Catalogue No. 92-395-XIE

Sampling and Weighting

2001 Census Technical Report

Statistics Canada Statistique Canada

Canada


ELECTRONIC PUBLICATIONS AVAILABLE AT
www.statcan.ca

2001 Census Technical Report

Sampling and Weighting

Introduction

The 2001 Census required the participation of the entire population of Canada, some 30 million people distributed over a territory of 9 million square kilometres. Although there are high quality standards governing the gathering and processing of the data, it is not possible to eliminate all errors. In order to help users assess the usefulness of census data for their purposes, the 2001 Census Technical Reports detail the conceptual framework and definitions used in conducting the census, as well as the data collection and processing procedures employed. Also, the principal sources of error, including where possible the size of these errors, are also described, as are any unusual circumstances which might limit the usefulness or interpretation of census data. With this information, users can determine the risks involved in basing conclusions or decisions on census data.

This *2001 Census Technical Report* deals with the method of sampling and weighting used in the 2001 Census as well as its effect on the results. Due to the fact that some information is collected on a sample basis and weighted to the full population level, bias and discrepancies can be observed in the final estimates. This report identifies these observed differences and explains the probable causes. This report has been prepared by Wesley Benjamin, Édith Hovington and Mike Bankier, with the support of staff from two divisions in Statistics Canada: the Social Survey Methods Division and the Census Operations Division.

Sampling is an accepted practice in many aspects of life today. The quality of produce in a market may be judged visually by a sample before a purchase is made; we form opinions about people based on samples of their behaviour; we form impressions about countries or cities based on brief visits to them. These are all examples of sampling in the sense of drawing inferences about the "whole" from information for a "part".

In a more scientific sense, sampling is used, for example, by accountants in auditing financial statements, in industry for controlling the quality of items coming off a production line, and by the takers of opinion polls and surveys in producing information about a population's views or characteristics. In general, the motivation to use sampling stems from a desire either to reduce costs or to obtain results faster, or both. In some cases, measurement may destroy the product (e.g., testing the life of light bulbs) and sampling is therefore essential. The disadvantage of sampling is that the results based on a sample may not be as precise as those based on the whole population. However, when the loss in precision (which may be quite small when the sample is large) is tolerable in terms of the uses to which the results are to be put, the use of sampling may be cost-effective.

The 2001 Census of Population made use of sampling in a variety of ways. It was used in ensuring that the quality of the census representative's work in collecting questionnaires met certain standards; it was used in the control of the quality of coding responses during processing; it was used in estimating both the amount of under-coverage and the amount of over-coverage; it was used in evaluating the quality of census data. However, the primary use of sampling in the census was during the field enumeration when all but the basic census data were collected only from a sample of households. This report describes this last use of sampling and evaluates the effect of sampling on the quality of census data.

Chapters 1 and 2 describe the data collection and data processing procedures. Chapter 3 reviews the history of the use of sampling in Canadian censuses and describes the sampling procedures used in the 2001 Census. Chapter 4 explains the procedures used for weighting up the sample data to the population level and provides operational and theoretical justifications for these procedures. In Chapter 5 the program of studies designed to evaluate the 2001 Census sampling and weighting procedures is presented, while Chapters 6 through 9 present the results of these studies. Chapter 10 presents some conclusions on the weighting procedures used in 2001.

Users will find additional information on census concepts, variables and geography in the *2001 Census Dictionary* (Catalogue No. 92-378-XIE), and an overview of the complete census process in the *2001 Census Handbook* (Catalogue No. 92-379-XIE).

1. Census Data Collection

1.1 General

The data collection stage of the 2001 Census process ensures that each of the 11.8 million households in Canada is enumerated on Census Day (Tuesday, May 15, 2001). The census enumerates the entire Canadian population, which consists of Canadian citizens (by birth and by naturalization), landed immigrants, and non-permanent residents. Non-permanent residents are persons living in Canada who have a Minister's permit, student or employment authorization, or who are claiming refugee status, and family members living with them.

The census also counts Canadian citizens and landed immigrants who are temporarily outside the country on Census Day, including federal and provincial government employees working outside Canada, Canadian embassy staff posted to other countries, members of the Canadian Armed Forces stationed abroad, and all Canadian crew members of merchant vessels. Because the census enumerates people where they usually or typically reside rather than where they physically happen to be on Census Day, the Census of Canada is considered a *de jure* census. This means that people outside the country on Census Day were enumerated if their usual or normal place of residence was back in Canada. Some countries conduct a *de facto* census. This type of census is based on where persons actually happen to be on Census Day and not necessarily where they live.

The Census of Canada uses different forms and questionnaires to collect data. The following forms are referred to in this report.

A Form 1 is called a Visitation Record (VR). The VR is used to list every occupied and unoccupied private dwelling or collective dwelling, agricultural operation and agricultural operator in the enumeration area. The VR serves as an address listing for field operations and control purposes for census collection.

The basic short questionnaire is called the 2A. The 2A questionnaire has ten questions and is distributed to every four in five households. The 2B is a longer questionnaire that collects the same information as the 2A plus additional information on a variety of topics. The 2B questionnaire is distributed to every one in five households. Each household that receives a 2A or 2B census questionnaire is asked to enumerate and provide information on all household members who fall into the census population.

A Form 4 is completed by census staff in situations where household occupants were absent or refused to respond. Information on private dwellings which were unoccupied on Census Day is recorded on a Form 2A or Form 2B.

A Form 3 (A and B) is used to enumerate persons in a collective dwelling (each person in the collective dwelling would complete a separate Form 3). It can also be used to enumerate usual residents in a private household who prefer to be enumerated on their own census questionnaire rather than be included on a 2A or 2B questionnaire.

Canadians stationed abroad (generally embassy or armed forces personnel) are given a Form 2C, which contains the same questions as the Form 2B except that housing questions are not included. However, questions about the person's usual place of residence in Canada are asked.

1.2 Collection Methods

To ensure the best possible collection coverage, Canada is divided into small geographic areas called enumeration areas (EAs). For collection purposes, each EA is under the responsibility of a census representative (CR). CRs are involved in mapping, listing, distribution and verification activities in their assigned EAs and they ensure that all questionnaires are returned to the processing centres. The number of households in an EA ranges from 175 in rural areas to 600 in urban areas. In the 2001 Census, there

were 42,851 enumeration areas in Canada. CRs work under the supervision of field census commissioners (CCs). The 2,917 CCs in 2001 were responsible for hiring CRs and for the planning and management of field collection activities in their designated area.

In 2001, approximately 98% of households were self-enumerated. Self-enumeration requires that a CR drop off a census questionnaire at each household during the two weeks before Census Day. An adult, or any other responsible member of the household, is asked to complete the questionnaire for all members of the household, and then return the questionnaire by mail in a pre-addressed envelope.

Approximately 2% of households were enumerated in the 2001 Census using the canvasser enumeration method. In this case, a CR visits the household and completes a questionnaire for the household by way of an interview. This method is normally used in remote and northern areas of the country, and on most Indian reserves. The canvasser enumeration method is also used in certain urban areas where it is considered highly likely that respondents would not return a questionnaire.

CRs and CCs are involved in a number of field-related collection activities. These include contacting a household to resolve problems that typically relate to the completeness or consistency of the information provided. They also deal with situations where no questionnaire is returned.

During the field collection operations, the CRs delivered a questionnaire to each dwelling within their EA, and wrote the person's name (if possible) and the address in their Visitation Records (VRs). At the same time, they copied down the unique identifiers that would later be captured and used to assign each household and dwelling to the correct geographic area. As well, they identified the block number for the dwelling from their EA map and copied the number into the VR and onto the questionnaire. These block numbers were later data-captured so that all the dwellings in Canada could be identified as belonging to a particular block.

2. Census Data Processing

2.1 Introduction

This part of the census process involved the processing of all the completed questionnaires. This encompasses everything from the key entry of the questionnaire data through to the creation of an accurate and complete retrieval database. Considered here are the steps of manual and automated data capture, questionnaire imaging, editing, error correction, coding, imputation and weighting. The final database was transferred to the Data Quality Measurement Project to determine the overall quality of the data, and to the Census Dissemination Project for the production and marketing of the 2001 Census products and services. In the remainder of this chapter, each data processing operation will be summarized.

An important innovation for the 2001 Census was to create an image retrieval system giving access to the images (pictures) of all the census questionnaires and Visitation Records (see Section 2.3). This would make it possible during subsequent processes to access original census questionnaires and forms without having to manually handle thousands of boxes and paper documents, as was required in past censuses.

2.2 Regional Processing

The Regional Processing team was responsible for the data capture of the questionnaire information into a machine-readable format for subsequent processing. This team was also responsible for the manual research and coding of the industry and occupation responses from 2B questionnaires. Given the number of census questionnaires and quantity of information to be captured (representing over four billion keystrokes), Regional Processing, since the 1981 Census, has been contracting this work out to Revenue Canada, now called the Canada Customs and Revenue Agency (CCRA). CCRA has used their network of systems, resources and staff to key and code census data. By using the staff and infrastructure already in place at CCRA, the census realized cost savings. Census data quality also benefits from the experience that CCRA has in processing past census questionnaires. For the 2001 Census, approximately 2,800 CCRA employees were sworn to secrecy under the Statistics Act to perform the census work. By this arrangement, CCRA employees work under the same rules and regulations as those which apply to the employees of Statistics Canada.

When the collection activities for a specific enumeration area (EA) were completed, the questionnaires, along with maps and Visitation Records, were shipped in EA boxes from the field collection units to one of eight designated CCRA tax centres across Canada.

The first processing step was to prepare completed questionnaires for data capture. This traditionally included the manual assignment of codes to the written answers provided by the respondents. For 2001, most of the written responses were converted to codes using automated systems (see Section 2.5). The only written responses that had to be manually coded for the 2001 Census were the questions on industry and occupation contained on the 2B questionnaires. Research into the automation of the coding of these questions has begun, and it is expected that an automated system will be operational for the 2006 Census.

The industry responses were coded at CCRA according to the North American Industry Classification System (NAICS), which was introduced as a standard within Statistics Canada a few years ago. NAICS is designed to provide a common framework for Canada, the United States and Mexico, which will enable the production of industry statistics under the North American Free Trade Agreement (NAFTA). This meant a change for industry coding from the last census where the type of industry was coded using the 1980 Standard Industrial Classification (SIC). In order to allow longitudinal comparisons, the 2001 industry question on the 2B questionnaire was also coded using the 1980 SIC during the Automated Coding phase (see Section 2.5).

Once the questionnaires were received and registered at one of the CCRA tax centres, and the industry and occupation codes assigned, the next step was to sort, label and batch the questionnaires in preparation for data capture. The labels affixed to each questionnaire contained a unique sequence number that was used to control the movement of the questionnaire throughout the CCRA operations. For the first time, the label also included a bar code to facilitate the scanning of the questionnaire in the imaging operation (see Section 2.3).

Data capture was then performed by traditional manual keying. Verification of the accuracy of the data capture operation was done by selecting a sample of questionnaires that were already key-entered and recapturing the data from the questionnaires in this sample. Quality control statistics were produced by comparing the two sets of captured data. As expected, the keying of data from the census questionnaires introduces some error. Errors occur for a variety of reasons, including inaccurate keying, poorly written or indicated responses on the questionnaires, and missed responses during key entry. The key verification process reduces keying error to a minimum.

As the data were keyed, they were transmitted in real time over dedicated communication lines to the CCRA computer in Ottawa. Within 24 hours, the data were then transferred to tape cartridges and transported by bonded carrier to Statistics Canada, where they were loaded into the mainframe computer. Questionnaires were reassembled into their EA boxes for shipment to the Statistics Canada 2001 processing site in Ottawa. After all the data were keyed, transferred to Statistics Canada and confirmed as being fully received by the Agency, no census data remained with the CCRA.

2.3 Imaging

In previous censuses, the remaining processing steps that required access to the questionnaires and Visitation Records (VRs) used the paper documents. For 2001, the need to handle the paper was eliminated by imaging (scanning) all the questionnaires and VRs as soon as they arrived at the 2001 processing site from the Canada Customs and Revenue Agency (CCRA) centres. Subsequent operations then had access to the questionnaires and VR images using an image retrieval system. This minimized the need to manage the original paper documents.

As the enumeration area (EA) boxes arrived at the 2001 processing site, they were registered. The documents were then prepared for imaging.

The 13 million documents (mainly questionnaires) were imaged using 15 high-volume scanners running five days a week, two shifts per day. The geographic identifier required to identify each document image was automatically assigned using the bar code on the label affixed during the data capture operations at CCRA (see Section 2.2). Quality control was performed to ensure that each document contained the correct number of pages, and that the number of questionnaires by form type was correct for each EA. A resolution operation resolved any difficulties that arose. Images were written to optical platters for subsequent access and archiving. They were also kept in magnetic storage for immediate access by the Interactive Verification activities.

2.4 Interactive Verification

The main objective of Interactive Verification was to identify and correct errors in the data, for which proper resolution required reference to the images of the questionnaires and/or Visitation Records. A detailed set of edit rules was applied to the captured data to identify possible errors, such as households with missing or duplicate persons, incorrect enumeration of foreign or temporary residents, questionnaires assigned to the wrong household, or misclassification of dwellings as occupied or unoccupied. A thorough review of the information on all relevant census forms was conducted to determine the appropriate corrective action for each edit failure. In some cases, this required adding and/or deleting persons or dwellings.

As the census data arrived on cartridges from the Canada Customs and Revenue Agency (CCRA), they were loaded into Statistics Canada's computers in preparation for the Interactive Verification activities. A series of automated "structural" edits were performed, mainly to verify the information filled out by the census representative (CR) on the front cover of the questionnaire. These edits included, among other things, matching questionnaire and household types, cross-checking the number of questionnaires and people enumerated, and verifying that the geographic identifiers were unique. Some edits were also performed on the income information on the 2B questionnaire, so that anomalies could be examined by income subject-matter specialists.

All edits were done by enumeration area (EA). Errors were flagged, and then corrected by referring to the images of the questionnaires and Visitation Record (VR) for that EA. The corrections were made to the electronic data using an interactive PC-based system. Some of the corrections were also electronically noted on the questionnaire images or on the VRs.

Once the EA editing work was completed, automated and manual processes were then used to verify the geographic identifiers that the CR had copied from the EA map onto the questionnaire and VR.

Interactive Verification also performed some special processing to ensure that Canadians living outside Canada on Census Day (people aboard coast guard and Canadian Armed Forces vessels, Canadian-registered merchant vessels, and diplomatic and military personnel) were enumerated properly.

As a final step in the Interactive Verification process, the data were reformatted and forwarded on for the final processing steps. These were the Automated Coding and Edit and Imputation phases.

2.5 Automated and Interactive Coding

Automated coding is the process of matching the write-in responses that were data-captured from the 2B questionnaires during Regional Processing (see Section 2.2) to entries in an automated reference file/classification structure containing a series of words or phrases and corresponding numerical codes. Although a large percentage of write-in responses can be coded in a purely automated manner, a number of responses always remain unmatched. Specially trained coding persons and subject-matter specialists reviewed all unmatched responses. Using the PC-based interactive coding systems and by examining responses to other questions on the questionnaire, sometimes relating to other members of the household, they assigned the appropriate numerical code. Automated coding was applied to write-in responses for the following questions on the 2B questionnaire:

- relationship to Person 1;
- language spoken at home;
- non-official languages;
- first language learned in childhood (mother tongue);
- language of work;
- place of birth;
- place of birth of parents;
- citizenship;
- ethnic origin (ancestry);
- population group;
- Indian Band/First Nation;
- place of residence 1 year ago;
- place of residence 5 years ago;

- major field of study;
- religion;
- place of work;
- industry (according to 1980 SIC).

As the responses for a particular variable were coded, the data for that variable were sent to the Edit and Imputation phase.

2.6 Edit and Imputation

The data collected in any survey or census contains omissions and inconsistencies. These errors can be the result of respondents answering the questions incorrectly or incompletely, or they can be due to errors generated during processing. For example, a respondent may be reluctant to answer a question, may fail to remember the right answer or may misunderstand the question. Census staff may code responses incorrectly or may make other mistakes during processing.

One of the first tasks of the Edit and Imputation project is to ensure that all dwellings classified as "occupied" have a household size. For those occupied dwellings for which a regular questionnaire (a Form 2A or 2B) was not completed, and for which only the dwelling non-response questionnaire (a Form 4) was received, the first job in Edit and Imputation was to ensure that the dwelling had a valid household size. For those dwellings where the household size was "unknown", the procedure was to impute the household size of the nearest neighbour. In addition, for 2001, a new procedure was introduced to reimpute the household size of some of these Forms 4 dwellings based on the Dwelling Classification Study described in Section 2.7.

The final clean-up of the data was done in Edit and Imputation and was, for the most part, fully automated. It applied a series of detailed edit rules that identified any missing or inconsistent responses. These missing or inconsistent responses were corrected most of the time by changing the values of as few variables as possible through imputation. Imputation invoked either **deterministic** or **minimum-change hot-deck** methods. For deterministic imputation, errors were corrected by inferring the appropriate response value from responses to other questions. For minimum-change hot-deck imputation, a record with a number of characteristics in common with the record in error was selected. Data from this "donor" record were borrowed and used to change the minimum number of variables necessary to resolve all the edit failures.

Two different automated systems were used to carry out this processing.

The **N**earest-neighbour **I**mputation **M**ethod (NIM), developed for the 1996 Census for performing Edit and Imputation for basic demographic characteristics such as age, sex, marital status, common-law status and relationship to Person 1, was expanded for 2001 and implemented in a system called CANCEIS (**CAN**adian **C**ensus **E**dit and **I**mputation **S**ystem) to include Edit and Imputation for such variables as industry, place of work, mode of transportation and mobility. As in 1996, CANCEIS continued to allow more extensive and exact edits to be applied to the response data, while preserving responses through minimum-change hot-deck imputation.

SPIDER (**S**ystem for **P**rocessing **I**nstructions from **D**irectly **E**ntered **R**equirements) was used to process the remaining census variables, such as mother tongue, dwelling and income. This tool translated subject-matter requirements, identified through decision logic tables, into computer-executable modules. SPIDER performed both deterministic and hot-deck imputation.

2.7 Coverage Adjustments for Unoccupied and Non-response Dwellings

The Dwelling Classification Study (DCS) takes a sample of dwellings reported as being either unoccupied or occupied during the collection process. Later, DCS interviewers return to these dwellings to determine if, on Census Day, they were occupied, unoccupied or should not have been listed because they did not meet the census definition of a dwelling.

If a dwelling was occupied, one of two separate adjustments was made to the census database. If the dwelling was listed as unoccupied in the census, then a technique called **random additions** was applied to add households and persons to the census database. In the 2001 Census, 111,628 households and 222,720 persons were added to the database to account for the estimated number of persons living in "unoccupied" dwellings. The second adjustment was concerned with occupied dwellings for which a completed census questionnaire was not received, i.e. non-response dwellings, and consisted in adjusting all such dwellings by creating a new household size for them on the census database. A total of 143,681 households with 317,587 persons were added to the census database through this adjustment.

2.8 Weighting

Data on age, sex, marital status, common-law status, mother tongue and relationship to Person 1 were collected from almost all Canadians. However, the bulk of the data gathered in the census came from the one-in-five, or 20%, sample of households which received a 2B questionnaire (see Section 1.1). Weighting, applied to the respondent data after Edit and Imputation, was used to adjust the census sample to represent the whole population.

The weighting method produces weights that are used to form estimates from the 20% sample data. For the 2001 Census, weighting employed a methodology known as calibration (or regression) estimation. Calibration estimation started with initial weights of approximately 5 and then adjusted them by the smallest possible amount needed to ensure closer agreement between the sample estimates (e.g. number of males, number of people aged 15 to 19) and the population counts for age, sex, marital status, common-law status and household size. This method is described in detail in Chapter 4.

3. Sampling in Canadian Censuses

In the context of a census of population, sampling refers to the process whereby certain characteristics are collected and processed only for a random sample of the dwellings and persons identified in the complete census enumeration. Tabulations that depend on characteristics collected only on a sample basis are then obtained for the whole population by scaling up the results for the sample to the full population level. Characteristics collected on all dwellings or persons in the census will be referred to as "basic characteristics" while those collected only on a sample basis will be known as "sample characteristics."

3.1 The History of Sampling in the Canadian Census

Sampling was first used in the Canadian census in 1941. A Housing Schedule was completed for every tenth dwelling in each census subdistrict. The information from 27 questions on the separate Housing Schedule was integrated with the data in the personal and household section of the Population Schedule for the same dwelling, thus allowing cross-tabulation of sample and basic characteristics. Also in the 1941 Census, sampling was used at the processing stage to obtain early estimates of earnings of wage-earners, of the distribution of the population of working age, and of the composition of families in Canada. In this case, a sample of every tenth enumeration area across Canada was selected and all Population Schedules in these areas were processed in advance.

Again in 1951, the Census of Housing was conducted on a sample basis. This time every fifth dwelling (those whose identification numbers ended in a 2 or 7) was selected to complete a housing document containing 24 questions. In the 1961 Census, persons 15 years of age and over in a 20% sample of private households were required to complete a Population Sample Questionnaire containing questions on internal migration, fertility and income. Sampling was not used in the smaller censuses of 1956 and 1966.

The 1971 Census saw several major innovations in the method of census-taking. The primary change was from the traditional canvasser method of enumeration to the use of self-enumeration for the majority of the population. This change was prompted by the results of several studies in Canada and elsewhere (Fellegi [1964]; Hansen et al. [1959]) that indicated that the effect of the enumerator was a major contribution to the variance of census figures in a canvasser census. Thus the use of self-enumeration was expected to reduce the variance[1] of census figures through reducing the effect of the enumerator, while at the same time giving the respondent more time and privacy in which to answer the census questions—factors which might also be expected to yield more accurate responses.

The second aspect of the 1971 Census that differentiated it from any earlier census was its content. The number of topics covered and the number of questions asked were greater than in any previous Canadian census. Considerations of cost, respondent burden, and timeliness versus the level of data quality to be expected using self-enumeration and sampling led to a decision to collect all but certain basic characteristics on a one-third sample basis in the 1971 Census. In all but the more remote areas of Canada, every third private household received the "long questionnaire" which contained all the census questions, while the remaining private households received the "short questionnaire" containing only the basic questions covering name, relationship to head, sex, date of birth, marital status, mother tongue, type of dwelling, tenure, number of rooms, water supply, toilet facilities, and certain coverage items. All households in pre-identified remote enumeration areas and all collective dwellings[2] received the long questionnaire. A more detailed description of the consideration of the use of sampling in the 1971 Census is given in *Sampling in the Census* (Dominion Bureau of Statistics [1968]).

[1] The "variance" of an estimate is a measure of its precision. Variance is discussed more fully in Chapter 9.

[2] A collective dwelling is a dwelling of a commercial, institutional or communal nature. Examples include hotels, hospitals, staff residences and work camps.

The content of the 1976 Census was considerably less than that of the 1971 Census. Furthermore, the 1976 Census did not include the questions that cause the most difficulty in collection (e.g., income) or that are costly to code (e.g., occupation, industry, and place of work). Therefore, the benefits of sampling in terms of cost savings and reduced respondent burden were less clear than for the 1971 Census. Nevertheless, after estimating the potential cost savings to be expected with various sampling fractions, and considering the public relations issues related to a reversion to 100% enumeration after a successful application of sampling in 1971, it was decided to use the same sampling procedure in 1976 as in 1971.

Most of the methodology used in the 1971 and 1976 censuses was kept for the 1981 Census, except that the sampling rate was reduced from every third occupied private household to every fifth. Studies done at the time showed that the resulting reduction in data quality (measured in terms of variance) would be tolerable, and would not be significant enough to offset the benefits of reduced cost and response burden, and improved timeliness (see Royce [1983]).The one-in-five sampling rate was maintained for the censuses of 1986, 1991, 1996 and 2001.

3.2 The Sampling Scheme Used in the 2001 Census

A wealth of information was collected from everyone in Canada on Census Day, May 15, 2001. The bulk of the information was acquired on a sample basis. In all self-enumeration areas, a one-in-five sample of private occupied households was selected to receive a long questionnaire (Form 2B) while the non-sample households received a short questionnaire (Form 2A). Basic questions on age, sex, marital status, mother tongue, relationship to the household reference person (Person 1) were asked of all respondents. Additional information on the dwelling, plus socio-economic questions, was asked on a sample basis.

All dwellings in those areas enumerated by the canvasser method (generally remote areas or Indian reserves) received the Form 2B. All collective dwellings also received the Form 2B. However, the following persons in collective dwellings were not asked the sample questions:

(a) inmates in correctional and penal institutions or jails;

(b) patients in general hospitals, special care homes and institutions for the elderly, and chronically ill or psychiatric institutions;

(c) children in orphanages and children's homes or young offenders facilities.

The basic drop-off or delivery procedure required the census representative to pre-plan a route covering all dwellings in his/her enumeration area (EA) and then to visit each dwelling and leave a census questionnaire. The selection of the sample, i.e., the decision as to which type of questionnaire to leave at each occupied dwelling, was facilitated by the Visitation Record (VR), the document in which the census representative listed each dwelling in his/her area. This document was printed so that every fifth line was shaded to signify that a Form 2B should be delivered. Those dwellings not in the sample received a short questionnaire (Form 2A). A random start was implemented by deleting either zero, one, two, three or four lines at the start of the VR according to whether the fifth, fourth, third, second or first dwelling in the EA was to be the first to receive the long form. Thereafter, the dwelling listed on each shaded line automatically received the long form. These procedures were spelled out in the Census Representative's Manual and emphasized in his/her training in order to minimize the risk of any deviation from the specified procedure for selecting the sample.

In sampling terminology, the census sample design can be described as a stratified systematic sample of private occupied dwellings using a constant one-in-five sampling rate in all strata (EAs). As a sample of persons, it can be regarded as a stratified systematic cluster sample with dwellings as clusters. For a more detailed description of the concepts and terminology of sampling, see Cochran (1977) or Sarndal, Swensson and Wretman (1992).

4. Estimation from the Census Sample

Any sampling procedure requires an associated estimation procedure for scaling sample data up to the population level. The choice of an estimation procedure is generally governed by both operational and theoretical constraints. From the operational viewpoint, the procedure must be feasible within the processing system of which it is a part, while from the theoretical viewpoint the procedure should minimize the sampling error of the estimates it produces. In the following two sections, the operational and theoretical considerations relevant to the choice of estimation procedures for the census sample are described.

4.1 Operational Considerations

Mathematically, an estimation procedure can be described by an algebraic formula that shows how the value of the estimator for the population is calculated as a function of the observed sample values. In small surveys that collect only one or two characteristics, or in cases where the estimation formula is very simple, it might be possible to calculate the sample estimates by applying the given formula to the sample data for each estimate required. However, in a survey or census in which a wide range of characteristics is collected, or in which the estimation formula is at all complex, the procedure of applying a formula separately for each estimate required is not feasible. In the case of a census, for example, every cell of every tabulation based on sample data at every geographic level represents a sample estimate which under this approach would require a separate application of the estimation formula. In addition, the calculation of each estimate separately would not necessarily lead to consistency between the various estimates made from the same census sample.

The approach taken in the census therefore (and in many sample surveys) is to split the estimation procedure into two stages: (a) the calculation of weights (known as the weighting procedure); (b) the summing of weights to produce estimated population counts. Any mathematical complexity is then contained in step (a) which is performed just once, while step (b) is reduced to a simple process of summing weights which takes place at the time a tabulation is retrieved. It should be noted that since the weight attached to each sample unit is the same for whatever tabulation is being retrieved, consistency between different estimates based on sample data is assured.

4.2 Theoretical Considerations

For a given sample design and a given estimation procedure, one can, from sampling theory, make a statement about the chances that a certain interval will contain the unknown population value being estimated. The primary criterion in the choice of an estimation procedure is minimization of the width of such intervals so that these statements about the unknown population values are as precise as possible. The usual measure of precision for comparing estimation procedures is known as the standard error. Provided that certain relatively mild conditions are met, intervals of plus or minus two standard errors from the estimate will contain the population value for approximately 95% of all possible samples.

As well as minimizing standard error, a second objective in the choice of estimation procedure for the census sample is to ensure, as far as possible, that sample estimates for basic (i.e., Form 2A) characteristics are consistent with the corresponding known population values. Fortunately, these two objectives are usually complementary in the sense that sampling error tends to be reduced by ensuring that sample estimates for certain basic characteristics are consistent with the corresponding population figures. However, while this is true in general, forcing sample estimates for basic characteristics to be consistent with corresponding population figures for very small subgroups can have a detrimental effect on the standard error of estimates for the sample characteristics themselves.

In the absence of any information about the population being sampled other than that collected for sample units, the estimation procedure would be restricted to weighting the sample units inversely to their probabilities of selection (e.g., if all units had a one-in-five chance of selection, then all selected units

would receive a weight of 5). In practice, however, one almost always has some supplementary knowledge about the population (e.g., its total size, and possibly its breakdown by a certain variable—perhaps by province). Such information can be used to improve the estimation formula so as to produce estimates with a greater chance of lying close to the unknown population value. In the case of the census sample, a large amount of very detailed information about the population being sampled is available in the form of the basic 100% data at every geographic level. We can take advantage of this wealth of population information to improve the estimates made from the census sample. However, this information can also be an embarrassment in the sense that it is impossible to make the sample estimates for basic characteristics consistent with all the population information at every geographic level. Differences between sample estimates and population values become visible when a cross-tabulation of a sample variable and a basic variable is produced. The tabulation has to be based on sample data with the result that the marginal totals for the basic variable are sample estimates that can be compared with the corresponding population figures appearing in a different tabulation based on 100% data. They will not necessarily agree.

4.3 Developing an Estimation Procedure for the Census Sample

Given that a weight has to be assigned to each unit (person, family or household) in the sample, the simplest procedure would be to give each unit a weight of 5 (because a one-in-five sample was selected). Such a procedure would be simple and unbiased[3] and, if nothing but the sample data were known, it might be the optimum procedure. However, although we know that the sample will contain almost exactly one-fifth of all households (excluding collective households and those in canvasser areas), one cannot be certain that it will contain exactly one-fifth of all persons, or one-fifth of each type of household, or one-fifth of all females aged 25 to 34, and so on. Therefore, this procedure would not ensure consistency even for the most important subgroups of the population. For large subgroups, these fractions should be very close to one-fifth, but for smaller subgroups they could differ markedly from one-fifth. The next most simple procedure would be to define certain important subgroups (e.g., age-sex groups within province) and, for each subgroup, to count the number of units in the population in the subgroup (N) and the number in the sample (n) and to assign to each sample unit in the subgroup a weight equal to N/n. These subgroups are often called **poststrata**.

For example, if there were 5,000 males aged 20 to 24 enumerated in Prince Edward Island, and 1,020 of these fell in the sample households, then a weight of 5,000/1,020 = 4.90 would be assigned to each male aged 20 to 24 in the sample in Prince Edward Island. This would ensure that whenever sex and age in five-year groups were cross-classified against a sample characteristic for Prince Edward Island, the marginal total for the male 20-24 age-sex group would agree with the population total of 5,000. This type of estimation procedure is known as **ratio estimation**. By contrast, note that if a simple weight of 5 was used, it would have resulted in a sample estimate of 5,100 (1,020 x 5).

Adjusting the simple weights of 5 by small amounts to achieve perfect agreement between estimates and population counts is known as **calibration**. Prior to 1991, calibration was achieved using a procedure called Raking Ratio Estimation. Household level estimates were generated using a household-level calibrated weight while the person-level estimates were generated using a person-level calibrated weight.

In 1991, the two step Generalized Regression (GREG Estimator) was introduced. It achieved a higher level of agreement between population counts and the corresponding estimates at the EA level than had been possible with Raking Ratio Estimation. In addition, a single household level calibrated weight was used to produce both the household and person level estimates. This eliminated inconsistencies that had been observed in some estimates prior to 1991.

With the GREG, the initial weights of approximately 5 were adjusted as little as possible for individual households such that there was perfect agreement between the estimates and the population counts for

[3] "Unbiased" means that the average of the estimates obtained by this procedure, over all possible samples, would equal the true population value.

as many of the basic characteristics as possible that are listed in Appendix B. (**These will be called constraints or auxiliary variables**.) It was required that this perfect agreement be achieved at the weighting area (WA) level. Each WA contained, on average, seven sampled EAs. More information on WAs is given in Section 7.1 of this report.

In 1996, each EA represented the work assignment for one census representative. Whole EAs were combined to form WAs. In 2001, EAs still represented the work assignments for census representatives but were sometimes made larger in urban areas. In 2001, a one-in-five systematic sample of households was still selected from each EA. A new geographic level, Dissemination Areas (DAs), however, was introduced. DAs were created to be similar in size to 1996 EAs, and whole DAs were combined to form WAs (approximately eight sampled DAs per WA).

4.4 The Two-step Generalized Regression Estimator

For five-year age ranges, marital status, common-law status, sex and household size (see Appendix B for the 32 auxiliary variables), the objectives for the 2001 Census weighting procedure are:

(a) To have **exact** population/estimate agreement at the WA level for as many of the 32 auxiliary variables as possible.

(b) To have **approximate** population/estimate agreement for the larger DAs for the 32 auxiliary variables.

In addition, it is required that:

(c) there be **exact** population/estimate agreement for "Total number of households" and "Total number of persons" for as many DAs as possible.

(d) final census weights be in the range 1–25 inclusive. In 1996, the final census weights could be in the range 0.01–25 inclusive. A lower bound of 1 was required for 2001 because it was felt that each sampled person should, at minimum, represent themselves.

(e) the method to generate weights be highly automated since the 6,141 WAs with households subject to sampling must be processed in a short period of time. This method must also adjust automatically for the different patterns of responses in WAs across the country.

Weights are calculated separately in each WA. The 2001 Census **initial** EA-level weights (which equal the number of private households in the population divided by the number in the sample) have either two or three weighting adjustment factors applied to them. First of all, households are sometimes poststratified at the WA level based on household size because small and large households are under-represented in the sample. A second adjustment is then applied to the weights to try to achieve approximate population/estimate agreement at the DA level, as is described in objective (b) above. Finally, a third adjustment is applied to achieve exact population/estimate agreement at the WA and DA levels, as is described in objectives (a) and (c) above. For simplification purposes, the dropping of constraints and the various reasons for this will only be discussed once the three adjustments have been described in more detail.

First, the households are sometimes **poststratified** based on household size (1, 2, 3, 4, 5, or 6+ persons) at the WA level. The initial weights are then multiplied by a factor to generate the poststratified weights. For example, based on the poststratified weights, the estimated number of one-person households for a WA would agree with the number of one-person households in the WA population. Very occasionally, a poststratified weight is truncated to ensure that it lies within the range 1–20 inclusive. An upper limit of 20 rather than 25 is used to give some "room" for further adjustment.

Secondly, a **first-step** regression weighting adjustment factor is calculated at the DA level. The 32 auxiliary variables (age, sex, marital status, household size) that are to be applied at the WA level in the second step are sorted in descending order based on the number of households they apply to in the

population at the DA level. On this ordered list, the first constraint, third constraint and so on, go into one group while the other 16 constraints go into a second group. The resulting weighting adjustment factors for each group of constraints are averaged together and applied to the poststratified weights (or the initial weights if poststratification was not done). Population/estimate differences at the DA level for the 32 constraints are usually reduced—but not eliminated—by using the first-step weights.

Finally, a **second-step** regression weighting adjustment factor is calculated at the WA level. The 32 auxiliary variables are applied at the WA level along with two auxiliary variables (number of households and number of persons) for each DA in the WA to determine the second-step weighting adjustment factors. These are applied to the first-step weights to generate the final weights. Population/estimate differences at the WA level for the 32 auxiliary variables are eliminated or reduced significantly using the final weights.

Constraints are discarded in the first and second steps because:

– they are **small** (they only apply to a few households in the population);
– they are **redundant** (also called linearly dependent [LD] constraints);
– they are **nearly redundant** (also called nearly linearly dependent [NLD] constraints); or
– they cause **outlier weights** (weights outside the range 1–25 inclusive) during the calculation of the weights.

For example, since the total number of females plus the total number of males equals the total number of persons, the total number of females can be dropped as a redundant or LD constraint since any two of the constraints being satisfied guarantees that the third will also be satisfied. If the “Marital status – widowed” constraint is dropped for being small (since there are very few widows in the WA), then the sum of the remaining marital status constraints (single, married, divorced, and separated) will nearly equal the total number of persons, suggesting that one constraint from this group of four could perhaps be dropped for being nearly redundant or NLD.

Initially, a check is done at the WA level for small, LD and NLD constraints, according to the following procedure:

(i) The size of a constraint is defined by the number of households in the population to which the constraint applies. A constraint whose size is SMALL or less (the SMALL parameter equalled 20, 30 or 40 households in 2001) is discarded since estimates, for small constraints, tend to be very unstable.

(ii) Next, LD constraints are discarded.

(iii) Following this, the condition number of the matrix being inverted to determine the weighting adjustment factors is lowered by discarding NLD constraints. The condition number (see Press et al., 1992) is the ratio of the largest eigenvalue to the smallest eigenvalue of the matrix being inverted. High condition numbers indicate near colinearity among the constraints, which could cause the estimates to be unstable. To lower the condition number, a forward-selection approach is used. The matrix is recalculated based only on the two largest constraints. If the condition number exceeds the COND parameter (which equalled 1,000, 2,000, 4,000, 8,000 or 16,000 in 2001, but always 1,000 in 1996), the second largest constraint is discarded. From here, the next largest constraint is added to the list of constraints being applied, the matrix is recalculated and its condition number determined. If the condition number increases by more than COND, the just-added constraint is discarded. This process continues until all constraints have been checked. If, after dropping these NLD constraints, the condition number exceeds the MAXC parameter (which equalled 10,000, 20,000, 40,000, 80,000 or 160,000 in 2001, but always 10,000 in 1996), additional constraints are dropped. Constraints are dropped in descending order, based on the amount by which they increased the condition number when they were initially included in the matrix. The condition number of the matrix is recalculated every time a constraint is dropped. When the condition number drops below MAXC, no more

constraints are dropped. It should be noted that in 2001, MAXC always equalled ten times the value of COND.

(iv) Any constraints dropped up to this point are not used in the weighting calculations.

Next, before calculating the first-step weighting adjustment factors for a DA, any remaining constraints which are small are dropped for that DA. Those that remain are partitioned into two groups, as was previously described. Then, for each group, any linearly dependent constraints are identified and dropped (constraints which are linearly dependent at the DA level may not be linearly dependent at the WA level). The first-step weighting adjustment factors are then calculated for the remaining constraints in each group. If any of the first-step adjusted weights fall outside the range 1–25 inclusive, additional constraints are dropped. A method similar to that used to discard NLD constraints is applied here except that a constraint is discarded if it causes outlier weights. In the interest of computational efficiency, the bisection method is used to identify which constraints should be dropped.

Next, the second-step weighting adjustment factors are calculated based on the constraints that were not discarded for being small, linearly dependent or nearly linearly dependent during the initial analysis of the matrix being inverted. If any of the second-step adjusted weights fall outside the range 1–25 inclusive, then additional constraints are dropped using the method outlined for the first-step adjustment.

The census weights are calculated independently in each WA. This makes it possible to use a different set of weighting system parameters for each WA (e.g. poststratify or not, SMALL, COND, MAXC, range of weights allowed). In 1996, an identical set of parameters was used for each WA in the country. In 2001, with the increased processing power achieved through running the weighting system on multiple personal computers (PCs), it was decided to calculate the weights for each WA with ten different sets of parameters. In each case, a statistic was calculated to determine which set of parameters minimized the differences between the population counts and the sample estimates for the constraints. The weights arrived at with this set of parameters were used for the corresponding WA. In order to retain certain important constraints, two WAs were weighted using ‘customized’ parameters that were unlike any of the other ten sets. This process of selecting the best weights on a WA-by-WA basis was called “cherry-picking” the parameters.

For more details on regression estimators see Bankier (2002) and Fuller (2002).

GREG weights are calculated only for sampled-EA private households which received the long census questionnaire (one-fifth of private dwellings were sampled; four-fifths were not). Sampled-EA private households which received a short questionnaire receive a weight of 0. All non-sampled EA private households receive a weight of 1 since 100% of the respondents in these areas provide information on the Form 2B. Collective households also receive a weight of 1. **In this report, the term “household” will refer to a private household unless otherwise specified**.

4.5 Two-pass Processing

For the 1996 and 2001 censuses, short-form (Form 2A) write-in responses to the relationship variables were not captured due to budgetary constraints. Instead, they were coded under the generic value ‘Other’. Long-form (Form 2B) write-in responses to the relationship variables were still captured and coded in the normal fashion.

During two-pass processing, the long-form data are processed in two stages. In the first stage—Pass 1—the long and short forms are processed together, representing 100% of the data. The captured long-form write-in responses for relationship are ignored and assigned the generic value ‘Other’ to coincide with the short-form write-in responses. Editing and imputation is performed the same way for both the long and short forms. In the second stage—Pass 2—only the long forms are processed; the short forms are not available during imputation. The captured long-form write-in responses for relationship are used rather

than the ‘Other’ responses. Because of the availability of the write-in responses, the quality of the results is assumed to be higher in Pass 2 than in Pass 1.

The weighting system uses the Pass 1 results for all households to calculate the household weights. While it might be possible to use the Pass 1 results for the short forms and Pass 2 results for the long forms, this method could bias the census estimates. This is because of differences in the distribution of the responses for the demographic variables between Pass 1 and Pass 2 as a result of the write-in responses for relationship being present in Pass 2. Published census estimates were produced using Pass 1 weights applied to Pass 2 long-form imputed results. The difference between the population counts (based on Pass 1 results) and Pass 2 estimates was small for most constraints. See Table 7.2.2.2 and Chart 7.2.2.3 in Section 7.2.2 for a comparison of Pass 1 and Pass 2 results.

5. The Sampling and Weighting Evaluation Program

The sampling and weighting evaluation program was designed to determine the effect of sampling and weighting on the quality of census sample data. Four studies in all were carried out to help measure the quality of the census sample data and estimates, and to provide information for the planning of future censuses. These studies involved:

(a) an examination of sampling bias;

(b) an evaluation of weighting procedures;

(c) an evaluation of sample estimate to population count consistency;

(d) a sampling variance evaluation for various 20% sample characteristics.

Each of these studies is described briefly below, with their results being presented in chapters 6 through 9.

Three factors explain why the counts provided in the following chapters do not exactly match the published counts. In the first place, only households subject to sampling were included in these studies. Secondly, Pass 1 rather than Pass 2 data were used (see Section 4.5) and, thirdly, no correction was made for “random additions” (see Section 2.7).

5.1 Sampling Bias

This study identified the characteristics which displayed large discrepancies between estimates based on initial weights and known population counts. These discrepancies are of interest for two reasons: first, their possible usefulness in identifying biases in the census household sample selected in the field; and second, their potential for showing the impact of non-response on census sample questions (long forms with no responses to sample questions are converted to short forms during census processing). These short-form biases caution against possible biases in long-form estimates. Biases in short-form characteristics are corrected through calibration. If long-form characteristics are correlated with short-form characteristics, their biases should also be reduced through calibration.

5.2 Evaluation of Weighting Procedures

The objective of this study was to evaluate the performance of the General Regression Estimator. This was done by examining the level of agreement between sample estimates and population counts for all the WA constraints for all of Canada, by trying to explain any inconsistencies through assessment of the number and type of constraints discarded at the WA level and of the reasons for their being discarded, and by taking a look at the distribution of census weights.

5.3 Sample Estimate and Population Count Consistency

This study examined the level of agreement between sample estimates and population counts for the basic characteristics used as constraints. This was done for various geographic areas.

5.4 Sampling Variance

The standard error (the square root of the variance) of an estimate is a measure of its precision. Estimates of standard errors for estimators using simple weights of 5 and assuming simple random sampling are relatively quick to calculate. However, estimates of standard errors for census estimators

taking into account the sample design and estimation techniques used are time consuming to calculate. Adjustment factors were calculated which represent the ratios of the estimates of the standard errors for census estimates to the simple estimates of the standard errors. An estimate of the standard error of a census estimate for any characteristic in any geographic area can then be obtained by multiplying the simple estimate of the standard error by the appropriate adjustment factor.

6. Sampling Bias

In this chapter, we will assess whether, following adjustments for non-response, the census sample is biased. This can be done by calculating the Z statistic

$$Z^{(0)} = \frac{\hat{X}^{(0)} - X}{\sqrt{V(\hat{X}^{(0)})}}$$

for short form characteristics such as Marital status – Single where the census population count X can be compared to the sample $\hat{X}^{(0)}$ based on initial weights. In the Z statistic, the difference between the estimate and the population count is divided by the square root of the variance of the estimate. If the sampling process is random, it can be shown that $Z^{(0)}$ will follow approximately a normal distribution with mean 0 and variance 1 (see Appendix C).

Table 6.1 and Chart 6.1 present Z statistics at the Canada level for 1996 and 2001 (along with the differences $\hat{X}^{(0)} - X$) for 32 characteristics closely resembling the constraints which were applied in generating the final census weights (see Appendix B). If $Z^{(0)}$ follows a normal distribution, the probability that $|Z^{(0)}| > 3$ is approximately 0.0026 for one characteristic. This suggests that, on average, $|Z^{(0)}| > 3$ for 0.0026 x 32 = 0.0832 of the 32 characteristics in Table 6.1. However, for the 2001 Census alone, 25 of the 32 characteristics have a Z statistic outside the range of –3 to 3. This provides strong evidence that the 2001 Census sample is biased. The large positive Z statistics for total number of persons, females, females ≥ 15 years, persons aged 5 to 14, persons aged 55+, married persons, 2-person households and 4-person households indicate that these characteristics are over-represented in the sample. The large negative Z statistics for males ≥ 15 years, persons aged 20 to 34, single persons, separated persons and 1-person households indicate that these characteristics are under-represented in the sample. Table 6.1 and Chart 6.1 also show that the absolute value of the Z statistic is often much larger in 2001 than in 1996.

Bias can originate from a variety of sources, including census representative errors (e.g., not selecting the sample according to specifications), non-response bias (e.g., young adult males are less likely to complete a long questionnaire than a short questionnaire), response bias (e.g., respondents answering differently on Form 2B than on Form 2A), processing errors, and so on. In terms of non-response bias, 1.3% of the households (both sampled and non-sampled) did not respond in 2001 (either because they refused or could not be contacted) compared to 0.8% in 1996. Such households are referred to as missed/refusal households. Furthermore, 0.7% of the sampled households in 2001 provided some responses to basic questions but didn't provide answers to the questions asked on a sample basis. This compares to 0.2% of the sampled households in 1996. During data processing, sampled households where there was complete non-response, either to all questions or to just the sampled ones, were converted from Form 2B to Form 2A households. As a result, they became non-sampled households and only the responses to the basic questions were imputed if required. This procedure of converting sampled households to non-sampled households is known as 2A/2B document conversion. It is possible that the missed/refusal households and those without sample question responses had different characteristics from other households. Converting Forms 2B to Forms 2A in this way could bias the sample. For example, it is known that the percentage of single-detached dwellings that are missed/refusal households is half what it is for the population as a whole.

Chart 6.1 shows that for many characteristics the Z statistic is larger in 2001 than in 1996. Z being a random variable, some of these differences may not be statistically significant. The 12 characteristics having statistically significant Z statistic differences are flagged with asterisks in Chart 6.1. They were identified by a W statistic, which is defined in Appendix C.

The geographic variation of the bias was also studied. The Z statistics for all 32 characteristics were calculated for the East, Quebec, Ontario and the West (including the three territories) regions in the same fashion as at the Canada level. The relative bias between these four regions is displayed for the 2001 and 1996 censuses in Chart 6.2 and 6.3 respectively. Again using the W statistic, regional differences which are statistically significant are flagged by placing the initials of the regions at the bottom of a chart. For example, QO QW indicates that there is a significant difference in the bias between Quebec and Ontario as well as between Quebec and the West.

Chart 6.2 shows that, for 2001, the only regions to exhibit a difference in the bias are Quebec-Ontario and Quebec-West. It is interesting to note that this holds for seven of the characteristics. The majority of the age characteristics show no differences between the regions, but the most noticeable of any is an over-representation of ages 15 to 19 in Quebec compared to an under-representation in Ontario and the West. There are more regional differences in the non-age characteristics, with the majority being present in the person characteristics. With the exception of 3-person households, which show a Quebec-Ontario difference, the household characteristics tend to agree across the regions.

If the 2001 Census regional biases are compared to those of the 1996 Census (see charts 6.2 and 6.3), some patterns remain the same between them (i.e. males, males >15 years, females >15 years, single persons, married persons).

Section 7.2.2 and Chapter 8 will show that these population/estimate differences are often significantly reduced by calibration of the census weights. As a result, the inferences based on calibrated estimates should be more accurate.

Table 6.1: Population/Estimate Differences Based on Initial Weights, 2001 and 1996 Censuses

	2001 Census						1996 Census					
Characteristic	**Count**	**Estimate[1]**	**Difference[2]**	**Disc.[3]**	**S.E.[4]**	**Z statistic[5]**	**Count**	**Estimate[1]**	**Difference[2]**	**Disc.[3]**	**S.E.[4]**	**Z statistic[5]**
Males	14,171,941	14,146,867	-25,074	-0.18	6,139	-4.08	13,717,654	13,694,786	-22,868	-0.17	5,752	-3.98
Females	14,699,518	14,772,915	73,397	0.50	5,940	12.36	14,176,680	14,222,665	45,985	0.32	5,552	8.28
Total	28,871,459	28,919,783	48,324	0.17	8,991	5.37	27,894,334	27,917,451	23,117	0.08	8,227	2.81
Males ≥ 15	11,340,286	11,295,995	-44,291	-0.39	4,747	-9.33	10,781,073	10,732,804	-48,269	-0.45	4,449	-10.85
Females ≥ 15	11,998,509	12,042,929	44,420	0.37	4,342	10.23	11,383,130	11,402,113	18,983	0.17	4,006	4.74
Age 0-4	1,636,092	1,641,720	5,628	0.34	2,986	1.88	1,858,332	1,874,111	15,779	0.85	3,073	5.14
Age 5-9	1,910,359	1,928,604	18,245	0.96	3,213	5.68	1,932,023	1,950,728	18,705	0.97	3,120	6.00
Age 10-14	1,986,213	2,010,534	24,321	1.22	3,271	7.44	1,939,776	1,957,694	17,918	0.92	3,125	5.73
Age 15-19	1,986,163	1,983,519	-2,644	-0.13	3,245	-0.81	1,903,023	1,907,732	4,709	0.25	3,074	1.53
Age 20-24	1,892,572	1,851,491	-41,081	-2.17	3,168	-12.97	1,840,654	1,816,301	-24,353	-1.32	3,013	-8.08
Age 25-29	1,835,744	1,810,124	-25,620	-1.40	3,077	-8.33	1,971,123	1,953,292	-17,831	-0.90	3,053	-5.84
Age 30-34	2,031,513	2,013,625	-17,888	-0.88	3,173	-5.64	2,405,559	2,401,580	-3,979	-0.17	3,317	-1.20
Age 35-39	2,452,299	2,446,624	-5,675	-0.23	3,427	-1.66	2,486,060	2,482,136	-3,924	-0.16	3,339	-1.18
Age 40-44	2,510,847	2,513,920	3,073	0.12	3,439	0.89	2,268,423	2,273,674	5,251	0.23	3,177	1.65
Age 45-49	2,273,676	2,283,700	10,024	0.44	3,286	3.05	2,050,229	2,059,233	9,004	0.44	3,040	2.96
Age 50-54	2,031,050	2,041,054	10,004	0.49	3,137	3.19	1,581,484	1,589,751	8,267	0.52	2,707	3.05
Age 55-59	1,549,675	1,567,071	17,396	1.12	2,758	6.31	1,271,221	1,269,086	-2,135	-0.17	2,448	-0.87
Age 60-64	1,234,930	1,249,389	14,459	1.17	2,469	5.86	1,157,926	1,160,459	2,533	0.22	2,338	1.08
Age 65-74	2,059,079	2,083,362	24,283	1.18	3,256	7.46	1,991,721	1,996,303	4,582	0.23	3,068	1.49
Age 75 and over	1,481,247	1,495,045	13,798	0.93	2,676	5.16	1,236,780	1,225,372	-11,408	-0.92	2,332	-4.89

Characteristic	2001 Census						1996 Census					
	Count	Estimate[1]	Difference[2]	Disc.[3]	S.E.[4]	Z statistic[5]	Count	Estimate[1]	Difference[2]	Disc.[3]	S.E.[4]	Z statistic[5]
Single	13,282,845	13,196,174	-86,671	-0.65	8,018	-10.81	12,779,218	12,741,878	-37,340	-0.29	7,320	-5.10
Married	11,750,092	11,906,204	156,112	1.33	6,678	23.38	11,537,475	11,628,813	91,338	0.79	6,076	15.03
Widowed	1,341,497	1,339,109	-2,388	-0.18	2,254	-1.06	1,303,304	1,291,501	-11,803	-0.91	2,130	-5.54
Divorced	1,794,079	1,784,704	-9,375	-0.52	2,824	-3.32	1,605,136	1,591,530	-13,606	-0.85	2,612	-5.21
Separated	702,946	693,591	-9,355	-1.33	1,749	-5.35	669,201	663,729	-5,472	-0.82	1,675	-3.27
Com.-law = yes	2,267,634	2,253,253	-14,381	-0.63	4,090	-3.52	1,770,338	1,768,774	-1,564	-0.09	3,568	-0.44
1-person hhlds	2,908,857	2,866,182	-42,675	-1.47	2,847	-14.99	2,584,348	2,558,041	-26,307	-1.02	2,524	-10.42
2-person hhlds	3,709,282	3,739,781	30,499	0.82	3,224	9.46	3,385,597	3,397,657	12,060	0.36	3,011	4.00
3-person hhlds	1,848,476	1,845,071	-3,405	-0.18	2,541	-1.34	1,804,304	1,809,076	4,772	0.26	2,435	1.96
4-person hhlds	1,812,783	1,826,921	14,138	0.78	2,481	5.7	1,813,493	1,825,159	11,666	0.64	2,378	4.91
5-person hhlds	714,618	719,013	4,395	0.61	1,664	2.64	737,751	740,921	3,170	0.43	1,640	1.93
6+-person hhlds	332,959	328,968	-3,991	-1.20	1,155	-3.46	334,207	327,786	-6,421	-1.92	1,124	-5.71

[1] Based on initial weights

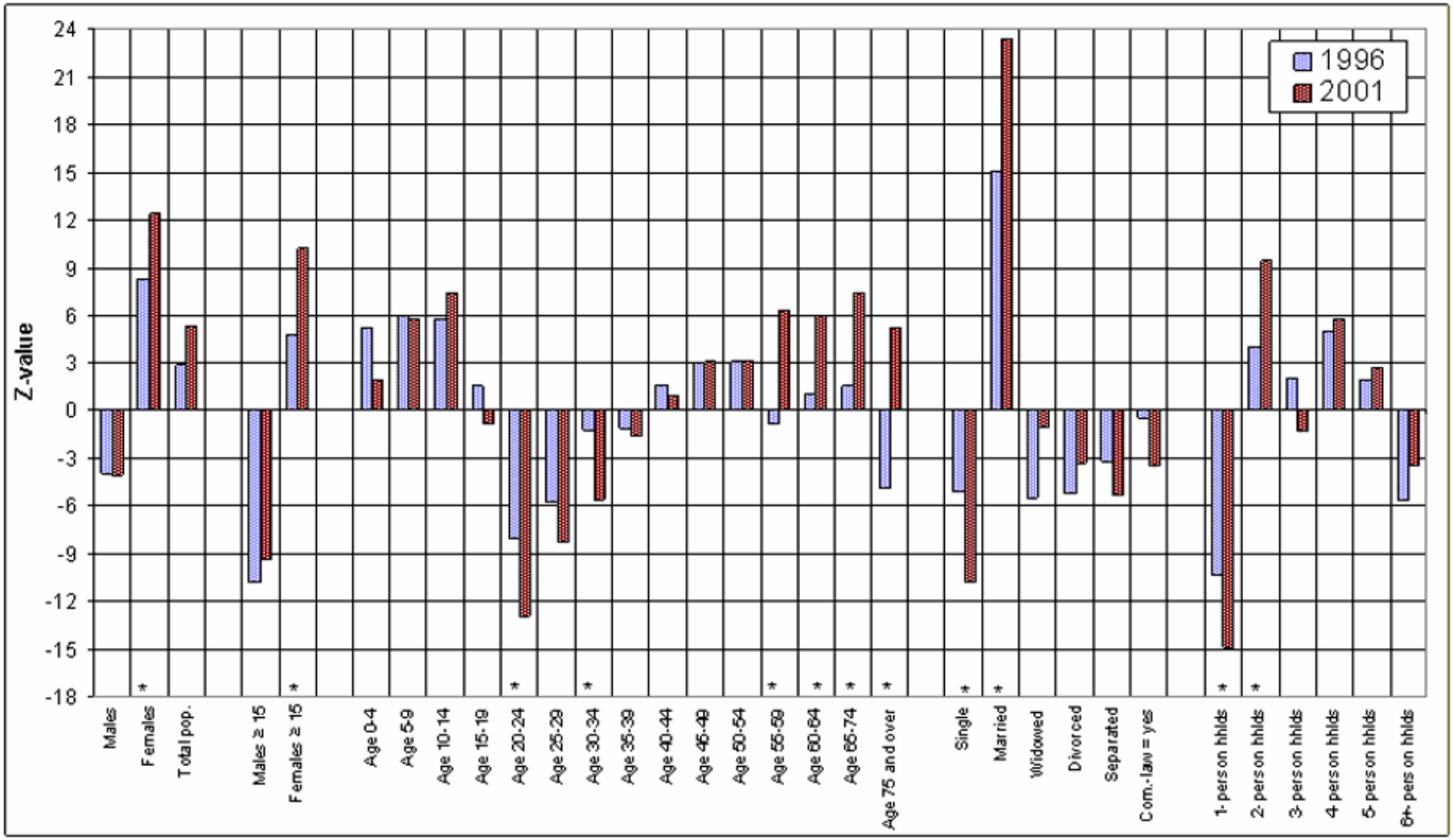
[2] Difference: estimate-count

[3] Disc.: discrepancy (100*[estimate-count]/count)

[4] S.E.: standard error of the initial weight estimate

[5] Z statistic: (estimate-count)/S.E.

Chart 6.1: Z Statistics for Population/Estimate Differences Based on Initial Weights, for Canada, 2001 and 1996 Censuses



* indicates a significant difference in the bias between the two censuses

Chart 6.2: Regional Z statistics in 2001

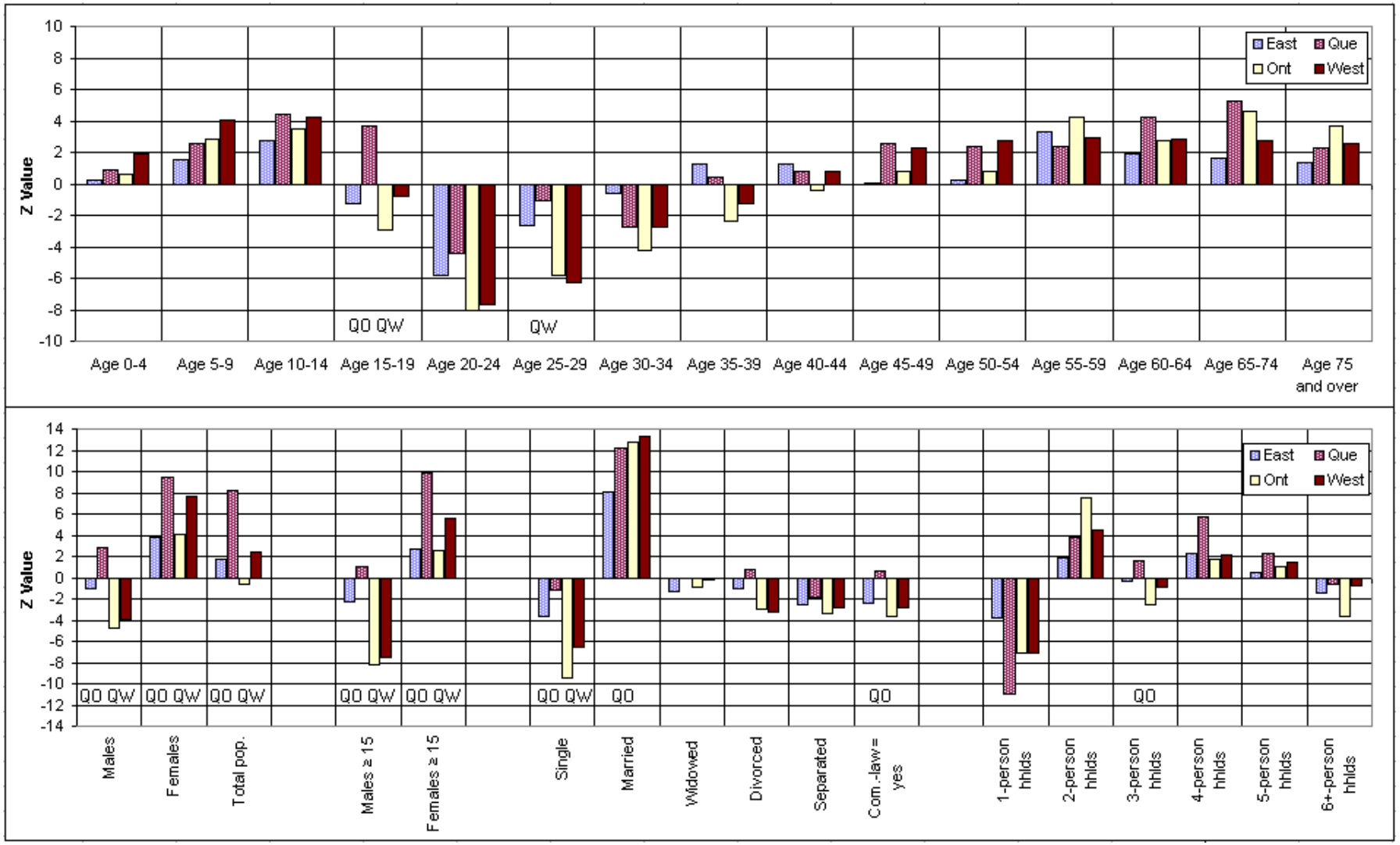
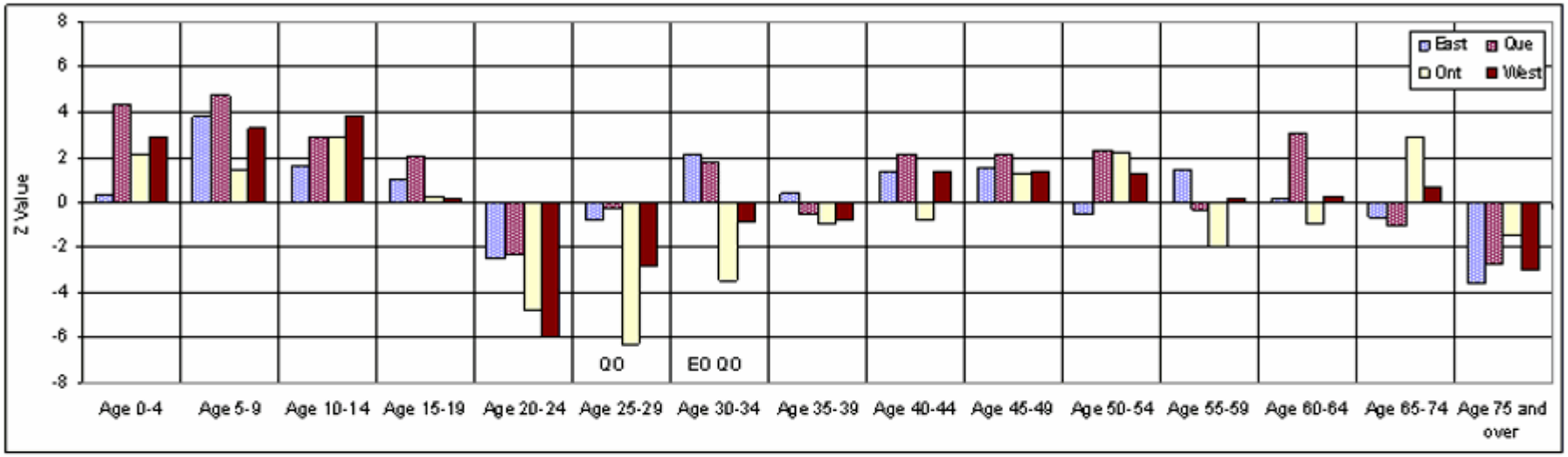
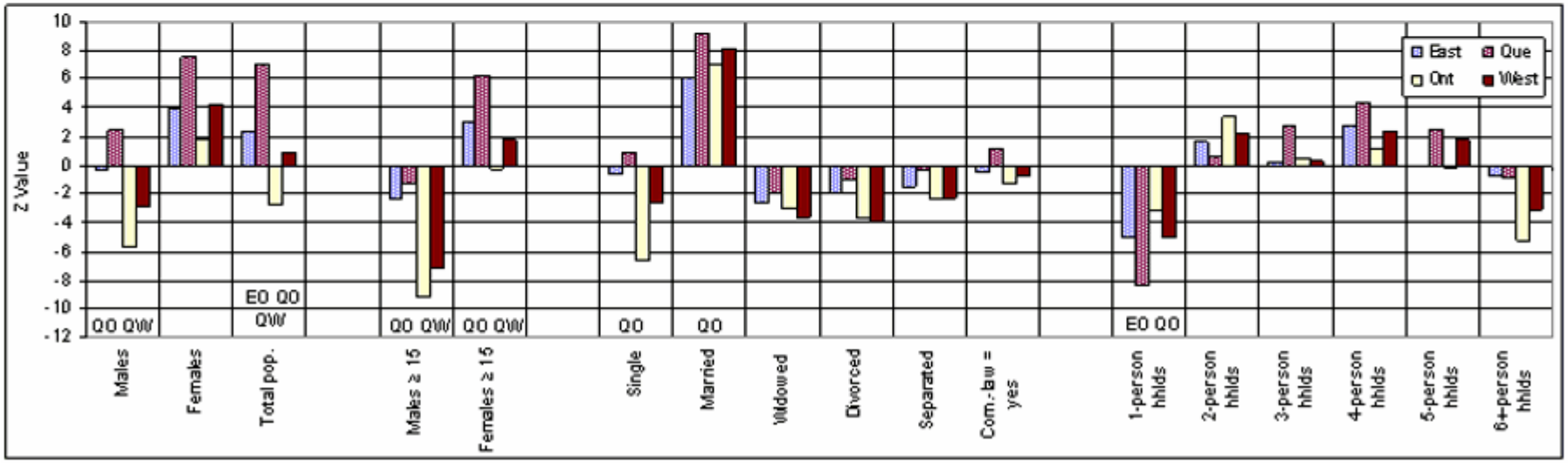


Chart 6.3: Regional Z statistics in 1996





7. Evaluation of Weighting Procedures

This chapter presents and evaluates certain aspects pertaining to census weighting procedures, such as weighting area formation and the size distribution of the weights. It also examines, for various characteristics, the discrepancies between population counts and sample estimates at the Canada level. Finally, it takes a look at the frequency at which constraints are discarded and the effect this has on these discrepancies.

7.1 Weighting Area (WA) Formation

In 2001, the country was partitioned into 6,148 WAs containing, on average, approximately eight whole DAs. The weighting program attempts to achieve agreement between certain sample estimates and the corresponding population counts for each WA. A WA was formed by grouping together DAs to adhere to the following conditions:

(a) A WA must respect the boundaries of census divisions (CDs).

(b) A WA should contain a population of between 1,000 and 3,000 households.

(c) A WA should, where possible, respect (in order of priority) census subdivision (CSD) boundaries, census tract (CT) boundaries and lastly federal electoral district (FED) boundaries.

(d) A WA should, where possible, be made up of contiguous DAs (i.e. not be in two or more parts or contain any “holes”) and it should be as compact as possible.

Table 7.1.1 below shows that 5,784 (94.2%) of the 2001 WAs are within the desired range of 1,000 to 3,000 households. A slightly larger percentage of WAs were within this range in 1996. The average number of dwellings per WA was 2,047. There were several WAs with a larger than average dwelling count, the largest having 17,043 dwellings. In 2001, there were seven WAs with zero population that are not included in Table 7.1.1. Table 7.1.1 also excludes those WAs where all the DAs were not subject to sampling. These include, for example, all the WAs in the Northwest Territories and Nunavut.

Agreement between sample estimates and population counts is ensured only for geographic areas which are made up of whole WAs. Table 7.1.2 looks at the relationship between 2001 Census CSD and CT boundaries and WA boundaries. For a given CSD, for example, the category ‘Geographic areas containing only part of one WA while the rest of the WA contains only complete geographic areas of the same kind’ indicates that the CSD is located entirely in one WA (i.e. it is not spread across two WAs), and that the WA contains only whole CSDs. These CSDs can represent a village or small town. The category ‘Geographic areas containing only part of one WA while the rest of the WA does not contain only complete geographic areas of the same kind’ is similar to the previous one except that the WA does not contain only entire CSDs (i.e. at least one CSD in the WA is spread between two or more WAs). A CSD belonging to the group ‘Geographic areas containing one or more whole WAs’ is a CSD (often a larger town or city) which covers one or more whole WAs, and for which each WA includes only one CSD or a portion of only one CSD. If the CSD falls in the group ‘Geographic areas that cross at least one WA boundary,’ it is spread between two or more WAs. The four groups of areas presented here are mutually exclusive and leave no areas unaccounted for. These definitions also apply to CTs.

According to the figures presented in Table 7.1.2, 12.8% of CSDs and 65.4% of CTs are made up of one or more whole WAs. It is here that the closest agreement between population counts and sample estimates is most likely to occur.

For more information about weighting areas and their delineation, see Kruszynski (1999).

Table 7.1.1: Size Distribution of Weighting Areas

	2001 Census		1996 Census	
Dwellings	**WA Count**	**Percentage**	**WA Count**	**Percentage**
1 - 999	1	0.0	4	0.1
1,000 - 1,499	1,132	18.4	1,686	28.4
1,500 - 1,999	2,248	36.6	2,213	37.2
2,000 - 2,499	1,622	26.4	1,417	23.9
2,500 - 3,000	786	12.8	560	9.4
3,001+	352	5.8	61	1.0
Total	6,141	100.0	5,941	100.0

Table 7.1.2: Number of CSDs and CTs that Respect WA Boundaries, 2001 Census

Description	**CSD**		**CT**	
	Number	**%**	**Number**	**%**
Geographic areas containing only part of one WA while the rest of the WA contains only complete geographic areas of the same kind	4,165	74.4	1,563	30.8
Geographic areas containing only part of one WA while the rest of the WA does not contain only complete geographic areas of the same kind	567	10.1	106	2.1
Geographic areas containing one or more whole WAs	717	12.8	3,313	65.4
Geographic areas that cross at least one WA boundary	151	2.7	87	1.7
Total	**5,600**		**5,069**	

7.2 Evaluation of the Census Weighting Methodology

7.2.1 Distribution of Weights

Chart 7.2.1.1 compares the 2001 final weight distribution to that of 1996. The distributions are very similar, however weights < 1 were not allowed in 2001. For 1996, the chart shows a higher percentage of households with smaller weights (< 2.99, including 0.7% with weights < 1) while in 2001, there is a higher percentage of households with weights in the range 3.00-5.99. There are only minor differences in the distribution of weights > 6.00.

Charts 7.2.1.2 to 7.2.1.4 compare the distributions of the 2001 Census initial weights, poststratified weights, first-step weights and final weights. The initial weights are tightly clustered around 5 as a result of a one-in-five sample of households being selected. The poststratified, first-step and final weight distributions become progressively more spread out as the constraints become more restrictive.

Chart 7.2.1.1: Comparison of 2001 and 1996 Final Household Weights

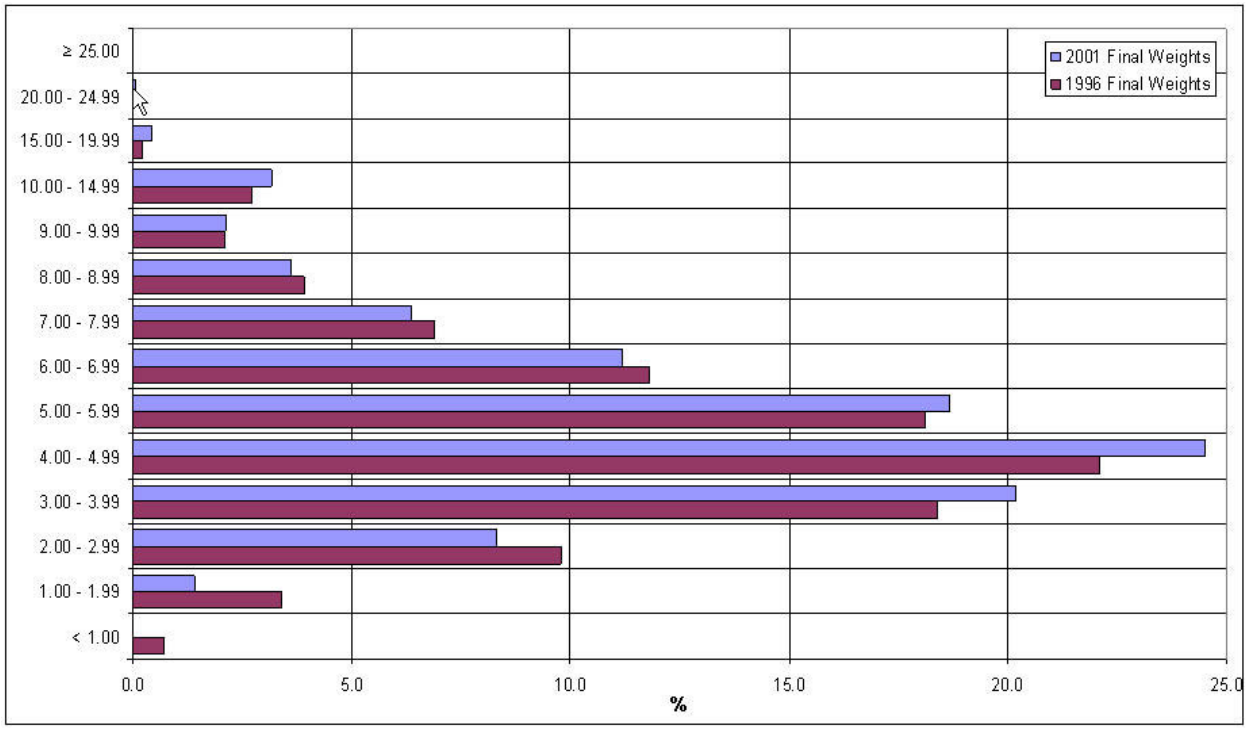


Chart 7.2.1.2: Comparison of 2001 Census Initial Weights and Poststratified Weights

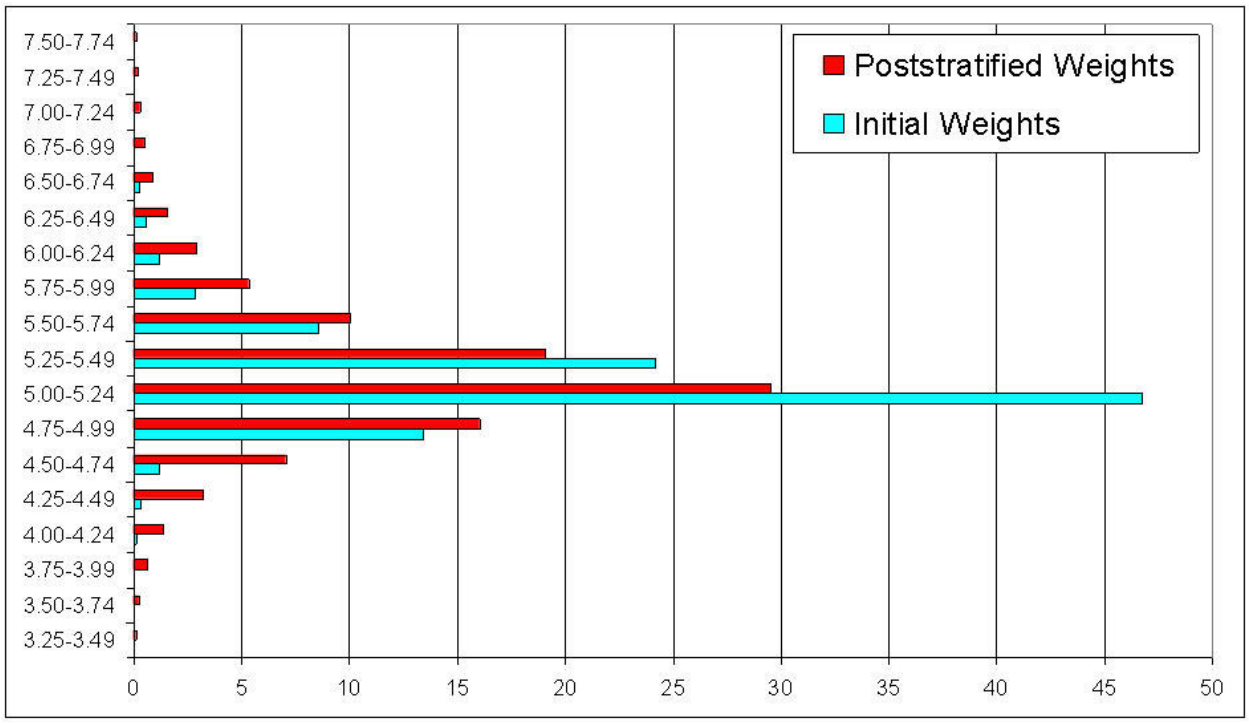


Chart 7.2.1.3: Comparison of 2001 Census Poststratified Weights and First-step Weights

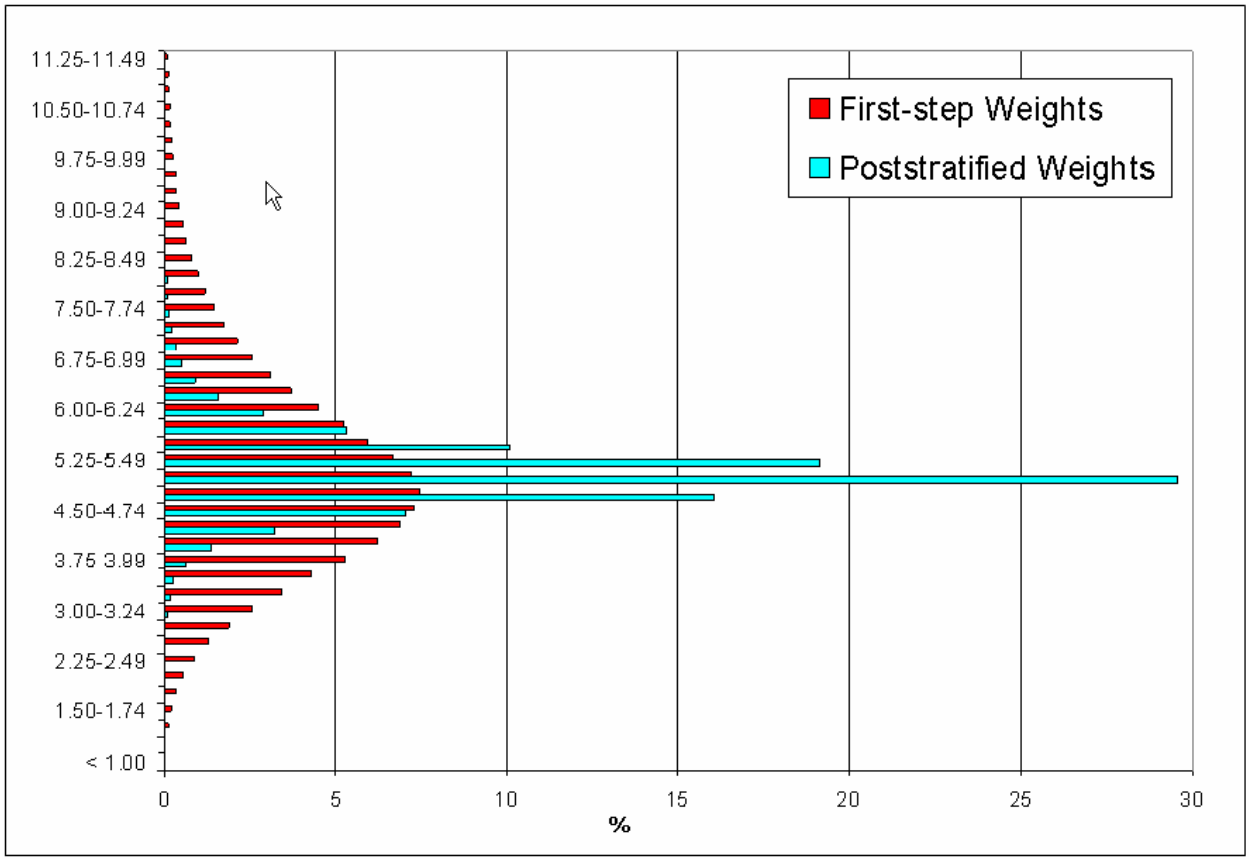
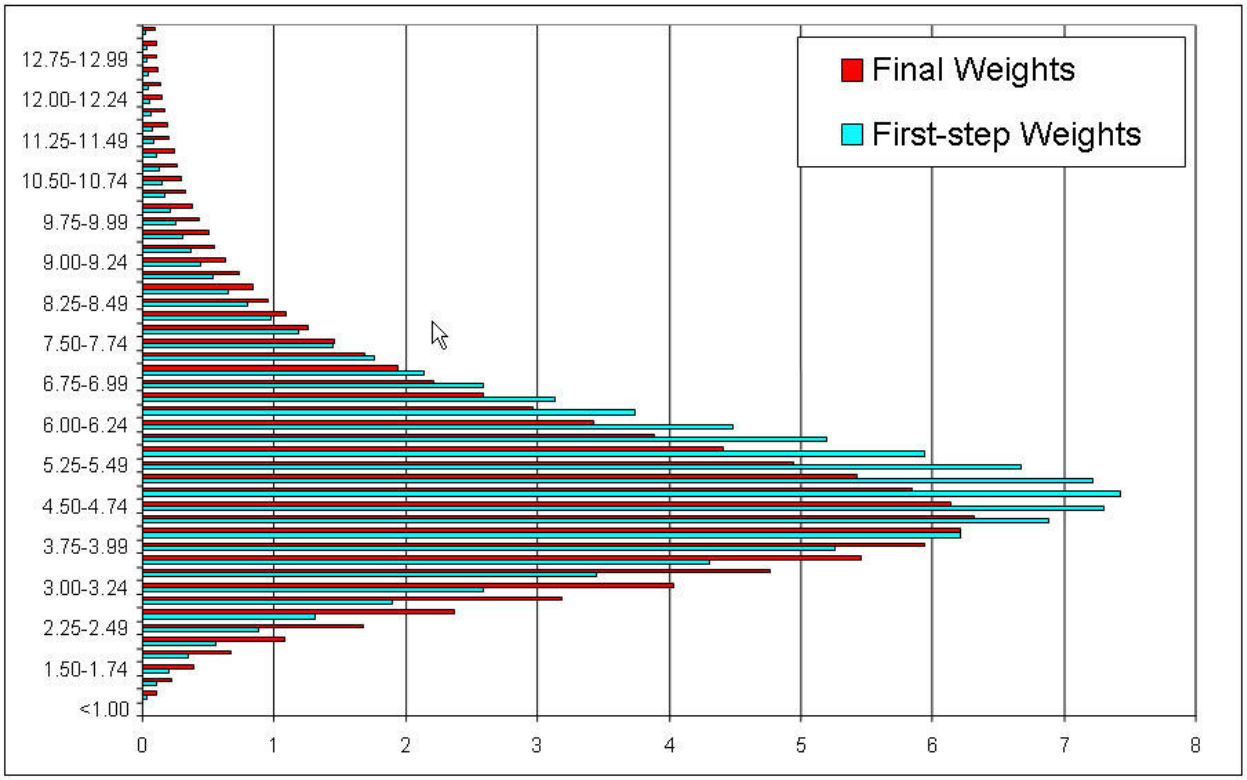


Chart 7.2.1.4: Comparison of 2001 Census First-step Weights and Final Weights



7.2.2 Discrepancies Between Population Counts and Sample Estimates

As discussed in Section 4.4, the final weights are chosen so as to reduce or eliminate discrepancies between the population counts and the corresponding sample estimates for 32 constraints at the WA level (see Appendix B). Some discrepancies remain, however, since constraints are sometimes discarded (see Sections 4.4 and 7.2.3). The population/estimate discrepancy is defined as

$$\textbf{population/estimate discrepancy} = \frac{\text{sample estimate - population count}}{\text{population count}} \times 100$$

The numerator in the above expression (sample estimate - population count) is referred to as the **"population/estimate difference."** The sample estimates and population counts are based on occupied dwellings from sampled EAs.

Table 7.2.2.1 and charts 7.2.2.1 and 7.2.2.2 show the 2001 and 1996 Canada-level population/estimate differences and discrepancies for the 32 WA-level constraints and for the initial and/or final weights. Because Chart 7.2.2.1 is similar to Chart 6.1, except for showing population/estimate discrepancies, rather than Z statistics, based on initial weights, and given that further explanations can be found in Chapter 6, this chart will not be examined in any detail. Overall, what it shows is that the discrepancies are generally much larger for 2001 than for 1996. Table 7.2.2.1 shows that, compared to 1996, the absolute value of the 2001 population/estimate discrepancies based on final weights are generally smaller for five-year age ranges and for most responses for marital status. For “Common-law status = yes” and some household sizes, the opposite tends to be true. Variations in the size of discrepancies between censuses usually result from a change in the number of constraints which were dropped, as will be discussed in Section 7.2.3. In comparing charts 7.2.2.1 and 7.2.2.2, it can be seen

that the 2001 population/estimate discrepancies based on final weights are dramatically smaller than those based on initial weights, with the exception of 5-person households. As discussed in Section 7.2.3, this is probably the result of this constraint being discarded frequently for causing outlier weights and, to a lesser extent, for being nearly linearly dependent.

Table 7.2.2.2 and Chart 7.2.2.3 show the 2001 population/estimate differences and discrepancies based on final weights for the 32 WA-level constraints, represented for Pass 1 and Pass 2 results, for Canada. We observe that Pass 1 discrepancies are smaller due to the fact that the census weights were calculated based on Pass 1 results. See Section 4.5 for further information on two-pass processing.

Table 7.2.2.1: Comparison of 1996 and 2001 Population/Estimate Discrepancies for Canada

Characteristic	2001 Census			1996 Census		
	Initial Weights	Final Weights		Initial Weights	Final Weights	
	Difference	Difference	Discrepancy	Difference	Difference	Discrepancy
Males	-25,074	-	0	-22,868	15	0.00
Males ≥ 15	-44,291	51	0	-48,269	-276	0.00
Persons ≥ 15	130	-	0	-29,285	3	0.00
Total households	-1,040	-	0	1,060	-	0.00
Total population	48,324	-	0	23,117	-	0.00
Age 0-4	5,628	559	0.03	15,779	-208	-0.01
Age 5-9	18,245	-792	-0.04	18,705	-258	-0.01
Age 10-14	24,321	234	0.01	17,918	462	0.02
Age 15-19	-2,644	779	0.04	4,709	1,853	0.10
Age 20-24	-41,081	-504	-0.03	-24,353	803	0.04
Age 25-29	-25,620	-785	-0.04	-17,381	105	0.01
Age 30-34	-17,888	7	0.00	-3,979	361	0.02
Age 35-39	-5,675	-556	-0.02	-3,924	320	0.01
Age 40-44	3,073	100	0.00	5,251	366	0.02
Age 45-49	10,024	687	0.03	9,004	971	0.05
Age 50-54	10,004	-87	0.00	8,267	993	0.06
Age 55-59	17,396	81	0.01	-2,135	254	0.02
Age 60-64	14,459	933	0.08	2,533	3,847	0.33
Age 65-74	24,283	271	0.01	4,582	-662	-0.03
Age 75 and over	13,798	-926	-0.06	-11,408	-9,207	-0.74
Single	-86,671	-53	0.00	-37,340	115	0.00
Married	156,112	-57	0.00	91,338	73	0.00
Widowed	-2,388	557	0.04	-11,803	-1,387	-0.11
Divorced	-9,375	206	0.01	-13,606	1,209	0.08

Characteristic	2001 Census			1996 Census		
	Initial Weights	Final Weights		Initial Weights	Final Weights	
	Difference	Difference	Discrepancy	Difference	Difference	Discrepancy
Separated	-9,355	-653	-0.09	-5,472	-10	0.00
Com.-law = yes	-14,381	4,115	0.18	-1,404	2,415	0.14
1-person hhlds	-42,675	-4,175	-0.14	-	-4,750	-0.18
2-person hhlds	30,499	-906	-0.02	12,060	-1,666	-0.05
3-person hhlds	-3,405	-5,010	-0.27	4,772	871	0.05
4-person hhlds	14,138	2,414	0.13	11,666	1,694	0.09
5-person hhlds	4,395	8,818	1.23	3,170	5,576	0.76
6+-person hhlds	-3,991	-1,142	-0.34	-	-1,725	-0.52

Chart 7.2.2.1: 1996 and 2001 Population/Estimate Discrepancies Based on Initial Weights

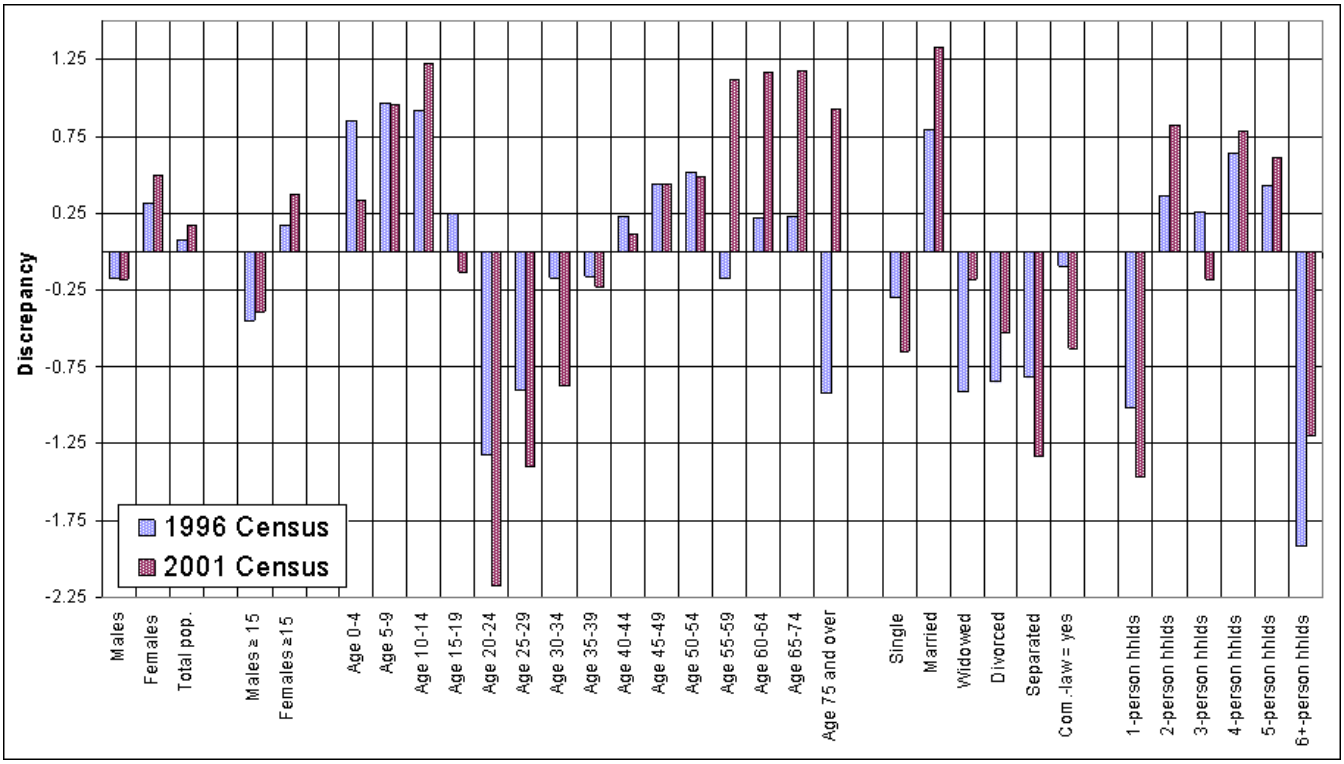


Chart 7.2.2.2: 1996 and 2001 Population/Estimate Discrepancies Based on Final Weights

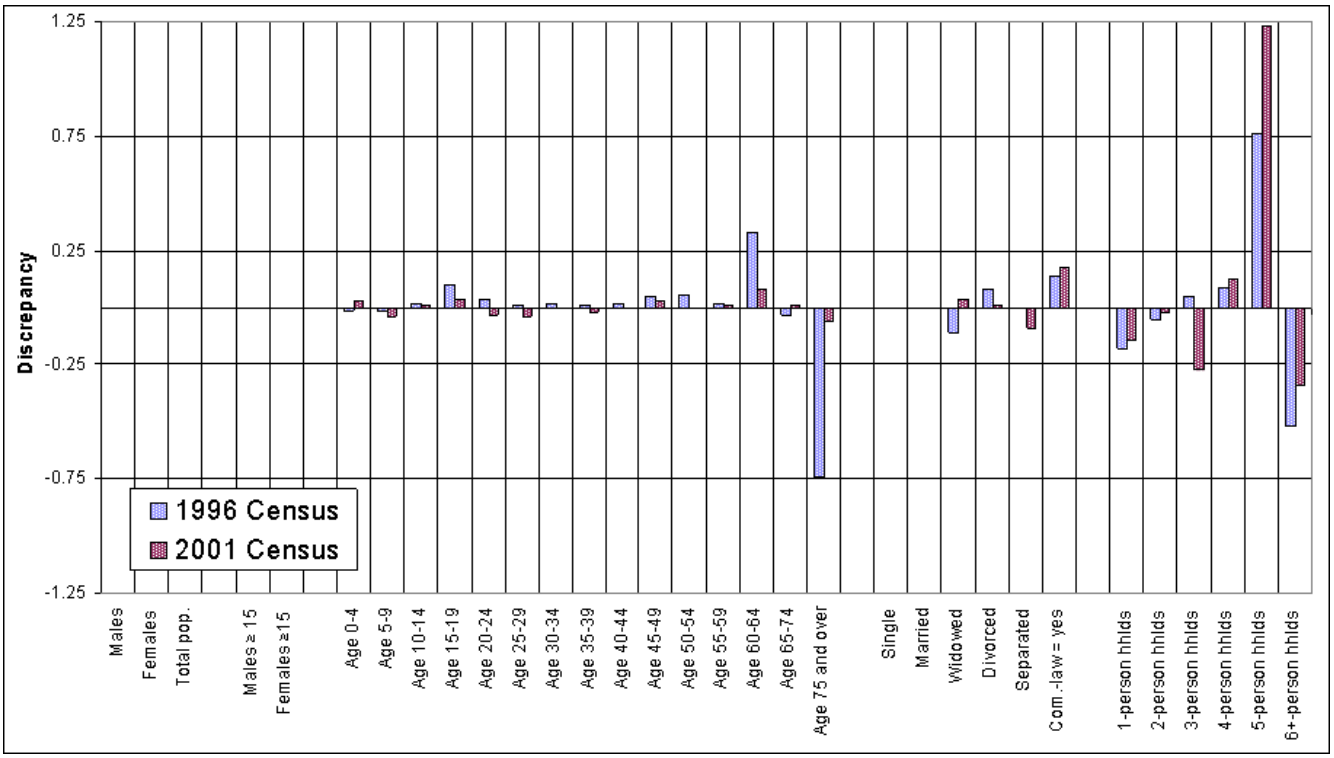


Table 7.2.2.2: Comparison of Pass 1 and Pass 2 Population/Estimate Discrepancies Based on Final Weights, for Canada, 2001 Census

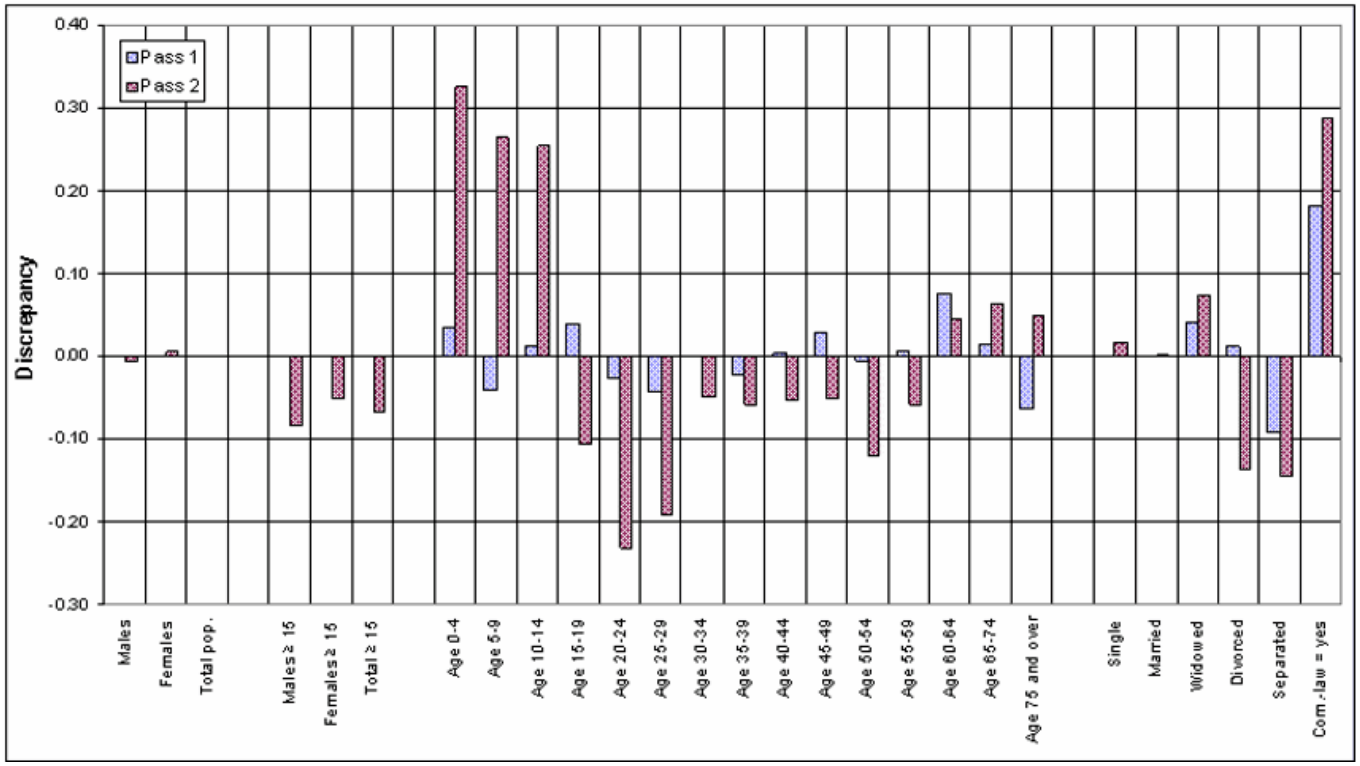
Characteristic	2001 Census – Pass 1				2001 Census – Pass 2				Pass 2 – Pass 1	
	Count	Estimate	Difference	Disc.	Count	Estimate	Difference	Disc.	Difference	Disc.
Males	14,171,941	14,171,941	0	0.00	14,393,344	14,392,459	-885	-0.01	-885	-0.01
Females	14,699,518	14,699,518	0	0.00	14,911,511	14,912,396	885	0.01	885	0.01
Total	28,871,459	28,871,459	0	0.00	29,304,855	29,304,855	0	0.00	0	0.00
Males ≥ 15	11,340,286	11,340,337	51	0.00	11,487,144	11,477,463	-9,681	-0.08	-9,732	-0.08
Females ≥ 15	11,998,509	11,998,458	-51	0.00	12,139,636	12,133,442	-6,194	-0.05	-6,144	-0.05
Total ≥ 15	23,338,795	23,338,795	0	0.00	23,626,780	23,610,904	-15,876	-0.07	-15,876	-0.07
Age 0-4	1,636,092	1,636,651	559	0.03	1,682,077	1,687,571	5,494	0.33	4,935	0.29
Age 5-9	1,910,359	1,909,567	-792	-0.04	1,960,872	1,966,069	5,197	0.27	5,990	0.31
Age 10-14	1,986,213	1,986,447	234	0.01	2,035,126	2,040,311	5,185	0.25	4,951	0.24
Age 15-19	1,986,163	1,986,942	779	0.04	2,026,860	2,024,694	-2,166	-0.11	-2,945	-0.15
Age 20-24	1,892,572	1,892,068	-504	-0.03	1,922,977	1,918,522	-4,455	-0.23	-3,951	-0.21
Age 25-29	1,835,744	1,834,959	-785	-0.04	1,866,784	1,863,210	-3,574	-0.19	-2,789	-0.15
Age 30-34	2,031,513	2,031,520	7	0.00	2,063,738	2,062,711	-1,027	-0.05	-1,034	-0.05
Age 35-39	2,452,299	2,451,743	-556	-0.02	2,484,983	2,483,560	-1,423	-0.06	-867	-0.03
Age 40-44	2,510,847	2,510,947	100	0.00	2,540,694	2,539,345	-1,349	-0.05	-1,449	-0.06
Age 45-49	2,273,676	2,274,363	687	0.03	2,297,674	2,296,514	-1,160	-0.05	-1,847	-0.08
Age 50-54	2,031,050	2,030,963	-87	0.00	2,051,231	2,048,768	-2,463	-0.12	-2,376	-0.12
Age 55-59	1,549,675	1,549,756	81	0.01	1,564,428	1,563,521	-907	-0.06	-988	-0.06
Age 60-64	1,234,930	1,235,863	933	0.08	1,246,010	1,246,568	558	0.04	-375	-0.03
Age 65-74	2,059,079	2,059,350	271	0.01	2,073,468	2,074,803	1,335	0.06	1,065	0.05
Age 75 and over	1,481,247	1,480,321	-926	-0.06	1,487,933	1,488,687	754	0.05	1,680	0.11

Characteristic	2001 Census – Pass 1				2001 Census – Pass 2				Pass 2 – Pass 1	
	Count	Estimate	Difference	Disc.	Count	Estimate	Difference	Disc.	Difference	Disc.
Single	13,282,845	13,282,792	-53	0.00	13,576,338	13,578,613	2,275	0.02	2,328	0.02
Married	11,750,092	11,750,035	-57	0.00	11,853,964	11,854,210	246	0.00	303	0.00
Widowed	1,341,497	1,342,054	557	0.04	1,353,562	1,354,561	999	0.07	442	0.03
Divorced	1,794,079	1,794,285	206	0.01	1,807,982	1,805,493	-2,489	-0.14	-2,695	-0.15
Separated	702,946	702,293	-653	-0.09	713,009	711,977	-1,032	-0.14	-379	-0.05
Com.-law = yes	2,267,634	2,271,749	4,115	0.18	2,322,437	2,329,084	6,647	0.29	2,532	0.10
1-person hhlds	2,908,857	2,904,682	-4,175	-0.14	2,932,655	*	*	*	*	*
2-person hhlds	3,709,282	3,708,376	-906	-0.02	3,736,957	*	*	*	*	*
3-person hhlds	1,848,476	1,843,466	-5,010	-0.27	1,868,996	*	*	*	*	*
4-person hhlds	1,812,783	1,815,197	2,414	0.13	1,833,471	*	*	*	*	*
5-person hhlds	714,618	723,436	8,818	1.23	729,190	*	*	*	*	*
6+-person hhlds	332,959	331,817	-1,142	-0.34	352,349	*	*	*	*	*

* Data not available

Note: Pass 2 counts and estimates include persons enumerated on Forms 2C (persons enumerated outside Canada) while Pass 1 counts and estimates do not.

Chart 7.2.2.3: Comparison of Pass 1 and Pass 2 Population/Estimate Discrepancies Based on Final Weights, for Canada, 2001 Census



7.2.3 Discarding Constraints

For the 2001 Census, the parameters of the weighting system were adjusted (see Section 4.4) so that fewer constraints were dropped compared to the 1996 Census, as will be shown in this section. This resulted in smaller population/estimate discrepancies in 2001 compared to 1996, as was shown in Section 7.2.2.

Table 7.2.3.1 shows how often each of the 32 constraints was discarded in the 6,141 sampled WAs in 2001 and the 5,941 sampled WAs in 1996. The reason a constraint was dropped (i.e. for being small, linearly dependent, nearly linearly dependent or causing outlier weights [see Section 4.4]) can help explain why certain constraints had large population/estimate discrepancies in Chart 7.2.2.2. This discussion will focus on the 2001 results. First, it should be noted that a constraint such as “Age 0-4” can be discarded frequently for being linearly dependent (which means it is redundant) and still have a small population/estimate difference. If a constraint is discarded frequently for causing outlier weights (such as “Common-law status = yes” or “5-person households”) or for being nearly linearly dependent (such as for 1-, 3- or 4-person households), this can cause large population/estimate discrepancies, as was observed in Chart 7.2.2.2.

Table 7.2.3.2 summarizes the information found in Table 7.2.3.1. In the former, we note that the number of linearly dependent constraints dropped in 1996 is adjusted upward by 2. This is to account for the constraints “Separated” and “6+-person households” not being used in 1996 due to the fact that they were linearly dependent on other constraints (see Appendix B). In 2001, the SMALL parameter was increased for some WAs. As a result, we note in Table 7.2.3.2 that the number of constraints eliminated for being small increased from 0.1 in 1996 to 0.4 in 2001. In addition, the constraints COND and MAXC were made larger for some WAs in 2001. Hence, Table 7.2.3.2 shows that the number of constraints eliminated for being nearly linearly dependent decreased from 1.6 in 1996 to 1.0 in 2001.

Table 7.2.3.3 summarizes information on the frequency of discarding the DA-level constraints on number of households and number of persons. If a WA contained eight DAs, for example, it would have 16 DA-level constraints. Table 7.2.3.3 shows that 0.7 of these constraints were dropped for being nearly linearly dependent in 2001 compared to 2.2 constraints in 1996. This is the result of COND and MAXC parameters being made larger for some WAs in 2001. Because no information was available for the 1996 Census on the number of DA-level constraints which were dropped, the numbers in Table 7.2.3.3 were approximated by running the weighting system with 2001 Census data and the 1996 weighting parameters.

Table 7.2.3.1: Frequency of Discarding WA-level Constraints in 1996 and 2001 Final Weight Adjustment

Characteristic	2001 Census					1996 Census				
	Small	LD	NLD	Outlier	Total	Small	LD	NLD	Outlier	Total
Males	0	0	0	0	0	0	0	0	1	1
Females**	-	-	-	-	-	-	-	-	-	-
Total population	0	0	0	0	0	0	0	0	0	0
Males ≥ 15	0	4	24	27	55	0	1	136	3	140
Persons ≥ 15	0	0	0	0	0	0	0	1	0	1
Total households	0	0	0	0	0	0	0	0	0	0
Age 0-4	29	4,286	2	124	4,441	6	3,071	20	57	3,154
Age 5-9	68	406	4	251	729	30	709	77	135	951
Age 10-14	79	1,359	2	141	1,581	35	2,110	33	61	2,239
Age 15-19	18	492	6	131	647	6	514	27	96	643
Age 20-24	2	243	15	125	385	1	216	133	119	469
Age 25-29	3	877	9	94	983	1	347	108	82	538
Age 30-34	3	158	5	83	249	1	29	23	42	95
Age 35-39	3	6	1	35	45	1	0	6	31	38
Age 40-44	2	0	0	19	21	1	3	13	45	62
Age 45-49	2	2	3	41	48	1	4	9	50	64
Age 50-54	2	7	1	38	48	2	157	67	83	309
Age 55-59	3	238	7	79	327	2	636	213	147	998
Age 60-64	5	1,751	65	130	1,951	3	1,122	973	128	2,226
Age 65-74	5	2	32	49	88	4	3	214	81	302
Age 75 and over	42	2,308	8	38	2,396	36	2,864	100	60	3,060
Single	1	0	0	2	3	0	0	1	3	4
Married	1	1	0	2	4	0	0	0	4	4
Widowed	6	593	15	128	742	2	0	174	345	521
Divorced	3	15	11	94	123	1	1	213	252	467

Characteristic	2001 Census					1996 Census				
	Small	LD	NLD	Outlier	Total	Small	LD	NLD	Outlier	Total
Separated*	20	5,510	3	34	5,567	-	-	-	-	-
Com.-law = yes	16	0	0	278	294	23	0	1	272	296
1-person hhlds	2	194	1,869	22	2,087	1	12	4,583	4	4,600
2-person hhlds	1	2	310	15	328	0	0	1,154	12	1,166
3-person hhlds	7	40	2,537	42	2,626	2	22	189	47	260
4-person hhlds	50	187	1,102	98	1,437	23	145	52	37	257
5-person hhlds	401	1,206	143	281	2,031	193	997	865	92	2,147
6+-person hhlds*	1,941	3,960	121	9	6,031	-	-	-	-	-

* Indicates the characteristic was not used as a constraint in 1996 because it was redundant.

** Indicates the characteristic was not used as a constraint in 1996 or 2001 because it was redundant.

Small = small constraint

LD = linearly dependent constraint

NLD = nearly linearly dependent constraint

Outlier = caused outlier weights

Table 7.2.3.2: Frequency of Discarding WA-level Constraints in 1996 and 2001 Final Weight Adjustment – Summary Statistics

	2001 Census					1996 Census				
	Small	**LD**	**NLD**	**Outlier**	**Total**	**Small**	**LD**	**NLD**	**Outlier**	**Total**
Total dropped constraints	2,715	23,847	6,295	2,410	35,267	375	12,963	9,385	2,289	25,012
Constraints dropped per WA	0.4	3.9	1.0	0.4	5.7	0.1	2.2	1.6	0.4	4.2
Adjusted total for two constraints not used in 1996 because LD						375	24,845	9,385	2,289	36,894
Constraints dropped per WA						0.1	4.2	1.6	0.4	6.2
Combined totals	Small + LD	26,562	NLD + Outlier	8,705	35,267	Small + LD	25,220	NLD + Outlier	11,674	36,894
Constraints dropped per WA		4.3		1.4	5.7		4.2		2.0	6.2

Small = small constraint

LD = linearly dependent constraint

NLD = nearly linearly dependent constraint

Outlier = caused outlier weights

Table 7.2.3.3: Frequency of Discarding DA-level Constraints in 1996 and 2001 Final Weight Adjustment – Summary Statistics

	2001 Census					1996 Census**				
	Small	**LD**	**NLD**	**Outlier**	**Total**	**Small**	**LD**	**NLD**	**Outlier**	**Total**
Total dropped constraints	1,354	357	4,191	917	6,819	1,082	393	12,973	1,069	15,517
Constraints dropped per WA	0.2	0.1	0.7	0.1	1.1	0.2	0.1	2.2	0.2	2.6
Combined totals	Small + LD	1,711	NLD + Outlier	5,108	6,819	Small + LD	1,475	NLD + Outlier	14,042	15,517
Constraints dropped per WA		0.3		0.8	1.1		0.2		2.4	2.6

** 1996 Census information is recreated using 2001 data with 1996 system parameters

Small = small constraint

LD = linearly dependent constraint

NLD = nearly linearly dependent constraint

Outlier = caused outlier weights

8. Sample Estimate and Population Count Consistency

In Chapter 7 (see Table 7.2.2.1), the discrepancies at the Canada level between the population counts and corresponding sample estimates based on final weights were studied where

$$\textbf{population/estimate discrepancy} = \frac{\text{sample estimate - population count}}{\text{population count}} \times 100$$

The sample estimates and population counts are based on occupied dwellings from sampled EAs.

In this chapter, these population/estimate discrepancies from both the 1996 and 2001 censuses will be examined for the following geographic levels:

(a) dissemination areas (DAs);

(b) weighting areas (WAs);

(c) census subdivisions (CSDs);

(d) census tracts (CTs);

(e) census divisions (CDs).

At the WA level, we observe that zero population/estimate discrepancies are guaranteed for constraints that are retained by the weighting system. In general, geographic areas made up of whole WAs have small population/estimate discrepancies. A look at Table 7.1.2 reveals that 12.8% of CSDs and 65.4% of CTs consist of one or more whole WAs. In addition, because of the way in which WAs are formed, 100% of CDs consist of whole WAs. For geographic areas smaller than WAs (such as DAs), population/estimate differences are usually larger.

The charts and tables in this chapter provide the percentiles of the population/estimate discrepancies for 31 characteristics which, except in a few cases, are identical to the 32 WA-level constraints applied to the census weights (see Appendix B). Let us define the term **percentile** by way of an example. For instance, Table 8.2.1 shows a 2001 percentile of -6.07% for "6+-person households." This means that 10% of the WAs have discrepancies of -6.07% or less. A 90th percentile of 7.98% means that 10% of the WAs have discrepancies of 7.98% or more. Population/estimate discrepancies for geographic areas having a population count less than or equal to 50 for a given characteristic are excluded from the tables and charts in this chapter. These discrepancies were found to be relatively large and could have significantly altered the percentiles presented in this chapter.

WA-level percentiles for all characteristics and percentiles for the "Total number of households" constraint were not easily obtainable for the 1996 Census. Rough estimations of the 1996 results were generated by running the census weighting system on 2001 Census data for the 2001 constraints listed in Appendix B with all other parameters being the same as in 1996.

It will be shown below that, at the Canada, CD and WA levels, the 2001 population/estimate discrepancies were generally smaller than those of 1996 while, at the DA and CT levels, they were somewhat larger. This was consistent with the 2001 objective of achieving smaller discrepancies at higher geographic levels while always having weights greater than or equal to 1.

8.1 Dissemination Areas

Canada is divided into 52,993 DAs, of which 47,933 were subject to sampling. Each DA has a population of 400 to 700 persons.

In comparing charts 8.1.1 and 8.1.2 to the other charts in this chapter, it is obvious that the population/estimate discrepancies are somewhat higher at the DA level than at the WA, CSD, CT or CD levels. This is not surprising given WAs are made up of whole DAs and that WAs are the lowest level at which sample estimates will agree with population counts for most characteristics.

The dissemination area (DA) was introduced for the 2001 Census (see Section 4.2). In 1996, its role was played by the enumeration area (EA). This explains why the 1996 percentiles in charts 8.1.1 and 8.1.2 are presented at the EA level while the 2001 percentiles are presented at the DA level. For almost all characteristics, the 2001 DA ranges are somewhat larger than the 1996 EA ranges between both the 10th and 90th percentiles and the 25th and 75th percentiles. This is probably because the SMALL parameter (see Section 4.4) was set to 20 in 1996 while in 2001, it was set to either 30 or 40 for a significant number of WAs. Allowing this larger value for the SMALL parameter in 2001 resulted in more constraints being dropped and generated larger discrepancies at the DA-level first-step adjustment. Contrary to 1996, this tended to increase the post-second-step-adjustment size of the discrepancies for the 32 DA-level constraints.

Three characteristics in Chart 8.1.2 warrant further discussion. The ranges between the 10th and 90th percentiles and the 25th and 75th percentiles for the "Marital status = separated" characteristic are smaller in 2001 than in 1996. This is because this characteristic was used as a weighting constraint for 2001, but not for 1996 (see Appendix B). The range between the 10th and 90th percentiles was zero in 2001 for "Total persons," while in 1996 it was non-zero. This can be explained by the fact that many fewer DA-level constraints were discarded at the second step in 2001 for being nearly linearly dependent (refer to Table 7.2.3.3) Also, the 1996 MAXC parameter (see Section 4.4) was set to 10,000 while the 2001 MAXC was generally in the range 20,000–160,000 as a means of retaining more constraints. Finally, the ranges between the 10th and 90th percentiles and the 25th and 75th percentiles for the "Common-law status = yes" characteristic are much larger in 2001. Table 7.2.2.1 shows that the Canada-level 2001 and 1996 population/estimate discrepancies based on initial weights for "Common-law status = yes" were -14,381 and -1,404 respectively. The reason for this increase in the size of the discrepancy in 2001 is not known. The Canada-level population/estimate discrepancy based on final weights was reduced to 4,115 in 2001 and to 2,415 in 1996. Given these patterns at the Canada level, it is no wonder that the ranges for this constraint are larger at the 2001 DA level than for 1996. Nevertheless, the extent of the increase in these ranges remains surprising.

8.2 Weighting Areas

Canada (excluding the Northwest Territories and Nunavut) is divided into 6,148 WAs, of which 6,141 are sampled WAs. On average, each WA has a population of 4,701 persons and is composed of eight whole DAs. WAs are used for calculating census weights but no results are published at this level.

Table 8.2.1 shows that, for both the 2001 and 1996 censuses, the 10th, 25th, 50th, 75th and 90th percentiles are zero for almost all person characteristics. For the household characteristics, most of the 25th, 50th, and 75th percentiles are also zero while some of the 10th and 90th percentiles are non-zero. These results are not surprising given that WAs consist of the lowest level at which sample estimates are forced to agree with population counts for the weighting constraints. It should be noted that the 1996 figures are approximated using 2001 data and the same weighting system parameters as in 1996.

8.3 Census Subdivisions

Canada is divided into 5,600 CSDs. CSDs correspond to municipalities or to areas deemed to be equivalent to municipalities for the purposes of statistical reporting (e.g. an Indian reserve). They have an average population of 5,400 persons, but can range anywhere in size from a very small town to a very large city. Table 7.1.2 shows that 12.8% of CSDs consist of one or more whole WAs.

Charts 8.3.1 and 8.3.2 summarize the population/estimate discrepancies for all sampled CSDs in Canada. For the 2001 Census, the CSD-level ranges between the 10th and 90th percentiles are smaller for most constraints but similar in magnitude to the ranges observed for the 10th and 90th percentiles at the DA level. The presumed reason for this is that 84.5% of CSDs make up only part of one WA (see Table 7.1.2); hence, exact population/estimate agreement would not be expected for most constraints. In contrast, the ranges observed for the 25th and 75th percentiles at the CSD level are much smaller than the corresponding ones at the DA level. This is likely a result of some of the constraints being applied to larger municipalities, which can be aggregations of primarily whole WAs.

Some discrepancies were smaller in 2001 than in 1996 while others were larger. Characteristics which were noticeably improved for 2001 include "Age 75+," "Marital status = widowed," "Marital status = separated," and "Marital status = divorced." Characteristics which were worse for 2001 include 3-person and 6+-person households.

8.4 Census Tracts

CTs are only located in large urban centres having an urban core population of 50,000 or more. There are 4,798 CTs in Canada. CTs usually have a population ranging from 2,500 to 8,000 persons, with the average being approximately 4,400 persons. Table 7.1.2 shows that 65.4% of CTs consist of one or more whole WAs.

Chart 8.4.1 summarizes the population/estimate discrepancies for all sampled CTs in Canada. Because 32.9% of CTs make up only part of one WA (see Table 7.1.2), it is not surprising that for 2001 the 10th and 90th percentiles are relatively large. What is surprising however is how much larger the 2001 percentiles are than the 1996 ones. This may be due in part to the 2001 DA discrepancies being somewhat larger than the 1996 DA discrepancies (see charts 8.1.1 and 8.1.2). The 25th and 75th percentiles for the discrepancies are generally zero (presumably because 65.4% of the CTs consist of whole WAs). As a result, they are not included in the charts.

8.5 Census Divisions

Canada is divided into 288 CDs. CDs have an average population of approximately 104,000 persons. A CD might correspond to a county, regional municipality, regional district, or any other area established by provincial/territorial law.

Table 8.5.1 summarizes the 2001 and 1996 Census population/estimate discrepancies for the sampled CDs. All CDs consist of complete WAs. Thus characteristics that are weighting constraints and which were rarely discarded have perfect or nearly perfect consistency at the CD level[4]. For other characteristics, as a general rule, the 2001 percentiles are smaller than the 1996 percentiles for person characteristics while the reverse holds true for household characteristics. This is consistent with what was observed in Table 7.2.2.1 with the population/estimate discrepancies at the Canada level.

[4] Even for characteristics with perfect consistency, published tabulations of basic characteristics based on sample data will not agree exactly with tabulations of the same characteristics based on 100% data. This can be attributed to the use of Pass 2 results with the sample data and Pass 1 results with the 100% data (see Section 4.5). In addition, tabulations of characteristics based on 100% data include institutional residents (see Section 3.2) while tabulations based on sample data do not.

Chart 8.1.1: Percentiles of Population/Estimate Discrepancies for DAs (2001 Census) and EAs (1996 Census) for Age Groups

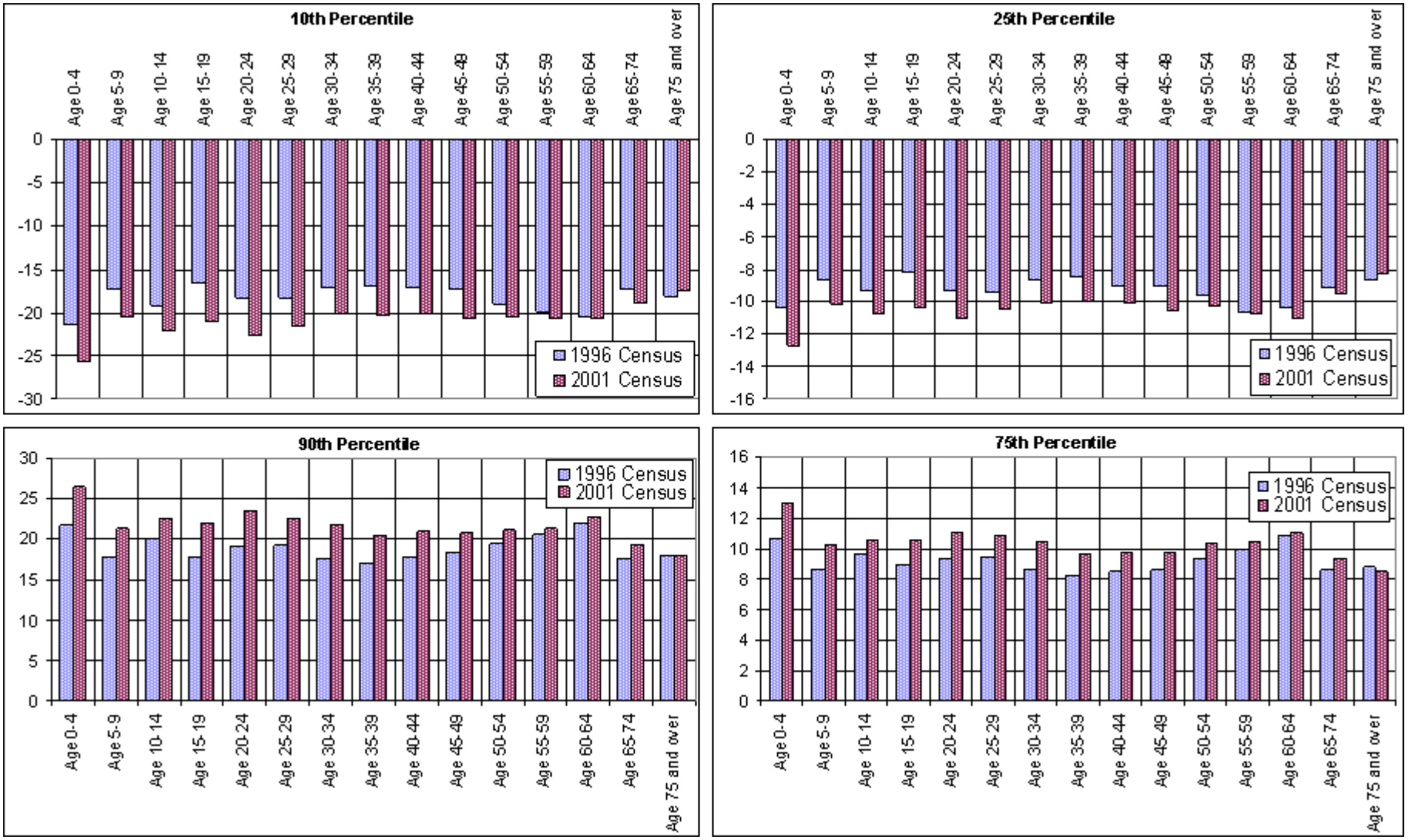
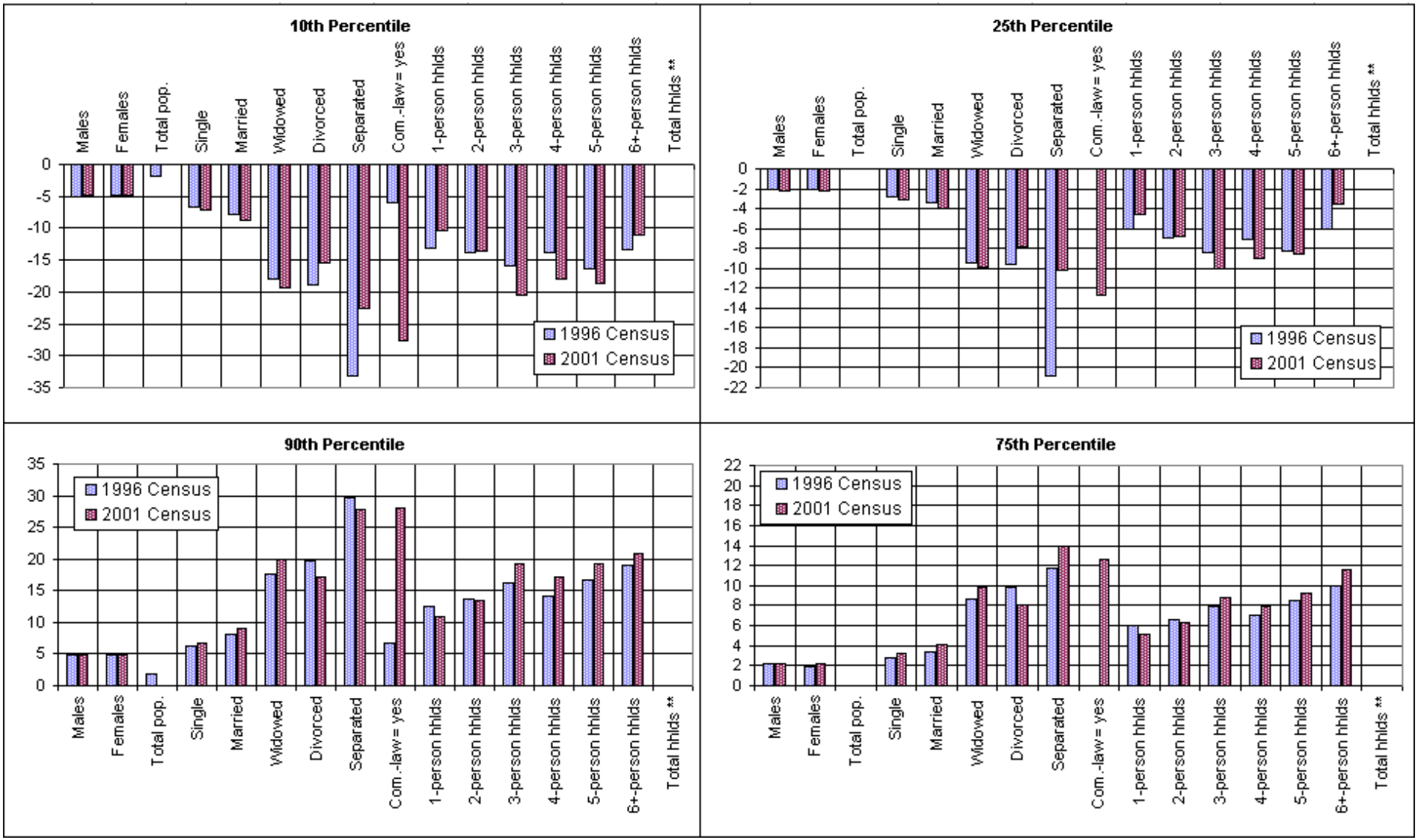


Chart 8.1.2: Percentiles of Population/Estimate Discrepancies for DAs (2001 Census) and EAs (1996 Census) for Other Population Characteristics and Household Characteristics



** Total household percentiles for 1996 are estimated with 2001 data.

Table 8.2.1: Percentiles of Population/Estimate Discrepancies for WAs

Characteristics	2001 Percentiles					1996 Percentiles **				
	10th	25th	50th	75th	90th	10th	25th	50th	75th	90th
Person characteristics										
Males	0	0	0	0	0	0	0	0	0	0
Females	0	0	0	0	0	0	0	0	0	0
Total population	0	0	0	0	0	0	0	0	0	0
Age 0-4	0	0	0	0	0	0	0	0	0	0
Age 5-9	0	0	0	0	0	0	0	0	0	0
Age 10-14	0	0	0	0	0	0	0	0	0	0
Age 15-19	0	0	0	0	0	0	0	0	0	0
Age 20-24	0	0	0	0	0	0	0	0	0	0
Age 25-29	0	0	0	0	0	0	0	0	0	0
Age 30-34	0	0	0	0	0	0	0	0	0	0
Age 35-39	0	0	0	0	0	0	0	0	0	0
Age 40-44	0	0	0	0	0	0	0	0	0	0
Age 45-49	0	0	0	0	0	0	0	0	0	0
Age 50-54	0	0	0	0	0	0	0	0	0	0
Age 55-59	0	0	0	0	0	0	0	0	0	0
Age 60-64	0	0	0	0	0	0	0	0	0	1.31
Age 65-74	0	0	0	0	0	0	0	0	0	0
Age 75 and over	0	0	0	0	0	-1.98	0	0	0	0
Single	0	0	0	0	0	0	0	0	0	0
Married	0	0	0	0	0	0	0	0	0	0
Widowed	0	0	0	0	0	0	0	0	0	0
Divorced	0	0	0	0	0	0	0	0	0	0
Separated	0	0	0	0	0	0	0	0	0	0
Com.-law = yes	0	0	0	0	0	0	0	0	0	0
Household characteristics										
1-person hhlds	-1.11	0	0	0	0.03	-1.33	-0.53	0	0.04	0.57
2-person hhlds	0	0	0	0	0	-0.24	0	0	0	0
3-person hhlds	-1.8	-0.18	0	0	0.4	0	0	0	0	0
4-person hhlds	-0.06	0	0	0	0.16	0	0	0	0	0
5-person hhlds	0	0	0	0	7.89	-0.57	0	0	0	7.4
6+-person hhlds	-6.07	-1.57	1.16	4.63	7.98	-4.89	-1.59	0.91	3.62	6.18
Total hhlds	0	0	0	0	0	0	0	0	0	0

** 1996 percentiles are estimated with 2001 data.

Chart 8.3.1: Percentiles of Population/Estimate Discrepancies for CSDs for Age Groups

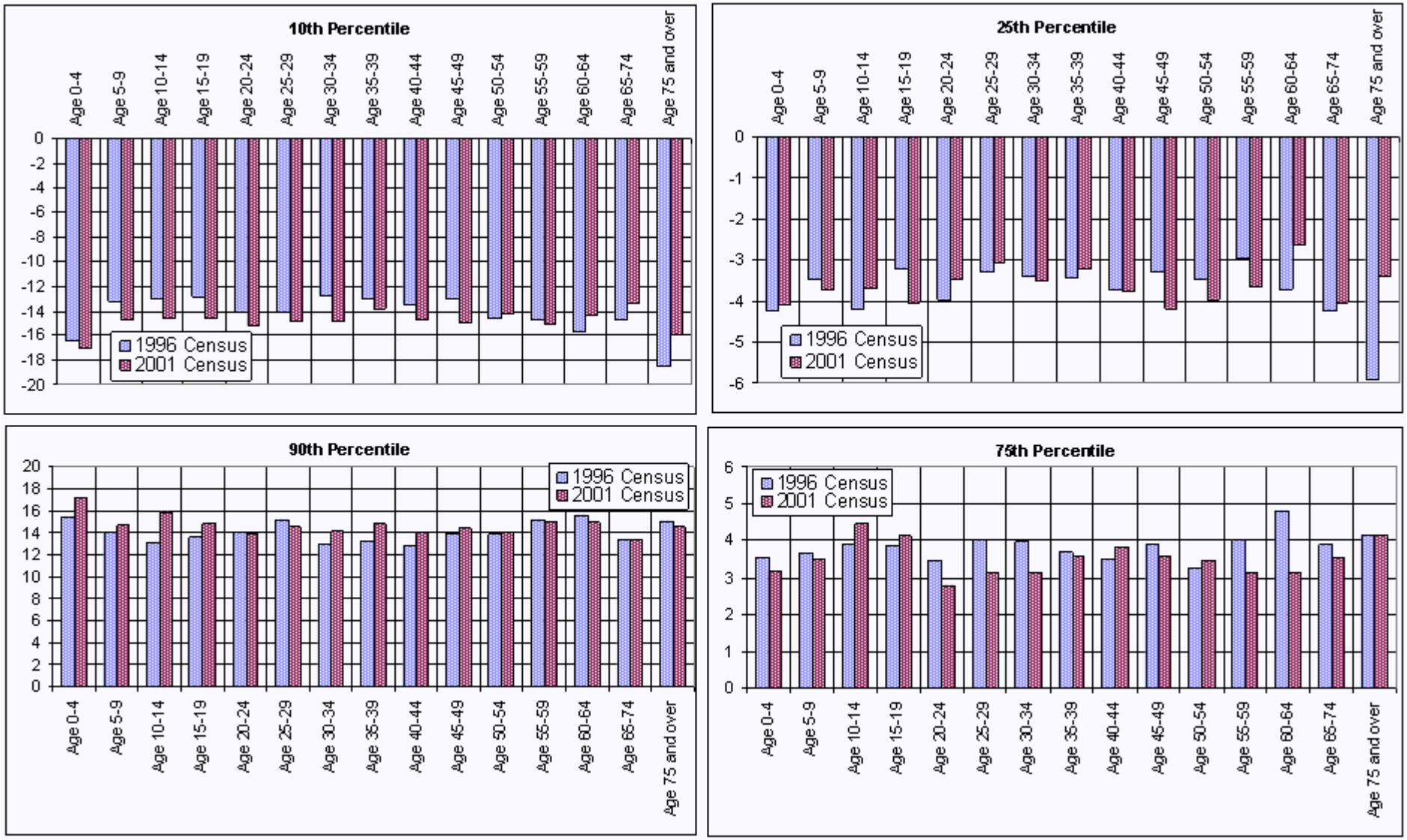
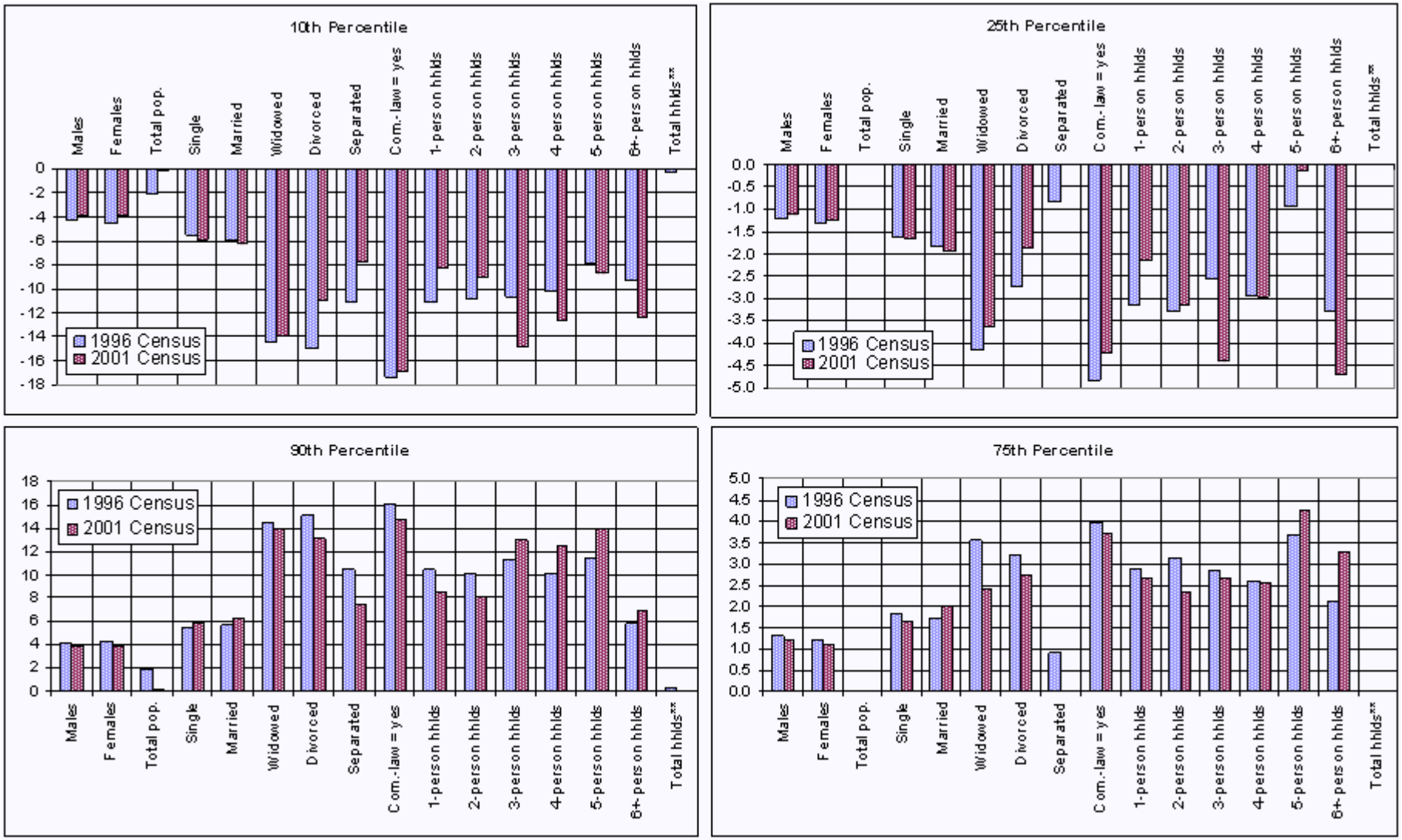
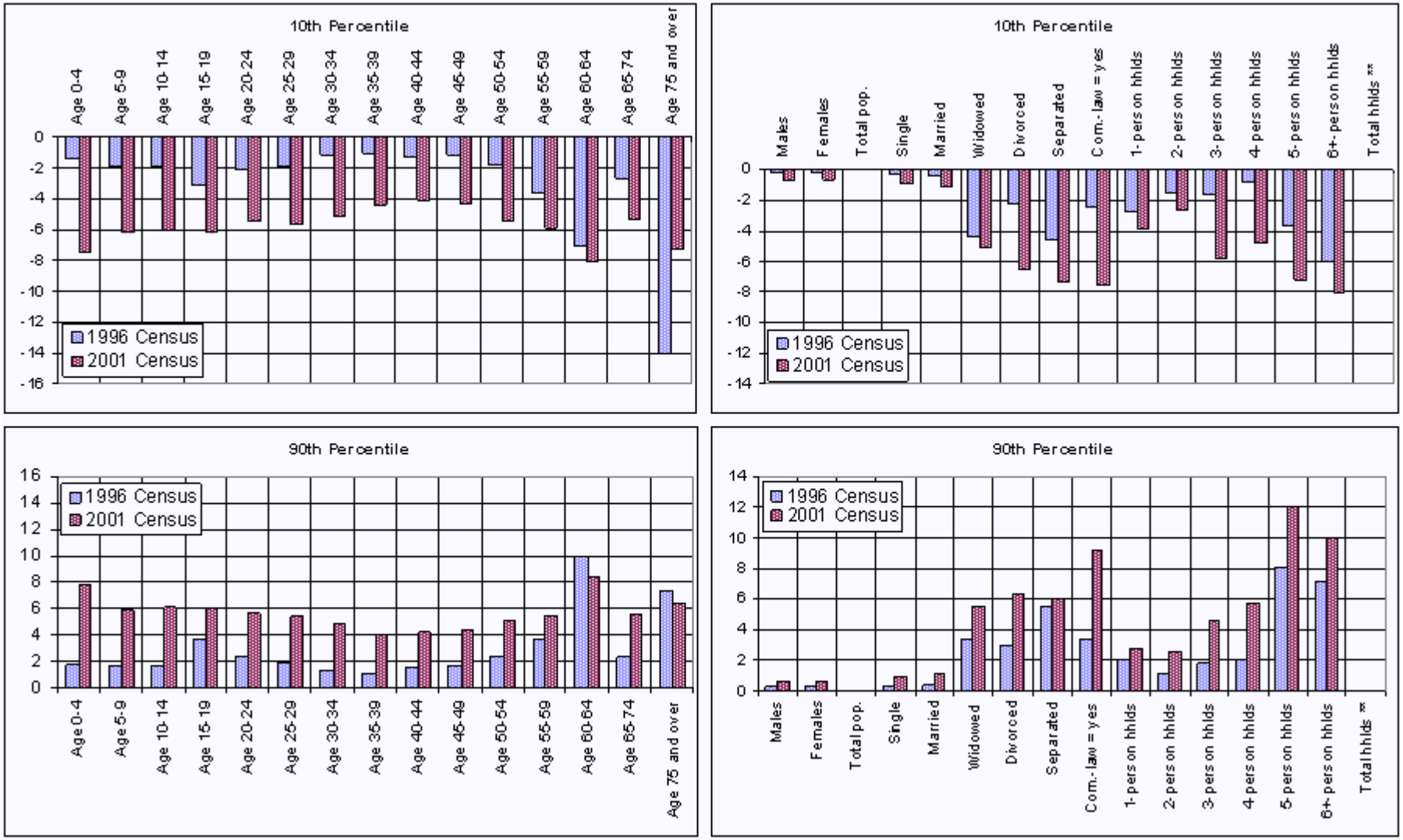


Chart 8.3.2: Percentiles of Population/Estimate Discrepancies for CSDs for Other Population Characteristics and Household Characteristics



** Total household percentiles for 1996 are estimated with 2001 data.

Chart 8.4.1: Percentiles of Population/Estimate Discrepancies for CTs



** Total household percentiles for 1996 are estimated with 2001 data.

Table 8.5.1 Percentiles of Population/Estimate Discrepancies for CDs

Characteristics	2001 Percentiles					1996 Percentiles				
	10th	25th	50th	75th	90th	10th	25th	50th	75th	90th
Person characteristics										
Males	0	0	0	0	0	0	0	0	0	0
Females	0	0	0	0	0	0	0	0	0	0
Total population	0	0	0	0	0	0	0	0	0	0
Age 0-4	-0.81	0	0	0	0.72	-0.44	0	0	0	0.05
Age 5-9	-0.51	0	0	0	0.23	-0.10	0	0	0	0.21
Age 10-14	-0.20	0	0	0	0.36	0	0	0	0	0.04
Age 15-19	-0.05	0	0	0	0.06	0	0	0	0	0.43
Age 20-24	-0.41	0	0	0	0.29	-0.89	0	0	0.09	1.17
Age 25-29	-0.54	0	0	0	0.44	-0.95	0	0	0	1.32
Age 30-34	-0.04	0	0	0	0.08	0	0	0	0	0
Age 35-39	0	0	0	0	0	0	0	0	0	0
Age 40-44	0	0	0	0	0	0	0	0	0	0
Age 45-49	0	0	0	0	0	0	0	0	0	0
Age 50-54	0	0	0	0	0	-0.17	0	0	0	0.21
Age 55-59	0	0	0	0	0	-1.05	0	0	0	0.85
Age 60-64	-0.18	0	0	0	0.49	-1.51	-0.01	0	0.92	2.52
Age 65-74	0	0	0	0	0	-0.25	0	0	0	0.06
Age 75 and over	-0.65	0	0	0	0.28	-3.70	-1.94	-0.17	0	0.81
Single	0	0	0	0	0	0	0	0	0	0
Married	0	0	0	0	0	0	0	0	0	0
Widowed	-0.14	0	0	0	0.04	-1.53	-0.33	0	0.09	1.60
Divorced	-0.08	0	0	0	0.31	-1.00	0	0	0.34	2.04
Separated	-0.96	0	0	0	0.94	-7.29	-2.25	0	1.15	4.30
Com.-law = yes	-0.21	0	0	0	0.84	-0.86	0	0	0	1.36

Characteristics	2001 Percentiles					1996 Percentiles				
	10th	25th	50th	75th	90th	10th	25th	50th	75th	90th
Household characteristics										
1-person hhlds	-0.34	-0.15	-0.04	0	0.01	-0.57	-0.36	-0.20	-0.10	0.10
2-person hhlds	-0.04	0	0	0	0	-0.10	0	0	0	0.01
3-person hhlds	-1.05	-0.59	-0.23	0	0.29	0	0	0	0	0.10
4-person hhlds	-0.27	-0.07	0	0.16	0.67	0	0	0	0	0.16
5-person hhlds	-0.79	0	0.99	2.47	5.16	-0.59	0	0.60	1.81	3.09
6+-person hhlds	-9.17	-3.86	-0.76	1.65	3.87	-6.55	-3.39	-0.84	0.99	2.65
Total hhlds **	0	0	0	0	0	0	0	0	0	0

** Total household percentiles for 1996 are estimated with 2001 data.

9. Sampling Variance

Sampling error can be divided into two components: variance and bias. The variance measures the variability of the estimate about its average value in hypothetical repetitions of the survey process, while the bias is defined as the difference between the average value of the estimate in hypothetical repetitions and the true value being estimated. Chapter 6 presented results of the Sampling Bias Study, describing the nature and extent of bias in the census sample prior to weighting. Even with a perfectly unbiased sampling method, the results would still be subject to variance, simply because the estimates are based only on a sample. The variance may be estimated using the data collected by the sample survey[5]. The Sampling Variance Study was carried out to estimate the effect of the sampling and estimation procedures on those census figures that are based on sample data.

On the basis of the 2B sample data, thousands of tables are produced by Statistics Canada. Conceptually, a measurement of precision, the estimated sampling variance, can be associated with every estimate calculated in these tables. This measurement takes into account both the sample design and the estimation method. In practice, however, it cannot be calculated for every census estimate because of high data processing costs. Sampling variance is thus estimated for only a subset of census estimates. From this, the combined effect of the sample design and the estimation method on the sampling variance can be estimated. Simple estimates of sampling variance, which are inexpensive to calculate, can then be adjusted for this impact to produce estimates of sampling variance for any census estimates.

The square root of the sampling variance, known as the standard error, can be approximated using the data in Tables 9.1 and 9.2. Table 9.1 gives non adjusted (simple) standard errors of census sample estimates. The figures in this table were obtained by assuming that one-in-five simple random sampling, and simple weighting by 5 were used. The standard errors are expressed in Table 9.1 as a function of the size of both the census estimate and the geographic area. For example, for an estimate of 250 persons in a geographic area with a total of 1,000 persons, the non-adjusted standard error is 25.

Standard errors are given in Table 9.1 for only a limited number of values for the estimated total and the total number of persons, households, dwellings or families in the area. The following formula may be used to calculate the non-adjusted standard errors (NASE) for any estimated total for an area of any size:

$$\text{NASE} = \sqrt{\frac{4E\,(N - E)}{N}}$$

where NASE is the non-adjusted standard error, E is the estimated total and N is the total number of persons, households, dwellings or families in the area. For example, for an estimated total of 750 persons in an area with a total of 9,000 persons, the non-adjusted standard error would be:

$$\sqrt{\frac{4(750)(9,000 - 750)}{9,000}} = 52$$

Table 9.2 provides adjustment factors[6] by which the non adjusted standard errors should be multiplied to adjust for the combined effect of the sample design and the estimation procedure. To calculate these adjustment factors, sampling variance estimates were calculated for regression estimates for different

[5] Unfortunately, the sampling variance does not provide any indication of the extent of non-sampling error.
[6] The squares of the adjustment factors are commonly known as "design effects."

categories of all of the characteristics[7] given in Table 9.2. This was done for each sampled WA. The provincial- and national-level sampling variance estimates were obtained by summing up the WA-level estimates. The adjustment factors were calculated for each characteristic in each category by dividing the square roots of these estimates by the non-adjusted standard errors. Adjustment factors were calculated at the provincial and national levels for each characteristic by averaging the adjustment factors for all of its categories. For further information on how these adjustment factors were calculated, see Hovington (2004).

To estimate the standard error for a given census sample estimate, the user should determine from Table 9.2 the adjustment factor applying to the characteristic and multiply this factor by the non adjusted standard error selected in Table 9.1. If the characteristic is not identified in Table 9.2, the user should pick the adjustment factor of 1 shown for the "All other…" category. For each characteristic in Table 9.2, adjustment factors are given at the national and provincial levels, as well as at the WA level. Unless the area is smaller than a province, the "National or Provincial Factor" column should be selected. In Table 9.2, adjustment factors are given for different provinces only where they differ significantly from those at the national level; this only occurred for some of the language characteristics. It should be noted that since no sampling occurred in the Northwest Territories or Nunavut, the adjustment factors for all characteristics in these territories should be zero. Since sampling was done in the Yukon territory, the "Other provinces" adjustment factor should be used, when available. If an adjustment factor is needed for a census estimate associated with an area smaller than a province, then the percentiles of WA-level factors will provide a more accurate value. The percentiles give the spread of all the adjustment factors calculated in the study at the WA level for the different category characteristics. N% of the adjustment factors at the WA level are below the Nth percentile and 100 - N% are above the Nth percentile. For example, 90% of the WA-level adjustment factors are below the 90th percentile and 10% are above it. The choice of which percentile to use will depend on how conservative a standard error estimate is being sought. For example, the 99th percentile would provide a very conservative estimate, while the 75th percentile would provide a somewhat less conservative one.

The following rules should be followed when calculating adjusted standard errors:

(a) When determining the standard error of an estimate relating to families or households, the number of families or households in the area, not the number of persons, should be used for selecting the appropriate column in Table 9.1.

(b) Unless otherwise specified, family characteristics involving husband, wife, lone parent or family reference person have the same adjustment factors as population characteristics. For example, the adjustment factor for the characteristic "highest level of schooling of husband, wife, or lone parent of a census family" is the same as the population characteristic "highest level of schooling".

(c) For cross-classifications of two or more characteristics, the largest adjustment factor for those characteristics should be used.

(d) Standard error adjustment factors do not apply to dollar values, for example, but to estimates of the number of persons, households, dwellings, or families. This means for instance that the household income adjustment factors apply to estimates of the number of households where income falls within a certain dollar range, not to estimates such as average household income.

The following example illustrates how to calculate the adjusted standard errors. Suppose the estimate of interest is the number of persons who immigrated to Canada between 1991 and 2001. The 2001 estimate for this characteristic was 1,830,680. The 2001 Census count for the population of Canada was 29,639,030. Since neither number is very close to any of the values given in Table 9.1, the formula given to calculate the non-adjusted standard error should be used. In this case the result would be 2,621. From

[7] For example, “$10,000 – 19,999” was one of the categories for which estimates of sampling variance were calculated for the characteristic "Number of persons in total income intervals."

Table 9.2, the national-level adjustment factor for the characteristic "period of immigration" after 1980 is 1.88. Consequently, the adjusted standard error for this estimate is 2,621 x 1.88 = 4,928.

The sample estimate and its standard error may be used to construct an interval within which the unknown population value is expected to be contained with a prescribed confidence. The particular sample selected in this survey is one of a large number of possible samples of the same size that could have been selected using the same sample design. Estimates derived from the different samples would differ from each other. If intervals from two standard errors below the estimate to two standard errors above the estimate were constructed using each of the possible estimates, then approximately 19 out of 20 such intervals would include the value normally obtained in a complete census. Such an interval is called a 95% (19 ÷ 20 = 95%) confidence interval. In order to guarantee 95% confidence however, these intervals must be calculated using the true standard errors of the sample estimates. The adjusted standard errors calculated from tables 9.1 and 9.2 are only estimates of the true standard errors. For provincial- and national-level sample estimates however, the adjusted standard errors should be close enough to the true standard errors so as to produce approximate 95% confidence intervals of reasonable precision. Below the provincial level, the adjusted standard errors may not be accurate enough for this purpose.

Using the standard error calculated above, an approximate 95% confidence interval for the number of persons who immigrated to Canada between 1991 and 2001 would be 1,830,680 ± 2(4,928) or 1,830,680 ± 9,856.

Table 9.1: Non-adjusted Estimates of Standard Errors of Sample Estimates

Estimated number of persons, households or dwellings in the area for characteristic of interest	Estimated total number of persons, households or dwellings in the area							
	500	**1,000**	**2,500**	**5,000**	**10,000**	**25,000**	**50,000**	**100,000**
50	15	15	15	15	15	15	15	15
100	20	20	20	20	20	20	20	20
250	20	25	30	30	30	30	30	30
500	0	30	40	40	45	45	45	45
1,000		0	50	55	60	60	65	65
2,500			0	70	85	95	100	100
5,000				0	100	125	135	140
10,000					0	155	180	190
25,000						0	225	275
50,000							0	315
100,000								0

	250,000	**500,000**	**1,000,000**	**2,500,000**	**5,000,000**	**10,000,000**	**25,000,000**	**30,000,000**
50	15	15	15	15	15	15	15	15
100	20	20	20	20	20	20	20	20
250	30	30	30	30	30	30	30	30
500	45	45	45	45	45	45	45	45
1,000	65	65	65	65	65	65	65	65
2,500	100	100	100	100	100	100	100	100
5,000	140	140	140	140	140	140	140	140
10,000	195	200	200	200	200	200	200	200
25,000	300	310	310	315	315	315	315	315
50,000	400	425	435	445	445	445	445	445
100,000	490	565	600	620	625	630	630	630

	250,000	500,000	1,000,000	2,500,000	5,000,000	10,000,000	25,000,000	30,000,000
250,000	0	705	865	950	975	985	995	995
500,000		0	1,000	1,265	1,340	1,380	1,400	1,400
1,000,000			0	1,550	1,790	1,900	1,960	1,965
2,500,000				0	2,235	2,740	3,000	3,030
5,000,000					0	3,160	4,000	4,085
10,000,000						0	4,900	5,165
15,000,000							4,900	5,475

Table 9.2: Standard Error Adjustment Factors at National, Provincial and WA Levels

Characteristics	National or provincial factors	Percentiles of WA-level factors						
		1st	25th	50th	75th	90th	95th	99th
Population characteristics								
Age								
Age groups 0-4, 5-9, 10-14, 15-19, 20-24, 25-29	0.15	0.00	0.00	0.04	0.17	0.32	0.53	1.35
Age groups 30-34, 35-44, 45-54, 55-59, 60-64	0.10	0.00	0.00	0.00	0.10	0.16	0.20	0.40
Age groups 65+	0.11	0.00	0.00	0.00	0.18	0.34	0.60	1.28
Sex	0.07	0.00	0.01	0.05	0.09	0.12	0.14	0.19
Common-law status								
In common-law relationship	0.46	0.00	0.00	0.21	0.47	0.77	1.36	2.17
Not in common-law relationship	0.36	0.00	0.00	0.09	0.34	0.65	1.00	1.91
Marital status								
Single, married (excluding separated)	0.07	0.00	0.00	0.04	0.07	0.11	0.14	0.19
Separated, divorced, widowed	0.14	0.00	0.00	0.00	0.22	0.41	0.57	1.31
Highest level of schooling	1.20	0.64	1.09	1.20	1.30	1.41	1.49	1.73
Highest degree, certificate or diploma	1.18	0.63	1.08	1.19	1.29	1.39	1.48	1.75
Major field of study	1.18	0.83	1.11	1.19	1.28	1.37	1.44	1.63

Characteristics	National or provincial factors	Percentiles of WA-level factors						
		1st	25th	50th	75th	90th	95th	99th
Place of birth								
Born in Canada	1.29	0.29	1.23	1.40	1.54	1.68	1.76	1.98
Born outside Canada	1.00	0.59	1.00	1.11	1.21	1.31	1.39	1.60
Citizenship								
Canada, by birth	1.21	0.00	1.14	1.37	1.61	1.90	2.15	2.78
Other	1.48	0.60	1.27	1.47	1.71	1.96	2.13	2.60
Number of citizenships								
Canadian only	1.23	0.29	1.17	1.35	1.51	1.66	1.76	2.00
One or two other ones	1.68	0.02	1.27	1.53	1.81	2.07	2.28	2.86
Period of immigration								
Before 1950, 1951-1960, 1961-1970, 1971-1980	1.36	0.73	1.14	1.28	1.43	1.58	1.69	2.00
1981-1990, 1991-1995, 1996-2001	1.88	0.71	1.42	1.74	2.02	2.3	2.51	3.09
Age at immigration	1.24	0.81	1.19	1.30	1.42	1.54	1.64	1.96
Mobility status (1 year ago)								
Non-movers	1.68	0.56	1.43	1.69	1.92	2.11	2.21	2.44
Movers (migrant, non-migrants)	1.79	0.44	1.33	1.66	1.94	2.17	2.32	2.66
Mobility status (5 years ago)								
Non-movers	1.66	0.67	1.47	1.69	1.87	2.03	2.13	2.31
Movers (migrant, non-migrants)	1.78	0.60	1.54	1.79	2.01	2.23	2.38	2.90

Characteristics	National or provincial factors	Percentiles of WA-level factors						
		1st	25th	50th	75th	90th	95th	99th
Immigrant / Non-immigrant population								
Immigrant population	1.29	0.31	1.24	1.42	1.63	1.96	2.25	2.98
Non-immigrant population	1.17	0.00	1.01	1.26	1.44	1.59	1.68	1.92
Visible minority								
Chinese, Asian, Blacks, Filipino, Latin American, Arab, Korean, Japanese, Visible Minority n.i.e., Multiple Visible Minorities	2.16	0.00	1.38	1.87	2.34	2.77	3.09	4.02
Aboriginal	1.41	0.37	1.30	1.60	1.91	2.19	2.39	3.01
Other	1.52	0.00	1.20	1.63	1.97	2.22	2.35	2.65
Ethnic origin								
English, French	1.51	0.00	1.20	1.62	1.96	2.20	2.34	2.62
Other	1.99	0.00	1.23	1.64	2.17	2.63	2.95	3.85
Religious denomination	1.69	0.00	1.19	1.59	1.97	2.37	2.67	3.45
Home language – English								
New-Brunswick, British-Colombia, Ontario, Alberta	1.63	0.15	1.33	1.69	1.96	2.17	2.30	2.62
Quebec	1.63	0.73	1.53	1.83	2.06	2.26	2.44	2.86
Other provinces	1.16	0.00	0.71	1.27	1.73	2.06	2.21	2.65
Canada	1.15	-	-	-	-	-	-	-
Home language – French								
Nova-Scotia, Quebec	1.26	0.00	0.90	1.41	1.77	2.02	2.17	2.68
New-Brunswick	0.94	0.38	1.28	1.61	1.84	2.11	2.35	2.47

Characteristics	National or provincial factors	Percentiles of WA-level factors						
		1st	25th	50th	75th	90th	95th	99th
Other provinces	1.59	0.07	1.31	1.64	1.96	2.28	2.55	3.42
Canada	0.74	-	-	-	-	-	-	-
First official language spoken – English								
Quebec	1.48	0.71	1.32	1.60	1.84	2.03	2.13	2.53
Newfoundland	0.69	0.00	0.23	0.54	0.94	1.27	1.53	2.22
Other provinces	1.24	0.12	0.94	1.24	1.50	1.72	1.85	2.18
Canada	0.82	-	-	-	-	-	-	-
First official language spoken – French								
New-Brunswick	0.93	0.40	1.15	1.41	1.61	1.84	1.97	2.15
Other provinces	1.27	0.17	1.17	1.39	1.65	1.90	2.05	2.53
Canada	0.79	-	-	-	-	-	-	-
First official language spoken – Other								
Both English and French	1.69	0.00	1.23	1.51	1.79	2.07	2.30	2.87
Neither	1.49	0.00	1.21	1.42	1.65	1.90	2.14	2.77
Official language spoken – English								
Quebec	1.40	0.00	1.26	1.48	1.67	1.89	2.10	2.56
Other provinces	1.30	0.34	1.12	1.36	1.54	1.69	1.78	1.97
Canada	0.86	-	-	-	-	-	-	-
Official language spoken – French								
New-Brunswick, Quebec	1.16	0.49	1.15	1.34	1.47	1.59	1.67	1.89

Characteristics	National or provincial factors	Percentiles of WA-level factors						
		1st	25th	50th	75th	90th	95th	99th
Other provinces	1.46	0.00	1.05	1.30	1.57	1.93	2.26	3.09
Canada	0.87	-	-	-	-	-	-	-
Official language spoken – Other								
Both English and French	1.29	0.62	1.26	1.42	1.57	1.71	1.80	1.99
Neither	1.49	0.00	1.20	1.42	1.64	1.89	2.12	2.75
Mother tongue – English								
Ontario, Alberta, British-Columbia	1.37	0.00	1.12	1.40	1.65	1.84	1.95	2.12
Quebec	1.45	0.60	1.26	1.51	1.77	1.95	2.07	2.44
Other provinces	1.04	0.00	0.80	1.09	1.37	1.63	1.77	2.09
Canada	1.08	-	-	-	-	-	-	-
Mother tongue – French								
New-Brunswick	0.83	0.12	1.11	1.28	1.55	1.82	1.96	2.24
Quebec	1.09	0.00	0.81	1.19	1.52	1.76	1.90	2.21
Other provinces	1.35	0.63	1.22	1.40	1.62	1.86	2.02	2.45
Canada	0.72	-	-	-	-	-	-	-
Mother tongue – Other	1.84	0.00	1.08	1.40	1.90	2.41	2.75	3.62
Language of work – English								
Quebec	1.16	0.71	1.14	1.26	1.37	1.48	1.55	1.73
Other provinces	0.78	0.19	0.67	0.77	0.88	1.01	1.11	1.32
Canada	0.76	-	-	-	-	-	-	-

Characteristics	National or provincial factors	Percentiles of WA-level factors						
		1st	25th	50th	75th	90th	95th	99th
Language of work – French								
New-Brunswick, Quebec	0.88	0.38	0.76	0.88	1.06	1.27	1.36	1.54
Other provinces	1.25	0.00	1.07	1.23	1.41	1.64	1.84	2.44
Canada	0.73	-	-	-	-	-	-	-
Language of work – Other	1.14	0.00	0.85	1.10	1.43	1.76	2.02	2.67
Industry	1.44	0.65	1.08	1.21	1.32	1.43	1.52	1.75
Occupation	1.07	0.74	0.97	1.14	1.26	1.38	1.46	1.70
Work activity in 2000	1.10	0.68	1.10	1.21	1.31	1.41	1.48	1.65
Weeks worked in 2000	1.04	0.63	1.07	1.19	1.29	1.38	1.44	1.58
Hours worked in reference week	1.39	0.75	1.11	1.19	1.27	1.35	1.40	1.54
Full-time / Part-time work								
Full-time work	0.82	0.71	0.56	0.92	1.00	1.08	1.13	1.26
Part-time work	1.12	0.96	1.11	1.18	1.25	1.32	1.38	1.51
Year last worked								
In 2001, in 2000, before 2000	0.86	0.39	0.81	0.97	1.15	1.28	1.35	1.49
Never worked	1.24	0.69	1.13	1.25	1.36	1.46	1.54	1.75

Characteristics	National or provincial factors	Percentiles of WA-level factors						
		1st	25th	50th	75th	90th	95th	99th
Class of worker								
Paid workers	0.87	0.39	0.85	1.04	1.29	1.47	1.59	1.86
Self-employed, unincorporated, unpaid family workers	1.25	0.70	1.11	1.23	1.35	1.50	1.63	2.00
Unpaid housework	1.19	0.63	1.11	1.21	1.31	1.42	1.50	1.69
Labour force participation status								
Employed	0.90	0.00	0.87	1.05	1.22	1.36	1.45	1.74
Unemployed	1.24	0.00	1.04	1.19	1.36	1.56	1.72	2.18
Not in labour force	1.04	0.58	1.01	1.16	1.29	1.42	1.50	1.71
Mode of transport to work								
Driver, walk, transit	1.01	0.58	1.00	1.16	1.29	1.41	1.49	1.74
Bike, motorcycle, passenger, taxi	1.23	0.45	1.06	1.19	1.32	1.46	1.59	1.96
Others	1.21	0.36	1.04	1.18	1.35	1.54	1.71	2.21
Place of work – Provinces	0.71	0.00	0.87	1.04	1.26	1.51	1.72	2.26
Place of work – Statistical Area Classification (census metropolitan area and census agglomeration influenced zone [MIZ])								
Strong or moderated MIZ	1.00	0.13	1.04	1.18	1.31	1.50	1.66	2.14
Weak or not in MIZ	0.80	0.00	0.90	1.11	1.31	1.53	1.69	2.14
In a CA or a CMA	0.84	0.37	0.93	1.08	1.24	1.41	1.55	1.98
In territories	0.31	0.00	0.72	0.87	1.02	1.23	1.37	1.86

Characteristics	National or provincial factors	Percentiles of WA-level factors						
		1st	25th	50th	75th	90th	95th	99th
Place of work – Type of commuting								
Work in same CSD of residence	1.01	0.45	0.95	1.09	1.21	1.32	1.38	1.49
Work in a different CSD of residence	1.08	0.64	1.09	1.19	1.30	1.44	1.56	1.97
Place of work status								
Worked at home, no fixed workplace	1.25	0.69	1.16	1.26	1.36	1.47	1.54	1.72
Worked outside Canada	1.26	0.00	0.96	1.16	1.35	1.57	1.74	2.25
Usual place of work	0.94	0.41	0.85	0.94	1.03	1.12	1.17	1.28
Number of persons in total income intervals ($)								
0-9,999	0.70	0.00	0.55	0.70	0.82	0.94	1.01	1.17
10,000-19,999, 20,000-29,999, 30,000-39,999, 40,000-49,999, 50,000-59,999, 60,000-69,999, 70,000-79,999, 75,000 or more	1.14	0.89	1.09	1.15	1.22	1.28	1.33	1.45
Census family status								
Husband, wife	0.11	0.00	0.06	0.09	0.13	0.18	0.21	0.28
Child	0.49	0.24	0.40	0.48	0.58	0.68	0.74	0.89
Female lone parent, male lone parent, non-member of a census family	0.85	0.52	0.80	0.93	1.10	1.25	1.33	1.55
Other	0.34	0.00	0.18	0.90	1.35	1.74	2.02	2.74
In census family								
Husband, wife, common-law partner present								
Yes, husband or wife	0.19	0.00	0.10	0.16	0.23	0.30	0.36	0.47
Yes, same-sex partner	1.76	0.00	1.49	1.71	2.00	2.38	2.68	3.38

Characteristics	National or provincial factors	Percentiles of WA-level factors						
		1st	25th	50th	75th	90th	95th	99th
Yes, opposite-sex partner	0.59	0.00	0.19	0.45	0.72	1.06	1.47	2.09
No	1.28	0.65	1.07	1.22	1.38	1.59	1.75	2.13
Economic family status								
Husband, wife	0.26	0.04	0.19	0.27	0.57	1.27	1.49	2.03
Lone parent, child	0.63	0.33	0.65	0.84	1.04	1.18	1.28	1.51
Other family members	1.35	0.61	1.10	1.24	1.43	1.64	1.80	2.24
All other population characteristics	1.00	--	--	--	--	--	--	--
Household and dwelling characteristics								
Tenure	0.83	0.58	0.89	0.98	1.05	1.12	1.17	1.27
Period of construction	1.03	0.74	1.05	1.13	1.20	1.28	1.35	1.58
Number of rooms	1.09	0.77	1.04	1.11	1.18	1.24	1.29	1.44
Number of bedrooms	1.21	0.68	1.02	1.10	1.20	1.34	1.48	1.87
Structural type								
Single detached house, row house, apartment in a building, mobile home, other movable dwelling	0.73	0.30	0.83	0.98	1.07	1.18	1.27	1.52
Semi-detached or double house, Apartment/flat in a detached duplex, other single-attached house	1.02	0.34	0.99	1.09	1.19	1.32	1.44	1.80

Characteristics	National or provincial factors	Percentiles of WA-level factors						
		1st	25th	50th	75th	90th	95th	99th
Household size								
One-person household	0.09	0.00	0.00	0.00	0.08	0.18	0.26	0.54
Other	0.17	0.00	0.00	0.00	0.25	0.86	1.21	1.72
Primary household maintainer	0.00	-	-	-	-	-	-	-
Age of primary household maintainer								
20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59, 60-64	0.67	0.46	0.61	0.67	0.76	0.89	1.00	1.21
65+	0.47	0.31	0.42	0.48	0.55	0.65	0.74	0.93
Sex of primary household maintainer	0.67	0.45	0.61	0.68	0.78	0.87	0.92	1.01
Number of household maintainers								
One household maintainer	1.17	0.78	0.98	1.04	1.09	1.14	1.17	1.23
More than one household maintainers	1.18	0.00	1.01	1.11	1.28	1.51	1.71	2.36
Reference person is a household maintainer	1.14	0.85	1.08	1.14	1.20	1.27	1.31	1.40
Person who does not live here is a household maintainer	1.05	0.50	0.95	1.07	1.20	1.35	1.48	1.78
Number of households in gross rent intervals (intervals of 100$)	1.07	0.69	1.01	1.11	1.21	1.33	1.43	1.74

Characteristics	National or provincial factors	Percentiles of WA-level factors						
		1st	25th	50th	75th	90th	95th	99th
Number of households in gross rent as a percentage of household income intervals								
Less than 10%	0.50	0.00	0.42	0.57	0.69	0.78	0.83	0.91
Between 10 and 20%, 20 and 30%, 30 and 40%, 40 and 50%, more than 50%	0.87	0.59	0.83	0.88	0.94	1.00	1.05	1.21
Number of households in owner's major payment intervals (intervals of 200$)	1.05	0.85	1.04	1.11	1.18	1.25	1.31	1.50
Number of households in owner's major payment as a percentage of household income intervals								
Between 30 and 40%	2.89	2.08	2.71	2.89	3.04	3.22	3.36	3.68
Less than 10%, Between 10 and 20%, 20 and 30%, 40 and 50%, more than 50%	0.88	0.65	0.85	0.89	0.94	1.00	1.05	1.19
Person responsible for household payments								
Person is the first maintainer	0.00	-	-	-	-	-	-	-
Other maintainers	0.90	0.00	0.90	1.07	1.29	1.52	1.72	2.38
Number of households in household income intervals (intervals of 10,000$)	1.05	0.69	1.02	1.10	1.17	1.23	1.27	1.36
Number of households in dwelling value intervals	0.97	0.71	1.01	1.10	1.18	1.29	1.39	1.69

Characteristics	National or provincial factors	Percentiles of WA-level factors						
		1st	25th	50th	75th	90th	95th	99th
Registered condominium								
Part	0.87	0.56	0.91	1.03	1.14	1.28	1.40	1.82
Not part	0.80	0.52	0.85	0.96	1.04	1.11	1.16	1.30
Condition of dwelling								
Regular maintenance, major or minor repairs	0.88	0.73	0.86	0.90	0.94	0.97	1.00	1.07
All other household and dwelling characteristics	1.00	-	-	-	-	-	-	-
Census family characteristics								
Labour force activity of husband, wife or lone parent								
Husband or wife in labour force	0.63	0.30	0.50	0.59	0.68	0.76	0.82	0.93
Lone parent in labour force	1.29	0.72	0.97	1.07	1.19	1.32	1.42	1.69
Age groups of children at home	0.15	0.00	0.00	0.00	0.23	0.43	0.71	1.56
Work activity in 2000 of husband, wife or lone parent								
Worked in 2000	0.85	0.48	0.70	0.79	0.88	0.97	1.04	1.16
Did not work in 2000	0.79	0.54	0.69	0.74	0.80	0.86	0.90	1.01
All other census family characteristics	1.00	-	-	-	-	-	-	-

Characteristics	National or provincial factors	Percentiles of WA-level factors						
		1st	25th	50th	75th	90th	95th	99th
Economic family characteristics								
Number of households in self-employment income intervals ($)								
0-9,999	1.22	0.00	0.71	0.92	1.09	1.22	1.29	1.45
10,000-19,999, 20,000-29,999, 30,000-39,999, 40,000-49,999, 50,000-59,999, 60,000-69,999, 70,000-79,999, 75,000 or more	1.51	0.67	1.05	1.17	1.31	1.47	1.61	1.97
Low income status								
Above line	1.78	1.05	1.65	1.86	2.07	2.27	2.39	2.69
Below line	1.90	1.34	1.76	1.94	2.13	2.32	2.44	2.72
Property taxes included in mortgage payment								
Taxes included	1.07	0.91	1.05	1.11	1.16	1.22	1.26	1.37
Taxes not included								
Wages and salaries ($)								
0-1,999	0.75	0.00	0.61	0.77	0.91	1.02	1.08	1.18
2,000-4,999, 5,000-6,999, 7,000-9,999, 10,000-11,999, 12,000-14,999, 15,000-19,999, 20,000-24,000, 25,000-29,999, 30,000-34,999, 35,000-39,000, 40,000-44,999, 45,000-49,999, 50,000-59,999, 60,000-74,999, 75,000 or more	1.18	0.77	1.11	1.19	1.28	1.37	1.44	1.63
Mother tongue of family reference person – English								
Newfoundland	0.16	0.00	0.08	0.13	0.18	0.25	0.31	0.50
Nova-Scotia, Prince-Edward-Island, Yukon	0.33	0.05	0.24	0.30	0.38	0.49	0.55	0.86

Characteristics	National or provincial factors	Percentiles of WA-level factors						
		1st	25th	50th	75th	90th	95th	99th
Quebec	1.00	0.70	1.03	1.12	1.21	1.30	1.37	1.62
Other provinces	0.61	0.29	0.50	0.61	0.74	0.87	0.94	1.08
Canada	0.59	-	-	-	-	-	-	-
Mother tongue of family reference person – French								
Quebec	0.44	0.00	0.20	0.37	0.62	0.85	0.97	1.13
New-Brunswick	0.62	0.13	0.63	1.02	1.12	1.19	1.24	1.35
Other provinces	1.02	0.68	1.03	1.13	1.22	1.32	1.39	1.59
Canada	0.52	-	-	-	-	-	-	-
Mother tongue of family reference person – Other than English or French	0.11	0.00	0.90	1.07	1.23	1.42	1.58	1.99
All other economic family characteristics	1.00	-	-	-	-	-	-	-

10. Conclusion

Sampling is now an accepted and integral part of census-taking. Its use can lead to substantial reductions in costs and respondent burden associated with a census, or alternatively, can allow the scope of a census to be broadened at the same cost. The price paid for these advantages is the introduction of sampling error to census figures that are based on the sample. The effect of sampling is most important for small census figures, whether they are counts for rare categories at the national or provincial level or counts for categories in small geographic areas. It should be noted that response errors and processing errors also contribute to the overall error of census figures and it is the same small census figures that are particularly susceptible to the effects of these non-sampling errors. Therefore, even with a 100% census, many small figures would be of limited reliability. As a general rule of thumb for the 2001 Census, figures of size 100 or less that are based on sample data are of very low reliability, while figures up to size 500 tend to have standard errors in excess of 10% of their size.

For many of the characteristics, a certain amount of bias was detected in the sample. A small portion of the bias may have been introduced during data processing and edit and imputation. The rest of the bias would have been due to one or more factors such as non-response bias, response bias or the selection of a biased sample by the census representatives. The procedures for weighting the sample data up to the population level were carried out successfully, and generally achieved the levels of sample estimate and population count consistency anticipated thus adjusting for certain biases observed in the sample. The consistency that was achieved at the provincial and Canada levels was better than in 1996 for most characteristics.

Appendix A – Glossary of Terms

The definitions of census terms, variables and concepts are presented here as they appear in the *2001 Census Dictionary* (Catalogue No. 92-378-XIE). Users should refer to the *2001 Census Dictionary* for full definitions and additional remarks related to any concepts, such as information on direct and derived variables and their respective universe.

Census division (CD): General term for provincially legislated areas (such as county, municipalité régionale de comté and regional district) or their equivalents. Census divisions are intermediate geographic areas between the province level and the municipality (census subdivision).

Census subdivision (CSD): General term for municipalities (as determined by provincial legislation) or areas treated as municipal equivalents for statistical purposes (for example, Indian reserves, Indian settlements and unorganized territories).

Census tract (CT): Census tracts are mall, relatively stable geographic areas that usually have a population of 2,500 to 8,000. They are located in census metropolitan areas and in census agglomerations with an urban core population of 50,000 or more in the previous census.

Dissemination area (DA): The dissemination area is a small, relatively stable geographic unit composed of one or more blocks. It is the smallest standard geographic area for which all census data are disseminated. DAs cover all the territory of Canada.

Enumeration area (EA): An enumeration area is the geographic area canvassed by one census representative. An EA is composed of one or more adjacent blocks. All the territory of Canada is covered by EAs.

Household: Refers to a person or a group of persons (other than foreign residents) who occupy the same dwelling and do not have a usual place of residence elsewhere in Canada. It may consist of a family group (census family) with or without other non-family persons, of two or more families sharing a dwelling, of a group of unrelated persons, or of one person living alone. Household members who are temporarily absent on Census Day (e.g. temporary residents elsewhere) are considered as part of their usual household. For census purposes, every person is a member of one and only one household. Unless otherwise specified, all data in household reports are for private households only.

Marital status: Refers to the conjugal status of a person. The various responses are: married and common-law; separated, but still legally married; divorced; widowed; never legally married (single).

Occupied private dwelling: Refers to a private dwelling in which a person or a group of persons is permanently residing. Also included are private dwellings whose usual residents are temporarily absent on Census Day. Unless otherwise specified, all data in housing products are for occupied private dwellings, rather than for unoccupied private dwellings or dwellings occupied solely by foreign and/or temporary residents.

Private dwelling: Refers to a separate set of living quarters with a private entrance either from outside or from a common hall, lobby, vestibule or stairway inside the building. The entrance to the dwelling must be one that can be used without passing through the living quarters of someone else.

Private household: Refers to a person or a group of persons (other than foreign residents) who occupy a private dwelling and do not have a usual place of residence elsewhere in Canada.

Appendix B – WA- and DA-level Constraints Applied to 2001 and 1996 Census Weights

Person WA-level Constraints

– Total persons
– Total persons aged ≥ 15

– Males
– Males aged ≥ 15

– Persons aged 0 to 4
– Persons aged 5 to 9
– Persons aged 10 to 14
– Persons aged 15 to 19
– Persons aged 20 to 24
– Persons aged 25 to 29
– Persons aged 30 to 34
– Persons aged 35 to 39
– Persons aged 40 to 44
– Persons aged 45 to 49
– Persons aged 50 to 54
– Persons aged 55 to 59
– Persons aged 60 to 64
– Persons aged 65 to 74
– Persons aged ≥ 75

– Married persons
– Single persons
– Divorced persons
– Widowed persons
– Separated persons**
– Common-law status = yes

Household WA-level Constraints

– Households of size 1
– Households of size 2
– Households of size 3
– Households of size 4
– Households of size 5
– Households of size 6 or more **
– Total households

DA-level constraints (these were EA-level constraints in 1996)

– Total households in DA
– Total persons in DA

** Not used as constraints in 1996 due to the fact they were known to be redundant.

Appendix C– Statistics Used In Sampling Bias Study

In Chapter 6, it is stated that under random sampling,

$$Z^{(0)} = \frac{\hat{X}^{(0)} - X}{\sqrt{V(\hat{X}^{(0)})}}$$

should follow an approximately normal (0,1) distribution. A justification for this is given here. Sampling was done independently in each EA. Therefore $\hat{X}^{(0)}$ is the sum of *H* independent random variables, where *H* is the number of EAs in Canada. There are 35,883 sampled EAs in Canada, therefore H is very large. Thus, according to the central limit theorem, $(\hat{X}^{(0)} - E(\hat{X}^{(0)}))/\sqrt{V(\hat{X}^{(0)})}$ will follow an approximately Normal (0,1) distribution (see Kendall and Stuart [1963], p. 193) as will $Z^{(0)} = (\hat{X}^{(0)} - X)/\sqrt{V(\hat{X}^{(0)})}$ if $E(\hat{X}^{(0)}) = X$. $Z^{(0)}$, however, would not have a mean of 0 if the EA level samples of households were significantly biased for any reason.

An additional statistic will now be derived which allows us to test if the bias between two regions or two censuses is the same. Let $\hat{X}_1^{(0)}$ and $\hat{X}_2^{(0)}$ be estimators (based on initial weights) of the known population counts X_1 and X_2 for two distinct geographic areas or for two different censuses. Let $RB(\hat{X}_1^{(0)}) = (E(\hat{X}_1^{(0)}) - X_1)/X_1$ and $RB(\hat{X}_2^{(0)}) = (E(\hat{X}_2^{(0)}) - X_2)/X_2$ be the relative biases of $\hat{X}_1^{(0)}$ and $\hat{X}_2^{(0)}$. We wish to test if the null hypothesis $H_0 : RB(\hat{X}_1^{(0)}) = RB(\hat{X}_2^{(0)})$ is true. This can be done using the statistic

$$W = \frac{rb(\hat{X}_1^{(0)}) - rb(\hat{X}_2^{(0)})}{\sqrt{\frac{1}{X_1^2}V(\hat{X}_1^{(0)}) + \frac{1}{X_2^2}V(\hat{X}_2^{(0)})}}$$

where $rb(\hat{X}_1^{(0)}) = (\hat{X}_1^{(0)} - X_1)/X_1$ and $rb(\hat{X}_2^{(0)}) = (\hat{X}_2^{(0)} - X_2)/X_2$ are unbiased estimators of $RB(\hat{X}_1^{(0)})$ and $RB(\hat{X}_2^{(0)})$ respectively. Thus, if the null hypothesis H_0 above is true, the expectation of W is zero. Note also that the denominator of W is the standard error of the numerator of W (there is no covariance term because estimates from separate regions or from different censuses are independent) and hence W has a variance of 1. Now if $\hat{X}_1^{(0)}$ approximately follows a normal distribution (again based on the central limit theorem), $rb(\hat{X}_1^{(0)})$ will also approximately follow a normal distribution, as will $rb(\hat{X}_2^{(0)})$ and $rb(\hat{X}_1^{(0)}) - rb(\hat{X}_2^{(0)})$. Thus W follows approximately a normal (0,1) distribution if the null hypothesis H_0 is true.

Appendix D. 2001 Census Products and Services

The census is a reliable source for describing the characteristics of Canada's people and dwellings. The range of products and services derived from census information is designed to produce statistics that will be useful, understandable and accessible to all users. Sources, such as the *2001 Census Catalogue*, the Statistics Canada Web site (http://www.statcan.ca) and, specifically, the On-Line Catalogue, contain detailed information about the full range of 2001 Census products and services.

There are several new product and service features for the 2001 Census:

1. Media

- The Internet is the preferred medium for disseminating standard data products and reference products.
- More census data are available to the public free of charge via the Internet.

2. Content

- Data tables for the 2001 Census are released by **topics**, that is, groups of variables on related subjects.
- Wherever possible, the language and vocabulary used in 2001 Census products available on the Internet is simplified to make the information accessible to more people.
- Users are offered various methods of searching and navigating through **census standard products** (including **reference products** on the Internet).

3. Geography

- Geographic units such as dissemination areas, urban areas, designated places and metropolitan influenced zones were added to the standard products line. Some new units, such as dissemination areas, replace others.

4. Variables

- Information on the following new subjects was collected in the 2001 Census: birthplace of parents, other languages spoken at home and language of work. The 2001 questionnaire also included the question on religion, which is asked in every decennial census. The family structure variable was broadened to include same-sex couples.

Bibliography

Bankier, M. 2002. *2001 Canadian Census Weighting*. Proceedings of the Statistics Canada Symposium 2002. Ottawa. Statistics Canada.

Cochran, W. 1977. *Sampling Techniques*. 3rd Edition. Toronto. John Wiley and Sons.

Dominion Bureau of Statistics. 1968. *Sampling in the Census*. Internal report. Ottawa. Statistics Canada.

Fellegi, I.P. 1964. "Response Variance and its Estimation." *Journal of the American Statistical Association*. 59, December: 1016–1041.

Fuller, Wayne A. 2002. "Regression Estimation for Survey Samples." *Survey Methodology*. 28, 1: 5–23.

Hansen, M.H., W.N. Hurwitz and M.A. Bershad. 1959. "Measurement Errors in Censuses and Surveys." *Bulletin of the International Statistical Institute*. 38, 359–374.

Hovington, Édith. 2004. *Étude de l'effet de plan de la variance d'échantillonnage pour le recensement de 2001*. Internal report. Ottawa. Statistics Canada.

Kendall, Maurice G. and Alan Stuart. 1963. *The Advanced Theory of Statistics*. Volume 1. London. Charles Griffin and Company Limited.

Kruszynski, G. 1999. *Evaluation of the 1996 Weighting Areas*. Internal report of the Geography Division. Ottawa. Statistics Canada.

Press, W.H., S.A. Teukolsky, W.T. Vetterling and B.P. Flannery. 1992. *Numeric Recipes in C*. New York. Cambridge University Press.

Royce, Don. 1983. *The Use of Sampling in the 1981 Canadian Census*. Internal report. Ottawa. Statistics Canada.

Sarndal, C., B. Swensson and J. Wretman. 1992. *Model Assisted Survey Sampling*. New York. Springer-Verlag.

Statistics Canada. 2002a. *2001 Census Dictionary*. 2001 Census Reference Series. Catalogue No. 92-378-XIE. Ottawa.

———. 2002b. *2001 Census Handbook*. 2001 Census Reference Series. Catalogue No. 92-379-XIE. Ottawa.