file should you use?

The data file to use for your analysis depends on which variables you will be looking at. You should choose the file which gives you the greatest accuracy for the set of variables you are using. The HEALTH file contains detailed information about the selected person; the SUPPS file contains the same variables as the HEALTH file plus the supplementary variables. The main difference between the files is in the number of records. The HEALTH file contains records for all the selected persons who responded to the main NPHS survey. The SUPPS file contains records for those selected persons who were eligible for the supplement and also responded. Therefore THERE ARE FEWER RECORDS ON THE SUPPS FILE THAN ON THE HEALTH FILE.

If your analysis involves only variables collected in the main portion of the NPHS interview, you should use the HEALTH file, since it contains the largest number of records and therefore will give the most reliable result. If your analysis involves variables collected in the supplementary portion of the NPHS interview (either alone, or with variables from the main portion), you should use the SUPPS file. Because the samples are not exactly the same (because of non-response to the supplement), any estimates you produce from the SUPPS file using only variables from the main NPHS will not agree exactly with the results you would obtain from the HEALTH file, although they should be close.

NPHS PUBLIC USE MICRODATA DOCUMENTATION

Any questions about the data set or its use should be directed to:

Health Statistics Division, Statistics Canada
NPHS

Tunney's Pasture
Ottawa, Ontario
K1A 0T6
(613) 951-1746

2. Background

In the fall of 1991, the National Health Information Council (NHIC), recommended that an on-going national survey of population health be conducted. This recommendation was based on consideration of the economic and fiscal pressures on the health care system and the commensurate requirement for information with which to improve the health status of the population in Canada. Existing sources of health data are unable to provide a complete picture of the health status of the population and the myriad of factors that have an impact on health.

Commencing in April 1992, Statistics Canada received funding for development of a National Population Health Survey. The survey was designed to be flexible, and to produce valid, reliable and timely data. Also, it was to be responsive to changing requirements, interests and policies.

3. Objectives

The objectives of the NPHS are to:

- aid in the development of public policy by providing measures of the level, trend and distribution of the health status of the population;
- provide data for analytic studies that will assist in understanding the determinants of health;
- collect data on the economic, social, demographic, occupational and environmental correlates of health;
- increase the understanding of the relationship between health status and health care utilization, including alternative as well as traditional services;
- provide information on a panel of people who will be followed over time to reflect the dynamic process of health and illness;
- provide the provinces and territories and other clients with a health survey capacity that will permit supplementation of content or sample;
- allow the possibility of linking survey data to routinely collected administrative data such as vital statistics, environmental measures, community variables, and health services utilization.

4. Survey Content

These objectives provided only a broad direction for the NPHS, particularly concerning the type of information to be collected. Therefore, survey content was selected according to the following criteria:

NPHS PUBLIC USE MICRODATA DOCUMENTATION

- 1) Information should relate to, and help monitor, the health goals and objectives of the provinces and territories. Where health goals have not been established, for example, at the national level, policy and programs could be considered in the selection of survey content.
- 2) The information should not duplicate data available from other sources.
- 3) With a view to increasing the understanding of health and its determinants, information collected should provide new knowledge in areas that have not been adequately studied.
- 4) The survey should focus on behaviours or conditions amenable to prevention, treatment, or intervention.
- 5) The survey should collect information about conditions that impose the greatest burden, in terms of suffering or cost, on affected individuals, the general population, or the health care system.
- 6) The survey should collect information on factors related to good health, not just those related to illness.

In each household, some limited information was collected from all household members and one person, aged 12 years and over, in each household was randomly selected for a more in-depth interview. Reflecting these guidelines, the questionnaire included components on health status, use of health services, risk factors and demographic and socio-economic status. For example, health status was measured through questions on self-perception of health, functional ability, chronic conditions, and activity restriction. The use of health services was measured through questions on visits to health care providers, hospital care and drug use. Behavioural risk factors include smoking, alcohol use and physical activity. In addition, a special focus of the first survey was psycho-social factors that may influence health, such as stress, self-esteem and social support. Demographic and socio-economic information included age, sex, education, ethnicity, household income and labour force status. A list of the questions asked are provided in Appendix A.

4.1 Survey Content of NPHS Supplements

The supplementary content focussed on collecting more detailed information on six subjects: nutrition, safety and accident prevention, smoking, breast-feeding, alcohol use during pregnancy, sexually transmitted diseases and use of health care services. These supplementary questions were asked only of the one household member who was selected for the in-depth interview on the main NPHS survey, but only if that person was aged 12 or older and if the household was part of the NPHS core sample (as opposed to the additional sample bought in by some provinces). If the selected respondent had children under 12 years of age, he/she would be asked questions on safety pertaining to each child; female respondents with children under the age of 5 were asked about breast-feeding each child and about their use of alcohol during pregnancy with each child. Other topic areas pertained only to the selected respondent.

5. Sample Design

The target population of the NPHS includes household residents in all provinces, with the principal exclusion of populations on Indian Reserves, Canadian Forces Bases and some remote areas in Quebec and Ontario. The Supplement to the NPHS followed the sample design of the NPHS very closely and thus the target population for the Supplement was the same as that for the main survey.

Sample design for the household component of the NPHS

Four factors shaped the design of the household component sample:

- the targeted national and provincial/territorial sample sizes;
- the decision to select one member per household to make up the longitudinal panel;
- the choice of the redesigned Labour Force Survey (LFS) as a vehicle for selecting the sample; and
- the decision to integrate the NPHS with the National Longitudinal Survey of Children (NLSC).

The first three factors resulted, respectively, in the allocation of the sample, the application of a technique (the "rejective method," described later) to improve the sample's representativeness, and the selection of provincial samples outside Quebec.

5.1 Sample Allocation

The NPHS was budgeted for a sample size of 19,600 households. It was further agreed among national and provincial representatives that each province needed a minimum of 1,200 households. Subject to this restriction the provincial sample sizes were obtained by using a well known allocation scheme that balances the reliability requirements at national and regional levels (Kish, 1988). According to this scheme the sample was allocated proportionally to $\sqrt{(0.804W_h^2 + 1/12^2)}$, where W_h is the 1991 Census proportion of households in province h , $h=1,...,12$. This allocation determined the base sample size for each province. Four provinces chose to increase their allotted sample size through the buy-in of additional units.

Within provinces the sample was initially distributed proportionally to the population size. The provincial buy-in samples and the use of a rejective method, described below, affected the sub-provincial allocations. Ontario and Manitoba's buy-in samples imposed minimum requirements by health regions, while N.B. and B.C. paid for additional sample coverage of certain areas only. In B.C. most of the buy-in requirement was met using telephone interviews from a Random-Digit Dialling (RDD) sample of telephone numbers. In applying the rejective method, sample sizes were inflated by the number of households expected to be screened out of the sample.

Table 1 below gives the sample sizes expected by province. Numbers represent in-scope private occupied dwellings before non-response, which was anticipated to be near 10%.

NPHS PUBLIC USE MICRODATA DOCUMENTATION

Table 1: Sample Sizes for the NPHS

Household sample sizes

Province	Original Allocation	Buy-in sample	To interview	Screened out	Total
Nfld	1,220	-	1,221	171	1,392
P.E.I.	1,201	-	1,199	223	1,422
N.S.	1,270	-	1,270	246	1,516
N.B.	1,243	180	1,423	234	1,657
Que.	3,584	-	3,479	-	3,479
Ont.	4,817	2,183	7,001	1,021	8,022
Man.	1,307	493	1,800	324	2,124
Sask	1,287	-	1,288	257	1,545
Alta	1,674	-	1,674	305	1,979
BC (1)	1,996	61	2,057	448	2,505
BC (2)	-	788	788	-	788
TOTAL	19,599	3,705	23,200	3,229	26,429

(1) Excludes RDD portion.

(2) RDD portion.

5.2 The Rejective Approach

The survey content primarily focuses on one member in each sample household who is chosen at random to become the longitudinal panel respondent. The panel underrepresents persons coming from large households, typically parents and children, since they have less chance of being chosen and overrepresents persons coming from small households, often single people or the elderly.

A rejective approach has been adopted to increase the representation of parents and youths in the panel. A portion of the sample is pre-identified for screening. After their member roster is completed, screened households that have no member aged under 25 years drop out of the survey. In order to maintain the required sample sizes, the number of households visited in each province is increased by the anticipated number of households screened out in this way.

The rejective method with an under 25-year old rule was adopted as it performed better than other rejection rules considered. For cost and operational reasons the percentages of screened households was usually limited to 25-30% in Ontario, 37.5-40% in urban areas elsewhere and 25-30% in rural areas. As apartment strata had a high concentration of small households, their sample sizes were reduced instead of applying a rejective method. The rejective approach was also not applied in remote regions because of the high contact costs there, and its use was limited in areas where sample buy-in demands were substantial.

5.3 Sample Selection

The sample design considered for the household component of the NPHS was a stratified two-stage design. In the first stage homogeneous strata are formed and independent samples of clusters are drawn from each stratum. In the second stage dwelling lists are prepared for each cluster and dwellings, or households, are selected from the lists.

In all provinces except Quebec the NPHS used the multi-purpose sampling methodology developed for the redesign of the Labour Force Survey (LFS). That methodology provides general household surveys with clustered samples of dwellings, thus making the design very cost effective for the listing and collection of data.

The basic LFS design is a multi-stage stratified sample of dwellings selected within clusters. Each province is divided into three types of areas (Major Urban Centres, Urban Towns and Rural Areas) from which separate geographic and/or socio-economic strata are formed. In most strata six clusters, usually Census Enumeration Areas (EAs), are selected with Probability Proportional to Size (PPS). In a few cases where the population density is low an additional stage is added by first selecting 2 or 3 large Primary Sampling Units, dividing them into clusters, and drawing a sample of six clusters from each. The number six is used throughout the sample design to allow a one-sixth rotation of the sample every month for the LFS.

The sample of dwellings is obtained after listing operations in sample clusters are completed. As sampling rates are predetermined there are often differences between anticipated and obtained sample counts. Excessive sample yields are corrected by dropping a portion of the originally selected units. This is usually done at aggregated levels and is called sample stabilisation. Note also that sample sizes are inflated to represent dwellings rather than households as approximately 15% of the dwellings are expected to be vacant or otherwise out-of-scope.

The sample design is set up to yield about 60,000 households. Surveys needing smaller sample sizes usually "reserve" from 1 to 6 rotations per province, a rotation being one-sixth of the total sample. Sample stabilisation is used to maintain the sample at desired levels, as when two rotations are reserved but the sample size needed only represents 1.5 rotations.

Requirements specific to the NPHS led to two modifications to this sampling strategy. The number of "reserves" needed was specified at the stratum level rather than the provincial level in order to meet the specific sub-provincial sample size requirements. It was also required that the number of clusters selected per stratum be a multiple of four for variance estimation and seasonal representativity (this allowed strata to have two or more independent samples of four clusters each - one per collection period). As NPHS usually requested only between 2 and 6 clusters per LFS stratum, similar LFS strata were grouped to form larger NPHS strata with the required number of sample clusters.

As a result of these modifications, the NPHS sample of clusters can be considered as a stratified replicated sample where strata are groups of LFS strata and replicates are typically independent, identically distributed samples of 4 clusters each. There were exceptions, but they are not expected to have a significant impact on survey results.

5.4 Integration of NPHS with the National Longitudinal Survey of Children (NLSC)

The National Longitudinal Survey of Children (NLSC) is a household survey which will follow a sample of about 25,000 children under 12 over time. The two surveys are integrated, meaning that common data for children are collected from both and that the NPHS children's sample will be used in NLSC estimates. In the provinces the NPHS is expected to provide a sample of 4,500 to 5,000 children to the NLSC. To obtain this sample size NPHS households where a child is selected for the panel will have the detailed questionnaire for children administered to all children in his or her family (subject to a maximum of 4).

Scheduling constraints required that children not be selected for the NPHS panel before the third survey collection period (or quarter). This distorted the seasonal representativity of children in the panel and reduced their sample size. To increase the sample yield for children without harming the seasonal representation of other household members in the last two quarters it was decided to reassign part of the NPHS sample from the first two quarters to these quarters. As this decision was made after the sample operations described above, the shift was applied to households within clusters rather than to entire clusters.

Figure 1 illustrates how the sample distribution was revised for the integration. The square on the left represents a cluster assigned to quarter 1 or 2. That on the right represents a cluster assigned to quarter 3 or 4. Households are classified by type into: (I) households with children; (II) other households with youths ("Youth" meaning under 25 years old); and (III) households without children or youths. The sample is divided into an "Adult" sample and a "Children" sample. In "Adult" sample households only persons aged 12 or older can be selected for the panel. Panel membership is restricted to children in "Children" households. If there are no children present, then either the household is screened out or a member (non-child) is selected at random for the panel.

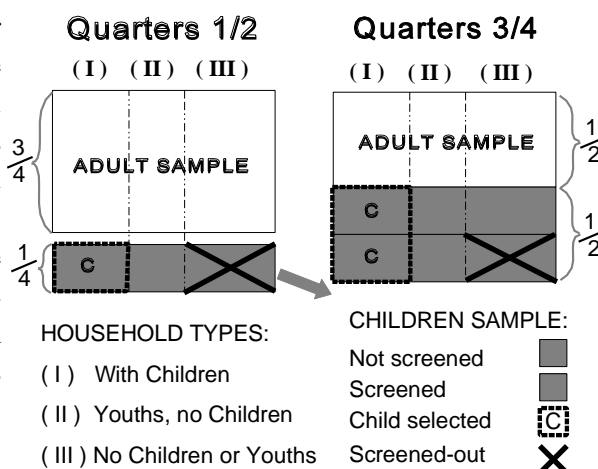


Figure 1: Repartition of the Sample

A quarter of the sample from quarters 1 and 2, and a half from quarters 3 and 4 are designated as "Children" households. "Children" households from quarter 1 or 2 will actually be visited in quarter 3 or 4, respectively. Outside of P.E.I. the rejective method can be applied strictly within the "Children" sample. When the screening rate is at 37.5% all "Children" households are screened. With lower rates some of them will not need to be screened. A 25% screening rate is illustrated in Figure 1. All the "Children" households from quarters 1 and 2 and one-half of those from quarters 3 and 4 are screened. With this method the number of non-children in the panel will be approximately the same in each quarter. However, there will be seasonal differences in sample yields within each household type.

For operational reasons there are no rejections and no shifting of collection periods in LFS apartment strata, high income strata and remote regions. In P.E.I. the number of available interviewers did not permit shifting the collection periods, and screening occurred in all quarters. The "Children" sample in these cases is selected strictly from clusters in quarters 3 and 4, resulting in a seasonal distortion of the sample for non-children. A programming error also caused no 12-year old to be selected for the panel in quarters 1 and 2. Selection probabilities for 12-year old were adjusted in quarters 3 and 4 to compensate this, but the result is that 12 year-old, just like children under 12, are not represented in the panel from the first two quarters.

5.5 Sample Design in Quebec

In Quebec the NPHS sample is selected from dwellings participating in a health survey organized by Santé Québec: the 1992-93 *Enquête sociale et de santé* (ESS). The survey sampled 16,010

dwelling using a two-stage design similar to that of the LFS. The province was divided geographically by crossing 15 Health Regions with four urban density classes (Montreal Census Metropolitan Area, regional capitals, small urban agglomerations and the rural sector). In each area clusters were stratified by socio-economic characteristics and selected using a PPS sample. Selected clusters were enumerated and random samples of their dwellings were drawn: 10 per cluster in major cities, 20 or 30 elsewhere.

Santé Québec provided non-confidential information which allowed the classification of their sample into 4 types of households: one-member households; households with children; other households with youths (persons aged under 25); and the rest (more than one member and no youth or child). A household type was determined by NPHS personnel for the ESS non-respondents.

The NPHS sample size was first allocated among the four urban density classes. To avoid having too much sample in Montreal the allocation was proportional to $\sqrt{(2W_h^2 + 1/4^2)}$, where W_h is the population share for class h , $h=1,2,3,4$. In each class an attempt was made to obtain a subsample from the ESS which, as far as the selected panel member was concerned, would be proportional to the populations for the 4 household types. This was done by drawing a sufficient number of households from the ESS to give the required yield for households with children (the most underrepresented group), and then removing excess sample from the other three household groups. An initial sample which was almost 50% higher than needed was thus selected. After removing from it 2/3 of the one-member households, 1/2 of the other households with no youths or children, and 1/6 of households with youths but no children, the objective was nearly attained.

Considerations for seasonal representation and variance estimation, and integration with the NLSC, affected the sub-sampling in Quebec as they did elsewhere. ESS strata were thus collapsed to allow the formation of replicates, with the clusters in each replicate covering all four quarters (two quarters are covered per cluster in the rural and small urban sectors as sample sizes are higher there). The sample of households with children was split into an "Adult" sample and a "Children" sample by a 3:2 ratio, the terms having the same meaning as in other provinces. "Children" sample households in quarters 1 and 2 were reassigned to quarters 3 and 4. As NPHS surveys the current occupants of dwellings selected for the ESS, and changes will have occurred in some of those dwellings, the samples of households without children for quarters 3 and 4 are also to be split, by a 2:3 ratio, into an "Adult" and a "Children" sample.

5.6 Sample design for the NPHS Supplement

The sample for the NPHS Supplement consisted of the core sample of the NPHS, that is, the non-RDD and non-buy-in portion of the survey. As such, the sample design of the supplement excludes the buy-in sample from New Brunswick, Ontario, Manitoba, and British Columbia as well as the RDD sample from British Columbia.

Table 2: Sample Sizes for the NPHS Supplement - Persons aged 12 and over

	Number of NPHS respondents in the NPHS Supplement
Newfoundland	918
Prince Edward Island	899
Nova Scotia	911
New Brunswick	974
Québec	2,581
Ontario	3,664
Manitoba	1,025
Saskatchewan	1,005
Alberta	1,310
British Columbia	1,499
CANADA	14,786

6. Data Collection

6.1 Questionnaire Design and Data Collection Method

The NPHS questions (including the supplements) were designed for Computer Assisted Interviewing(CAI), which meant that, as the questions were developed, the associated logical flow into and out of the questions were programmed. This included specifying the type of answer required, the minimum and maximum values, on-line edits associated with the question and what to do in case of item non-response.

With CAI, the interview can be monitored based on answers provided by the respondent. Some valuable controls include directing the skip patterns based on responses or fixing minimum and maximum values . On-screen prompts are shown when an invalid entry is recorded and thus immediate feedback is given to the respondent and/or the interviewer to correct inconsistencies. Other enhancements are the automatic insertion of reference periods based on current dates. Prefilling of text or data based on information gathered during the interview allows the interviewer to proceed without having to search back at previous answers. This type of prefill includes such things as using the correct name or gender within the questions themselves. Allowable ranges/answers based on data collected during the interview can also be programmed . In other words the questionnaire can be customized to the respondent - based on data collected at that time.

6.2 Tests

A number of tests were conducted before the main survey was implemented in the field.

Focus groups were held in the development stages of the questionnaires to verify various aspects of their content. The main objectives were to verify the clarity and the quality of the questions, respondent reactions to sections that were felt to be sensitive (mental health, alcohol, etc.), and to obtain approximate times for the length of the different sections.

Two field tests were also conducted. The tests involved four of Statistics Canada's Regional Offices and interviews were carried out by experienced Labour Force Survey interviewers. The main objectives of the two tests were again to observe respondent reaction to the survey, to obtain estimates of time for the various sections, and to see what kind of response rates could be obtained. Verifying response rates was especially important in Quebec where selected respondents had participated in the Santé Québec study a few months before. Field operations and procedures, interviewer training, and the computer programme application (questionnaire on computer) were also tested.

In addition to the field tests, the computer programme application was intensively tested in house to debug it and to ensure that all possible paths were being correctly followed. Computer

application testing was an ongoing operation up until the start of the main survey.

6.3 Interviewing

Collection operations were divided in four quarters (June, August and November 1994, and March 1995) and interviews were conducted by Statistics Canada Labour Force Survey (LFS) interviewers, who are part-time employees hired and trained specifically to carry out the LFS, using the computer-assisted interviewing method.

All respondents were first contacted in person except for a small sample in British Columbia that was conducted by telephone using the RDD approach. Many of the interviews, where started in person, were finished on the telephone either because the selected respondent was not available at the time of the initial visit or because the long interview time prevented the completion of the interview in one contact. The total interview took an average of one hour in each household.

In all dwellings, information about all household members is obtained from a knowledgeable household member - usually the person at home at the time of the interviewer visit. Such 'proxy' reporting, which accounts for approximately 55% of the information collected for this part of the interview, is used to avoid the high cost and extended time requirements that would be involved in repeat visits or calls necessary to obtain information directly from each respondent.

Proxy reporting was allowed for the selected respondent only for reasons of illness or incapacity. Such proxy reporting accounts for 4% of the information collected.

6.4 Supervision and Control

All LFS interviewers are under the supervision of a staff of senior interviewers who are responsible for ensuring that interviewers are familiar with the concepts and procedures of the LFS and its many supplementary surveys, and also for periodically monitoring their interviewers and reviewing their completed documents. The senior interviewers are, in turn, under the supervision of the LFS program managers, located in each of the eight Statistics Canada regional offices.

Some households were re-contacted by Senior Interviewers by telephone after the 3rd and 4th quarters to verify the quality of work of interviewers. At the time of the re-contact, the household composition was verified and an assessment of the interviewer's work was obtained.

6.5 Non-Response to the NPHS

Interviewers are instructed to make all reasonable attempts to obtain NPHS interviews with members of eligible households. For individuals who at first refuse to participate in the NPHS, a letter is sent from the Regional Office to the dwelling address stressing the importance of the

survey and the household's cooperation. This is followed by a second call (or visit) from the interviewer. For cases in which the timing of the interviewer's call (or visit) is inconvenient, an appointment is arranged to call back at a more convenient time. For cases in which there is no one home, numerous call backs are made. Under no circumstances are sampled dwellings replaced by other dwellings for reasons of non-response.

Each quarter, after all attempts to obtain interviews have been made, a small number of non-responding households remain.

In addition to household non-response, some selected respondents who completed the main NPHS interview did not respond to the NPHS supplementary questions.

6.6 Non-response follow-up

Many strategies were put in place to reduce the number of non-response cases. Before interviews started, a maximum recommended assignment size by interviewer was calculated based on test results to allow efficient follow-up of no contact cases (i.e. to avoid over burdening interviewers).

Interviewer procedures included ways of reducing the number of no-contacts by making visits at various times of the day or on the way to or from other dwellings, talking to neighbours or landlords to determine who lives in the dwelling and obtain telephone numbers, etc.

Refusals were followed-up by Senior Interviewers, Project Supervisors or by other Interviewers to try to convince respondents to participate in the survey.

In addition to the two official languages, the questionnaires were translated in Spanish, Portuguese, Chinese, Punjabee and Italian to try to reduce the number of non-interviews due to language problems.

To maximize the response rate, a large number of non-response cases were also followed-up in subsequent quarters.

7. Data Processing

7.1 Data Capture

Because NPHS used CAI, capture was part of the data collection process. The data collected during the interview were recorded directly onto a laptop computer. Each question is represented by a screen on the computer. After the answer to each question is entered, the next question appears automatically on the screen.

7.2 Editing

Some editing usually done at Head Office has been performed on-line in the (CAI) application and are performed during data collection. The editing to deal with out of range values and flow errors were controlled through the use of CAI. These types of errors were controlled by CAI by not allowing invalid values to be entered as responses, and by not allowing incorrect question paths to be followed. For example, CAI ensured that questions that did not apply to the respondent and therefore should not have been answered did not have responses in them. In other situations, warning messages were invoked, but no corrective action was taken if an interviewer entered contradictory responses between questions. Because no corrective action was taken in such instances, edits were developed to be performed after data collection at Head Office. Inconsistencies were usually corrected by setting one or both of the variables in question to "not stated". No imputation was performed.

7.3 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, diseases), the coding followed either the standard classification systems as used in the Census of the Population or in other Statistics Canada Surveys such as the Health and Activity Limitation Survey and General Social Survey-cycle 6.

7.4 Creation of Derived Variables

A number of variables on the file have been derived by using items found on the NPHS questionnaires in order to facilitate data analysis. 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. Appendix F provides the details on how these variables were derived.

7.5 Weighting

The principle behind estimation in a probability sample such as the NPHS is that each person in the sample "represents", besides himself or herself, several other persons not in the sample. For example, in a simple random 2% sample of the population, each person in the sample represents 50 persons in the population. In the terminology used here, it can be said that each person has a weight of 50.

The weighting phase is a step which calculates, for each person, what their associated weight is. This weight appears on the microdata file, and must be used to derive meaningful estimates from the survey. For example, if the number of individuals who smoke daily (see question SMOK-Q2 in section 9.2) is to be estimated, it is done by selecting the records referring to those individuals in the sample having that characteristic and summing the weights entered on those records.

Details of the method used to calculate these weights are presented in Chapter 11. Weighting for the NPHS supplements followed the same procedure as for the main NPHS weighting, but with an additional correction for those people who responded to the main survey but not to the supplement.

7.6 Suppression of Confidential Information

It should be noted that the 'Public Use' microdata files described above differ in a number of important respects from the survey 'master' files held by Statistics Canada. These differences are the result of actions taken to protect the anonymity of individual survey respondents. Users requiring access to information excluded from the microdata files may purchase custom tabulations. Estimates generated will be released to the user, subject to meeting the guidelines for analysis and release outlined in Section 9 of this document.

8. Data Quality

8.1 Response Rates

The calculation of response rates for the NPHS was complicated by two factors which are unique to this survey. The first problem occurred as a result of using the rejective method. Recall that a certain percentage of dwellings were designated as EFR (see Section 5.2 for more details). Households which were ultimately rejected do not contribute to the estimates, but are considered as household respondents since they provided the information that the NPHS requested. The EFR households that did not respond are considered to be non-respondents (as are non-EFR households that did not respond).

Secondly, the integration of the NPHS with the NLSC complicated the calculation of the response rate for the selected persons. Recall that in certain pre-determined dwellings, if at least one child under twelve years old was found, then a child was the selected person and he/she was administered the NLSC questionnaire. In these cases, there was no respondent to the NPHS selected person questionnaire. For this reason, these dwellings are considered to be out of scope for the purpose of calculating the NPHS selected person response rate.

The following is a description of how the Household response rate and the Selected Person response rate were calculated. It should be noted that out of scope dwellings (vacant or abandoned dwellings, dwellings under construction, or households not eligible for the sample) were not used in any of the calculations.

Household response rate

$$\text{HH response rate} = \frac{\text{\# of responding households including rejected households}}{\text{all in-scope households}}$$

A non-rejected responding household had *at least* one general component questionnaire completed for a member of the household. The household response rate at the Canada level for the NPHS was **88.7%**. At the provincial level, this rate varied from 85.2% in Ontario to 93.2% in Alberta.

Selected person response rate

The selected person response rate can be thought of as the number of health component questionnaires that *were* completed as compared to the number which *should have been* completed.

$$\text{SP response rate} = \frac{\text{\# of completed health component questionnaires}}{\text{\# of persons eligible to answer the health component questionnaire}}$$

NPHS PUBLIC USE MICRODATA DOCUMENTATION

where the number of persons eligible to answer the health component is equal to the number of non-rejected responding households, minus the number of dwellings where a child who was less than 12 years old was the selected person.

The selected person response rate for the NPHS was **96.1%** at the Canada level, and ranged from 94.7% in Nova Scotia to 97.6% in Saskatchewan.

It should be noted that because of the complications described above, multiplying the two rates together gives a meaningless value. The information that is used to calculate these rates is different in each case, and therefore a combined rate cannot be determined.

Relevant information for Calculation of Response Rates:

Number of respondents at the household level:	20725
Number of respondents at the selected person level:	17626

Number of rejected households:	3447
Number of dwellings where a child was selected:	2383

Number of non-respondents at the household level:	3091
Number of non-respondents at the selected person level:	716

Number of out of scope households:	4512
------------------------------------	------

Calculation of Household response rate:

$$\text{HH Rate} = \frac{20725 + 3447}{20725 + 3447 + 3091} = \frac{24172}{27263} = 88.7\%$$

Calculation of Selected Person response rate:

$$\text{SP Rate} = \frac{17626}{20725 - 2383} = \frac{17626}{18342} = 96.1\%$$

NPHS PUBLIC USE MICRODATA DOCUMENTATION

8.2 Response Rates to the NPHS Supplement

The following table summarizes the response rates to the NPHS Supplement. Since only certain respondents to the NPHS were asked the supplementary questions, this means that persons were selected, not households; therefore, the response rates below refer to person response rates. A person was eligible to receive the supplementary questions if he/she was the selected person for the in-depth NPHS questions, was aged 12 or over, was part of the “core” NPHS sample, and responded to the main NPHS questions.

	NPHS Respondents eligible for Supplement	Number of respondents in the NPHS Supplement	Person response rate to the NPHS Supplement
Newfoundland	918	824	89.8%
Prince Edward Island	899	807	89.8%
Nova Scotia	911	843	92.5%
New Brunswick	974	890	91.4%
Québec	2,581	2,362	91.5%
Ontario	3,664	3,269	89.2%
Manitoba	1,025	956	93.3%
Saskatchewan	1,005	928	92.3%
Alberta	1,310	1,179	90.0%
British Columbia	1,499	1,342	89.5%
CANADA	14,786	13,400	90.6%

Response rates to the NPHS Supplement were very high and very consistent across the country, ranging from 89.2% in Ontario to 93.3% in Manitoba.

8.2 Survey Errors

The survey produces estimates based on information collected from and about a sample of individuals. Somewhat different estimates might have been obtained if a complete census had been taken using the same questionnaire, interviewers, supervisors, processing methods, etc. as those actually used in the survey. The difference between the estimates obtained from the sample and those resulting from a complete count taken under similar conditions is called the sampling error of the estimate.

Errors which are not related to sampling may occur at almost every phase of a survey operation. Interviewers may misunderstand instructions, respondents may make errors in answering questions, the answers may be incorrectly entered 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 on CAI questionnaire or misunderstanding of instructions, procedures to ensure that data collection errors were minimized.

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. Partial non-response to NPHS was basically non-existent; once the questionnaire was started, it tended to be completed with very little non-response. Total non-response occurred because the interviewer was either unable to contact the respondent, no member of the household was able to provide the information, 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, could not recall the requested information, or could not provide non-proxy 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. This section of the documentation outlines the measures of sampling error which Statistics Canada commonly uses and which it urges users producing estimates from this microdata file to use also.

NPHS PUBLIC USE MICRODATA DOCUMENTATION

The basis for measuring the potential size of sampling errors is the standard error of the estimates derived from survey results.

However, because of the large variety of estimates that can be produced from a survey, the standard error of an estimate is usually expressed relative to the estimate to which it pertains. This resulting measure, known as the coefficient of variation (C.V) 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.

For example, suppose that, based upon the survey results, one estimates that 24% of Canadians aged 12 and over are daily cigarettes smokers is found to have standard error of .003. Then the coefficient of variation of the estimate is calculated as:

$$\left(\frac{.003}{.24} \right) \times 100\% = 1.25\%$$

9. Guidelines For Tabulation, Analysis And Release

This section of the documentation outlines the guidelines to be adhered to by users tabulating, analyzing, publishing or otherwise releasing any data derived from the survey microdata tapes. With the aid of these guidelines, users of microdata should be able to produce figures that are in close agreement with those produced by Statistics Canada and, at the same time, will be able to develop currently unpublished figures in a manner consistent with these established guidelines.

9.1 Rounding Guidelines

In order that estimates for publication or other release derived from these microdata tapes correspond to those produced by Statistics Canada, users are urged to adhere to the following guidelines regarding the rounding of such estimates:

- a) Estimates in the main body of a statistical table are to be rounded to the nearest hundred units using the normal rounding technique. 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, in normal rounding to the nearest 100, if the last two digits are between 00 and 49, they are changed to 00 and the preceding digit (the hundreds digit) is left unchanged. If the last digits are between 50 and 99 they are changed to 00 and the preceding digit is incremented by 1.
- b) 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 100 units using normal rounding.
- c) Averages, proportions, rates and percentages are to be computed from unrounded components (i.e. numerators and/or denominators) and then are to be rounded themselves to one decimal using normal rounding. In normal rounding to a single digit, if the final 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 increased by 1.
- d) Sums and differences of aggregates (or ratios) are to be derived from their corresponding unrounded components and then are to be rounded themselves to the nearest 100 units (or the nearest one decimal) using normal rounding.
- e) In instances where, due to technical or other limitations, a rounding technique other than normal rounding is used resulting in estimates to be published or otherwise released which differ from corresponding estimates published by Statistics Canada, users are urged to note the reason for such differences in the publication or release document(s).
- f) Under no circumstances are unrounded estimates to be published or otherwise released by users. Unrounded estimates imply greater precision than actually exists.

9.2 Sample Weighting Guidelines for Tabulation

The sample design used for the NPHS was not self-weighting. That is to say, the sampling weights are not identical for all individuals in the sample. When producing simple estimates, including the production of ordinary statistical tables, users must apply the proper sampling weight.

If proper weights are not used, the estimates derived from the microdata tapes cannot be considered to be representative of the survey population, and will not correspond to those produced by Statistics Canada.

Users should also note that some software packages may not allow the generation of estimates that exactly match those available from Statistics Canada, because of their treatment of the weight field.

9.2.1 Definitions of types of estimates: Categorical vs. Quantitative

Before discussing how the NPHS data can be tabulated and analyzed, it is useful to describe the two main types of point estimates of population characteristics which can be generated from the microdata file for the National Population Health Survey.

Categorical Estimates:

Categorical estimates are estimates of the number, or percentage of the surveyed population possessing certain characteristics or falling into some defined category. The number of individuals who smoke daily is an example of such an estimate. An estimate of the number of persons possessing a certain characteristic may also be referred to as an estimate of an aggregate.

Example of Categorical Question:

SMOK-Q2: At the present do/does ... smoke cigarettes daily, occasionally or not at all?

- ☐ Daily
- ☐ Occasionally
- ☐ Not at all

Quantitative Estimates:

Quantitative estimates are estimates of totals or of means, medians and other measures

of central tendency of quantities based upon some or all of the members of the surveyed population. They also specifically involve estimates of the form \hat{X} / \hat{Y} where \hat{X} is an estimate of surveyed population quantity total and \hat{Y} is an estimate of the number of persons in the surveyed population contributing to that total quantity.

An example of a quantitative estimate is the average number of cigarettes smoked per day by individuals who smoke daily. The numerator is an estimate of the total number of cigarettes smoked per day by individuals who smoke daily, and its denominator is an estimate of the number of individuals who smoke daily.

Example of Quantitative Question:

SMOK-Q4: How many cigarettes do/does you/he/she smoke each day now?

|_| Number of Cigarettes

9.2.2 Tabulation of Categorical Estimates

Estimates of the number of people with a certain characteristic can be obtained from the microdata file by summing the final weights of all records possessing the characteristic(s) of interest. Proportions and ratios of the form \hat{X} / \hat{Y} are obtained by:

- a) summing the final weights of records having the characteristic of interest for the numerator (\hat{X}),
- b) summing the final weights of records having the characteristic of interest for the denominator (\hat{Y}), then
- c) dividing the numerator estimate by the denominator estimate.

9.2.3 Tabulation of Quantitative Estimates

Estimates of quantities can be obtained from the microdata file by multiplying the value of the variable of interest by the final weight for each record, then summing this quantity over all records of interest. For example, to obtain an estimate of the total number of cigarettes smoked each day by individuals who smoke daily, multiply the value reported in question SMOK-Q4 by the final weight for the record, then sum this value over all records with a response of 'daily' to SMOK-Q2.

To obtain a weighted average of the form \hat{X} / \hat{Y} , the numerator (\hat{X}) is calculated as for a quantitative estimate and the denominator (\hat{Y}) is calculated as for a categorical estimate. For example, to estimate the average number of cigarettes smoked per day by individuals who smoke daily,

- a) estimate the total number of cigarettes smoked per day by individuals who smoke daily as described above,
- b) estimate the number of individuals who smoke daily by summing the final weights of all records with a response of 'daily' to SMOK-Q2, then
- c) divide estimate (a) by estimate (b).

9.3 Guidelines for Statistical Analysis

The National Population Health Survey is based upon a complex design, with stratification and multiple stages of selection, and unequal probabilities of selection of respondents. Using data from such complex surveys presents problems to analysts because the survey design and the selection probabilities affect the estimation and variance calculation procedures that should be used.

While many analysis procedures found in statistical packages allow weights to be used, the meaning or definition of the weight in these procedures 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, analysis of variance), a method exists which can make the application of standard packages more meaningful. If the weights on the records are rescaled so that the average weight is one (1), then the results 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. The rescaling can be accomplished by using in the analysis a weight which is equal to the original weight divided by the average of the original weights for the sampled units (people) contributing to the estimator in question.

In order to provide a means of assessing the quality of tabulated estimates, Statistics Canada has produced a set of Approximate Sampling Variability Tables (commonly referred to as "C.V. Tables") for the NPHS. These tables can be used to obtain approximate coefficients of variation for categorical-type estimates and proportions. See Chapter 10 for more details.

9.4 Release Guidelines

Before releasing and/or publishing any estimate from these microdata tapes, users should first determine the number of sampled respondents who contribute to the calculation of the estimate. If this number is less than 30, the weighted estimate should not be released regardless of the value of the coefficient of variation for this estimate. For weighted estimates based on sample sizes of 30 or more, users should determine the coefficient of variation of the rounded estimate and follow the guidelines below.

Sampling Variability Guidelines

Type of Estimate	cv (in %)	Guidelines
1. Unqualified	0.0 - 16.5	Estimates can be considered for general unrestricted release. Requires no special notation.
2. Qualified	16.6 - 25.0	Estimates can be considered for general unrestricted release but should be accompanied by a warning cautioning subsequent users of the high sampling variability associated with the estimates. Such estimates should be identified by the letter Q (or in some other similar fashion).
3. Confidential	25.1 - 33.3	Estimates can be considered for general unrestricted release only when sampling variabilities are obtained using an exact variance calculation procedure. Unless exact variances are obtained, such estimates should be deleted and replaced by dashes (---) in statistical tables.
4. Not for Release	33.4 or greater	Estimates cannot be released in any form under any release OR circumstances. In statistical tables, such estimates should be deleted and replaced by dashes(--)

10. APPROXIMATE SAMPLING VARIABILITY TABLES

In order to supply coefficients of variation which would be applicable to a wide variety of categorical estimates produced from this microdata file and which could be readily accessed by the user, a set of Approximate Sampling Variability Tables has been produced. These "look-up" tables allow the user to obtain an approximate coefficient of variation based on the size of the estimate calculated from the survey data.

The coefficients of variation (C.V) are derived using the variance formula for simple random sampling and incorporating a factor which reflects the multi-stage, clustered nature of the sample design. This factor, known as the design effect, was determined by first calculating design effects for a wide range of characteristics and then choosing from among these a conservative value to be used in the look-up tables which would then apply to the entire set of characteristics.

The table below shows the design effects, sample sizes and population counts which were used to produce the Approximate Sampling Variability Tables for the NPHS Supplements.

Input Data For Provincial, Regional and Canada Level Sampling Variability Tables For Selected Members (Ages 12 and Over, eligible for NPHS Supplements)

PROVINCE	DESIGN EFFECT	SAMPLE SIZE	POPULATION
Newfoundland	1.79	824	483,363
Prince Edward Island	1.59	807	109,603
Nova Scotia	1.44	843	763,944
New Brunswick	1.47	890	626,303
Quebec	1.68	2,362	6,029,670
Ontario	1.61	3,269	9,050,016
Manitoba	1.78	956	890,750
Saskatchewan	1.54	928	792,049
Alberta	1.54	1,179	2,166,102
British Columbia	1.37	1,342	3,036,798
Atlantic	1.87	3,364	1,983,215
Prairies	1.72	3,063	3,848,903
Canada	1.96	13,400	23,948,603

All coefficients of variation in the Approximate Sampling Variability Tables are approximate and,

therefore, unofficial. Estimates of actual variance for specific variables may be obtained from Statistics Canada on a cost-recovery basis. The use of actual variance estimates would allow users to release otherwise unreleaseable estimates, i.e. estimates with coefficients of variation in the 'confidential' range.

Remember: If the number of observations on which an estimate is based is less than 30, the weighted estimate should not be released regardless of the value of the coefficient of variation for this estimate. This is because the formulas used for estimating the variance do not hold true for small sample sizes.

10.1 How to use the C.V. tables for Categorical Estimates

The following rules should enable the user to determine the approximate coefficients of variation from the Sampling Variability Tables for estimates of the number, proportion or percentage of the surveyed population possessing a certain characteristic and for ratios and differences between such estimates.

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

The coefficient of variation depends only on the size of the estimate itself. On the appropriate Sampling Variability Table, locate the estimated number in the left-most column of the table (headed "Numerator of Percentage") and follow the asterisks (if any) across to the first figure encountered. This figure is the approximate coefficient of variation.

Rule 2: Estimates of Proportions or Percentages Possessing a Characteristic

The coefficient of variation of an estimated proportion or percentage depends on both the size of the proportion or percentage and the size of the total upon which the proportion or percentage is based. Estimated proportions or percentages are relatively more reliable than the corresponding estimates of the numerator of the proportion or percentage, when the proportion or percentage is based upon a sub-group of the population. This is due to the fact that the coefficients of variation of the latter type of estimates are based on the largest entry in a row of a particular table, whereas the coefficients of variation of the former type of estimators are based on some entry (not necessarily the largest) in that same row. (Note that in the tables the cv's decline in value reading across a row from left to right). For example, the estimated proportion of individuals who smoke daily out of those who smoke at all is more reliable than the estimated number who smoke daily.

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

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

Rule 3: Estimates of Differences Between Aggregates or Percentages

The standard error of a difference between two estimates is approximately equal to the square root of the sum of squares of each standard error considered separately. That is, the standard error of a difference ($\hat{d} = \hat{X}_2 - \hat{X}_1$) is:

$$\sigma_{\hat{d}} = \sqrt{(\hat{X}_1 \alpha_1)^2 + (\hat{X}_2 \alpha_2)^2}$$

where \hat{X}_1 is estimate 1, \hat{X}_2 is estimate 2, and α_1 and α_2 are the coefficients of variation of \hat{X}_1 and \hat{X}_2 respectively. The coefficient of variation of \hat{d} is given by $\sigma_{\hat{d}} / \hat{d}$. This formula is accurate for the difference between separate and uncorrelated characteristics, but is only approximate otherwise.

Rule 4: Estimates of Ratios

In the case where the numerator is a subset of the denominator, the ratio should be converted to a percentage and Rule 2 applied. This would apply, for example, to the case where the denominator is the number of individuals who smoke at all and the numerator is the number of individuals who smoke daily out of those who smoke at all.

Consider the case where the numerator is not a subset of the denominator, as for example, the ratio of the number of individuals who smoke daily or occasionally as compared to the number of individuals who do not smoke at all. The standard deviation of the ratio of the estimates is approximately equal to the square root of the sum of squares of each coefficient of variation considered separately multiplied by \hat{R} , where \hat{R} is the ratio of the estimates ($\hat{R} = \hat{X}_1 / \hat{X}_2$). That is, the standard error of a ratio is:

$$\sigma_{\hat{R}} = \hat{R} \sqrt{\alpha_1^2 + \alpha_2^2}$$

where α_1 and α_2 are the coefficients of variation of \hat{X}_1 and \hat{X}_2 respectively.

The coefficient of variation of \hat{R} is given by $\sigma_{\hat{R}} / \hat{R} = \sqrt{\alpha_1^2 + \alpha_2^2}$. The formula will tend to overstate the error, if \hat{X}_1 and \hat{X}_2 are positively correlated and understate the error if \hat{X}_1 and \hat{X}_2 are negatively correlated.

Rule 5: Estimates of Differences of Ratios

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

10.2 Examples of using the C.V. tables for Categorical Estimates

The following 'real life' examples are included to assist users in applying the foregoing rules.

** update these examples using NPHS supplement data and design effects **

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

Suppose that a user estimates that 20,374,240 individuals have ever seen cigarette packages with health warning messages on them in Canada. How does the user determine the coefficient of variation of this estimate?

- 1) Refer to the CANADA level cv table for the NPHS Supplement.
- 2) The estimated aggregate (20,374,240) does not appear in the left-hand column (the 'Numerator of Percentage' column), so it is necessary to use the figure closest to it, namely 20,000,000.
- 3) The coefficient of variation for an estimated aggregate (expressed as a percentage) is found by referring to the first non-asterisk entry on that row, namely, 0.4%.
- 4) So the approximate coefficient of variation of the estimate is 0.4%. The finding that there were 20,374,240 individuals who have seen cigarette packages with health warning messages on them is publishable with no qualifications.

Example 2 : Estimates of Proportions or Percentages Possessing a Characteristic

Suppose that the user estimates that $17,625,925/20,374,240=86.5\%$ of individuals in Canada who have ever seen cigarette packages with health warnings on them agree with the packages having the warnings. How does the user determine the coefficient of variation of this estimate?

- 1) Refer to the CANADA level cv table for NPHS Supplement.
- 2) Because the estimate is a percentage which is based on a subset of the total population (i.e., individuals who have ever seen health warnings on cigarette packages), it is necessary to use both the percentage (86.5%) and the numerator portion of the percentage (17,625,925) in determining the coefficient of variation.
- 3) The numerator, 17,625,925, does not appear in the left-hand column (the 'Numerator of Percentage' column) so it is necessary to use the figure closest to it, namely 15,000,000. Similarly, the percentage estimate does not appear as any of the column headings, so it is necessary to use the figure closest to it, 90.0%.

- 4) The figure at the intersection of the row and column used, namely 0.5% is the coefficient of variation (expressed as a percentage) to be used.
- 5) So the approximate coefficient of variation of the estimate is 0.5%. The finding that 86.5% of individuals who have ever seen cigarette packages with health warnings on them agree with the packages having the warnings can be published with no qualifications.

Example 3 : Estimates of Differences Between Aggregates or Percentages

Suppose that a user estimates that $7,275,450/20,374,240=35.7\%$ of those who have seen health warnings on cigarette packages have never read them (estimate 1) while $2,159,152/20,374,240=10.6\%$ of those who have seen health warnings on cigarette packages have read them more than once a day (estimate 2). Note that these estimates are based on the results of questions SSMOK-Q1 and SSMOK-Q4. How does the user determine the coefficient of variation of the difference between these two estimates?

- 1) Using the CANADA level cv table for the NPHS Supplement in the same manner as described in example 2 gives the cv for estimate 1 as 1.8% (expressed as a percentage), and the cv for estimate 2 as 4.0% (expressed as a percentage).
- 2) Using rule 3, the standard error of a difference ($\hat{d} = \hat{X}_2 - \hat{X}_1$) is:

$$\sigma_{\hat{d}} = \sqrt{(\hat{X}_1 \alpha_1)^2 + (\hat{X}_2 \alpha_2)^2}$$

where \hat{X}_1 is estimate 1, \hat{X}_2 is estimate 2, and α_1 and α_2 are the coefficients of variation of \hat{X}_1 and \hat{X}_2 respectively.

That is, the standard error of the difference $\hat{d} = (.357 - .106) = .251$ is:

$$\begin{aligned}\sigma_{\hat{d}} &= \sqrt{[(.357)(.018)]^2 + [(.106)(.040)]^2} \\ &= .008\end{aligned}$$

- 3) The coefficient of variation of \hat{d} is given by $\sigma_{\hat{d}} / \hat{d} = .008/.251 = 0.031$.
- 4) So the approximate coefficient of variation of the difference between the estimates is 3.1% (expressed as a percentage). This estimate can be published with no qualifications.

Example 4 : Estimates of Ratios

Suppose that the user estimates that 7,275,450 individuals never read the health warnings on

NPHS PUBLIC USE MICRODATA DOCUMENTATION

cigarette packages, while 4,300,945 read the health warnings on cigarette packages at least once a day. The user is interested in comparing the estimate of non-readers to daily readers of health warnings on cigarette packages in the form of a ratio. How does the user determine the coefficient of variation of this estimate?

- 1) First of all, this estimate is a ratio estimate, where the numerator of the estimate ($= \hat{X}_1$) is the number of individuals who never read the health warnings on cigarette packages. The denominator of the estimate ($= \hat{X}_2$) is the number of individuals who read the health warnings on cigarette packages at least once a day.
- 2) Refer to the CANADA level cv table for the NPHS Supplement.
- 3) The numerator of this ratio estimate is 4,300,945. The figure closest to it is 4,000,000. The coefficient of variation for this estimate (expressed as a percentage) is found by referring to the first non-asterisk entry on that row, namely, 2.6%.
- 4) The denominator of this ratio estimate is 7,275,450. The figure closest to it is 7,000,000. The coefficient of variation for this estimate (expressed as a percentage) is found by referring to the first non-asterisk entry on that row, namely, 1.9%.
- 5) So the approximate coefficient of variation of the ratio estimate is given by rule 4, which is,

$$\alpha_{\hat{R}} = \sqrt{\alpha_1^2 + \alpha_2^2}$$

where α_1 and α_2 are the coefficients of variation of \hat{X}_1 and \hat{X}_2 respectively.

That is ,

$$\begin{aligned}\alpha_{\hat{R}} &= \sqrt{(.026)^2 + (.019)^2} \\ &= 0.032\end{aligned}$$

The obtained ratio of non-readers to daily readers of health warning messages on cigarette packages is 4,300,945/7,275,450, which is 0.59:1. The coefficient of variation of this estimate is 3.2% (expressed as a percentage), which is releasable with no qualifications.

10.3 How to use the C.V. tables to obtain 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 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, \hat{X} , are generally expressed as two numbers, one below the estimate and one above the estimate, as $(\hat{X}-k, \hat{X}+k)$ where k is determined depending upon the level of confidence desired and the sampling error of the estimate.

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

$$CI_X = [\hat{X} - t \hat{X} \alpha_{\hat{X}}, \hat{X} + t \hat{X} \alpha_{\hat{X}}]$$

where α_X is the determined coefficient of variation of \hat{X} , 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.

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.

10.4 Example of using the C.V. tables to obtain confidence limits

A 95% confidence interval for the estimated proportion of individuals who have ever seen

cigarette packages with health warnings on them and who agree with the packages having the warnings (from example 2, section 10.2) would be calculated as follows.

$$\hat{X} = .865$$

$$t = 2$$

$\alpha_X = .005$ is the coefficient of variation of this estimate as determined from the tables.

$$CI_X = \{.865 - (2) (.865) (.005), .865 + (2) (.865) (.005)\}$$

$$CI_X = \{.856, .874\}$$

10.5 How to use the C.V. tables to do a 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 on the difference $\hat{X}_1 - \hat{X}_2$ be σ_d .

If $t = \frac{\hat{X}_1 - \hat{X}_2}{\sigma_d}$ 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 0.05 level.

10.6 Example of using the C.V. tables to do a t-test

Let us suppose we wish to test, at 5% level of significance, the hypothesis that there is no difference between the proportion of individuals who have ever seen health warnings on cigarette packages and never read them AND the proportion of those who have ever seen health warnings on cigarette packages and read them more than once a day. From example 3, section 10.2, the standard error of the difference between these two estimates was found to be = .008. Hence ,

$$t = \frac{\hat{X}_1 - \hat{X}_2}{\sigma_d} = \frac{.357 - .106}{.008} = \frac{.251}{.008} = 31.4$$

Since $t = 31.4$ is greater than 2, it must be concluded that there is a significant difference between the two estimates at the 0.05 level of significance.

10.7 Exact Variances/ Coefficients of Variation

All coefficients of variation in the Approximate Sampling Variability Tables (ASV Tables) are indeed approximate and, therefore, unofficial. However, exact coefficients of variation for specific variables may be obtained from Statistics Canada on a cost-recovery basis. The types of estimates supported include aggregates, proportions, ratios, differences between aggregates, proportions or ratios, as well as more sophisticated types of analyses such as estimates of coefficients from linear regressions and logistic regressions, among others. The exact coefficients of variation are obtained via an exact variance program, which uses a technique called "jackknifing". This technique involves dividing the records on the microdata files into subgroups (or replicates) and determining the variation in the estimates from replicate to replicate. There are a number of reasons why a user may require an exact variance. A few are given below.

Firstly, if a user desires estimates at a geographic level smaller than the province (for example, at the urban/rural level), then the ASV tables provided are not adequate. Coefficients of variation of these estimates may be obtained using "domain" estimation techniques through the exact variance program.

Secondly, should a user require more sophisticated analyses such as estimates of coefficients from linear regressions or logistic regressions, the ASV tables will not provide correct associated coefficients of variation. Although some standard statistical packages allow sampling weights to be incorporated in the analyses, the variances that are produced often do not take into account the stratified and clustered nature of the design properly, whereas the exact variance program would do so.

Thirdly, for estimates of quantitative variables, separate tables are required to determine their sampling error. Since most of the variables for the National Population Health Survey are primarily categorical in nature, this has not been done. Thus, users wishing to obtain coefficients of variation for quantitative variables can do so through the exact variance program. As a general rule, however, the coefficient of variation of a quantitative total will be larger than the coefficient of variation of the corresponding category estimate (i.e., the estimate of the number of persons contributing to the quantitative estimate). If the corresponding category estimate is not releasable, the quantitative estimate will not be either. For example, the coefficient of variation of the estimate of the total number of cigarettes smoked each day by individuals who smoke daily would be greater than the coefficient of variation of the corresponding estimate of the number of individuals who smoke daily. Hence if the coefficient of variation of the latter is not releasable, then the coefficient of variation of the corresponding quantitative estimate will also not be releasable.

Lastly, should a user find himself/ herself in a position where he/she can use the ASV tables, but this renders a coefficient of variation in the "confidential" range (25.1% - 33.3%), the user should not release the associated estimate unless the coefficient of variation is recalculated through the exact variance program and it is found that the estimate is in fact releasable. The reason for this is that the coefficients of variation produced by the ASV tables are based on a wide range of

NPHS PUBLIC USE MICRODATA DOCUMENTATION

variables and are therefore considered crude, whereas the exact variance program would give an exact coefficient of variation associated with the variable in question.

The exact variance/ coefficient of variation program will be available in November of 1995 and any user interested in this service should contact Diane Stukel (613-951-2244) from the Health Statistics Methods Section within Household Survey Methods Division at Statistics Canada. Although there will be no charge for any computer time required, there will be a fee charged for any consultation time required to set up the request as well as for any time required to set up the associated computer runs. The daily consultation rate, based on a 7.5 hour day, is \$477.88; this rate may be broken down into an appropriate number of hours or minutes, if required. Naturally, the length of the consultation will vary from request to request and will depend upon the complexity of the analysis, the number of variables to be analyzed, etc.

NPHS PUBLIC USE MICRODATA DOCUMENTATION

10.8 Release cut-off's for the NPHS Supplement

The minimum cut-offs for estimates of totals at the provincial and Canada levels for the NPHS supplements are specified in the table below. Estimate sizes smaller than the minimum given in the "Confidential" column may not be released under any circumstances.

**Table of Release Cut-offs for Totals Based on Provincial/Regional/Canada
Level Estimates for Selected Members
(Ages 12 and Over, eligible for NPHS Supplements)**

Province	Unqualified	Qualified	Confidential
Newfoundland	35,500	16,000	9,500
Prince Edward Island	7,500	3,500	2,000
Nova Scotia	45,000	20,500	11,500
New Brunswick	36,000	16,000	9,000
Quebec	153,500	68,000	38,500
Ontario	161,000	71,000	40,000
Manitoba	57,000	26,000	14,500
Saskatchewan	45,500	20,500	11,500
Alberta	99,000	44,500	25,000
British Columbia	110,000	49,000	27,500
Atlantic	39,500	17,500	10,000
Prairies	78,000	34,500	19,500
CANADA	128,000	56,000	31,500

11. Weighting

The sample for the NPHS Supplement was the core sample of the NPHS, that is, the non-RDD and non-buy-in part of the survey. Thus, the weighting for the NPHS Supplement follows the weighting for the NPHS very closely.

As described previously, the household component of the National Population Health Survey has two basic designs: one for the nine provinces outside of Quebec, and one for Quebec. In the nine provinces outside of Quebec, the NPHS uses the design of the Labour Force Survey (LFS) with many modifications, to generate a sample of its own. For this reason, the derivation of weights is tied to the weighting procedure used for the LFS. In Quebec, however, a two-phase sample design was implemented, where the first phase was drawn by the Enquête Sociale et de Santé (ESS) in 1992-93, and the second phase sample was drawn by the NPHS. Thus, in Quebec, the derivation of the weights is tied to the weighting procedure used by the ESS. See the section entitled "Sample Design" for more details. In section 11.1 below, the weighting procedure for the provinces outside of Quebec is described, and in section 11.2, the weighting procedure for Quebec is outlined.

11.1 Weighting Procedure for the Provinces Outside of Quebec

To begin, the basic LFS weighting procedure is briefly described below, followed by a description of a number of other multiplicative weight adjustments that are necessary in the formation of final weights. For the NPHS Supplements, only those dwellings which were selected as part of the "core" sample (as opposed to the sample buy-in which occurred in some provinces) were included in the weighting.

11.1.1 LFS Basic Weights

The LFS uses a stratified multi-stage design (mainly 2-stage, but in some cases, 3-stage).

For example, in those places where a 2-stage design is used, at the first stage, clusters are selected using either probability proportional-to-size, size systematic without replacement (PPS systematic) sampling or randomized PPS systematic sampling or the random group method. For more information on these methods, contact Diane Stukel in the Health Statistics Methods Section within Household Survey Methods Division at Statistics Canada. An LFS "cluster weight" is then calculated as the inverse probability of selecting a cluster, in accordance with the above sample selection schemes.

At the second stage, dwellings are selected within sampled clusters using systematic sampling. A "dwelling weight" is calculated as the inverse probability of selecting a dwelling given that the cluster which contains it is selected.

An "LFS basic weight" is then given by the product of the cluster weight and the dwelling weight.

11.1.3 Weight Adjustments to the Basic Weights

All of the weight adjustments which follow are applied to the LFS basic weight, to compensate for design features specific to the LFS.

Adjustment 1: Rotation Group Weight Adjustment

The full LFS sample is comprised of 6 "rotation groups" (although in some remote areas and apartment strata this number differs from 6). In the LFS, the rotation group that has been in the sample for six months is rotated out of the sample, and a fresh rotation group is rotated in to replace it. This serves to reduce the respondent burden. The NPHS requests sample from the LFS in terms of integral numbers of rotation groups (between 1 and 6), although a fractional number may actually be required to fulfil sample size needs. For example, the NPHS may require 2.3 rotations. It will request 3 from the LFS and later "stabilize out" the .7 rotations that are not needed (see Stabilization Weight Adjustment). Thus, the first multiplicative weight adjustment, which compensates for the integral number requested, is given by:

$$\frac{\text{number of rotations in an LFS stratum used by LFS (usually 6)}}{\text{integral number of rotations in an LFS stratum requested by NPHS}}.$$

In Winnipeg there are two instances and in Vancouver one instance where, in each case, three LFS strata were collapsed together before selecting clusters within the collapsed strata. For these three cases, an extra adjustment is made to the corresponding LFS basic weights to reflect the fact that clusters were selected via (randomized) PPS systematic sampling within the collapsed strata rather than within the usual LFS strata.

Adjustment 2: Cluster Growth Weight Adjustment

There may be clusters which experience growth between the time when a Census enumeration of the cluster takes place and the time when the cluster is listed for the LFS. The cluster selection probability is based on the Census enumeration figure, which may be out of date. This has the effect that the number of dwellings in the LFS sample increases very slightly with moderate growth in the housing stock. Substantial growth can be tolerated in an isolated cluster before the additional sample represents a field collection problem. However, if growth takes place in more than one cluster in an interviewer assignment, the cumulative effect of all the increases may create a workload problem. In clusters where substantial growth has taken place, subsampling is used as a means of keeping interviewer assignments manageable. The NPHS also institutes a similar subsampling of clusters which have experienced moderate growth, albeit the growth is not great enough so that the cluster is considered to be a growth cluster by the LFS. Thus,

NPHS PUBLIC USE MICRODATA DOCUMENTATION

the second multiplicative weight adjustment is given by the inverse of this subsampling ratio in clusters where subsampling has occurred for either the LFS or the NPHS.

Adjustment 3: Stabilization Weight Adjustment

Stabilization is a means of capping the sample size within a stabilization area to prevent the associated costs from becoming too prohibitive. A "stabilization area" consists of clusters in the high-income and apartment frame, and consists of groups of strata in the regular frame. "Stabilization" addresses the problem of growth that occurs within a stabilization area. The growth is large enough to be a concern even after cluster growth adjustment, although no single cluster contributes to the growth substantially enough to be considered the root of the problem. This problem is remedied through subsampling within the stabilization area. In addition to regular stabilization, it is at this point that the fractional part of a rotation requested of the LFS but not required by the NPHS, is "stabilized out" through subsampling (see Rotation Group Weight Adjustment). Thus, the third multiplicative weight adjustment is given by:

$$\frac{\text{number of dwellings selected by the LFS within a cluster}}{\text{number of dwellings actually used by the NPHS within a cluster}}.$$

Adjustment 4: Multiples Weight Adjustment

It sometimes happens that an interviewer discovers that a listing that was thought to constitute single private occupied dwelling in fact constitutes two or more private occupied dwellings. This may happen, for instance, when a basement apartment is attached to a dwelling but has its own separate entrance. In this case, since interviewing takes place in only one of the two private occupied dwellings (selected at random), the weight associated with that dwelling is boosted up by a factor of two. Thus, the fourth multiplicative weight adjustment is given by the number of private occupied dwellings that the listing in question actually constitutes. For most listings, this adjustment factor will be one.

Adjustment 5: Household Non-response Weight Adjustment

Despite all the attempts made by the interviewers, some non-response at the household level is inevitable. Non-response encompasses any of the following situations: refusal, special circumstance, language barrier, no one at home, temporarily absent or computer problem. Non-response is compensated for by proportionally adjusting the weights of responding households. This fifth adjustment is given by:

$$\frac{\text{sum of weights for all sampled households in an NPHS stratum/season combination}}{\text{sum of weights for respondent households in an NPHS stratum/season combination}}.$$

Note that this adjustment is made at the NPHS stratum level for each season. Here, NPHS strata are groups of LFS strata. The adjustment was made at this level since it was the smallest geographic level which ensured stability (i.e., adjustments less than or equal to 2.5). The adjustment was calculated separately for each season since the non-response rate was significantly different for each season. Here, the first two quarters of collection constitute the summer "season" while the last two quarters of collection constitute the winter "season". For those few cases where the nonresponse adjustment exceeded 2.5, the adjustment was recalculated at the NPHS stratum level rather than at the NPHS stratum/season level. The "weights" referred to above are the LFS basic weight multiplied by all the adjustments to this point (i.e., weight adjustments one through four). The adjustment is based on the assumption that the households that were actually interviewed represent the characteristics of those that should have been interviewed. To the extent that this assumption is not true, the estimates are somewhat biased. Note that some non-respondents in a given collection period were successfully recontacted in a later period. These cases were treated as if they had responded in the collection period in which they were intended to be interviewed. Given the number of such cases, it is not expected that this will have a significant effect on seasonal data.

Adjustment 6: Rejective Method Weight Adjustment

As discussed in the section entitled "Sample Design", in the last two quarters of data collection a portion of the sampled households are screened out or rejected from the sample after determining that there are no youths or children residing within (i.e., no one under the age of 25). These "rejected" households come from that portion of the "Children Sample" that are "screened" for household composition. This methodology was implemented to compensate for an over-representation in the sample of members of small sized households and an under-representation of members of large sized households. The latter type of household tends to consist of parents and their children while the former type tends to consist of single people, older people or couples without children. Since some of the households containing no youths or children are screened out or "rejected", representation in the sample of households of this type comes solely from the "Adult Sample" and from the non-screened portion of the "Children Sample". Thus, to compensate for the "rejected" part of the sample, the weights for those households containing no youths or children from the "Adult Sample" and from the non-screened portion of the "Children Sample" are boosted by another multiplicative weight adjustment. This sixth adjustment is given by the inverse of one minus the overall screening rate within a stratum. Note that in P.E.I., this adjustment was implemented a little differently since, among other reasons, the rejective method was applied in all four quarters of data collection rather than in the last two quarters only. Also note that this

adjustment was not applied in apartment strata, high income strata and remote strata, since the rejective method was not implemented there.

11.1.5 Further Weight Adjustments For Selected Members

Data from the NPHS supplementary questionnaire is obtained for only one member aged 12 or more from a sampled household; that is, it is collected from the selected member for the NPHS. The associated final weight for each individual on this file is obtained as follows. First, the LFS basic weight is multiplied by weight adjustments one through six, as well as weight adjustments 7B through 10B given below, to form an LFS "intermediate" weight. Then, the non-response adjustment 10C for the NPHS Supplement is carried out. Finally, benchmarking is performed upon the final sample.

Adjustment 7B: NLSC Integration Weight Adjustment

In the last two quarters of data collection, the NPHS selects potential respondents for both the NPHS and NLSC selected member questionnaire. In some sampled households, a maximum of 4 children (aged less than 12) are selected, but are administered the NLSC questionnaire. Their data does not reside on the main NPHS microdata file. In other households, the one respondent aged 12 or older is selected and administered the NPHS questionnaire. Only the data for these respondents resides on the NPHS supplement microdata file. For more details on integration with the NLSC, see the section entitled "Sample Design". The respondents aged 12 or more from households containing children are selected from the "Adult Sample" only. To compensate for the fact that households containing children coming from the "Children Sample" do not contribute to the estimates for individuals aged 12 or more, the weights for those households containing children sampled in the last two quarters that come from the "Adult Sample" are boosted by the multiplicative weight adjustment given by the inverse of the proportion of the total sample which is assigned to the "Adult Sample". For those individuals aged greater than 12, one adjustment is made at the cluster level. On the other hand, for those aged 12, a separate adjustment is made for groups of LFS strata (which usually correspond to NPHS strata), to be consistent with Adjustment 9B, which is also made at this level.

Adjustment 8B: Selected Member Inverse Selection Probability

As mentioned above, one member aged 12 or more from each sampled household is chosen as the selected member. A weight adjustment must be made to reflect the selection and is given by the inverse selection probability. The original intention was that each member aged 12 or more would be selected with equal probability given by the inverse of the number of members in the household aged 12 or more. However, due to an error made in the CAPI application, no 12 year olds were selected in the first two quarters. To

compensate, in the last two quarters, instead of each member of a household being selected with the same probability, 12 year olds were given a larger probability of selection. In P.E.I., 12 year olds were twice as likely to be selected as any other member aged 13 or more, and elsewhere in Canada, 1.75 times as likely to be selected as any other member aged 13 or more .

Adjustment 9B: Twelve Year Old Weight Adjustment

Due to the error mentioned above, twelve year olds were only selected in the last two quarters of data collection. In order to obtain an accurate representation of twelve year olds, their weights had to be adjusted to account for the first two quarters when they had no probability of being selected. This adjustment is made for groups of LFS strata which usually correspond to NPHS strata, except for the cases of remote and high income strata. In households with children, twelve year olds could be selected from the "Adult Sample" in all quarters, but were actually only selected from the "Adult Sample" in the last two quarters. Since, within most NPHS strata, 40% of the "Adult Sample" occurred in the last two quarters, the weights of twelve year olds selected in these two quarters were boosted by the inverse of this rate, or by 2.5. On the other hand, in households with youths but no children, twelve year olds could be selected from both the "Adult Sample" and the "Children Sample". However, in the first two quarters, they were not selected from the "Adult Sample" as they should have been. Thus, in households with youths but no children, the weights of twelve year olds were boosted by a multiplicative factor given by the ratio of the percentage of the total sample within an NPHS stratum where they should have been selected to the percentage of the total sample where they were actually selected, or by 1.6. Finally, in households with no youths or children, twelve year olds could never be selected, so no adjustment was made to the weights of twelve year olds in this household type. Note that the rates differ somewhat in P.E.I., apartment strata, high income strata and remote strata.

Adjustment 10B: NPHS Selected Member Non-response Weight Adjustment

It may happen that, although a household is considered to be "responding", the information for the selected member of the household was not completed. The members for which this is true are considered to be selected member non-respondents and a weight adjustment is made to responding selected members in the same age-sex group and province, to compensate. The multiplicative adjustment is given by:

$$\frac{\text{sum of weights of all sampled selected members in a province-age-sex category}}{\text{sum of weights of respondent selected members in a province-age-sex category}}.$$

The "weights" referred to above are the LFS basic weights multiplied by all the adjustments to this point (i.e., weight adjustments 1 through 6 as well as 7B through 9B). The age categories used for each of the two sexes are: 12-24, 25-44, 45-64, and 65+ since only those aged 12 or more are administered the selected member questionnaire.

Adjustment 10C: NPHS Supplement Non-response Weight Adjustment

As in adjustment 10B, it may happen that, although a household is considered to be "responding", the information for the NPHS supplement was not completed for the NPHS selected member of the household. The members for which this is true are considered to be NPHS supplement non-respondents and a weight adjustment is made to NPHS supplement respondents in the same age-sex group and province, to compensate. The multiplicative adjustment is given by:

$$\frac{\text{sum of weights of all NPHS Supp. sampled members in a province-age-sex category}}{\text{sum of weights of NPHS Supp. respondents in a province-age-sex category}}$$

The "weights" referred to above are the LFS basic weights multiplied by all the adjustments to this point (i.e., weight adjustments 1 through 6 as well as 7B through 10B). The age categories used for each of the two sexes are: 12-24, 25-44, 45-64, and 65+ since only those aged 12 or more are administered the selected member questionnaire.

Adjustment 12B: Benchmarking Weight Adjustment

Independent estimates in the form of population projections are available monthly for various age and sex groups by province. The population projections are based on the most recent Census data, as well as records of births and deaths, and estimates of migration. In the final step, this auxiliary information is used to transform the weights to this point into final weights. The Benchmarking Weight Adjustment ensures that the final weights sum to the population projections mentioned above for the auxiliary variables in question, that is, for the following age categories for both males and females: 0-11, 12-24, 25-44, 45-64, 65+. This weight adjustment is given by the following formula:

$$\frac{\text{population projection for a province-age-sex category}}{\text{sum of weights of respondent selected members in a province-age-sex category}}$$

Here, the age categories used for both females and males are given by: 12-24, 25-44, 45-64, and 65+. The "weights" referred to in the above equation involve the LFS weights calculated previously. This means that the LFS basic weights were multiplied by weight adjustments one through six as well as by weight adjustments 7B through 10C in this

section. Since the NPHS and supplement data was collected over four quarters, the population projections used in the benchmarking are an average of the projections for the four months in which the survey took place. At the last step, the final weights are formed by multiplying the "weights" to this point by the above Benchmarking Weight Adjustment.

11.2 Weighting Procedures for Quebec

The National Population Health Survey used a subsample of the *Enquête sociale et de santé* (ESS) in its design (see "Sample Design" section for more details). For this reason, the calculation of NPHS weights is tied to the weighting procedures used for the ESS. As well, since the NPHS Supplement was administered to NPHS respondents, the design for the supplement is closely tied to that of the NPHS. The following sections describe the ESS weighting procedures and the steps required to produce weights for NPHS members and ultimately for respondents to the NPHS Supplement.

11.2.1 ESS Weights

The ESS contribution to the weights is calculated as follows:

ESS Cluster Weights

The ESS used a stratified multi-stage design. After several levels of stratification, clusters were selected from each stratum using probability proportional to size (PPS). The size measure used was the household count in the cluster based upon the 1986 Census. An "**ESS cluster weight**" can be calculated as the inverse probability of selecting a cluster.

ESS Dwelling Weights

After selecting a cluster, a fixed number of dwellings were allocated to be selected from the cluster. Each dwelling in the cluster had an equal chance of being selected. The "**ESS dwelling weight**" is then the inverse of the probability of selecting the dwelling within the cluster multiplied by the ESS cluster weight.

11.2.2 NPHS Basic Dwelling Weights

There were two major steps to selecting the NPHS sample. First the subset of ESS clusters to be used in the NPHS had to be identified. Second the subset of ESS dwellings within each retained cluster had to be selected.

Probability of Retaining an ESS Cluster for NPHS

As ESS strata were sometimes very small, NPHS strata were defined as comprising of one or more ESS strata. A fixed number of clusters were allocated to be retained from each NPHS stratum. In cases where the NPHS stratum consisted of more than one ESS stratum, the allocation of clusters to ESS strata was proportional to the number of households in each ESS stratum in order to produce a PPS sample of clusters in each NPHS stratum. Fractional sample sizes were randomly rounded up or down to the next

integer. Once the number of clusters to be retained from an ESS stratum had been determined, each cluster within the ESS stratum had the same probability of retention in most cases. The exceptions were clusters in which the number of dwellings grew by more than 150% between the 1986 Census and the 1992-93 ESS cluster listing. These clusters were given a higher probability of retention (either 100% or 40% greater probability of retention).

Probability of Retaining an ESS Dwelling for NPHS

In clusters that were retained for the NPHS, only dwellings that were selected for ESS were eligible to be selected for NPHS. Those dwellings which were out of scope for ESS (businesses, collectives, demolished or abandoned) had a probability of one of being retained. From the ESS in scope dwellings, a fixed number of dwellings within each cluster were initially retained for the NPHS. A further sub-group of these selected dwellings were dropped because of their ESS household composition. The probabilities that a dwelling would be retained due to its household composition are shown in the following table.

Probability of Retaining an Initially Selected NPHS Dwelling

ESS Household Composition	Probability of Retention
Households with children (under 12 years old)	1
One person households	1/3
Other households with at least one youth (aged 12-24)	5/6
Other households	1/2

The "**basic dwelling weight**" is the ESS dwelling weight times the inverse of the product of the ESS cluster retention probability and the ESS dwelling retention probability. The ESS dwelling retention probability includes both the probability of a dwelling being initially retained for NPHS and the probability of being retained due to its household composition.

11.2.3 Further Weight Adjustments to the Basic Weights

Multiples Weight Adjustment

Sometimes when an interviewer visited a dwelling, he/she found an extra dwelling that was missed during cluster listing. An example of this might be a basement apartment. In this case each dwelling is known as a multiple. When this occurred, one dwelling was selected at random and interviewed. The weight of the selected dwelling is then adjusted by a multiplicative factor equal to the number of multiples.

Cluster Growth Weight Adjustment

NPHS PUBLIC USE MICRODATA DOCUMENTATION

In a few cases, clusters were relisted by NPHS. If there was a growth of 15-30% between ESS counts and NPHS counts, then a multiplicative weight adjustment of

$$\frac{NPHS \text{ count}}{ESS \text{ count}}$$

is made to each selected dwelling within the cluster. If the growth was less than 15% then the growth is assumed to be negligible and this adjustment is set to one. For all of these dwellings, the multiples and cluster growth adjustments are multiplied by the basic dwelling weight to give a "**preliminary weight**".

If the growth was over 30% then extra dwellings were selected for NPHS from the extra dwellings listed within the cluster. For these selected extra dwellings, the "**preliminary weight**" is the inverse of the product of the ESS cluster selection probability and NPHS cluster retention probability multiplied by

$$\frac{\text{number of extra dwellings listed}}{\text{number of extra dwellings selected}}$$

and the multiples adjustments. Since none of these dwellings were interviewed by ESS, there is no way to categorize them into one of the ESS household composition categories.

Household Nonresponse Weight Adjustment

Nonresponse is inevitable in almost all surveys and NPHS is no exception. To adjust for total nonresponding households, the following adjustment is made

$$\frac{\text{sum of weights for respondent and nonrespondent households}}{\text{sum of weights for respondent households}}$$

The weight in this case is the preliminary weight.

A separate adjustment is done within a nonresponse weighting area. For the ESS in scope dwellings the nonresponse weighting areas are defined as an intersection of an NPHS stratum and ESS household type (the four ESS household composition categories described in Table 1) by quarter. If this produces a high adjustment factor (greater than 2.5), then household types are systematically collapsed together until the factor is less

than 2.5.

For the dwellings which were added because the cluster had greater than 30% growth during NPHS relisting, the weighting area consists of the added dwellings within the cluster by quarter.

The ESS out of scope dwellings are grouped into two non-response weighting areas by quarter for non-response adjustment purposes. The first group contains all of the dwellings which had an ESS response code of 10 (demolished, vacant, abandoned). The second contains all of the dwellings which had an ESS response code of 18 (collective or business).

Multiplying the preliminary weight by the household nonresponse weight adjustment produces the "**demographic weight**".

11.2.5 Further Weight Adjustments for Selected Members

One member from each NPHS responding household is designated as the selected longitudinal member. If this person is a child under twelve years of age who lives in a "Children" sample dwelling (see "Sample Design" section for the definition of "Children" sample dwelling) then all of the children in the household to a maximum of four are administered the National Longitudinal Survey of Children (NLSC) questionnaire. Otherwise the selected member (aged twelve and over) is asked an additional set of NPHS questions. Several adjustments have to be made to account for this design and the nonresponse to this questionnaire, as well as nonresponse to the NPHS supplementary questionnaire.

NLSC Integration Weight Adjustment

In a "Children" sample household where a child is found, one child is chosen to be the selected member for the NPHS longitudinal panel. This child, as well as all other children in the household, to a maximum of four, is administered the NLSC selected member questionnaire. This data does not reside on the present microdata file. An adjustment has to be made to account for the adults and youths in these dwellings who had no chance of being the selected member. This adjustment is only applied to adults and youths that are selected for the longitudinal panel in "Adult" dwellings where children were found by NPHS.

The adjustment is equal to the inverse subsampling rate for the "Adult" sample. The adjustment depends upon which combination of the following categories the dwelling fell into

- 1) Did ESS find children in the dwelling? (yes or no)
- 2) Which ESS urban density class does the dwelling belong to?

A separate adjustment is generated for dwellings where ESS found children and dwellings where ESS did not find children because the subsampling rate was different for these two categories. In the ESS Montreal and regional capitals classes, the adjustment is made at the cluster level while in the ESS smaller urban agglomerations and rural sector classes, it is made at the NPHS stratum level. For an exception to this rule see "Twelve Year Old Weight Adjustment" later in this section.

Selected Member Inverse Selection Probability

In a dwelling belonging to the "Children" sample in which there were no children under the age of twelve or a dwelling belonging to the "Adult" sample, every member aged 12 or more was originally intended to have an equal probability of being the selected longitudinal member. However, due to a software error, twelve year olds were not selected in the first two quarters. To compensate for this they were given double the probability of being selected in quarters 3 and 4. A weight adjustment equal to the inverse probability of an individual within the household being the selected member is applied.

Twelve Year Olds Weight Adjustment

In order to get an accurate representation of twelve year olds, their weight has to be increased to account for households where they were not eligible to be selected as a result of the software error. This adjustment is equal to the inverse probability that a twelve year old was eligible to be selected from a dwelling where a person twelve or over was intended to be the longitudinal respondent.

Recall that in the Montreal and regional capitals classes, clusters are only covered in one quarter. In quarters 1 and 2 a twelve year old was not eligible to be selected. Therefore, in order for the weight adjustment to account for these ineligible twelve year olds, it must be done at the NPHS stratum level rather than the cluster level. For consistency, both the integration and twelve year old weight adjustment are calculated at the NPHS stratum level for twelve year olds regardless of the ESS class.

NPHS Selected Member Nonresponse Weight Adjustment

This adjustment compensates for selected individuals within NPHS responding households (i.e., households for which the demographic questionnaire was completed) who do not respond to the selected member questionnaire. The adjustment is equal to

NPHS PUBLIC USE MICRODATA DOCUMENTATION

$$\frac{\text{sum of weights for NPHS sampled individuals in an age-sex category}}{\text{sum of weights for respondent individuals in an age-sex category}}$$

The weight in this case is the demographic weight multiplied by all of the previous adjustments made in the selected members weight adjustment section. The age-sex categories are 12-24, 25-44, 45-64, 65 and over - males and females. The adjustment is made at the provincial level.

NPHS Supplement Nonresponse Weight Adjustment

This adjustment compensates for selected individuals within NPHS responding households (i.e., households for which the selected member questionnaire was completed) who do not respond to the NPHS supplementary questionnaire. The adjustment is equal to

$$\frac{\text{sum of weights for NPHS Supp. sampled individuals in an age-sex category}}{\text{sum of weights for NPHS Supp. respondent individuals in an age-sex category}}$$

The weight in this case is the demographic weight multiplied by all of the previous adjustments made in the selected members weight adjustment section. The age-sex categories are 12-24, 25-44, 45-64, 65 and over - males and females. The adjustment is made at the provincial level.

Final population Benchmarking Weight Adjustment

The formula for this adjustment is as follows:

$$\frac{\text{population projections for an age-sex category}}{\text{sum of weights for respondent individuals in an age-sex category}}$$

The weights are the demographic weights multiplied by all of the previous adjustments made in this section including the supplementary nonresponse weight adjustment. The age-sex categories are 12-24, 25-44, 45-64, 65 and over - males and females. Note that adjustments to the benchmarks have been made to account for the out of scope northern regions.

The **"final weight"** is calculated by multiplying the demographic weight by all of the adjustments made in this section.