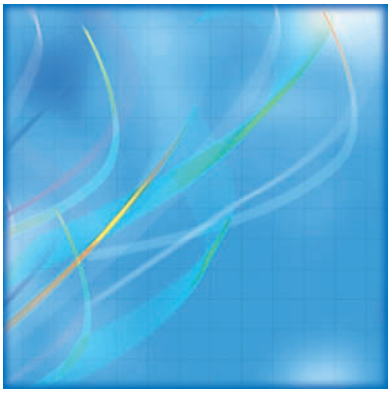
Catalogue no. 92-568-X

2006 Census Technical Report: Sampling and Weighting



Census year 2006

Statistics Canada Statistique Canada

Canada

How to obtain more information

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

Statistics Canada's National Contact Centre

Toll-free telephone (Canada and United States):

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

Local or international calls:

Inquiries line	1-613-951-8116
Fax line	1-613-951-0581

Depository Services Program

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

To access this product

This product, Catalogue no. 92-568-X, is available free in electronic format. To obtain a single issue, visit our website at www.statcan.gc.ca and select "Publications."

Standards of service to the public

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

Statistics Canada
Social Survey Methods Division

2006 Census Technical Report: Sampling and Weighting

Census year 2006

Published by authority of the Minister responsible for Statistics Canada

August 2009

Catalogue no. 92-568-X

ISBN 978-1-100-11975-5

Frequency: Occasional

Ottawa

Cette publication est galement disponible en français.

Note of appreciation

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

Sampling and Weighting

Table of contents

Page

Table of contents (continued)

Page

Table of contents (continued)

List of tables

Page

Table of contents (continued)

List of charts

Introduction

The 2006 Census required the participation of the entire population of Canada, over 31 million people distributed over a territory of 9 million square kilometres. Although there are high quality standards governing the collection 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 2006 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 *2006 Census Technical Report* deals with the method of sampling and weighting used in the 2006 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, Darryl Janes, 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. A 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 is often cost-effective.

The 2006 Census of Population made use of sampling in a variety of ways. It was used in ensuring that the quality of the enumerator'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 undercoverage and the amount of overcoverage; 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 2006 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 an overview of the studies designed to evaluate the 2006 Census sampling and weighting procedures is presented, while Chapters 6, 7, 8 and 9 present the results of these studies. Chapter 10 presents some conclusions on the weighting procedures used in 2006.

Users will find additional information on census concepts, variables and geography and an overview of the complete census process in the *2006 Census Dictionary* (Catalogue no. 92-566-XWE).

1 Census data collection

1.1 General

The data collection stage of the 2006 Census process ensured that each of the 13.5 million dwellings in Canada received a census questionnaire. The census enumerated the entire population of Canada, which consists of Canadian citizens (by birth and by naturalization), landed immigrants, and non-permanent residents together with family members living with them. Non-permanent residents are persons living in Canada who have a Work or Study Permit, or who are claiming refugee status, and family members living with them.

The census also counted Canadian citizens and landed immigrants who were temporarily outside the country on Census Day. This included federal and provincial government employees working outside Canada, Canadian embassy staff posted to other countries, members of the Canadian Forces stationed abroad and all Canadian crew members of merchant vessels.

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 collection unit. The VR serves as an address listing for field operations and control purposes for census collection.

The basic short questionnaire is called the 2A. It is distributed to four in five private dwellings. The 2B is a longer questionnaire that collects the same information as the 2A as well as additional information on a variety of topics. The 2B questionnaire is distributed to every one in five private dwellings. 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 2C is mainly used to enumerate people posted outside Canada, including Canadian government employees (federal and provincial) and their families, and members of the Canadian Forces and their families. The Form 2C contains the same questions as the 2B questionnaire with the exception of housing questions.

A Form 2D contains the same questions as the Form 2B but is used to enumerate northern areas and most Indian reserves, Indian settlements, Indian government districts and '*terres réservées.*' In canvasser areas, it is also used to enumerate usual residents of a Hutterite colony.

A Form 3 is an individual census questionnaire 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. Form 3A is the short version of the questionnaire, and Form 3B is the long version.

1.2 Collection methods

To ensure the best possible coverage, the country was divided into small geographic areas called collection units (CUs). In the 2006 Census, there were approximately 50,000 collection units.

About 98% of households were enumerated using self-enumeration. Starting May 2, Canada Post delivered a census questionnaire to about 70% of households, with the remaining 30% receiving their questionnaire

from an enumerator. Householders were asked to complete the questionnaire for themselves and for members of their household and return it either online or in the postage paid yellow envelope by May 16, Census Day.

About 2% of households were enumerated using the canvasser method. An enumerator visited a household and completed a questionnaire for the household by a personal interview. This method was normally used in remote and northern areas of the country and on most Indian reserves. It is also used in large urban downtown areas where residents are transient.

For the first time, the 2006 Census offered all households in Canada the option of completing their questionnaire online. Each paper questionnaire had a unique Internet access code printed on the front along with the 2006 Census website address (www.census2006.ca). Respondents needed this access code to complete their questionnaire online. If a questionnaire was completed and returned online, the information was directly submitted into the data processing centre system and was verified for completeness. Approximately 18% of the population responded via the internet.

Some households were enumerated through the Census Help Line (CHL), a free, nationwide, multilingual service that was available to all respondents. The Census Help Line collected census information through a computer-assisted telephone interviewing (CATI) system. Households from which a questionnaire had not been received within an acceptable time frame were contacted individually by enumerators in order to collect their information. CATI was also used when enumerators contacted households for missing responses on their questionnaire.

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 capture of questionnaire data through to the creation of an accurate and complete census database. Considered here are the steps of questionnaire registration, data capture, questionnaire imaging, editing, error correction, coding, imputation and weighting. In the remainder of this chapter, each data processing operation will be summarized.

Automated processes, implemented for the 2006 Census, had to be monitored to ensure that all Canadian residences were enumerated once and only once. The Master Control System was built to control and monitor the process flow. The Master Control System held a master listing of all the dwellings in Canada (each dwelling was identified with a unique identifier and about two-thirds of the dwellings also had an address). This system was updated on a daily basis with information of each dwelling's status in the census process flow (i.e., delivered, received, processed, etc.). Reports were generated and accessible online to the census managers to ensure that operations were efficient and effective.

2.2 Receipt and registration

Respondents completing paper questionnaires mailed them back to a centralized data processing centre. Canada Post registered their receipt automatically by scanning the barcode on the front of the questionnaire through the transparent portion of the return envelope. The envelopes were then transported to the Data Processing Centre along with a compact disk containing the list of all of the identifiers for the registered questionnaires.

Responses received through the Internet or the Census Help Line telephone interview were received directly by the Data Processing Centre and their receipt registered automatically.

The registration of each returned questionnaire was flagged on the Master Control System at Statistics Canada. About 10 days after Census Day, a list of all of the dwellings for which a questionnaire had not been received was generated by the Master Control System and then transmitted to Field Operations for follow-up. Registration updates were sent to Field Operations on a daily basis to prevent follow-up on households which had subsequently completed their questionnaire, either by telephone or through the Internet.

2.3 Imaging and data capture from paper questionnaires

The 2006 Census was Canada's first census to capture data using automated capture technologies rather than manual keying. There were 5 steps in the imaging process:

- **Document preparation:** mailed-back questionnaires were removed from envelopes and foreign objects, such as clips and staples, were detached in preparation for scanning. Forms that were in a booklet format were separated into single sheets by cutting off the spine.

- **Scanning:** scanning, using 18 high-speed scanners, converted the paper to digital images (pictures).

- **Automated image quality assurance:** an automated system verified the quality of the scanning. Images failing this process were flagged for rescanning or keying from paper.

- **Automated data capture:** optical mark recognition and optical character recognition technologies were used to extract respondents' data from the images. Where the systems could not recognize the handwriting with sufficient accuracy, data repair was done by an operator.

- **Check-out:** as soon as the questionnaires were processed successfully through all of the above steps, the paper questionnaires were checked out of the system. Check-out is a quality assurance process that ensures the images and captured data are of sufficient quality that the paper questionnaires are no longer required for subsequent processing. Questionnaires that had been flagged as containing errors were pulled at check-out and reprocessed.

2.4 Coverage edits

At this stage, a number of automated edits were performed on the respondent data. These edits were designed to detect cases where invalid persons may have been created either due to respondent error or data capture error. Examples include data erroneously entered in a blank person column, crossed off data that was captured in error, or data provided for the same person more than once, usually due to the receipt of duplicate forms (e.g., a husband completed the Internet version and his wife filled in the paper form and mailed it back). The edits were also designed to detect the possible absence of usual residents, when data are not provided for every household member listed at the beginning of the questionnaire.

Data from questionnaires that failed the edits were forwarded to processing clerks for verification. An interactive system enabled the clerks to examine the captured data and compare them with the image if available (online questionnaires would not have an image). Edit failures were resolved by manually deleting invalid or duplicate persons and adding missing ones (i.e., creating blank person records), as necessary and appropriate.

2.5 Completion edits and failed edit follow-up

Following the coverage edits, another set of automated edits was run to detect cases where there were either too many missing responses, or there were indications that data may not have been provided for all usual residents in the household. Households failing these edits were transmitted to the Census Help Line for follow-up. An interviewer telephoned the respondent to resolve any coverage issues and to fill in the missing information, using a computer-assisted telephone interviewing application. The data were then sent back to the Data Processing Center for reintegration into the system for subsequent processing.

2.6 Coding

The long-form questionnaires (2B, 2C, 2D and 3B) contained questions where answers could be checked off against a list, as well as questions requiring a written response from the respondent in the boxes provided. These written responses underwent automated coding to assign each one a numerical code, using Statistics Canada reference files, code sets and standard classifications. Reference files for the automated match process were built using actual responses from past censuses. Specially trained coders and subject-matter specialists resolved cases where a code could not be automatically assigned. The variables for which coding applied were: Relationship to Person 1, Place of birth, Citizenship, Non-official languages, Home language, Mother tongue, Ethnic origin, Population group, Indian band/First Nation, Place of residence 1 year ago, Place of residence 5 years ago, Major field of study, Location of study, Place of birth of parents, Language at work, Industry, Occupation and Place of work.

About 37 million write-ins were coded from the 2006 long-form questionnaires. An average of about 82% of these were coded automatically.

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

2.7 Classification and non-response adjustments for unoccupied and non-response dwellings

The Dwelling Classification Survey (DCS) was used to estimate the error rates in classifying dwellings in the self-enumerated collection areas as occupied or unoccupied in the field. Based on this information, adjustments were made to the census database. The DCS selected a random sample of 1,405 self-enumerated CUs that were revisited in July and August 2006 to reassess the occupancy status as of census day for each dwelling for which no response had been received. The DCS found that 17.4% of the 934,564 dwellings classified as unoccupied were actually occupied and that 29.1% of the 366,527 dwellings with no responses that were classified as occupied or with occupancy status classified as unknown were actually unoccupied. Estimates based on the DCS sample were used to adjust the occupancy status for individual dwellings. This resulted in an increase of 3.6% in the number of occupied dwellings, and a decrease of 5.2% in the number of unoccupied dwellings at the Canada level.

After this adjustment of the occupancy status by the DCS, occupied dwellings with total non-response had the number of usual residents (if not known) and all the responses to the census questions imputed by borrowing the unimputed responses from another household within the same CU that had its type of questionnaire (long or short). This process, called whole household imputation (WHI), imputed 96% of the total non-response households. The other 4% of the total non-response households where no donor household was found under the WHI process were imputed as part of the main edit and imputation (E & I) process. Utilizing a single donor under WHI was more efficient computationally and was less likely to produce implausible results than using several donors as part of the main E & I process, as was done in 2001.

More details on the DCS and the whole household imputation procedure can be found in the 2006 Census Technical Report on Coverage, Catalogue no. 92-567-XWE.

2.8 Edit and imputation

The data collected in any survey or census contain some omissions or inconsistencies. For example, a respondent might be unwilling to answer a question, fail to remember the right answer, or misunderstand the question. Also, census staff may code responses incorrectly or make other mistakes during processing.

The final clean-up of data, done in the edit and imputation process, was for the most part fully automated. Two types of imputation were applied. The first type, called 'deterministic imputation,' involved assigning specific values under certain conditions. Detailed edit rules were applied to identify these conditions, and then the variables involved in the rules would be assigned a pre-determined value. The second type of imputation, called 'minimum-change donor imputation,' applied a series of detailed edit rules that identified any missing or inconsistent responses. These missing or inconsistent responses were corrected by changing as few variables as possible. For minimum-change donor 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 missing or inconsistent responses. The CANadian Census Edit and Imputation System (CANCEIS) was the automated system used for nearly all deterministic and hot-deck donor imputation in 2006.

2.9 Weighting

Questions on age, sex, marital status, mother tongue and relationship to Person 1 were asked of 100% of the population, as in previous censuses. However, the bulk of census information was acquired on a 20%

sample basis, using the additional questions on the 2B questionnaire. Weighting was used to project the information gathered from the 20% sample to the entire population.

For the 2006 Census, weighting employed the same methodology used in the 2001 Census, known as calibration estimation. This began by first assigning initial weights of approximately 5 to the sampled households. These weights were then adjusted by the smallest possible amount needed to ensure closer agreement between the sample estimates and the population counts for a number of characteristics related to age, sex, marital status, common-law status and household size (e.g., number of males, number of people aged 15 to 19). 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[1] of census figures in a canvasser census. Thus the use of self-enumeration was expected to reduce the variance 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 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,

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

relationship to head of household, sex, date of birth, marital status, mother tongue, type of dwelling, tenure, number of rooms, water supply, toilet facilities, and certain census 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]).

The content of the 1976 Census was considerably less than that of the 1971 Census. Furthermore, the 1976 questionnaire 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, response burden, and improved timeliness (see Royce [1983]). The one-in-five sampling rate has been maintained for every census since 1981.

3.2 The sampling scheme used in the 2006 Census

A wealth of information was collected from everyone in Canada on Census Day, May 16, 2006. The bulk of the information was acquired on a sample basis. In all self-enumeration areas, a one-in-five sample of occupied private dwellings was selected to receive a long questionnaire (Form 2B) while the non-sampled occupied private dwellings received a short questionnaire (Form 2A). Basic questions on age, sex, marital status, common-law status, mother tongue, relationship to the household reference person (Person 1) were asked of all respondents, as well as the type of dwelling. Additional information on the dwelling, plus socio-economic questions, was asked on a sample basis. The usual residents of an occupied dwelling are called a household, so the terms household and occupied dwelling will be used interchangeably in this report.

All dwellings in those areas enumerated in person by the canvasser method (generally remote areas or Indian reserves) received the Form 2D.

Most persons in collective dwellings received a long form (usually a Form 3B except for those in Hutterite colonies and senior units that received a Form 2B). The following persons in collective dwellings, however, were not asked the sample questions (i.e., a form 3A was used):

(a) inmates in correctional and penal institutions or jails

2. A collective dwelling is a dwelling used for commercial, institutional or communal purpose, such as hotels, hospitals and work camps.

(b) patients in general hospitals, special care homes, and collective dwellings or institutions for senior citizens, the chronically ill or psychiatric institutions

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

(d) people in shelters.

A senior unit is an accommodation within a collective residence for senior citizens that contains one or more senior citizens judged capable of completing a census form. In the 2001 Census, as described in point (b) above, these persons were not asked the sample questions. New for 2006, a random sample of 1 in 5 of these senior units was selected and they were provided with a long form to complete. The other senior units were provided with a short form to complete. These senior units were treated as if they were private dwellings that had been subject to sampling by the 2006 Census weighting system, and are included in the various tables provided in this report. There were 40,755 senior units containing 47,540 persons in 2006. Note, however, that a senior unit is not a household or a dwelling; the seniors' residence facility as a whole is one household and one dwelling. Furthermore, these numbers are lower than the true counts because of problems during collection and processing, where some residences for senior citizens were misclassified as nursing homes. As a result, data quality for senior units alone is poor, and should be used with caution.

Section 1.2 discusses the various methods of census collection. Each dwelling, regardless of collection method, was assigned a Visitation Record (VR) number. The VR number was used to determine which type of census questionnaire (i.e., Form 2A or 2B) would be delivered. Every fifth dwelling was selected to be part of the census sample, and was given a long form. The remaining four-fifths were given a short form.

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 (CUs). 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. Sections 4.1 and 4.2 describe the operational and theoretical considerations relevant to the choice of estimation procedures for the census sample. Sections 4.3 and 4.4 discuss some of the methodology used in developing the census weights. The remaining sections introduce the data universes used in the weighting process, and briefly discuss why discrepancies may occur between population counts and weighted estimates.

4.1 Operational considerations

Mathematically, an estimation procedure can be described by an algebraic formula, or estimator, that shows how the estimate 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. For example, a separate application of the estimation formula would be required for every cell of every published census tabulation based on sample data. In addition, the calculation of each estimate separately would not necessarily lead to consistency between the various estimates made from the same census sample.

Therefore, the approach taken in the census (and in many sample surveys) is to split the estimation procedure into two steps: (a) the calculation of weights (known as the weighting procedure) and (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., 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. These differences are discussed further in Section 4.6 of this report.

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 dwellings (excluding collective dwellings 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 'post-strata.'

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

For example, if there were 5,000 males aged 20 to 24 enumerated in Prince Edward Island, and if 1,020 of these fell in the sample of dwellings, 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 the 1991 Census, 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 estimator (GREG) was introduced. It achieved a higher level of agreement between population counts and the corresponding estimates at the enumeration area (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. The two-step GREG estimator was also used in 1996.

In 2001 and 2006, a pseudo-optimal regression estimator was used because it typically gave slightly better agreement between the population counts and estimates than the GREG, while ensuring the calibrated weights were all equal to one or more. See Bankier (2002) for a more detailed comparison of the regression estimators.

With the Pseudo-optimal Regression estimator, the initial weights of approximately 5 were adjusted as little as possible for individual dwellings such that there was perfect agreement between the estimates and the population counts for 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. More information on WAs is given in Section 7.1 of this report.

In 2006, Canada was divided into approximately 50,000 collection units to be used in the collection of census data. The collection unit (CU) is similar in size and has similar attributes to the enumeration area (EA) used prior to the 2006 Census. A one-in-five systematic sample of dwellings was selected from most CUs to be used in the census weighting process. Dissemination areas (DAs) are another geographic level similar in size to CUs. Entire DAs were combined to form WAs. On average there are eight DAs and seven sampled CUs in a WA.

4.4 The two-step Pseudo-optimal Regression estimator

There are 34 auxiliary variables used in the regression process. These include five-year age ranges, marital status, common-law status, sex, household size, and dwelling type. See Appendix B for the 34 auxiliary variables. The objectives for the 2006 Census weighting procedure are:

(a) To have **exact** population/estimate agreement at the WA level for as many of the 34 auxiliary variables as possible.

(b) To have **approximate** population/estimate agreement for the larger DAs for the 34 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 to 25 inclusive. A lower bound of 1 is required because it is felt that each sampled person should, at minimum, represent themselves

(e) the method to generate weights be highly automated since the 6,602 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 by using an automated weighting system. For each WA being processed, a set of user-defined parameters are passed to the system. An initial weight is assigned to each sampled private household in the WA, and these weights have either two or three weighting adjustment factors applied to them. First of all, households are sometimes post-stratified 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 initial weights and the three adjustments have been described in more detail.

First, an initial CU-level weight is assigned to each private household in the WA. The weight is equal to the number of private households in the CU divided by the number of private households that were sampled in that CU. Since approximately one in five households would be sampled, initial weights tend to be near five. In 2001, senior units were not part of the census weighting process, and were excluded from the sampling process. However, in 2006, senior units were treated similarly to private households, so they made up part of the sampling frame. Since the proportion of senior units in any CU was usually very small, they typically had little effect on the weighting results. However, for a small number of CUs where there were a high proportion of senior units, the private households and the senior residences were treated as two distinct populations, so two sets of initial weights were calculated for each of these CUs in order to reduce sampling bias. Once the initial weights were created, senior units were treated no differently than private households throughout the remainder of the weighting process. When the standard error adjustment factors of Chapter 9 were calculated, however, a CU where the private households and senior units were treated as two distinct populations were considered as two sampling strata rather than one.

In the first adjustment step, households are sometimes post-stratified based on household size (1, 2, 3, 4, 5, or 6+ persons) at the WA level. The initial weights are multiplied by a factor to generate the post-stratified weights. For example, based on the post-stratified 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 post-stratified weight is constrained to ensure that it lies within the range 1 to 20 inclusive. An upper limit of 20 rather than 25 is used to give some 'room' for further adjustment.

Next, a first-step regression weighting adjustment factor is calculated at the DA level. The 34 auxiliary variables (age, sex, marital status, household size, and dwelling type) that are to be applied at the WA level in the second adjustment 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 17 constraints go into a second group. The resulting weighting adjustment factors for each group of constraints are averaged together and applied to the post-stratified weights (or the initial weights if post-stratification was not done). Population/estimate differences at the DA level for the 34 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 34 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 34 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)
- they cause outlier weights (weights outside the range 1 to 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:

- 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 less than or equal to the SMALL parameter (which equalled 20, 30 or 40 households in 2006) is discarded since estimates, for small constraints, tend to be very unstable.

- Next, LD constraints are discarded.

- 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 collinearity 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 2006), 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 2006), 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 2006, MAXC always equaled 10 times the value of COND.

- 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 to 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 (see Press et al., 1992) is used to identify which constraints should be dropped.

Finally, 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 to 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., SMALL, COND, MAXC, whether to post-stratify, whether to use dwelling type constraints). In 1996, an identical set of parameters was used for every WA in the country. In 2001 and 2006, 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 several different sets of parameters. Two dwelling type constraints were introduced in 2006 due to large discrepancies observed for these characteristics in certain regions in 2001. These constraints were single

detached dwellings and apartments in buildings with less than five storeys. Although the introduction of new constraints may reduce the discrepancies for these characteristics, it may result in other constraints being dropped in their place, which would result in a larger discrepancy for those other characteristics. Therefore, the use of the dwelling type constraints was parameterized so they would only be used in WAs where they had an overall positive effect on the discrepancies. Twenty different sets of parameters were used to calculate the weights in each WA in 2006. These represented the 10 sets of parameters used in 2001 with the dwelling type constraints excluded (as in 2001) and included. A statistic was calculated for each set of parameters to determine which set 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. 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).

Regression weights are calculated only for sampled-CU private households and sampled senior units that have received the long census questionnaire (one-fifth of these were sampled; four-fifths were not). Sampled-CU private households and senior units that received a short questionnaire receive a weight of 0 because they contain no information on sample variables. All non-sampled CU private households and senior units receive a weight of 1 since 100% of the respondents in these areas provide information on Form 2B or 2D. Collective households also receive a weight of 1. In this report, the term 'household' will refer to a private household or a senior unit unless otherwise specified.

4.5 Two-pass processing

For the 1996, 2001 and 2006 censuses, short 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 (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, called '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, called '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 2 data for the sampled population and Pass 1 results for the remaining 80% of the population) and Pass 2 estimates is small for most constraints. See Table 7.2.2.2, Chart 7.2.2.3, and Chart 7.2.2.4 in Section 7.2.2 for a comparison of Pass 1 and Pass 2 results.

4.6 Differences between population counts and final weighted estimates

Final household weights are generated such that the population counts match the weighted estimates for as many characteristics as possible. Characteristics available from both the long and short form for which consistency is attempted include five-year age ranges, sex, marital status, common-law status, household size and dwelling type. The weighting process attempts to control the population/estimate differences at the weighting area (WA) level where WAs typically contain 1,000 to 3,000 dwellings that are subject to sampling.

There are a few reasons why sample estimates may be different from population counts, particularly for small areas. The main ones are listed below:

(1) Constraints dropped during the regression process: As described in Section 4.4, constraints can be dropped for generating outlier weights, having small counts, or by being linearly dependent or nearly linearly dependent. Constraints which are dropped are not controlled on, and will usually have some difference between the population counts and the estimates.

(2) Sub-WA areas: The weighting area is the smallest geographic area for which the weighting system attempts to have agreement between the population counts and weighted estimates for as many auxiliary variables as possible. Therefore, small areas that are contained within WAs (such as DAs or very small municipalities) will usually see discrepancies between the population counts and the weighted estimates.

4.7 Different universes

There are three separate universes for which the census data may be observed:

(1) Private dwellings – Consists of private households and senior units that were subject to sampling. These households were used in the creation of the final household level weights. The majority of the information that is presented in this publication represents the private dwelling universe.

(2) Private dwellings and non-institutional collectives – Consists of sampled private households and senior units, non-institutional collectives, and also private households and senior units from non-sampled CUs. The additional persons in this universe all received a long form questionnaire, and therefore have 2B data present. This universe is used for all census publications related to sampled variables.

(3) Private and collective dwellings – Consists of all private households and senior units (sampled and non-sampled) and all collectives (institutional and non-institutional). Residents of institutional collectives answer the short form questionnaire and therefore do not have sampled data available. For this reason, this complete universe is used for all census publications related to basic variables (questions asked on both the short and long forms) but cannot be used for sampled publications.

The institutional collective's population represents some of the differences that will be observed by someone comparing a 2B publication to a 2A publication. The counts and estimates for the three universes discussed above can be found in Table 7.2.2.3.

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) determining the standard errors for various 20% sample characteristics.

Each of these studies is described briefly below, with their results being presented in Chapters 6, 7, 8 and 9.

5.1 Sampling bias

This study assessed differences between estimates based on initial weights and known population counts. These differences are of interest for two reasons: first, their possible usefulness in identifying biases in the census household sample selected in the field; and second, they may indicate a possible negative impact of non-response on census sample questions. Biases in short form characteristics are corrected through calibration during the weighting procedure. 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 Pseudo-optimal Regression estimator. This was done by examining the level of agreement between sample estimates (based on the final weights) and population counts for all the WA-level constraints. Any inconsistencies were explained through assessment of the number and type of constraints discarded at the WA level and the reasons for their being discarded. In addition, the distribution of census weights was studied.

5.3 Sample estimate and population count consistency

This study examined the level of agreement between sample estimates (based on the final weights) 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
for short form characteristics such as 'Marital status = single,' where the census population count X can be compared to the sample estimate $\hat{X}^{(0)}$ 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 and unbiased, 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 the 2001 and 2006 censuses (along with the differences $\hat{X}^{(0)} - X$) for 34 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 $\left|Z^{(0)}\right| > 3$ is approximately 0.0026 for one characteristic. This suggests that, on average, we would expect to see 0.0026 x 34 = 0.0884 of the 34 characteristics having $\left|Z^{(0)}\right| > 3$. However, according to Table 6.1, 22 of the 34 characteristics in 2006 have a Z statistic outside the range of –3 to 3. This provides strong evidence that the 2006 Census sample is biased. Similarly in 2001, 25 of the 32 characteristics were outside that range.

Chart 6.1 shows that for many characteristics the Z statistic is much different between 2001 and 2006. Since Z is a random variable, some of these differences may not be statistically significant. A W statistic, which is defined in Appendix C, was calculated for each characteristic to determine whether or not the Z statistics from 2001 and 2006 were significantly different. The W statistic, its p-value, and 17 characteristics with statistically significant differences in the Z values (because their p-values are less than 0.05) are identified in Table 6.1 as well as in Chart 6.1.

In Chart 6.1, it can be seen that the downward bias in the sample in 2006 increased significantly (as flagged by an asterisk) for males, males age 15 and over, persons aged 15 to 19, single persons, 3-person households and 6+-person households while the upward bias in the sample in 2006 increased significantly for 5-person households. In addition, it can be seen that the downward bias in the sample in 2006 decreased significantly for separated persons and 1-person households while the upward bias in the sample in 2006 decreased significantly for females, females aged 15 and over, persons aged 10 to 14, married persons and 4-person households. Finally, the upward bias in the sample in 2006 changed significantly to a downward bias for the total population count for the age group 5 to 9 and the age group 40 to 44.

Chart 6.1 also shows a consistent downward bias in the 2001 and 2006 census samples of persons aged 20 to 39 and a consistent upward bias in the 2001 and 2006 census samples of persons aged 45 and over and for 2-person households.

Chart 6.1 also shows a very large upward bias in 2006 for single-detached dwellings and a somewhat smaller downward bias for apartments of less than 5 storeys.

Bias in the sample can originate from a variety of sources, including enumerator 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 they would respond on Form 2A), processing errors, and so on.

The large biases in the 2006 sample for 5-person and 6+-person households were the result of reducing the number of persons on the long form from six persons to five persons because more space was required to allow the automated data capture of write-in responses. The number of persons on the short form remained at six. In 2006, sometimes households with more than five persons who received a long form did not request a second long form and only listed five persons as living in the household. This caused the large increase in the upward bias in the sample for five-person households and a corresponding large increase in the downward bias in the sample for six-person households. The weighting calibration process was only able to partially correct for these biases, and these biases also made it more difficult for the calibration to correct for other biases.

Another possible source of bias in the census sample was non-response. The percentage of households with no responses at the end of field operations was 2.8% in 2006 compared to 1.6% in 2001. After adjustments were done to the occupancy status by the Dwelling Classification Survey (see Section 2.7), the percentage of occupied dwellings with no responses was 3.5% in 2006 compared to 2.0% in 2001. In 2006, whole household imputation was used to impute for 96% of these total non-response households with 18.6% of them becoming long forms. In 2001, whole household imputation was not done. In both 2001 and 2006, long forms with total non-response to the questions asked on a sample basis were converted to short forms. This process was called 'Document Conversion.' In 2006, 12,638 long-form households were converted to short forms. In 2001, 17,692 long-form households with some short form responses, but no long-form responses were converted to short forms, while 144,282 total non-response households (of which approximately 20% would be expected to have originally been long forms) become short forms. The much smaller number of long forms converted to short forms in 2006 was the result of most total non-response households being dealt with by whole household imputation.

This change was made because the 2001 approach may have introduced significant biases into the sample. For example, in 2001 it was known that the percentage of single-detached dwellings that were total non-response households was half that of the population as a whole. See Section 7.2.2 for a more detailed discussion of the impact on sampling bias of the introduction of whole household imputation in 2006. The discussion in Section 7.2.2 casts some doubt on the utility of using the W statistic above to determine whether or not the Z statistics from 2001 and 2006 were significantly different. This is because many of the differences appear to be the result of the introduction of whole household imputation rather than because of sampling variability. The use of the W statistic to test for regional differences in the bias for 2006 below, however, is not affected by this concern.

A third possible source of bias comes from errors that were either made by the respondent or introduced by the data capture process. Some of the inconsistencies that resulted were detected and corrected by the edit and imputation process described in Section 2.8.

The geographic variation of the bias was also studied. The Z statistics for all 34 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 2006 and 2001 censuses in Chart 6.2 and Chart 6.3 respectively. Again, using the W statistic, regional differences which are statistically significant are flagged by placing the initials of the regions at either the bottom or the top of the chart. For example, WQ and OQ indicate that there is a significant difference in the bias between the West and Quebec as well as between Ontario and Quebec.

Comparing Chart 6.2 to Chart 6.3, it can be seen that there were more significant regional differences in 2006 than in 2001. It is interesting to note for 2006 that the downward bias for the total population count is much larger for the West and Ontario than for Quebec and the East. It is also interesting to note that for females there is a downward bias in the sample for the West and Ontario and an upwards bias for Quebec and the East.

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 in 2006 and 2001 censuses based on initial weights

	2006 Census						2001 Census						2006 vs 2001		
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]	W	p-value	Significant bias difference[6]
Male	15,041,551	14,962,170	-79,381	-0.53	6,273	-12.65	14,171,941	14,146,867	-25,074	-0.18	6,139	-4.08	-5.83	0.000	Y
Female	15,653,041	15,653,964	923	0.01	6,095	0.15	14,699,518	14,772,915	73,397	0.50	5,940	12.36	-8.79	0.000	Y
Total persons	30,694,592	30,616,134	-78,458	-0.26	9,227	-8.50	28,871,459	28,919,783	48,324	0.17	8,991	5.37	-9.77	0.000	Y
Male ≥ 15 years old	12,263,445	12,187,117	-76,328	-0.62	4,917	-15.52	11,340,286	11,295,995	-44,291	-0.39	4,747	-9.33	-4.00	0.000	Y
Female ≥ 15 years old	13,005,067	13,007,175	2,108	0.02	4,557	0.46	11,998,509	12,042,929	44,420	0.37	4,342	10.23	-7.03	0.000	Y
Age 0 to 4	1,640,859	1,639,505	-1,354	-0.08	2,987	-0.45	1,636,092	1,641,720	5,628	0.34	2,986	1.88	-1.65	0.098	N
Age 5 to 9	1,760,005	1,756,879	-3,126	-0.18	3,067	-1.02	1,910,359	1,928,604	18,245	0.96	3,213	5.68	-4.68	0.000	Y
Age 10 to 14	2,025,216	2,025,458	242	0.01	3,296	0.07	1,986,213	2,010,534	24,321	1.22	3,271	7.44	-5.24	0.000	Y
Age 15 to 19	2,083,373	2,070,265	-13,108	-0.63	3,315	-3.95	1,986,163	1,983,519	-2,644	-0.13	3,245	-0.81	-2.18	0.030	Y
Age 20 to 24	2,029,449	1,978,067	-51,382	-2.53	3,267	-15.73	1,892,572	1,851,491	-41,081	-2.17	3,168	-12.97	-1.56	0.120	N
Age 25 to 29	1,940,880	1,905,221	-35,659	-1.84	3,142	-11.35	1,835,744	1,810,124	-25,620	-1.40	3,077	-8.33	-1.90	0.058	N
Age 30 to 34	1,976,478	1,956,416	-20,062	-1.02	3,131	-6.41	2,031,513	2,013,625	-17,888	-0.88	3,173	-5.64	-0.60	0.545	N
Age 35 to 39	2,161,430	2,148,991	-12,439	-0.58	3,224	-3.86	2,452,299	2,446,624	-5,675	-0.23	3,427	-1.66	-1.68	0.092	N
Age 40 to 44	2,559,477	2,541,482	-17,995	-0.70	3,467	-5.19	2,510,847	2,513,920	3,073	0.12	3,439	0.89	-4.29	0.000	Y
Age 45 to 49	2,571,429	2,579,715	8,286	0.32	3,462	2.39	2,273,676	2,283,700	10,024	0.44	3,286	3.05	-0.60	0.548	N
Age 50 to 54	2,313,657	2,327,389	13,732	0.59	3,305	4.15	2,031,050	2,041,054	10,004	0.49	3,137	3.19	0.48	0.631	N
Age 55 to 59	2,045,868	2,060,482	14,614	0.71	3,139	4.66	1,549,675	1,567,071	17,396	1.12	2,758	6.31	-1.74	0.082	N
Age 60 to 64	1,558,145	1,570,085	11,940	0.77	2,751	4.34	1,234,930	1,249,389	14,459	1.17	2,469	5.86	-1.52	0.129	N
Age 65 to 74	2,229,023	2,248,564	19,541	0.88	3,371	5.80	2,059,079	2,083,362	24,283	1.18	3,256	7.46	-1.38	0.167	N
Age 75+	1,799,303	1,807,615	8,312	0.46	2,927	2.84	1,481,247	1,495,045	13,798	0.93	2,676	5.16	-1.93	0.053	N
Single	14,170,280	13,989,389	-180,891	-1.28	8,347	-21.67	13,282,845	13,196,174	-86,671	-0.65	8,018	-10.81	-7.40	0.000	Y
Married	12,291,457	12,406,423	114,966	0.94	6,942	16.56	11,750,092	11,906,204	156,112	1.33	6,678	23.38	-4.91	0.000	Y
Widowed	1,435,852	1,431,006	-4,846	-0.34	2,321	-2.09	1,341,497	1,339,109	-2,388	-0.18	2,254	-1.06	-0.68	0.494	N
Divorced	2,044,164	2,040,145	-4,019	-0.20	2,989	-1.34	1,794,079	1,784,704	-9,375	-0.52	2,824	-3.32	1.52	0.129	N
Separated	752,839	749,172	-3,667	-0.49	1,800	-2.04	702,946	693,591	-9,355	-1.33	1,749	-5.35	2.45	0.014	Y
Common-law = yes	2,725,161	2,703,240	-21,921	-0.80	4,419	-4.96	2,267,634	2,253,253	-14,381	-0.63	4,090	-3.52	-0.70	0.483	N
1-person hhld	3,338,596	3,329,891	-8,705	-0.26	3,017	-2.89	2,908,857	2,866,182	-42,675	-1.47	2,847	-14.99	9.06	0.000	Y
2-person hhld	4,153,415	4,182,506	29,091	0.70	3,368	8.64	3,709,282	3,739,781	30,499	0.82	3,224	9.46	-1.02	0.305	N

Table 6.1 Population/estimate differences in 2006 and 2001 censuses based on initial weights (continued)

	2006 Census						2001 Census						2006 vs 2001		
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]	W	p-value	Significant bias difference[6]
3-person hhld	1,963,201	1,951,120	-12,081	-0.62	2,611	-4.63	1,848,476	1,845,071	-3,405	-0.18	2,541	-1.34	-2.25	0.024	Y
4-person hhld	1,843,987	1,848,858	4,871	0.26	2,506	1.94	1,812,783	1,826,921	14,138	0.78	2,481	5.70	-2.67	0.007	Y
5-person hhld	713,994	735,811	21,817	3.06	1,659	13.15	714,618	719,013	4,395	0.61	1,664	2.64	7.42	0.000	Y
6+-person hhld	338,241	302,192	-36,049	-10.66	1,157	-31.16	332,959	328,968	-3,991	-1.20	1,155	-3.46	-19.42	0.000	Y
Single-detached dwelling	6,769,581	6,812,477	42,896	0.63	2,554	16.80	..	..	..	..	..	..	..	..	..
Apartment ≤ 5 storeys	2,285,965	2,258,899	-27,066	-1.18	2,119	-12.77	..	..	..	..	..	..	..	..	..

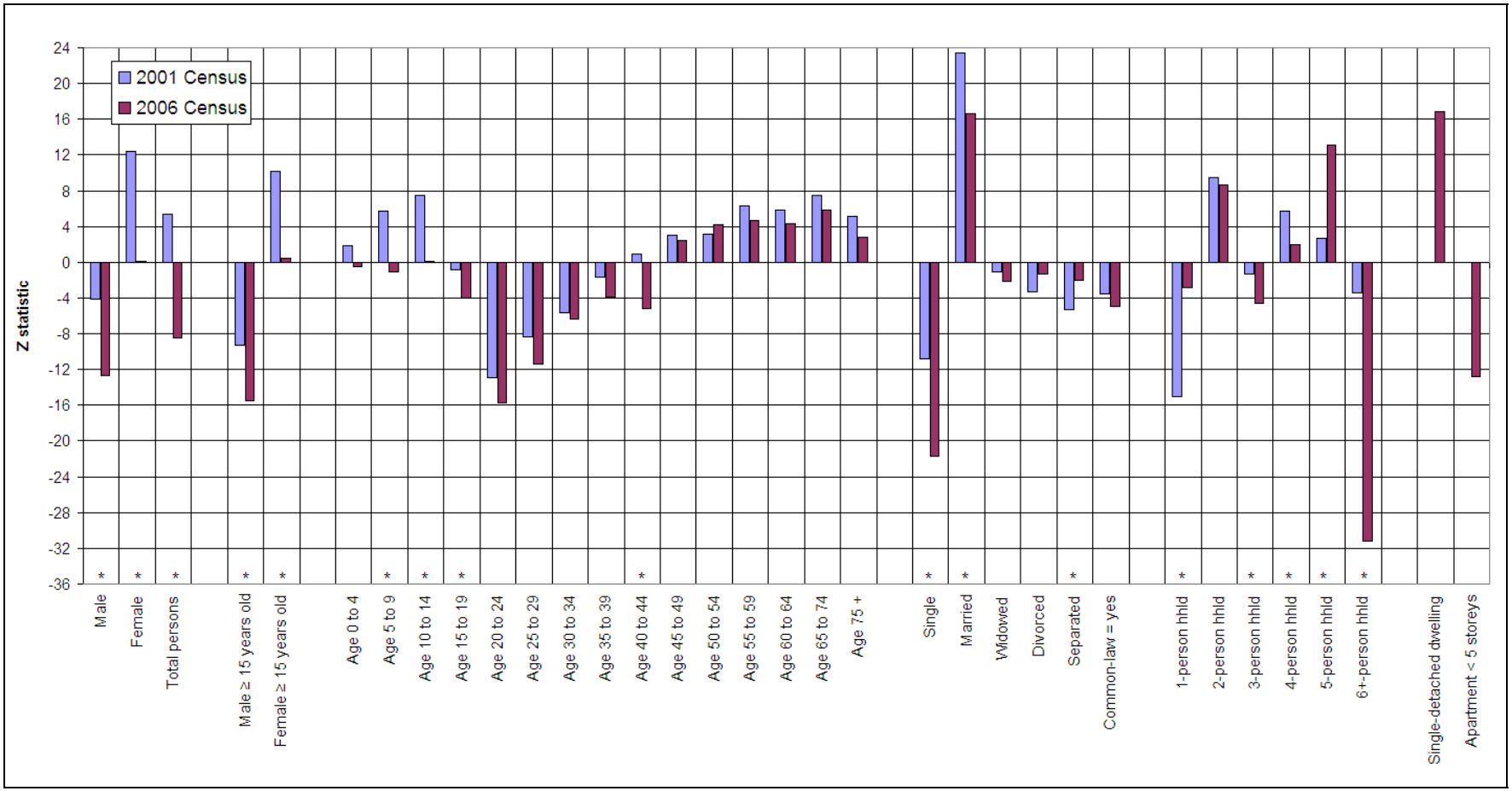
.. not available for a specific reference period

Notes:

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.
6. Indicator of whether initial biases are significantly different between 2006 and 2001 (based on p-value < 0.05).

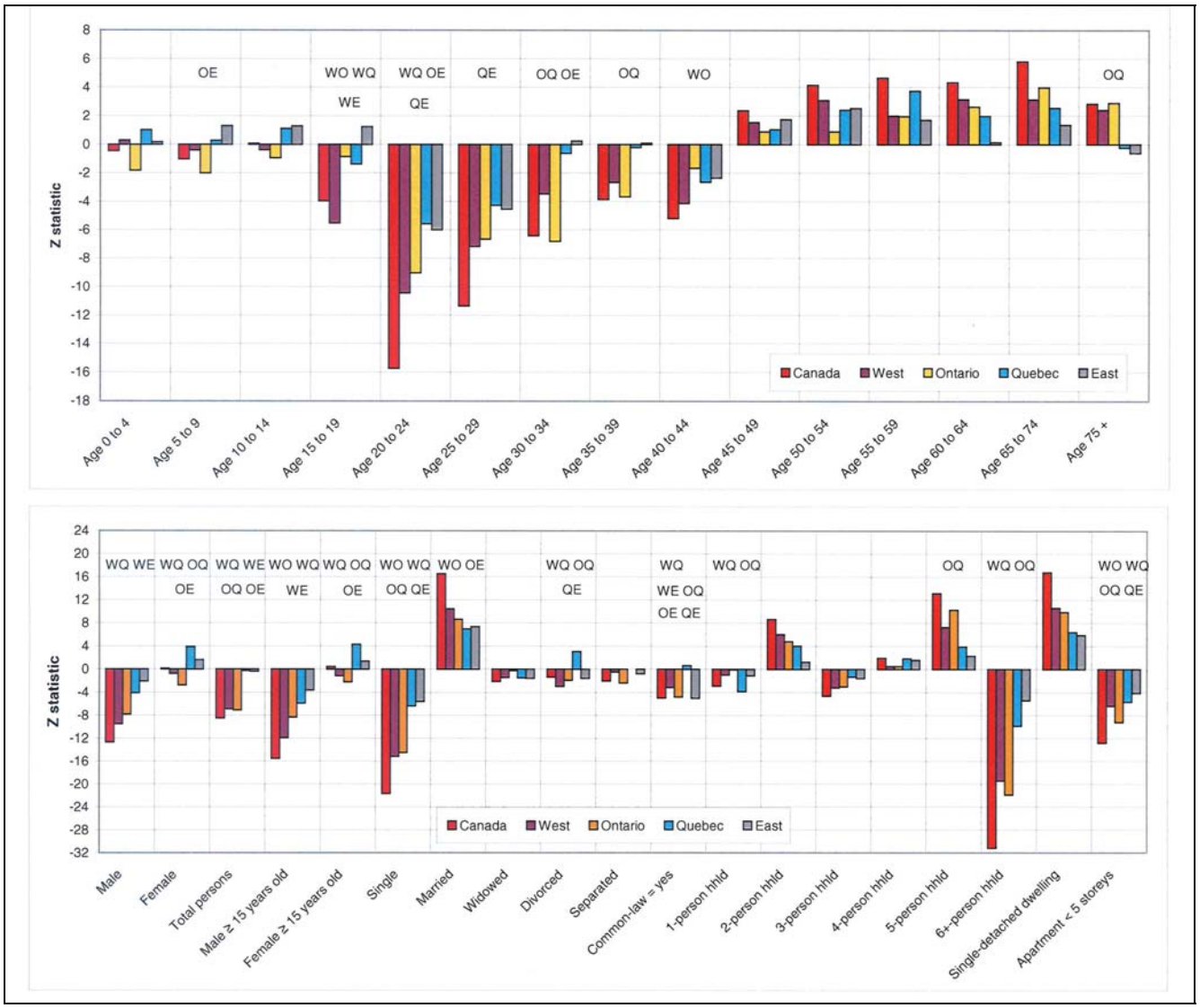
Sources: Statistics Canada, 2006 and 2001 censuses.

Chart 6.1 Z statistics for population/estimate differences based on initial weights, for Canada, 2006 and 2001 censuses



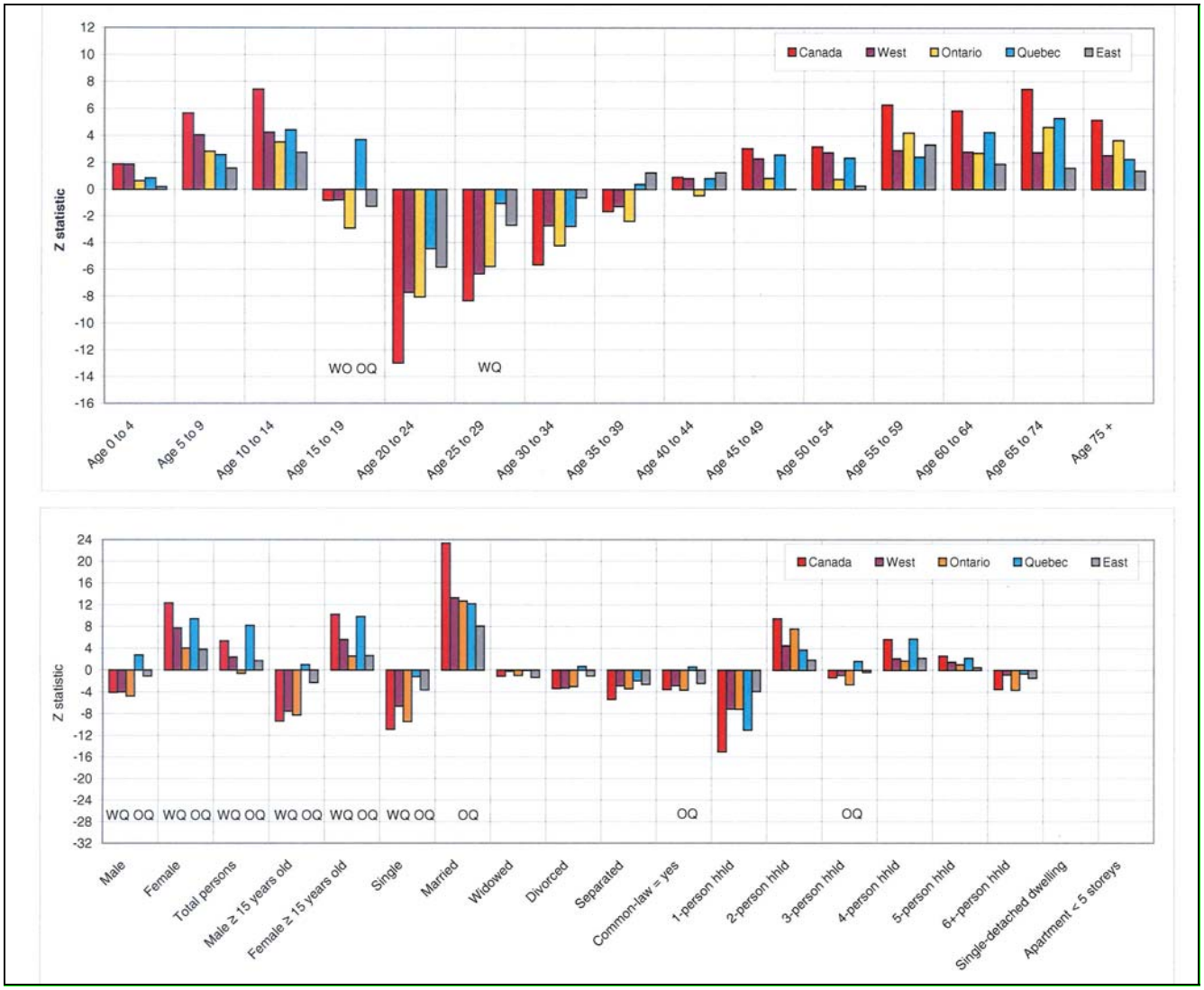
* Indicates a statistically significant difference in the bias between 2006 and 2001.
Sources: Statistics Canada, 2006 and 2001 censuses.

Chart 6.2 Regional Z statistics in 2006



Source: Statistics Canada, 2006 Census.

Chart 6.3 Regional Z statistics in 2001



Source: Statistics Canada, 2001 Census.

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. Also, it examines, for various characteristics, the discrepancies between population counts and sample estimates at the Canada level. It also discusses Pass 1 versus Pass 2 results and the different data universes for which census data may be presented. Finally, it takes a look at the frequency that constraints are discarded and the effect this has on these discrepancies.

7.1 Weighting area (WA) formation

In the 2006 Census, the country was partitioned into 6,607 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 and census tract (CT) 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 shows that 6,559 (99.3%) of the 2006 WAs are within the desired range of 1,000 to 3,000 households in the 2006 Census. This is considerably better than in 2001 when only 94.2% of WAs were within the range. The algorithm that was used to generate WAs in 2006 was the same as in 2001, so the automated results were similar. However, the improvement is due to many more manual adjustments being made at the end of the process in 2006 than were made in 2001. Many of the abnormal WAs were either split, amalgamated, or realigned to better follow the conditions mentioned above.

The average number of dwellings per WA was 1,869. The largest WA contained 4,820 dwellings, an improvement from 2001 when the largest WA contained 17,043 dwellings. In 2006, there were five WAs with zero population. In these cases, the WAs contained DAs that were not subject to sampling. These WAs with zero population are in Labrador, 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. These areas include provinces and CDs, as well as CSDs and CTs in which no WA within them makes up part of another CSD or CT. Table 7.1.2 looks at the relationship between 2006 Census CSD and CT boundaries and WA boundaries. There are four mutually exclusive scenarios possible:

1. 'Geographic areas containing only part of one WA while the rest of the WA contains only complete geographic areas of the same kind' – This means that the CSD or CT was small enough to fit entirely within a WA, and that the same WA only consisted of whole CSDs or CTs. None of the CSDs or CTs

in that WA crossed into a different WA. Therefore condition (c) was satisfied. This scenario occurs frequently for CSDs because there are many very small municipalities such as reservations and villages that contribute little or no population that is subject to sampling.

2. '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' – This means that the CSD or CT was small enough to fit entirely with a WA, but a different CSD or CT within that same WA was shared by a different WA. Condition (c) is not satisfied.

3. 'Geographic areas containing one or more whole WAs' – This means that the CSD or CT was large enough to contain whole WAs. None of the WAs crossed into a different CSD or CT. Therefore, condition (c) was satisfied. This scenario occurs frequently for CTs because CTs occur in urban areas, which are usually subject to sampling, and CTs are designed to be larger than WAs in general.

4. 'Geographic areas that cross at least one WA boundary' – This means that the CSD or CT is shared by at least 2 WAs. Condition (c) is not satisfied.

According to the figures presented in Table 7.1.2, 13.2% of CSDs and 67.0% 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. The results in Table 7.1.2 are very similar to the results from 2001 because the same automated algorithm was used in both censuses.

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

Table 7.1.1 Size distribution of weighting areas

Dwellings	2006 Census		2001 Census	
	WA count	Percentage (%)	WA count	Percentage (%)
0 to 499	7	0.1	7	0.1
500 to 999	2	0.0	1	0.0
1,000 to 1,499	1,705	25.8	1,132	18.4
1,500 to 1,999	2,443	37.0	2,248	36.6
2,000 to 2,499	1,768	26.8	1,622	26.4
2,500 to 2,999	643	9.7	786	12.8
3,000 to 3,499	30	0.5	352	5.7
3,500+	9	0.1		
Total	**6,607**	**100.0**	**6,148**	**100.0**

Sources: Statistics Canada, 2006 and 2001 censuses.

Table 7.1.2 Number of census subdivisions and census tracts that respect weighting area boundaries, 2006 Census

Description	CSD		CT	
	Number	Percentage (%)	Number	Percentage (%)
Geographic areas containing only part of one WA while the rest of the WA contains only complete geographic areas of the same kind	4,255	78.5	1,466	28.8
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	337	6.2	115	2.3
Geographic areas containing one or more whole WAs	714	13.2	3,409	67.0
Geographic areas that cross at least one WA boundary	112	2.1	99	1.9
Total	**5,418**	**100.0**	**5,089**	**100.0**

Source: Statistics Canada, 2006 Census.

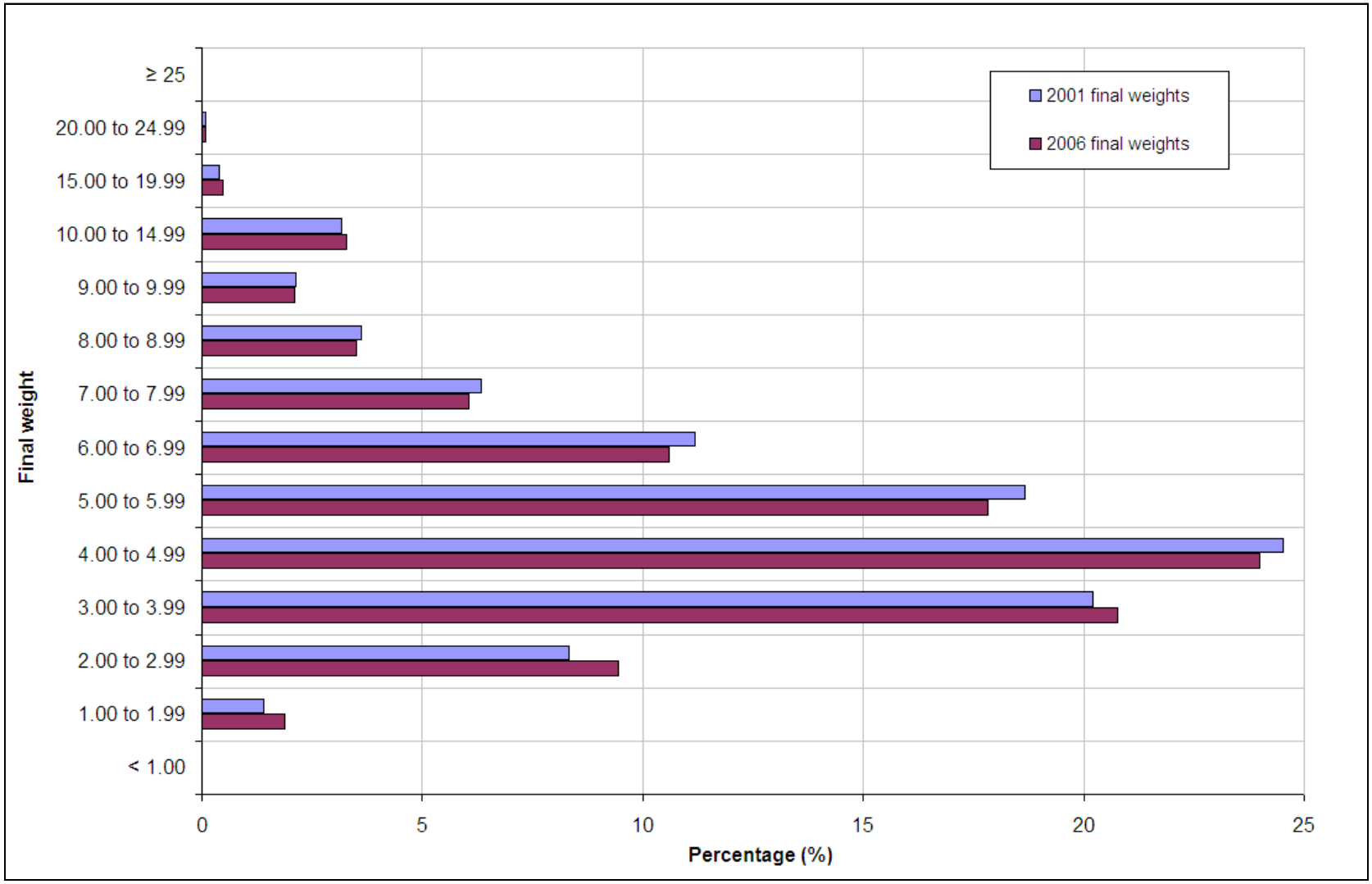
7.2 Evaluation of the census weighting methodology

7.2.1 Distribution of weights

Chart 7.2.1.1 compares the 2006 final weight distribution to that of 2001. The distributions are almost identical, but the chart shows that there were slightly more households with weights less than 4 in 2006 than there were in 2001. Conversely, there were fewer households with weights between 4 and 9 in 2006 than in 2001.

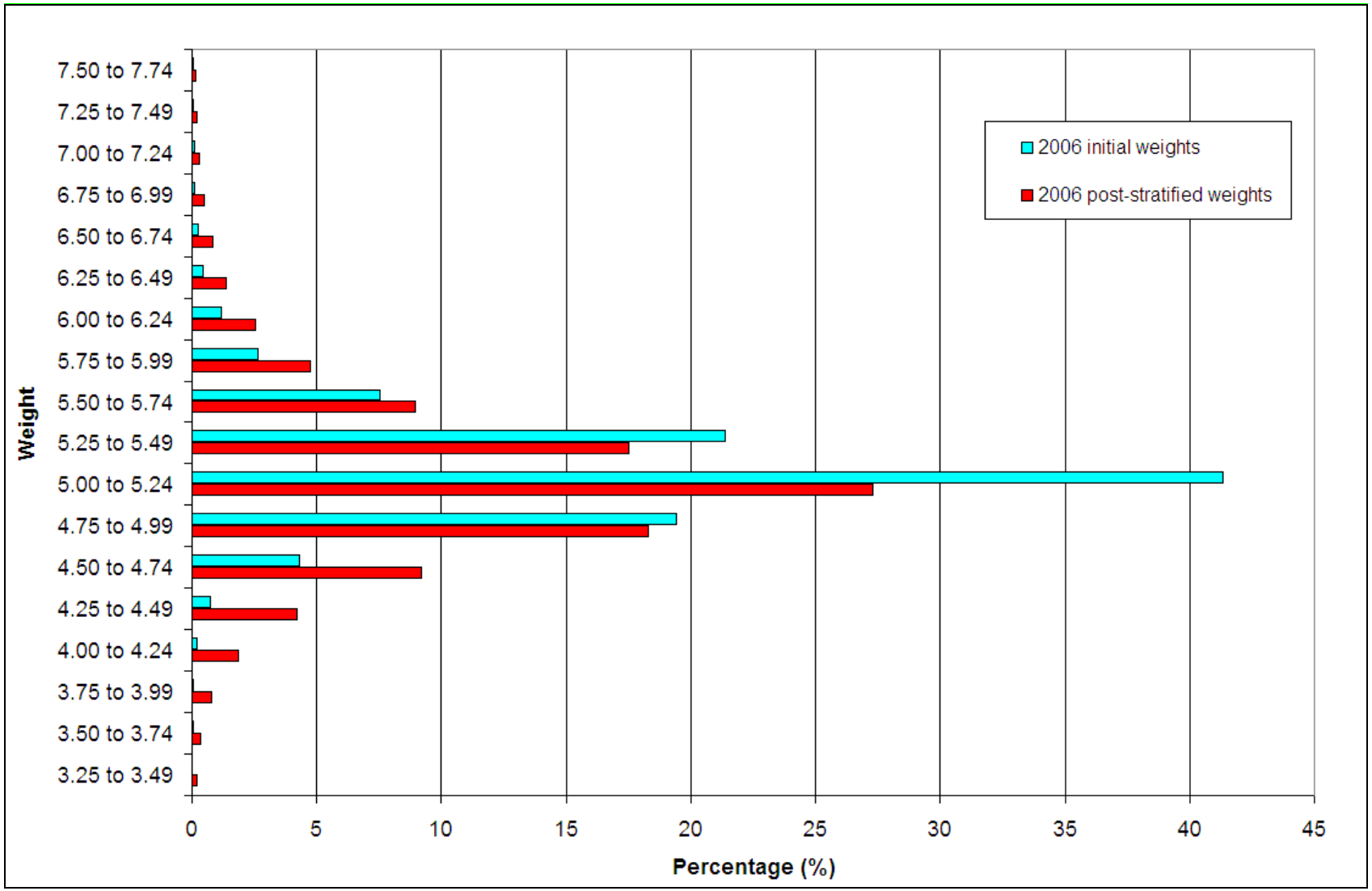
Charts 7.2.1.2, 7.2.1.3 and 7.2.1.4 compare the distributions of the 2006 Census initial weights, post-stratified weights, first-step weights and final weights. The initial weights are tightly clustered around 5 as a result of a one-in-five sample of households being selected. The post-stratified, first-step and final weight distributions become progressively more spread out as the constraints become more restrictive.

Chart 7.2.1.1 Comparison of 2006 and 2001 final household weights



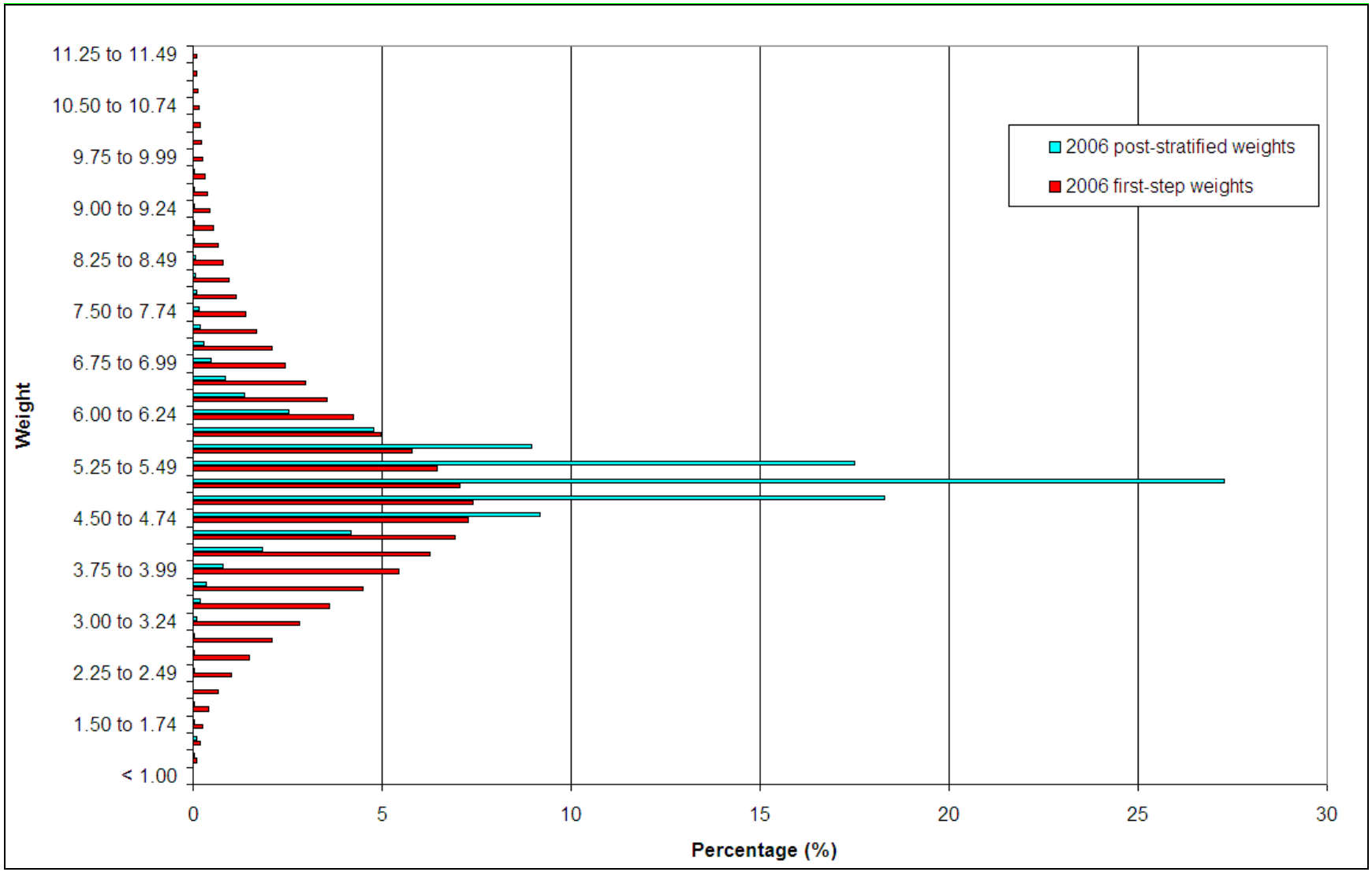
Sources: Statistics Canada, 2006 and 2001 censuses.

Chart 7.2.1.2 Comparison of initial weights and post-stratified weights, 2006 Census



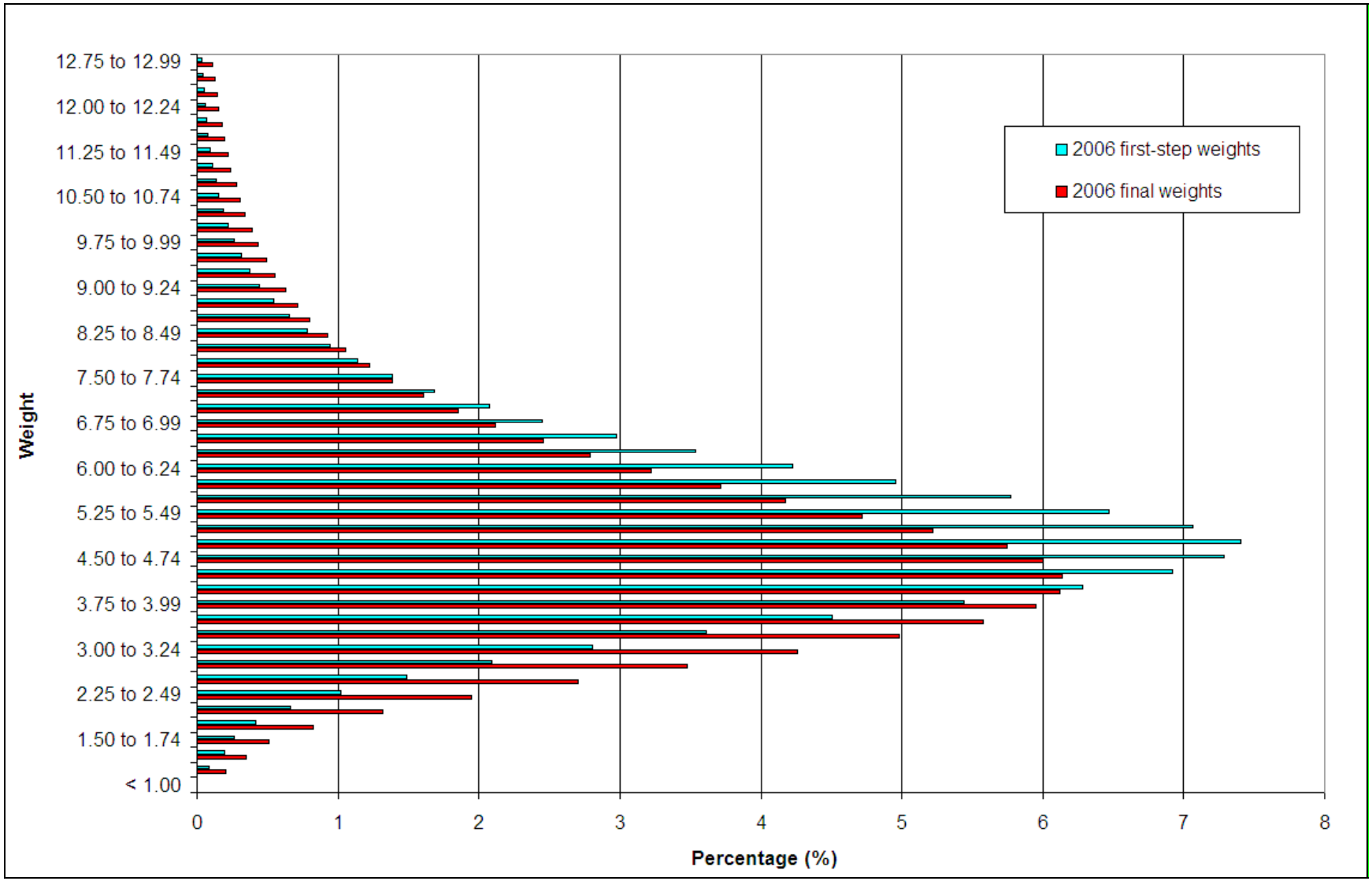
Source: Statistics Canada, 2006 Census.

Chart 7.2.1.3 Comparison of post-stratified weights and first-step weights, 2006 Census



Source: Statistics Canada, 2006 Census.

Chart 7.2.1.4 Comparison of first-step weights and final weights, 2006 Census



Source: Statistics Canada, 2006 Census.

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 34 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

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

The numerator in the above expression (sample estimate − population count) is referred to as the 'population/estimate difference.' The comparison between sample estimates and population counts is based on occupied private dwellings from sampled CUs.

Table 7.2.2.1 and Charts 7.2.2.1 and 7.2.2.2 show the 2006 and 2001 Canada-level population/estimate differences and discrepancies for the 34 WA-level constraints using either the initial or the final weights. Chart 7.2.2.1 is similar to Chart 6.1 in that it is based on initial weights, but it shows population/estimate discrepancies rather than Z statistics, so much of the discussion of Chart 6.1 is applicable to Chart 7.2.2.1.

In Table 7.2.2.1 and Chart 7.2.2.1, it is also shown what the sampling bias would have been in 2006 if the 2001 approach of applying document conversion rather than whole household imputation (WHI) had been used to deal with total non-response long forms. To determine this, long forms with whole household imputation applied were treated as short forms and the initial weights were recalculated at the CU level to reflect this. The recalculated initial weights were applied to the reduced sample to generate new population/estimate differences that appear in the column labelled 'Without WHI' in Table 7.2.2.1. These differences also appear as discrepancies in Chart 7.2.2.1 and are labelled as '2006 without WHI' in the legend. The population/estimate differences under whole household imputation using the original initial weights and the unreduced sample are placed in the column labelled 'With WHI' in Table 7.2.2.1 and '2006 with WHI' in the legend of Chart 7.2.2.1.

In general, it can be seen that the 2006 'Without WHI' differences in Table 7.2.2.1 are much more like the 2001 differences than the 2006 'With WHI' differences. Also, the population/estimate differences are frequently smaller for 'With WHI' than for 'Without WHI,' in 2006 (e.g., this is the case for female, persons aged less than 15 years or those aged 45 years and over; marital status married, widowed, divorced or separated; households of size 1, 2, 4 and 5; and, single-detached dwelling type and apartments less than five storeys). Thus, the introduction of whole household imputation in 2006 to deal with total non-response households was generally beneficial.

While not shown in Table 7.2.2.1 and Chart 7.2.2.1, the initial weights for the reduced sample were calculated a second time separately for 1, 2, 3, 4, 5 and 6+ households at the CU level. Under this approach, the 2006 'Without WHI' differences were much more similar to the 2006 'With WHI' differences. This suggests that whole household imputation gives similar results to what would have been achieved by document conversion and initial weights if the initial weights had been post-stratified by household size.

Table 7.2.2.1 shows that, compared to 2001, the absolute value of the 2006 population/estimate discrepancies based on final weights are noticeably larger for the age ranges 15 to 19 and 25 to 34, but similar or smaller for the other age ranges. The absolute discrepancies in 2006 are also larger for households with 4, 5, and 6+-persons. As discussed in Chapter 6, the fact that the number of persons on the 2B paper questionnaire was reduced from 6 to 5 in 2006 is likely a major cause for this. In comparing Charts 7.2.2.1 and 7.2.2.2, it can be seen that the 2006 population/estimate discrepancies based on final weights are dramatically smaller than those based on initial weights, with the exception of the 5-person and 6+-person households. This is likely due to the difficulty of correcting for such large initial biases while still correcting for the remaining constraints at the same time. The discrepancies for these two constraints are still significantly reduced with the final weights compared to those with the initial weights. It should also be noted that the discrepancies based on the final weights for the two dwelling type characteristics (single detached dwellings and apartments < 5 storeys) have been noticeably reduced from those based on the initial weights despite the fact that these were not controlled on in all WAs. The reduction in the discrepancy for these characteristics likely resulted in an increase in the discrepancy for other characteristics that were dropped in their stead. The exact impact on the other characteristics cannot be observed due to the many factors at play. Chart 7.2.2.2-A is the same as Chart 7.2.2.2, but it has been rescaled so that the discrepancies are more easily seen for the other constraints. Chart 7.2.2.2-A shows that aside from household size constraints, the 'common law status = yes' constraint has the largest discrepancy.

Table 7.2.2.2 and Chart 7.2.2.3 show the 2006 population/estimate differences and discrepancies based on final weights for the 34 WA-level constraints, based on Pass 1 and Pass 2 results, for Canada (see Section 4.5). The Pass 1 discrepancies are smaller than the Pass 2 discrepancies, due to the fact that the census weights were calculated based on Pass 1 results. Chart 7.2.2.4 examines the difference between Pass 1 and Pass 2 results for both the 2006 and 2001 censuses. It shows that, with the exception of the common law, widowed, and separated constraints, the difference between Pass 1 and Pass 2 estimates is much lower in 2006 than in 2001. This may be partially due to the whole household imputation process which may have resulted in more consistency between the Pass 1 and Pass 2 data than in 2001.

Table 7.2.2.2 shows that there is no population/estimate difference for the total number of persons with both Pass 1 and Pass 2 results. It should be noted that this represents a combination of persons from both private households and senior units. However, when the Pass 1 or Pass 2 results for these two universes are observed separately, then the total population for private households is overestimated by 1,982 persons and the total population for senior units is underestimated by the same amount.

Table 7.2.2.3 presents the counts and estimates for the three separate universes for which the census data may be observed. These were discussed in more detail in Section 4.7. This weighting report focuses on data coming from the Private universe. Table 7.2.2.3 shows the difference in population counts and estimates when collectives and institutions are considered since these are included in published census tabulations.

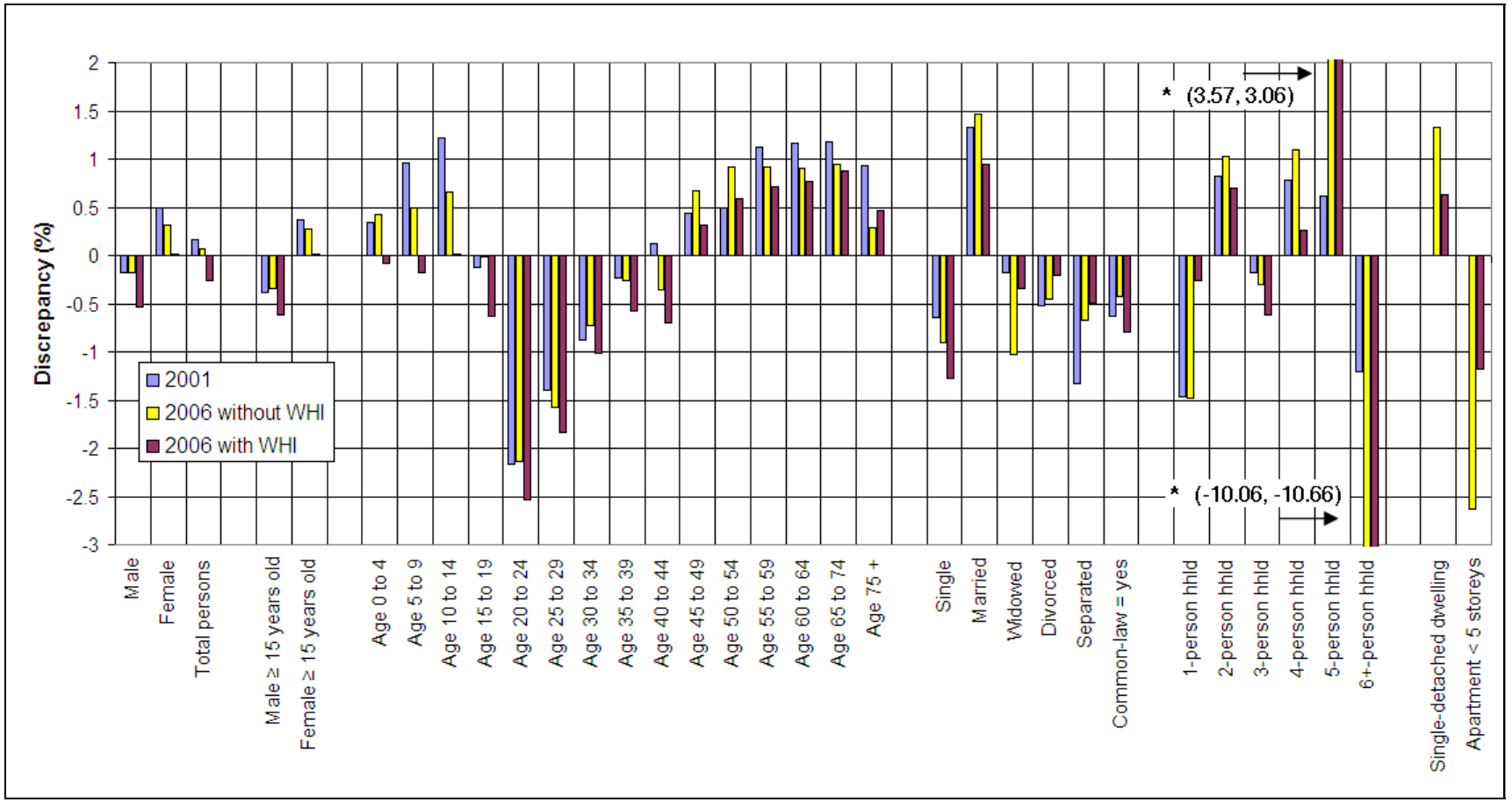
Table 7.2.2.1 Comparison of 2001 and 2006 population/estimate discrepancies for Canada

	2006 Census				2001 Census		
	Initial weight differences**		Final weights		Initial weights	Final weights	
Characteristic	With WHI	Without WHI	Difference	Discrepancy	Difference	Difference	Discrepancy
Male	-79,381	-27,123	-49	0.00	-25,074	0	0.00
Female	923	50,176	49	0.00	73,397	0	0.00
Male ≥ 15 years old	-76,328	-41,932	116	0.00	-44,291	51	0.00
Person ≥ 15 years old	-74,220	-5,817	243	0.00	130	0	0.00
Total households	-1,056	-1,245	0	0.00	-1,040	0	0.00
Total persons	-78,458	23,053	0	0.00	48,324	0	0.00
Age 0 to 4	-1,354	6,915	-68	0.00	5,628	559	0.03
Age 5 to 9	-3,126	8,560	-360	-0.02	18,245	-792	-0.04
Age 10 to 14	242	13,395	186	0.01	24,321	234	0.01
Age 15 to 19	-13,108	-319	1,856	0.09	-2,644	779	0.04
Age 20 to 24	-51,382	-43,328	-226	-0.01	-41,081	-504	-0.03
Age 25 to 29	-35,659	-30,595	-1,359	-0.07	-25,620	-785	-0.04
Age 30 to 34	-20,062	-14,450	-359	-0.02	-17,888	7	0.00
Age 35 to 39	-12,439	-5,644	391	0.02	-5,675	-556	-0.02
Age 40 to 44	-17,995	-8,963	348	0.01	3,073	100	0.00
Age 45 to 49	8,286	17,325	328	0.01	10,024	687	0.03
Age 50 to 54	13,732	21,134	342	0.01	10,004	-87	0.00
Age 55 to 59	14,614	18,772	249	0.01	17,396	81	0.01
Age 60 to 64	11,940	14,123	-522	-0.03	14,459	933	0.08
Age 65 to 74	19,541	20,930	300	0.01	24,283	271	0.01
Age 75+	8,312	5,197	-1,105	-0.06	13,798	-926	-0.06
Single	-180,891	-127,391	-42	0.00	-86,671	-53	0.00
Married	114,966	179,374	-184	0.00	156,112	-57	0.00
Widowed	-4,846	-14,702	-252	-0.02	-2,388	557	0.04
Divorced	-4,019	-9,166	141	0.01	-9,375	206	0.01
Separated	-3,667	-5,061	336	0.04	-9,355	-653	-0.09
Common-law = yes	-21,921	-11,523	2,995	0.11	-14,381	4,115	0.18
1-person hhld	-8,705	-49,455	-3,370	-0.10	-42,675	-4,175	-0.14
2-person hhld	29,091	42,615	-1,658	-0.04	30,499	-906	-0.02
3-person hhld	-12,081	-5,989	-3,727	-0.19	-3,405	-5,010	-0.27
4-person hhld	4,871	20,108	6,025	0.33	14,138	2,414	0.13
5-person hhld	21,817	25,493	15,862	2.22	4,395	8,818	1.23
6+-person hhld	-36,049	-34,017	-13,132	-3.88	-3,991	-1,142	-0.34
Single-detached dwelling	42,896	89,944	8,648	0.13	..	..	..
Apartment < 5 storeys	-27,066	-60,002	-1,089	-0.05	..	..	..

** Whole household imputation was applied in the calculation of 2006 Census weights.
.. not available for a specific reference period

Sources: Statistics Canada, 2006 and 2001 censuses.

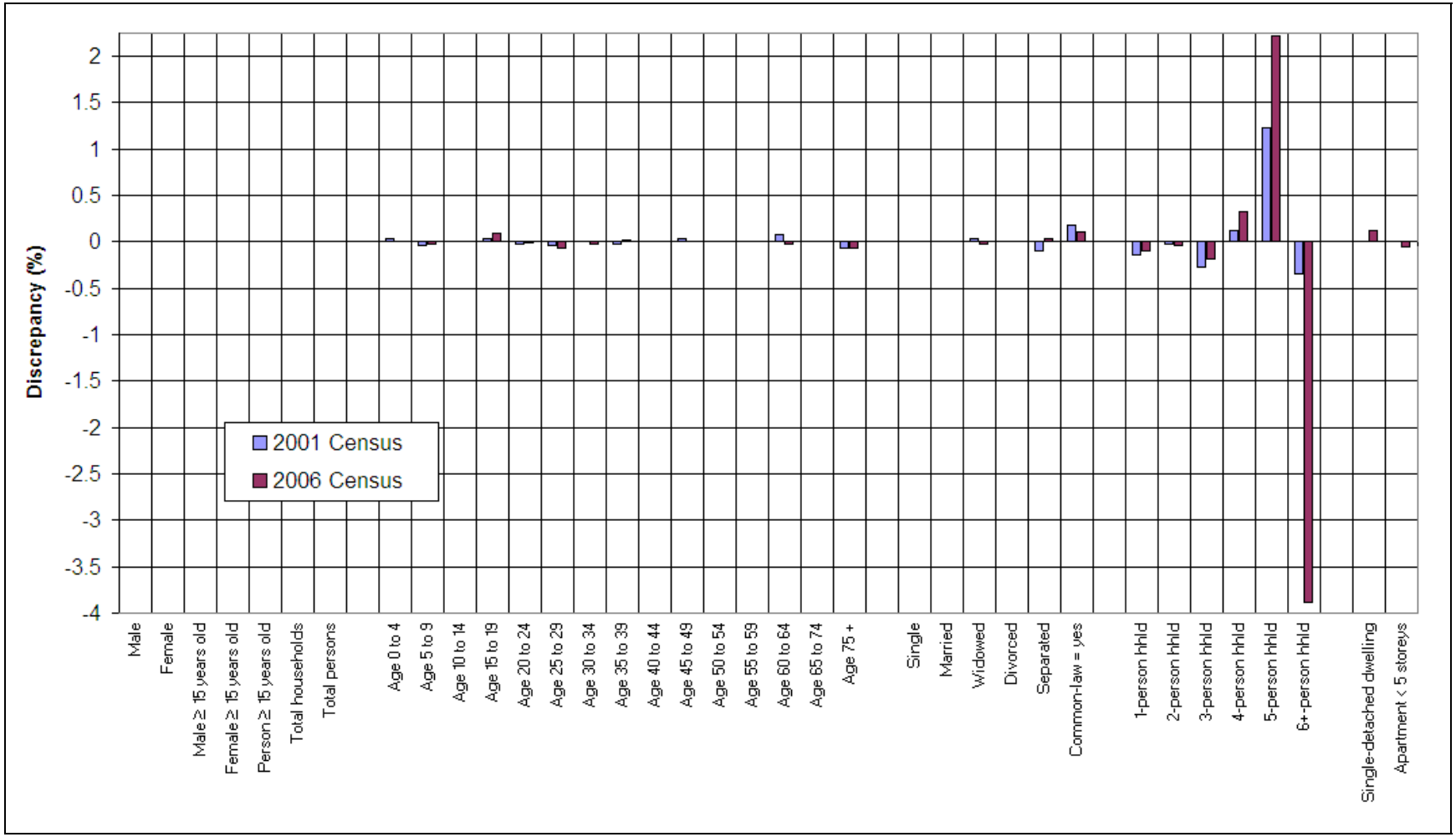
Chart 7.2.2.1 Comparison of initial weight discrepancies with and without whole household imputation



* Values inset in the chart are for bars that exceed the limit of the chart. Values are for bars from left to right.

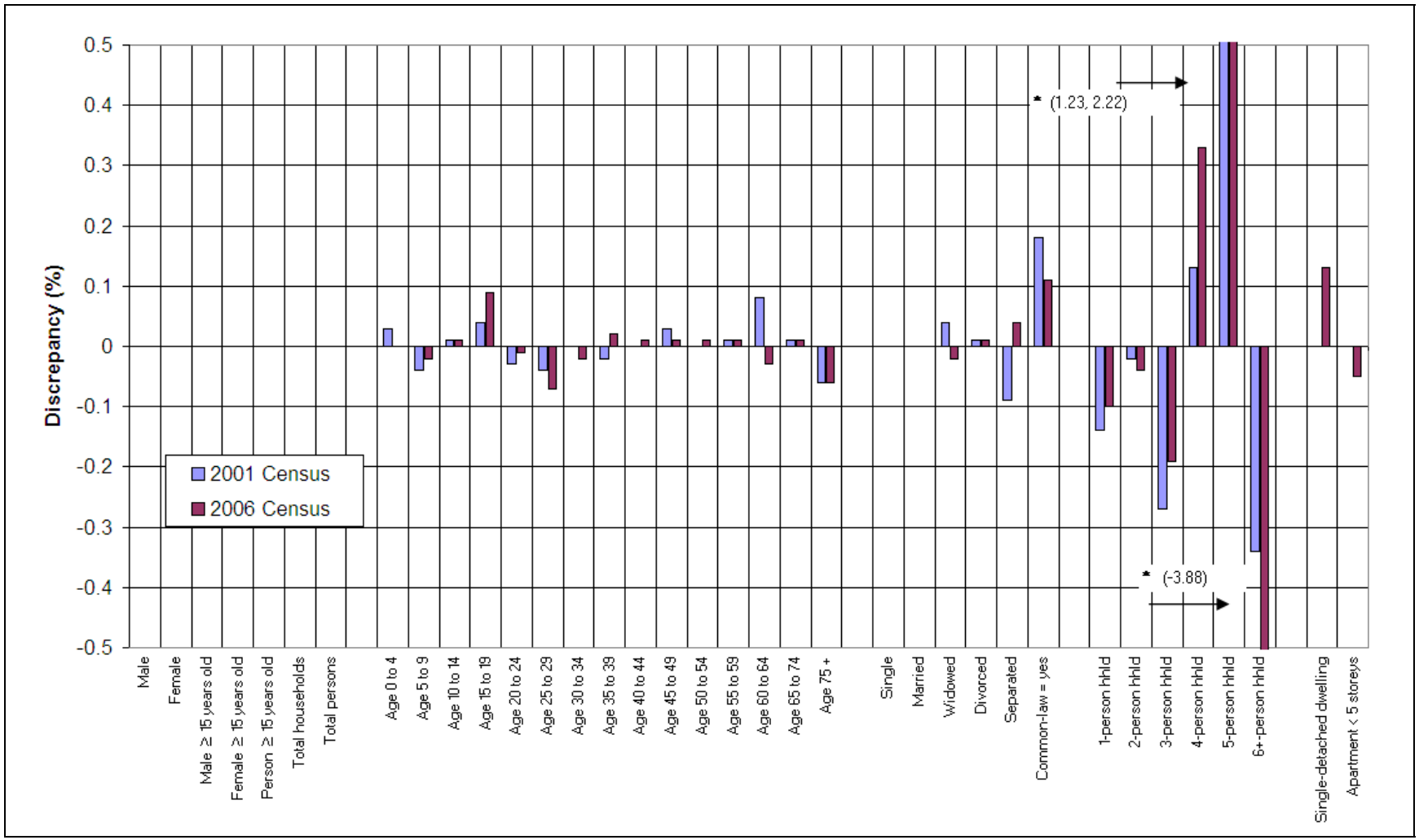
Sources: Statistics Canada, 2006 and 2001 censuses.

Chart 7.2.2.2 Population/estimate discrepancies based on final weights



Sources: Statistics Canada, 2006 and 2001 censuses.

Chart 7.2.2.2-A Population/estimate discrepancies based on final weights (rescaled)



* Values inset in the chart are for bars that exceed the limit of the chart. Values are for bars from left to right.

Sources: Statistics Canada, 2006 and 2001 censuses.

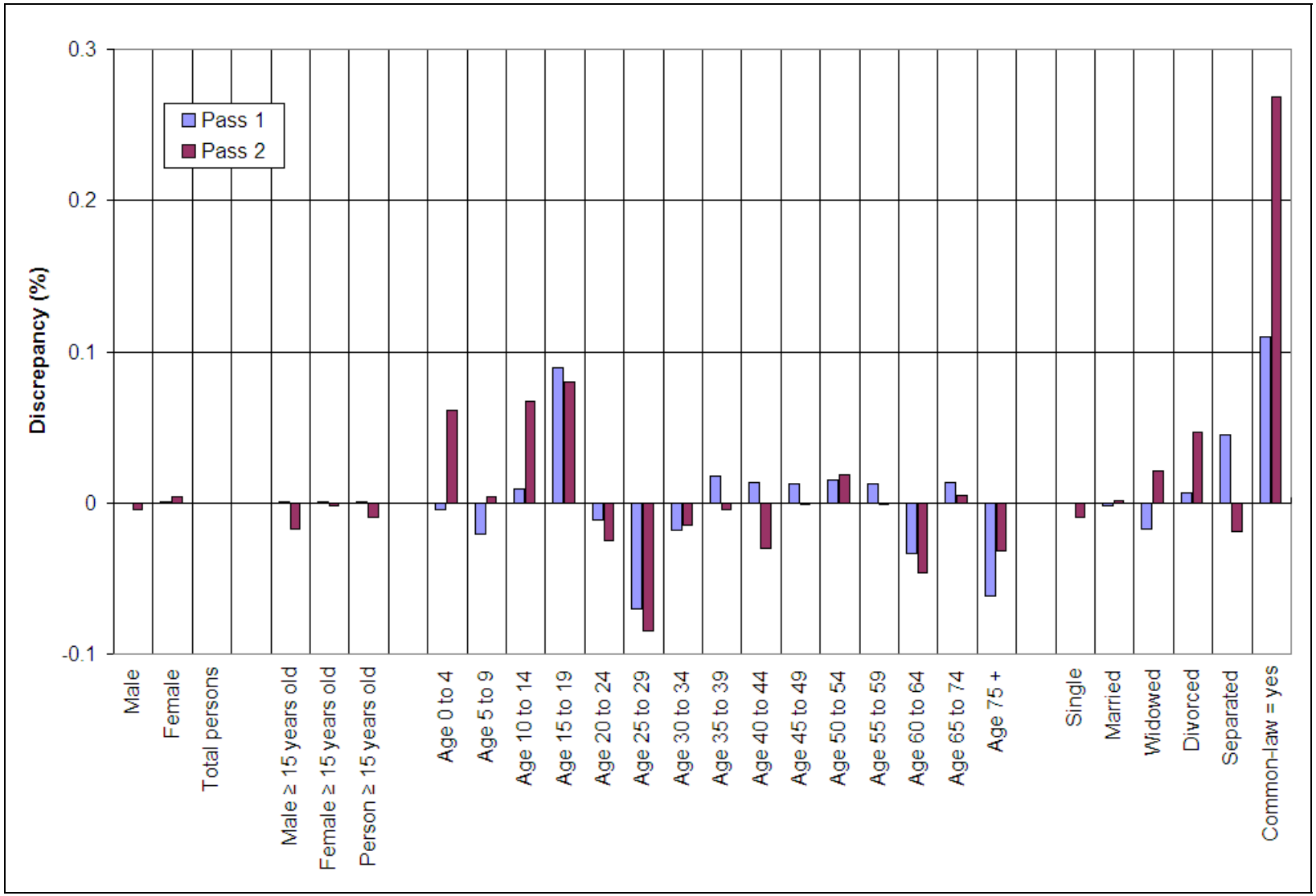
Table 7.2.2.2 Comparison of Pass 1 and Pass 2 population/estimate discrepancies based on final weights, for Canada, 2006 Census

	2006 Census: Pass 1				2006 Census: Pass 2				Pass 2 – Pass 1	
Characteristic	Count	Estimate	Difference[1]	Disc.[2]	Count	Estimate	Difference[1]	Disc.[2]	Difference[3]	Disc.[4]
Male	15,041,551	15,041,502	-49	0	15,041,422	15,040,736	-686	0	-637	0
Female	15,653,041	15,653,090	49	0	15,653,170	15,653,856	686	0	637	0
Total persons	30,694,592	30,694,592	0	0	30,694,592	30,694,592	0	0	0	0
Male ≥ 15 years old	12,263,445	12,263,561	116	0	12,263,025	12,260,877	-2,148	-0.02	-2,264	-0.02
Female ≥ 15 years old	13,005,067	13,005,193	126	0	13,004,908	13,004,626	-282	0	-408	0
Person ≥ 15 years old	25,268,512	25,268,754	242	0	25,267,933	25,265,503	-2,430	-0.01	-2,672	-0.01
Age 0 to 4	1,640,859	1,640,791	-68	0	1,641,102	1,642,103	1,001	0.06	1,069	0.07
Age 5 to 9	1,760,005	1,759,645	-360	-0.02	1,760,149	1,760,221	72	0	432	0.02
Age 10 to 14	2,025,216	2,025,402	186	0.01	2,025,408	2,026,764	1,356	0.07	1,170	0.06
Age 15 to 19	2,083,373	2,085,229	1,856	0.09	2,083,289	2,084,962	1,673	0.08	-183	-0.01
Age 20 to 24	2,029,449	2,029,223	-226	-0.01	2,029,402	2,028,899	-503	-0.02	-277	-0.01
Age 25 to 29	1,940,880	1,939,521	-1,359	-0.07	1,940,768	1,939,125	-1,643	-0.08	-284	-0.01
Age 30 to 34	1,976,478	1,976,119	-359	-0.02	1,976,505	1,976,221	-284	-0.01	75	0
Age 35 to 39	2,161,430	2,161,821	391	0.02	2,161,366	2,161,265	-101	0	-492	-0.02
Age 40 to 44	2,559,477	2,559,825	348	0.01	2,559,271	2,558,507	-764	-0.03	-1,112	-0.04
Age 45 to 49	2,571,429	2,571,757	328	0.01	2,571,359	2,571,322	-37	0	-365	-0.01
Age 50 to 54	2,313,657	2,313,999	342	0.01	2,313,669	2,314,099	430	0.02	88	0
Age 55 to 59	2,045,868	2,046,117	249	0.01	2,045,821	2,045,799	-22	0	-271	-0.01
Age 60 to 64	1,558,145	1,557,623	-522	-0.03	1,558,054	1,557,336	-718	-0.05	-196	-0.01
Age 65 to 74	2,229,023	2,229,323	300	0.01	2,229,016	2,229,129	113	0.01	-187	-0.01
Age 75+	1,799,303	1,798,198	-1,105	-0.06	1,799,413	1,798,841	-572	-0.03	533	0.03
Single	14,170,280	14,170,238	-42	0	14,170,125	14,168,822	-1,303	-0.01	-1,261	-0.01
Married	12,291,457	12,291,273	-184	0	12,291,559	12,291,750	191	0	375	0
Widowed	1,435,852	1,435,600	-252	-0.02	1,435,992	1,436,293	301	0.02	553	0.04
Divorced	2,044,164	2,044,305	141	0.01	2,044,209	2,045,165	956	0.05	815	0.04
Separated	752,839	753,175	336	0.04	752,707	752,563	-144	-0.02	-480	-0.06
Common-law = yes	2,725,161	2,728,156	2,995	0.11	2,726,070	2,733,383	7,313	0.27	4,318	0.16
1-person hhld	3,338,596	3,335,226	-3,370	-0.10	3,338,596	3,335,226	-3,370	-0.10	0	0
2-person hhld	4,153,415	4,151,757	-1,658	-0.04	4,153,415	4,151,757	-1,658	-0.04	0	0
3-person hhld	1,963,201	1,959,474	-3,727	-0.19	1,963,201	1,959,474	-3,727	-0.19	0	0
4-person hhld	1,843,987	1,850,012	6,025	0.33	1,843,987	1,850,012	6,025	0.33	0	0
5-person hhld	713,994	729,856	15,862	2.22	713,994	729,856	15,862	2.22	0	0
6+-person hhld	338,241	325,109	-13,132	-3.88	338,241	325,109	-13,132	-3.88	0	0
Single-detached dwelling	6,769,581	6,778,229	8,648	0.13	6,769,581	6,778,229	8,648	0.13	0	0
Apartment < 5 storeys	2,285,965	2,284,876	-1,089	-0.05	2,285,965	2,284,876	-1,089	-0.05	0	0

Notes:
1. Difference: estimate – count.
2. Disc.: discrepancy (100 * [estimate - count]/count).
3. Difference: difference Pass 2 – difference Pass 1.
4. Disc.: discrepancy (100 * [difference Pass 2 – difference Pass 1]/difference Pass 1).

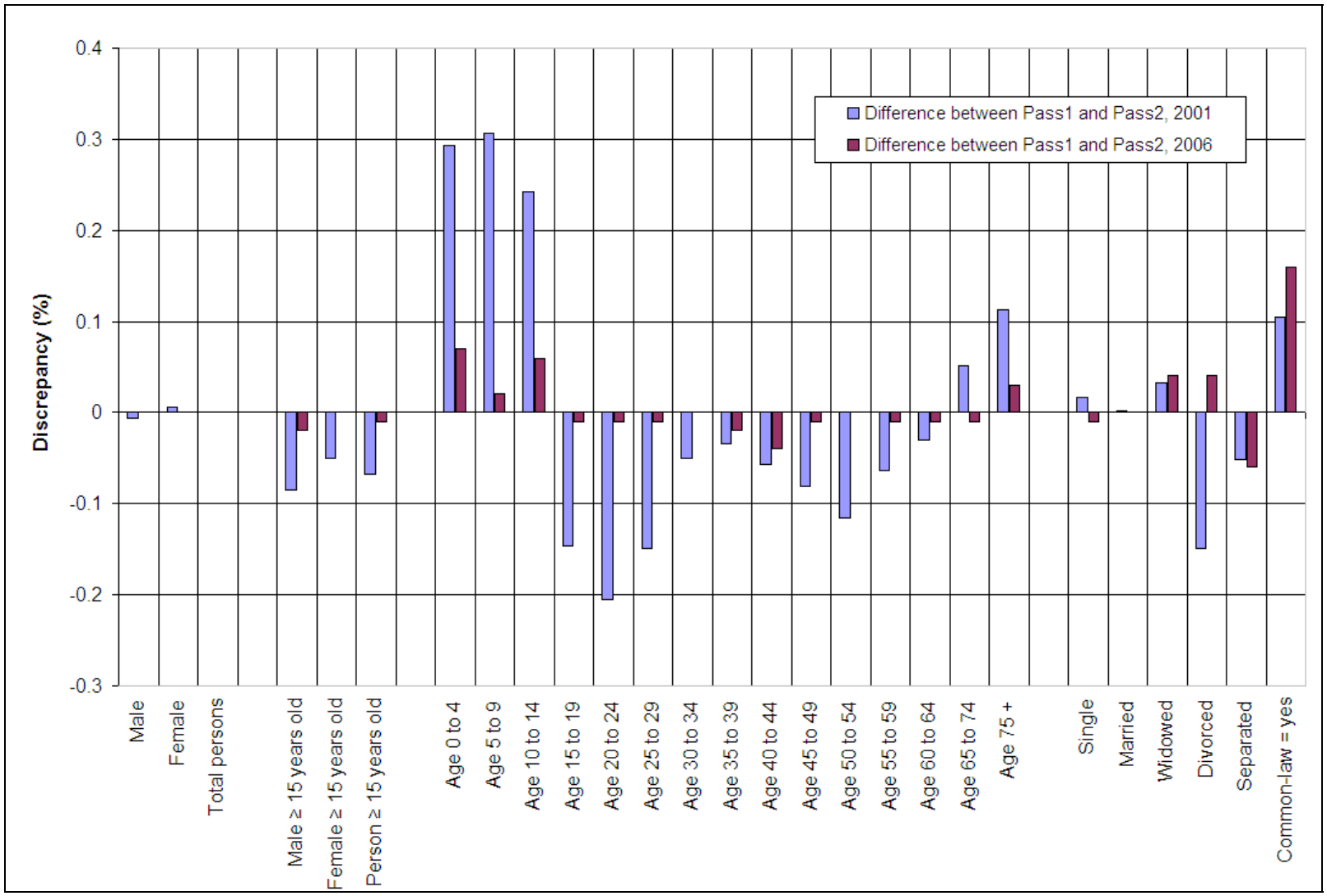
Source: Statistics Canada, 2006 Census.

Chart 7.2.2.3 Comparison of Pass 1 and Pass 2 population/estimate discrepancies based on final weights, for Canada, 2006 Census



Source: Statistics Canada, 2006 Census.

Chart 7.2.2.4 Comparison of population/estimate discrepancies in Pass 1 and Pass 2 differences, 2006 and 2001 censuses



Sources: Statistics Canada, 2006 and 2001 censuses.

Table 7.2.2.3 Comparison of universes – Population counts and estimates, 2006 Census

Characteristic	Pass 2 – Privates		Pass 2 – Privates/Collectives		Pass 2 – Privates/Collectives/Institutions		Institutions
	Count	Estimate	Count	Estimate	Count	Estimate	Count
Male	15,041,422	15,040,736	15,326,954	15,326,268	15,475,970	15,475,284	149,016
Female	15,653,170	15,653,856	15,914,076	15,914,762	16,136,927	16,137,613	222,851
Total persons	30,694,592	30,694,592	31,241,030	31,241,030	31,612,897	31,612,897	371,867
Male ≥ 15 years old	12,263,025	12,260,877	12,472,933	12,470,785	12,618,649	12,616,501	145,716
Female ≥ 15 years old	13,004,908	13,004,626	13,193,720	13,193,438	13,414,410	13,414,128	220,690
Person ≥ 15 years old	25,267,933	25,265,503	25,666,653	25,664,223	26,033,059	26,030,629	366,406
Age 0 to 4	1,641,102	1,642,103	1,689,395	1,690,396	1,690,539	1,691,540	1,144
Age 5 to 9	1,760,149	1,760,221	1,808,205	1,808,277	1,809,373	1,809,445	1,168
Age 10 to 14	2,025,408	2,026,764	2,076,777	2,078,133	2,079,926	2,081,282	3,149
Age 15 to 19	2,083,289	2,084,962	2,134,246	2,135,919	2,140,493	2,142,166	6,247
Age 20 to 24	2,029,402	2,028,899	2,072,397	2,071,894	2,080,384	2,079,881	7,987
Age 25 to 29	1,940,768	1,939,125	1,977,418	1,975,775	1,985,580	1,983,937	8,162
Age 30 to 34	1,976,505	1,976,221	2,011,584	2,011,300	2,020,228	2,019,944	8,644
Age 35 to 39	2,161,366	2,161,265	2,197,813	2,197,712	2,208,273	2,208,172	10,460
Age 40 to 44	2,559,271	2,558,507	2,597,151	2,596,387	2,610,458	2,609,694	13,307
Age 45 to 49	2,571,359	2,571,322	2,606,788	2,606,751	2,620,598	2,620,561	13,810
Age 50 to 54	2,313,669	2,314,099	2,344,230	2,344,660	2,357,304	2,357,734	13,074
Age 55 to 59	2,045,821	2,045,799	2,072,074	2,072,052	2,084,621	2,084,599	12,547
Age 60 to 64	1,558,054	1,557,336	1,578,195	1,577,477	1,589,868	1,589,150	11,673
Age 65 to 74	2,229,016	2,229,129	2,255,529	2,255,642	2,288,363	2,288,476	32,834
Age 75+	1,799,413	1,798,841	1,819,228	1,818,656	2,046,889	2,046,317	227,661
Single	14,170,125	14,168,822	14,541,272	14,539,969	14,666,870	14,665,567	125,598
Married	12,291,559	12,291,750	12,415,528	12,415,719	12,470,398	12,470,589	54,870
Widowed	1,435,992	1,436,293	1,451,805	1,452,106	1,612,819	1,613,120	161,014
Divorced	2,044,209	2,045,165	2,066,245	2,067,201	2,087,387	2,088,343	21,142
Separated	752,707	752,563	766,180	766,036	775,423	775,279	9,243
Common-law = yes	2,726,070	2,733,383	2,789,627	2,796,940	2,789,627	2,796,940	0
1-person hhld	3,338,596	3,335,226	3,367,367	3,363,997	3,367,367	3,363,997	0
2-person hhld	4,153,415	4,151,757	4,186,669	4,185,011	4,186,669	4,185,011	0
3-person hhld	1,963,201	1,959,474	1,985,694	1,981,967	1,985,694	1,981,967	0
4-person hhld	1,843,987	1,850,012	1,865,954	1,871,979	1,865,954	1,871,979	0
5-person hhld	713,994	729,856	730,325	746,187	730,325	746,187	0
6+-person hhld	338,241	325,109	370,002	356,870	370,002	356,870	0
Single-detached dwelling	6,769,581	6,778,229	6,871,318	6,879,966	6,871,318	6,879,966	0
Apartment < 5 storeys	2,285,965	2,284,876	2,289,388	2,329,459	2,289,388	2,329,459	0

Source: Statistics Canada, 2006 Census.

7.2.3 Discarding constraints

For the 2006 Census, 20 sets of parameter combinations were examined in the weighting system for each weighting area (WA), and the set of parameters with the best results in any given WA was chosen (see Section 4.4).

Appendix B gives a complete list of the 34 constraints being used. Thirty-two of these constraints were part of each test involving the different parameter sets. Two of these, the single-detached dwellings and apartments in buildings with less than 5 storeys, were new in 2006, and were only added as constraints for certain parameter combinations.

Table 7.2.3.1 shows how often each of the 34 constraints was discarded in the 6,602 sampled WAs in 2006 and the 6,141 sampled WAs in 2001. 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 2006 Census results. First, it should be noted that a constraint such as 'Age 0 to 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), it can cause large population/estimate discrepancies, as was observed in Chart 7.2.2.2.

The two dwelling type constraints (single-detached dwellings and apartments < 5 storeys) were new in 2006 and treated differently than the 32 constraints also used in 2001. The level of non-response for the dwelling type variable was analysed at the DA level. These two constraints were automatically dropped for 399 WAs that contained a DA that was determined to have a significant level of non-response for this variable that would make the estimates for these characteristics unreliable. For the remaining 6,203 WAs, the use of these constraints was included as a parameter. Ten of the twenty parameter combinations for which the WAs were processed attempted to control on these characteristics. In the 'cherry-picking' process, 3,688 WAs had the final weights selected from a parameter combination which attempted to control on these two constraints. This means that, for example, while the constraint single-detached dwelling was only dropped by the weighting system for 304 WAs, it was still only controlled on in 3,384 WAs.

Table 7.2.3.2 summarizes the information found in Table 7.2.3.1. The total number of constraints dropped is higher in 2006 because there are more WAs (6,602 WAs in 2006, 6,141 WAs in 2001), but the average number of WA-level constraints dropped per WA is fairly consistent between 2001 and 2006.

Table 7.2.3.2 also summarizes information on the frequency of discarding DA-level constraints on the number of households and the number of persons. If a WA contained 8 DAs, for example, it would have 16 DA-level constraints. Overall there was a decrease in the average number of DA-level constraints being dropped (0.8 in 2006, 1.1 in 2001). The most notable decrease appears in the SMALL category, where only 248 constraints were dropped in 2006, compared to 1,354 in 2001. This is partially due to having more sets of parameters to choose from.

Table 7.2.3.1 Frequency of discarding weighting area-level constraints in 2001 and 2006 in final weight adjustment

Characteristic	2006 Census					2001 Census				
	Small	LD	NLD	Outlier	Total	Small	LD	NLD	Outlier	Total
Male	0	0	0	1	1	0	0	0	0	0
Female	0	0	0	0	0	0	0	0	0	0
Total persons	0	0	0	1	1	0	0	0	0	0
Male ≥ 15 years old	0	3	22	51	76	0	4	24	27	55
Person ≥ 15 years old	0	1	0	3	4	0	0	0	0	0
Total households	0	0	0	0	0	0	0	0	0	0
Age 0 to 4	16	4,342	6	191	4,555	29	4,286	2	124	4,441
Age 5 to 9	60	1,217	11	342	1,630	68	406	4	251	729
Age 10 to 14	66	961	5	279	1,311	79	1,359	2	141	1,581
Age 15 to 19	12	678	7	142	839	18	492	6	131	647
Age 20 to 24	2	228	18	150	398	2	243	15	125	385
Age 25 to 29	2	1,190	7	110	1,309	3	877	9	94	983
Age 30 to 34	1	657	9	169	836	3	158	5	83	249
Age 35 to 39	2	121	3	139	265	3	6	1	35	45
Age 40 to 44	0	6	2	48	56	2	0	0	19	21
Age 45 to 49	0	4	1	41	46	2	2	3	41	48
Age 50 to 54	1	5	1	41	48	2	7	1	38	48
Age 55 to 59	0	10	2	62	74	3	238	7	79	327
Age 60 to 64	2	1,431	42	108	1,583	5	1,751	65	130	1,951
Age 65 to 74	5	33	34	60	132	5	2	32	49	88
Age 75+	28	2,201	10	52	2,291	42	2,308	8	38	2,396
Single	0	0	0	5	5	1	0	0	2	3
Married	0	0	0	10	10	1	1	0	2	4
Widowed	11	615	25	183	834	6	593	15	128	742
Divorced	0	12	14	106	132	3	15	11	94	123
Separated	33	5,933	12	30	6,008	20	5,510	3	34	5,567
Common-law = yes	9	0	0	318	327	16	0	0	278	294
1-person hhld	0	175	1,716	49	1,940	2	194	1,869	22	2,087
2-person hhld	0	1	320	26	347	1	2	310	15	328
3-person hhld	2	46	2,828	45	2,921	7	40	2,537	42	2,626
4-person hhld	40	264	1,416	157	1,877	50	187	1,102	98	1,437
5-person hhld	473	1,377	162	473	2,485	401	1,206	143	281	2,031
6+-person hhld	2,228	4,133	118	24	6,503	1,941	3,960	121	9	6,031
Single-detached dwelling**	256	3	3	42	304	..	..	..	..	..
Apartment < 5 storeys**	550	2	0	141	693	..	..	..	..	..

** Only 3,688 of the 6,602 WAs used this constraint in the weight calculations.

.. not available for a specific reference period

Notes:

Small: Small constraint.

LD: Linearly dependent.

NLD: Nearly linearly dependent.

Outlier: Caused outlier weights.

Sources: Statistics Canada, 2006 and 2001 censuses.

Table 7.2.3.2 Frequency of discarding constraints at the weighting area and dissemination area levels in 2001 and 2006 in final weight adjustment – Summary statistics

	2006 Census					2001 Census				
	Small	LD	NLD	Outlier	Total	Small	LD	NLD	Outlier	Total
WA level constraints										
Total constraints dropped	2,993	25,644	6,791	3,416	38,844	2,715	23,847	6,295	2,410	35,267
Constraints dropped per WA	0.5	3.9	1.0	0.5	5.9	0.4	3.9	1.0	0.4	5.7
DA level constraints										
Total constraints dropped	248	190	4,154	869	5,461	1,354	357	4,191	917	6,819
Constraints dropped per WA	0.0	0.0	0.6	0.1	0.8	0.2	0.1	0.7	0.1	1.1

Notes:
Small: Small constraint
LD: Linearly dependent
NLD: Nearly linearly dependent
Outlier: Caused outlier weights

The dwelling type constraints (single detached, apartments less than 5 floors) were not included in these counts since they were not included in all WAs.

Sources: Statistics Canada, 2006 and 2001 censuses.

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

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

The comparison between sample estimates and population counts is based on occupied private dwellings from sampled CUs.

In this chapter, these population/estimate discrepancies from both the 2001 and 2006 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. Table 7.1.2 reveals that 13.2% of CSDs and 67.0% 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.

The charts and tables in this chapter provide the percentiles of the population/estimate discrepancies for 33 characteristics which, except in a few cases, are identical to the 34 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 10th percentile of -11.12% for '6+-person households' in 2006. This means that 10% of the WAs have discrepancies of -11.12% or less. A 90th percentile of 6.75% means that 10% of the WAs have discrepancies of 6.75% 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.

In the next few sections, the 2006 discrepancies will be compared to those in 2001 for various levels of geography.

8.1 Dissemination areas

Canada is divided into 54,626 DAs, of which 52,448 contained sampled households in the 6,602 WAs in the weighting process. A DA, on average, will have a population of 580 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 that 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.

For the most part, the distribution of the discrepancies at the DA level is similar in 2006 compared to 2001 with them sometimes being slightly larger and sometimes being slightly smaller. The discrepancies are marginally higher, however, for most age range constraints in 2006 compared to 2001.

8.2 Weighting areas

Canada is divided into 6,607 WAs, of which 6,602 are sampled WAs. On average, each WA has a population of 4,785 persons and is composed of 8 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 2006 and 2001 censuses, the 10th, 25th, 50th, 75th and 90th percentiles are zero for 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. The most noticeable difference is the larger discrepancies for the 5-person and 6+-person households, which is to be expected based on the Canada level discrepancy for these constraints. The two dwelling type constraints have non-zero percentiles at the WA level because controlling on them was parameterized and not included in every WA.

8.3 Census subdivisions

Canada is divided into 5,418 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 5,859 persons, but can range anywhere in size from a very small town to a very large city. Table 7.1.2 shows that 13.2% 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 most part, the distribution of the discrepancies at the CSD level is similar in 2006 compared to 2001. The discrepancies are marginally higher, however, for age range constraints for the 10th, and 90th percentiles in 2006 compared to 2001. The trend does not hold for the 25th and 75th percentiles of the age range constraints. There are also some large discrepancies in 2006 for the 5-person and 6+-person households. This is not surprising given the large discrepancies, based on the initial weights, seen in Chart 6.1.

8.4 Census tracts

CTs are only located in large urban centres having an urban core population of 50,000 or more. There are 5,089 CTs in Canada. CTs usually have a population ranging from 1,500 to 8,000 persons, with the average being approximately 4,500 persons. Table 7.1.2 shows that 67.0% of CTs consist of one or more whole WAs.

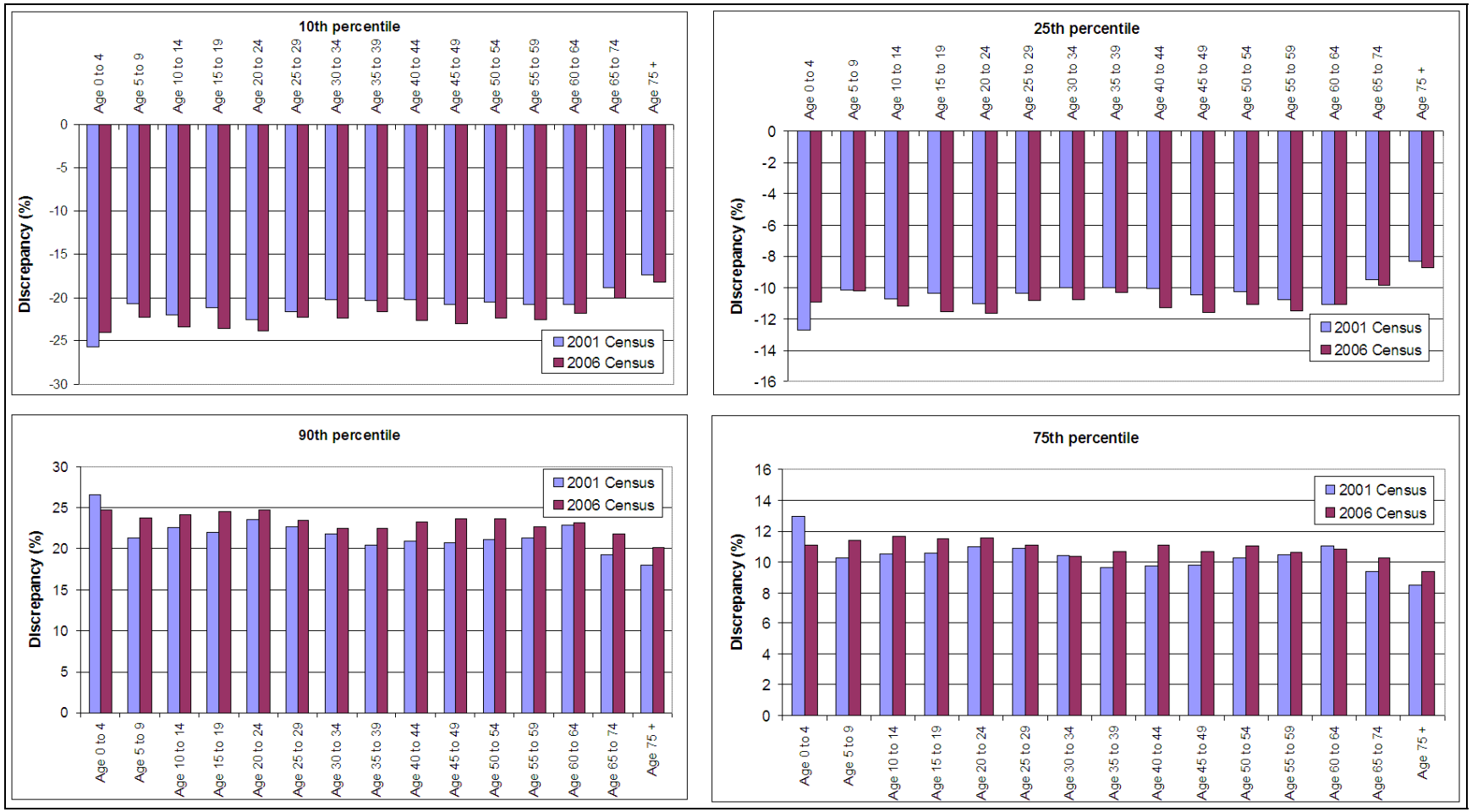
Chart 8.4.1 summarizes the population/estimate discrepancies for all sampled CTs in Canada. It is not surprising that the discrepancies are similar between 2001 and 2006 for most characteristics. Just like with the CSDs, the 5-person and 6+-person households have large discrepancies at the 10th and 90th percentiles.

8.5 Census divisions

Canada is divided into 288 CDs. CDs have an average population of approximately 110,000 persons. A CD might correspond to a county, regional municipality, regional district, or any other area established by provincial or territorial law.

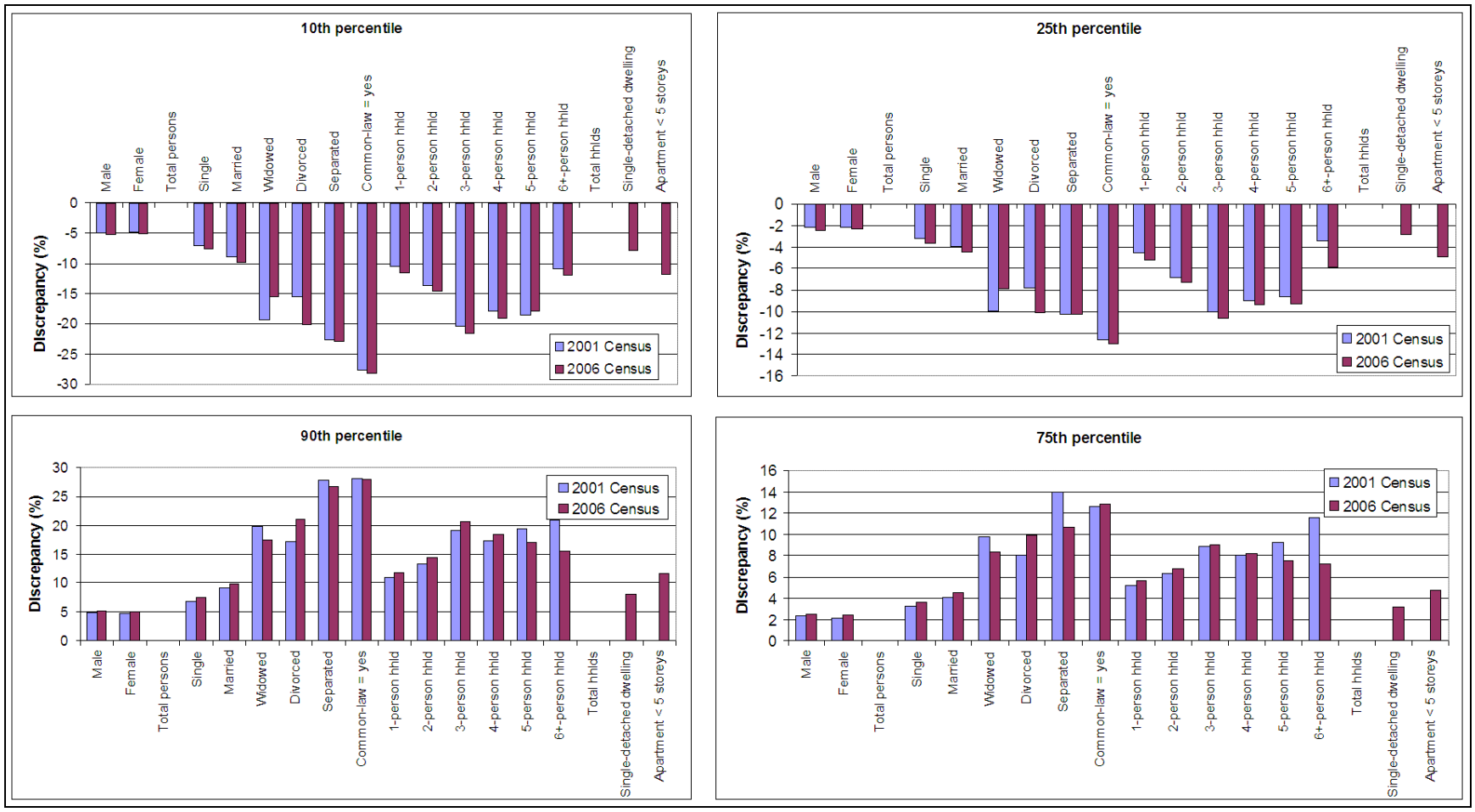
Table 8.5.1 summarizes the 2006 and 2001 Census population/estimate discrepancies for the sampled CDs. All CDs consist of complete WAs. Thus, characteristics that were rarely discarded have perfect or nearly perfect consistency at the CD level. With the exception of the 5-person and 6+ person household characteristics, the size of discrepancies for characteristics that were discarded more frequently is still very small.

Chart 8.1.1 Percentiles of population/estimate discrepancies for 2006 and 2001 dissemination areas (age characteristics)



Sources: Statistics Canada, 2006 and 2001 censuses.

Chart 8.1.2 Percentiles of population/estimate discrepancies for 2006 and 2001 dissemination areas (non-age characteristics)



Sources: Statistics Canada, 2006 and 2001 censuses.

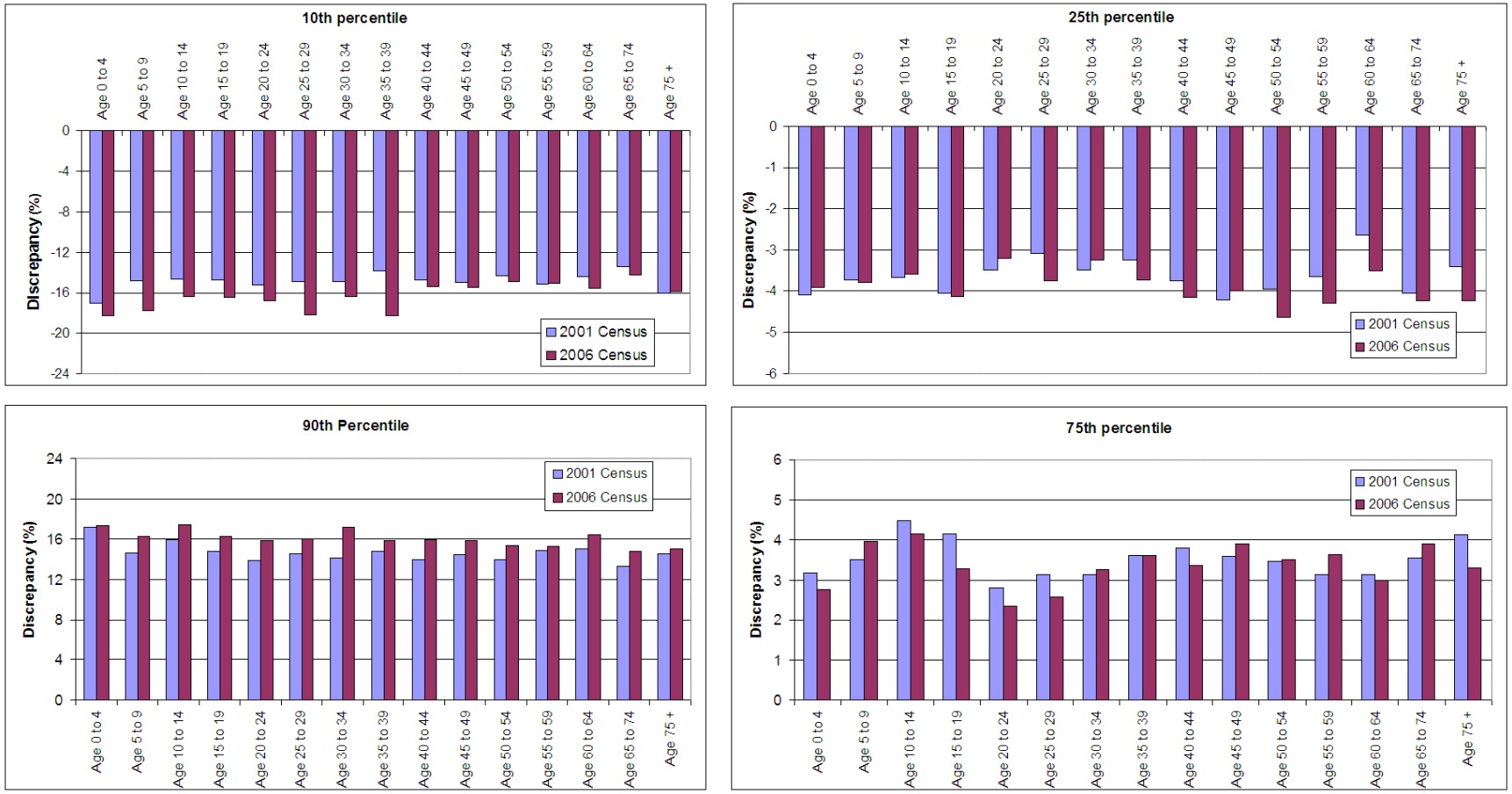
Table 8.2.1 Percentiles of population/estimate discrepancies for weighting areas

Characteristic	2006 percentiles					2001 percentiles				
	10th	25th	50th	75th	90th	10th	25th	50th	75th	90th
Male	0	0	0	0	0	0	0	0	0	0
Female	0	0	0	0	0	0	0	0	0	0
Total persons	0	0	0	0	0	0	0	0	0	0
Age 0 to 4	0	0	0	0	0	0	0	0	0	0
Age 5 to 9	0	0	0	0	0	0	0	0	0	0
Age 10 to 14	0	0	0	0	0	0	0	0	0	0
Age 15 to 19	0	0	0	0	0	0	0	0	0	0
Age 20 to 24	0	0	0	0	0	0	0	0	0	0
Age 25 to 29	0	0	0	0	0	0	0	0	0	0
Age 30 to 34	0	0	0	0	0	0	0	0	0	0
Age 35 to 39	0	0	0	0	0	0	0	0	0	0
Age 40 to 44	0	0	0	0	0	0	0	0	0	0
Age 45 to 49	0	0	0	0	0	0	0	0	0	0
Age 50 to 54	0	0	0	0	0	0	0	0	0	0
Age 55 to 59	0	0	0	0	0	0	0	0	0	0
Age 60 to 64	0	0	0	0	0	0	0	0	0	0
Age 65 to 74	0	0	0	0	0	0	0	0	0	0
Age 75+	0	0	0	0	0	0	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
Common-law = yes	0	0	0	0	0	0	0	0	0	0
1-person hhld	-0.79	0	0	0	0.12	-1.11	0	0	0	0.03
2-person hhld	0	0	0	0	0	0	0	0	0	0
3-person hhld	-2.04	-0.14	0	0	0.65	-1.80	-0.18	0	0	0.40
4-person hhld	-0.28	0	0	0	1.24	-0.06	0	0	0	0.16
5-person hhld	0	0	0	0	15.35	0	0	0	0	7.89
6+-person hhld	-11.12	-4.36	0	3.06	6.75	-6.07	-1.57	1.16	4.63	7.98
Total hhlds	0	0	0	0	0	0	0	0	0	0
Single-detached dwelling	-2.02	-0.75	0.02	0.96	2.71	..	..	..	..	..
Apartment < 5 storeys	-6.84	-2.35	0.04	2.56	7.09	..	..	..	..	..

.. not available for a specific reference period

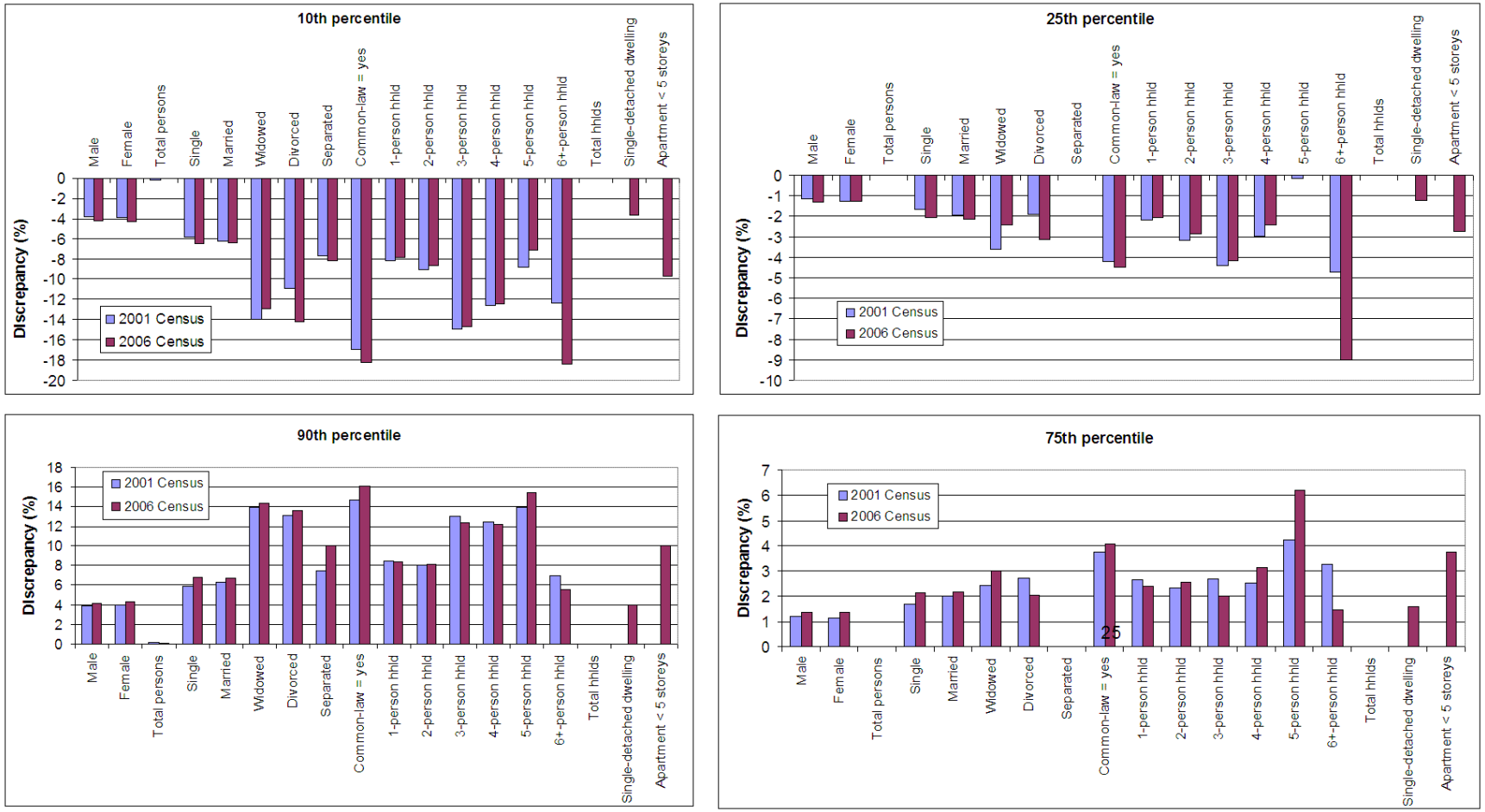
Sources: Statistics Canada, 2006 and 2001 censuses.

Chart 8.3.1 Percentiles of population/estimate discrepancies for census subdivisions (age characteristics)



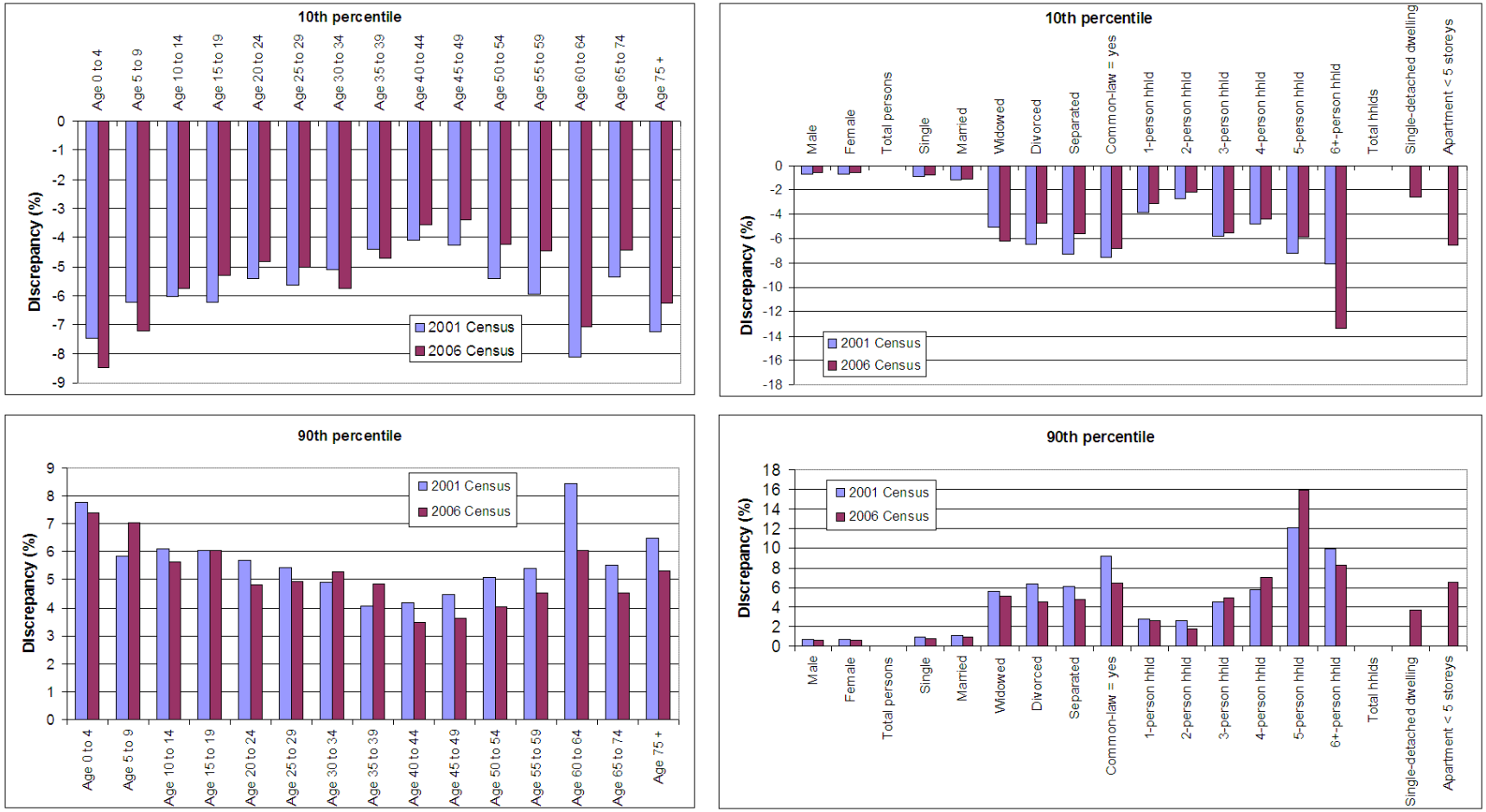
Sources: Statistics Canada, 2006 and 2001 censuses.

Chart 8.3.2 Percentiles of population/estimate discrepancies for census subdivisions (non-age characteristics)



Sources: Statistics Canada, 2006 and 2001 censuses.

Chart 8.4.1 Percentiles of population/estimate discrepancies for census tracts



Sources: Statistics Canada, 2006 and 2001 censuses.

Table 8.5.1 Percentiles of population/estimate discrepancies for census divisions

Characteristic	2006 percentiles					2001 percentiles				
	10th	25th	50th	75th	90th	10th	25th	50th	75th	90th
Male	0	0	0	0	0	0	0	0	0	0
Female	0	0	0	0	0	0	0	0	0	0
Total persons	0	0	0	0	0	0	0	0	0	0
Age 0 to 4	-2.07	-0.17	0	0.23	2.28	-0.81	0	0	0	0.72
Age 5 to 9	-1.67	-0.20	0	0	1.06	-0.51	0	0	0	0.23
Age 10 to 14	-0.51	0	0	0	0.99	-0.20	0	0	0	0.36
Age 15 to 19	-0.03	0	0	0	0.76	-0.05	0	0	0	0.06
Age 20 to 24	-1.57	0	0	0	0.84	-0.41	0	0	0	0.29
Age 25 to 29	-2.65	-0.46	0	0	0.81	-0.54	0	0	0	0.44
Age 30 to 34	-1.28	0	0	0	0.75	-0.04	0	0	0	0.08
Age 35 to 39	-0.18	0	0	0	0.48	0	0	0	0	0
Age 40 to 44	0	0	0	0	0	0	0	0	0	0
Age 45 to 49	0	0	0	0	0	0	0	0	0	0
Age 50 to 54	0	0	0	0	0	0	0	0	0	0
Age 55 to 59	0	0	0	0	0	0	0	0	0	0
Age 60 to 64	-0.25	0	0	0	0.50	-0.18	0	0	0	0.49
Age 65 to 74	0	0	0	0	0	0	0	0	0	0
Age 75+	-0.48	0	0	0	0.21	-0.65	0	0	0	0.28
Single	0	0	0	0	0	0	0	0	0	0
Married	0	0	0	0	0	0	0	0	0	0
Widowed	-0.41	0	0	0	0.48	-0.14	0	0	0	0.04
Divorced	0	0	0	0	0.05	-0.08	0	0	0	0.31
Separated	-1.72	0	0	0	2.21	-0.96	0	0	0	0.94
Common-law = yes	-0.74	0	0	0	1.10	-0.21	0	0	0	0.84
1-person hhld	-0.40	-0.13	-0.01	0	0.03	-0.34	-0.15	-0.04	0	0.01
2-person hhld	-0.05	0	0	0	0	-0.04	0	0	0	0
3-person hhld	-1.24	-0.66	-0.21	0	0.39	-1.05	-0.59	-0.23	0	0.29
4-person hhld	-0.53	-0.10	0.03	0.55	1.29	-0.27	-0.07	0	0.16	0.67
5-person hhld	-0.63	0.17	2.42	5.16	8.55	-0.79	0	0.99	2.47	5.16
6+-person hhld	-18.92	-10.78	-4.98	-2.22	1.53	-9.17	-3.86	-0.76	1.65	3.87
Total hhlds	0	0	0	0	0	0	0	0	0	0
Single-detached dwelling	-0.62	-0.22	0.06	0.38	0.71	..	..	..	..	..
Apartment < 5 storeys	-3.35	-1.30	0.01	1.42	3.92	..	..	..	..	..

.. not available for a specific reference period

Sources: Statistics Canada, 2006 and 2001 censuses.

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, 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.[4] The sampling variance was studied 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, the estimated sampling variance is a measurement of precision and 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 1 in 5 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:

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

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:

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

4. Unfortunately, the sampling variance does not provide any indication of the extent of non-sampling error.

Table 9.2 provides adjustment factors[5] 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, estimates of the sampling variances were calculated for regression estimates for different categories of all of the characteristics[6] given in Table 9.2. This was done for each sampled WA. The estimates of sampling variance at the provincial and national levels were obtained by summing up the WA-level estimates. The adjustment factors for each characteristic in each category were calculated 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 example, the adjustment factors for 'Sex' are the average of those for the categories Male and Female. The majority of characteristics have their categories grouped based on similar adjustment factors, and the factor from the appropriate group should be used for each category. In cases where a table references multiple categories, the largest adjustment factor involved should be used. 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' should be selected. Adjustment factors for different provinces are given in Table 9.2 only for cases 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 Nunavut, the adjustment factors for all characteristics in this territory should be zero. Since sampling was done in the Yukon Territory and the Northwest Territories, the 'Other provinces' adjustment factor should be used, if 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 categories of a characteristic. 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 adjustment factors at the WA level are below the 90th percentile and 10% are above it. The choice of which percentile to use will depend on how conservative the estimate of the standard error is desired to be. For example, using the 99th percentile would provide a very conservative estimate, while using the 75th percentile would provide a somewhat less conservative estimate.

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

5. The squares of the adjustment factors are commonly known as 'design effects.'
6. For example, '$10,000 to $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.'

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) All the standard error adjustment factors are for estimates of the number of persons, households, dwellings, or families, as opposed to, for example, dollar values. For example, the household income adjustment factors are for estimates of the number of households whose income falls in a certain dollar range, and not for 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 1996 and 2006. The 2006 Census estimate for this characteristic was 1,954,605. The 2006 Census count for the population of Canada for sampled variables was 31,241,030. 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 2,707. From Table 9.2, the national-level adjustment factor for the characteristic 'period of immigration' after 1990 is 1.67. Consequently, the adjusted standard error for this estimate is 2,707 x 1.67 = 4,520.

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 different possible estimates, then approximately 19 out of 20 of such intervals would include the value that would normally be 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 sample estimates at the provincial and national level, however, the adjusted standard errors should be close enough to the true standard errors to calculate 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 1996 and 2006 would be 1,954,605 $\pm$ 2(4,520) or 1,954,605 $\pm$ 9,040.

It should be noted that estimates in small areas can be unreliable, as demonstrated with the following example. A community with a population of 500 persons that had an estimate of 50 for the number of persons who immigrated to Canada between 1996 and 2006 would have a standard error of 15 based on Table 9.1. Since this population is smaller than the provincial level, a WA level adjustment factor must be selected from Table 9.2. Taking the most conservative figure from the 99th percentile would result in an adjusted standard error of 15 x 2.46 = 36.9. This would result in an approximate 95% confidence interval of 50 $\pm$ 2(36.9) or 50 $\pm$ 73.8. That is to say that the actual population value in this community of persons who immigrated to Canada between 1996 and 2006 could be anywhere in the range from 0 to 123 with 95%

confidence. Even a somewhat less conservative figure using the 75th percentile adjustment factor (1.52) results in a 95% confidence interval that ranges from 5 to 95.

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	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

… not applicable

Source: Statistics Canada, 2006 Census.

Table 9.2 Standard error adjustment factors at national or provincial and weighting area levels

Characteristic	National or provincial factors	Percentiles of WA level factors						
		1	25	50	75	90	95	99
Population characteristic								
Age								
0 to 4, 5 to 9, 10 to 14, 15 to 19, 20 to 29, 30 to 34 years old	0.17	0.00	0.00	0.00	0.13	0.32	0.60	0.78
35 to 44, 45 to 54, 55 to 59, 60 to 64, 65+ years old	0.12	0.00	0.00	0.00	0.11	0.18	0.27	0.71
Sex	0.07	0.00	0.00	0.04	0.07	0.10	0.13	0.18
Common-law status								
In common-law relationship	0.19	0.00	0.00	0.00	0.09	0.23	0.65	1.03
Not in common-law relationship	0.16	0.00	0.00	0.00	0.04	0.19	0.39	0.87
Marital status								
Single, married	0.07	0.00	0.00	0.00	0.05	0.10	0.13	0.20
Separated, divorced, widowed	0.13	0.00	0.00	0.00	0.11	0.22	0.32	0.68
Highest degree, certificate or diploma	1.12	0.62	1.03	1.15	1.25	1.36	1.43	1.63
Major field of study	1.10	0.29	0.86	0.98	1.14	1.26	1.34	1.55
Place of birth								
Born in Canada	1.30	0.24	0.98	1.27	1.48	1.64	1.74	1.97
Born outside of Canada	0.98	0.56	0.97	1.07	1.19	1.31	1.38	1.55
Citizenship								
Canada, by birth	1.21	0.19	0.90	1.17	1.41	1.59	1.74	2.00
Other	1.16	0.06	0.78	1.10	1.37	1.61	1.73	2.00
Number of citizenships								
Canadian only	1.26	0.26	1.10	1.30	1.48	1.64	1.76	2.02
One or two other ones	1.67	0.00	1.17	1.38	1.64	1.92	2.09	2.52
Period of immigration								
Before 1950, 1951 to 1960, 1961 to 1970, 1971 to 1980, 1981 to 1990	1.37	0.44	0.91	1.06	1.26	1.43	1.55	1.79
1991 to 1995, 1996 to 2001, 2001 to 2006	1.67	0.00	0.91	1.18	1.52	1.87	2.09	2.46
Age at immigration	1.20	0.61	0.95	1.15	1.32	1.46	1.54	1.74
Mobility status (1 year ago)								
Non-mover	1.70	0.51	1.32	1.57	1.80	1.97	2.08	2.36
Mover (migrant, non-migrant)	1.86	0.32	1.46	1.71	1.94	2.16	2.29	2.67
Mobility status (5 years ago)								
Non-mover	1.58	0.59	1.32	1.51	1.70	1.86	1.96	2.10
Mover (migrant, non-migrant)	1.75	0.63	1.49	1.70	1.90	2.12	2.28	2.73

Table 9.2 Standard error adjustment factors at national or provincial and weighting area levels (continued)

Characteristic	National or provincial factors	Percentiles of WA level factors						
		1	25	50	75	90	95	99
Immigrant or non-immigrant population								
Immigrant population	1.30	0.79	1.29	1.44	1.60	1.76	1.87	2.11
Non-immigrant population	1.21	0.19	0.89	1.15	1.39	1.58	1.71	1.97
Visible minority								
Chinese, South Asian, Black, Filipino, Latin American, Arab, Korean, Japanese, Visible minority, n.i.e., Multiple visible minority	2.04	0.00	1.16	1.45	1.88	2.32	2.59	3.16
Aboriginal	1.41	0.54	1.36	1.61	1.85	2.07	2.21	2.52
Other	1.53	0.22	1.03	1.41	1.75	2.04	2.20	2.48
Ethnic origin	1.56	0.76	1.43	1.63	1.79	1.95	2.04	2.22
Home language – English								
New-Brunswick, British-Colombia, Ontario, Alberta	1.59	0.17	1.31	1.68	1.93	2.11	2.24	2.48
Quebec	1.61	0.79	1.57	1.80	2.01	2.18	2.33	2.62
Other provinces	1.24	0.00	0.82	1.37	1.73	2.01	2.15	2.49
Canada	1.16	…	…	…	…	…	…	…
Home language – French								
Nova-Scotia, Quebec	1.28	0.00	0.99	1.48	1.78	1.99	2.13	2.82
New-Brunswick	0.99	0.49	1.38	1.65	1.90	2.21	2.32	2.63
Other province	1.64	0.00	1.37	1.67	1.96	2.26	2.45	2.88
Canada	0.77	…	…	…	…	…	…	…
First official language spoken – English								
Quebec	1.43	0.77	1.32	1.55	1.75	1.92	2.02	2.25
Newfoundland	0.68	0.00	0.29	0.61	1.02	1.39	1.65	2.28
Other province	1.21	0.18	0.92	1.22	1.45	1.64	1.76	2.01
Canada	0.79	…	…	…	…	…	…	…
First official language spoken – French								
New-Brunswick	0.95	0.57	1.23	1.44	1.60	1.79	1.97	2.32
Other province	1.28	0.23	1.21	1.42	1.65	1.88	2.01	2.31
Canada	0.76	…	…	…	…	…	…	…
First official language spoken – other								
English or French	1.64	0.00	1.18	1.43	1.70	1.99	2.24	2.62
Neither	1.03	0.00	0.74	0.90	1.09	1.27	1.39	1.66
Official language spoken – English								
Quebec	1.36	0.00	1.24	1.41	1.58	1.74	1.87	2.25
Other province	1.28	0.37	1.11	1.33	1.50	1.64	1.74	1.90
Canada	0.84	…	…	…	…	…	…	…

Table 9.2 Standard error adjustment factors at national or provincial and weighting area levels (continued)

Characteristic	National or provincial factors	Percentiles of WA level factors						
		1	25	50	75	90	95	99
Official language spoken – French								
New-Brunswick, Quebec	1.25	0.56	1.23	1.42	1.56	1.68	1.76	1.94
Other province	1.62	0.00	1.16	1.38	1.68	2.05	2.29	2.91
Canada	0.94	…	…	…	…	…	…	…
Official language spoken - other								
Bilingual	1.22	0.72	1.23	1.36	1.48	1.60	1.70	1.88
Neither	1.44	0.00	1.04	1.26	1.52	1.77	1.94	2.34
Mother tongue – English								
Ontario, Alberta, British-Columbia	1.34	0.00	1.11	1.39	1.62	1.82	1.93	2.14
Quebec	1.39	0.57	1.24	1.45	1.66	1.85	1.96	2.21
Other province	1.10	0.00	0.92	1.19	1.44	1.66	1.79	2.02
Canada	1.07	…	…	…	…	…	…	…
Mother tongue – French								
New-Brunswick	0.85	0.00	1.04	1.31	1.52	1.75	1.87	2.31
Quebec	1.14	0.05	1.00	1.34	1.58	1.77	1.90	2.16
Other province	1.35	0.54	1.21	1.39	1.60	1.83	1.98	2.30
Canada	0.75	…	…	…	…	…	…	…
Mother tongue – other than English or French	1.76	0.00	1.02	1.24	1.56	1.95	2.24	2.87
Language of work – English								
Quebec	1.15	0.74	1.13	1.24	1.34	1.45	1.52	1.68
Other province	0.78	0.24	0.67	0.77	0.88	1.01	1.10	1.26
Canada	0.75	…	…	…	…	…	…	…
Language of work – French								
New-Brunswick, Quebec	0.88	0.38	0.75	0.87	10.60	1.25	1.33	1.48
Other province	1.22	0.00	1.07	1.20	1.34	1.52	1.67	1.97
Canada	0.72	…	…	…	…	…	…	…
Language of work – other	1.22	0.00	0.75	1.01	1.25	1.54	1.81	2.26
Industry	1.23	0.66	0.94	1.11	1.24	1.36	1.45	1.66
Occupation	1.17	0.69	0.87	0.97	1.12	1.24	1.31	1.47
Work activity in 2005	1.08	0.68	1.06	1.17	1.26	1.36	1.42	1.57
Weeks worked in 2005	1.02	0.57	0.92	1.07	1.20	1.29	1.35	1.48
Hours worked in reference week	1.09	0.51	0.86	0.98	1.12	1.22	1.29	1.44

Table 9.2 Standard error adjustment factors at national or provincial and weighting area levels (continued)

Characteristic	National or provincial factors	Percentiles of WA level factors						
		1	25	50	75	90	95	99
Full-time or part-time work								
Full-time work	0.82	0.35	0.71	0.81	0.89	0.98	1.03	1.13
Part-time work	1.09	0.58	0.98	1.07	1.15	1.22	1.27	1.39
Year last worked								
In 2006, in 2005, before 2005	0.86	0.39	0.82	0.97	1.12	1.24	1.30	1.44
Never worked	1.20	0.66	1.09	1.19	1.28	1.38	1.46	1.61
Class of worker								
Paid worker	0.89	0.43	0.85	1.02	1.25	1.41	1.50	1.71
Self-employed, unincorporated, unpaid family worker	1.26	0.66	1.11	1.23	1.35	1.47	1.57	1.85
Unpaid housework	1.20	0.65	1.10	1.20	1.30	1.39	1.45	1.62
Labour force status participation								
Employed	0.91	0.00	0.83	0.98	1.16	1.29	1.37	1.54
Unemployed	1.22	0.00	1.02	1.15	1.28	1.42	1.53	1.85
Not in labour force	1.07	0.64	1.02	1.16	1.28	1.41	1.49	1.68
Mode of transport to work								
Driver, walk, transit, passenger, taxi	0.77	0.39	0.74	0.85	0.93	1.03	1.09	1.25
Bike, motorcycle	1.21	0.00	1.01	1.13	1.24	1.36	1.46	1.77
Other	1.21	0.60	1.03	1.15	1.28	1.42	1.54	1.95
Place of work – province	0.56	0.00	0.70	0.85	1.02	1.22	1.37	1.74
Place of work – Statistical Area Classification (metropolitan influence zone [MIZ])								
Strong or moderated MIZ	0.95	0.48	0.99	1.09	1.19	1.31	1.40	1.67
Weak or not in MIZ	0.79	0.00	0.99	1.16	1.34	1.54	1.68	2.05
In a CA or a CMA	0.88	0.49	0.99	1.14	1.28	1.41	1.51	1.81
In the territories	0.49	0.00	1.09	1.28	1.49	1.79	2.00	2.30
Place of work – type of commuting								
Work in same CSD of residence	1.06	0.45	0.96	1.11	1.24	1.36	1.43	1.56
Work in a different CSD of residence	1.08	0.67	1.06	1.15	1.25	1.35	1.44	1.65
Place of work status								
Worked at home, no fixed workplace	1.27	0.62	1.13	1.24	1.35	1.46	1.53	1.67
Worked outside Canada	1.23	0.00	0.96	1.12	1.27	1.43	1.55	1.85
Usual place of work	0.99	0.46	0.85	0.95	1.05	1.14	1.19	1.31
Number of persons in total income intervals ($)								
0 to 9,999	0.73	0.16	0.56	0.67	0.77	0.87	0.92	1.05

Table 9.2 Standard error adjustment factors at national or provincial and weighting area levels (continued)

Characteristic	National or provincial factors	Percentiles of WA level factors						
		1	25	50	75	90	95	99
10,000 to 19,999, 20,000 to 29,999, 30,000 to 39,999, 40,000 to 49,999, 50,000 to 59,999, 60,000 to 69,999, 70,000 to 74,999, 75,000 or more	1.12	0.73	0.87	0.95	1.09	1.17	1.22	1.31
Census family status								
Husband, wife	0.12	0.00	0.05	0.08	0.11	0.15	0.17	0.23
Child	0.49	0.23	0.37	0.43	0.50	0.58	0.64	0.75
Female lone parent, male lone parent, non-member of a census family	0.80	0.44	0.71	0.83	1.02	1.16	1.23	1.41
Other	0.46	0.00	0.21	1.01	1.33	1.66	1.87	2.40
In census family								
Husband, wife, common-law partner present								
Yes, husband or wife	0.16	0.00	0.06	0.10	0.15	0.19	0.22	0.28
Yes, same-sex partner	1.63	0.00	1.40	1.55	1.73	1.92	2.02	2.39
Yes, same-sex married	1.28	0.00	1.04	1.21	1.37	1.51	1.61	2.12
Yes, opposite-sex partner	0.31	0.00	0.07	0.17	0.28	0.40	0.70	1.07
No	0.94	0.43	0.81	0.90	1.00	1.10	1.18	1.41
Economic family status								
Husband, wife (including same sex)	0.16	0.00	0.08	0.13	0.23	0.81	1.01	1.28
Lone parent, child	0.58	0.16	0.51	0.61	0.75	0.85	0.90	1.02
Other family member	1.26	0.00	0.83	1.03	1.28	1.47	1.60	1.88
All other population characteristics	1.00	…	…	…	…	…	…	…
Household and dwelling characteristic								
Tenure	0.77	0.55	0.80	0.90	0.99	1.07	1.11	1.19
Period of construction	1.03	0.80	1.04	1.11	1.17	1.23	1.28	1.39
Number of rooms	1.09	0.78	1.04	1.10	1.17	1.23	1.27	1.39
Number of bedrooms	1.02	0.47	0.95	1.05	1.12	1.20	1.25	1.38
Structural type								
Single-detached dwelling, apartment in a building	0.47	0.20	0.38	0.66	0.84	0.94	1.01	1.24
Other structural type	0.89	0.30	0.89	1.03	1.15	1.25	1.34	1.55
Household size								
1-person, 2-person household	0.06	0.00	0.00	0.00	0.00	0.05	0.10	0.24
Other	0.32	0.00	0.00	0.00	0.09	1.05	1.26	1.73
Primary household maintainer	0.00	…	…	…	…	…	…	…

Table 9.2 Standard error adjustment factors at national or provincial and weighting area levels (continued)

Characteristic	National or provincial factors	Percentiles of WA level factors						
		1	25	50	75	90	95	99
Age of primary household maintainer								
15 to 19	1.12	0.00	0.76	0.85	0.97	1.09	1.16	1.38
20 to 24, 25 to 29, 30 to 34, 35 to 39, 40 to 44, 45 to 49, 50 to 54, 55 to 59, 60 to 64	0.68	0.46	0.61	0.67	0.75	0.85	0.93	1.11
65+	0.36	0.23	0.31	0.35	0.39	0.44	0.47	0.58
Sex of primary household maintainer	0.70	0.47	0.64	0.71	0.81	0.89	0.93	1.01
Number of household maintainers								
One household maintainer	1.02	0.84	0.98	1.04	1.09	1.15	1.18	1.26
More than one household maintainers	1.04	0.00	1.00	1.08	1.20	1.36	1.47	1.85
Reference person is a household maintainer	1.89	1.03	1.61	1.82	2.00	2.18	2.28	2.51
Person who does not live here is a household maintainer	1.60	0.35	1.21	1.46	1.73	2.04	2.24	2.79
Number of households in gross rent intervals (intervals of $100)	1.09	0.56	0.87	0.99	1.11	1.21	1.28	1.43
Number of households in gross rent as a percentage of households income intervals								
Less than 10%, between 10 and 20%, 20 and 30%, 30 and 40%, 40 and 50%, more than 50%	1.10	0.62	0.84	0.90	0.96	1.02	1.06	1.17
Number of households in owner's major payment intervals (intervals of $200)	1.02	0.70	0.87	0.95	1.07	1.15	1.19	1.28
Number of households in owner's major payment as a percentage of household income intervals								
Less than 10%, between 10 and 20%, 20 and 30%, 30 and 40%, 40 and 50%, more than 50%	1.09	0.71	0.85	0.90	0.95	1.00	1.04	1.14
Person responsible for household payments								
Person is the first maintainer	0.00	…	…	…	…	…	…	…
Other maintainer	0.89	0.00	0.88	1.00	1.19	1.36	1.47	1.85
Number of households in household income intervals (intervals of $10,000)	1.05	0.63	0.84	0.90	0.97	1.07	1.14	1.25
Number of households in dwelling's value intervals	0.92	0.66	0.97	1.07	1.15	1.22	1.27	1.40
Registered condominium								
Part	0.88	0.62	0.96	1.06	1.16	1.27	1.34	1.51

Table 9.2 Standard error adjustment factors at national or provincial and weighting area levels (continued)

Characteristic	National or provincial factors	Percentiles of WA level factors						
		1	25	50	75	90	95	99
Not part	0.72	0.47	0.79	0.90	0.99	1.07	1.11	1.19
Condition of dwelling								
Regular maintenance, major or minor repair	1.09	0.90	1.06	1.12	1.17	1.23	1.26	1.35
All other household and dwelling characteristics	1.00	…	…	…	…	…	…	…
Census family characteristic								
Labour force activity of husband, wife or lone parent								
Husband or wife in labour force	0.61	0.36	0.56	0.65	0.73	0.80	0.85	0.93
Lone parent in labour force	0.99	0.69	0.91	1.01	1.11	1.22	1.29	1.47
Age of children at home	0.37	0.00	0.00	0.10	0.28	0.68	0.98	2.53
Work activity in 2005 of husband, wife or lone parent								
Worked in 2005	1.06	0.80	1.06	1.16	1.26	1.37	1.45	1.67
Did not work in 2005	0.97	0.72	0.94	1.06	1.22	1.37	1.46	1.73
All other census family characteristics	1.00	…	…	…	…	…	…	…
Economic family characteristics								
Number of households in self-employment income intervals ($)								
0 to 9,999	1.01	0.06	0.75	0.94	1.10	1.22	1.30	1.43
10,000 to 19,999, 20,000 to 29,999, 30,000 to 39,999, 40,000 to 49,999, 50,000 to 59,999, 60,000 to 69,999, 70,000 to 74,999, 75,000 or more	1.24	0.55	0.83	0.93	1.07	1.24	1.34	1.56
Low income status								
Above line	1.80	1.02	1.57	1.77	1.97	2.15	2.28	2.54
Below line	1.93	1.38	1.73	1.89	2.09	2.27	2.39	2.64
Property taxes included in mortgage payment								
Taxes included, Taxes not included	1.06	0.89	1.03	1.09	1.15	1.21	1.24	1.34
Wages and salaries ($)	1.11	0.70	0.88	1.01	1.15	1.25	1.31	1.44
Mother tongue of family reference person – English								
Newfoundland, Nova-Scotia, Prince-Edward-Island, Yukon Territory[1]	0.35	0.13	0.26	0.34	0.43	0.51	0.56	0.83
Quebec	1.01	0.72	1.02	1.10	1.18	1.28	1.34	1.50

Table 9.2 Standard error adjustment factors at national or provincial and weighting area levels (continued)

Characteristic	National or provincial factors	Percentiles of WA level factors						
		1	25	50	75	90	95	99
Other province	0.63	0.33	0.52	0.65	0.80	0.93	1.01	1.13
Canada	0.62	…	…	…	…	…	…	…
Mother tongue of family reference person – French								
Quebec	0.48	0.16	0.31	0.46	0.71	0.90	1.00	1.14
New-Brunswick, Yukon Territory[1]	0.65	0.29	0.74	1.02	1.13	1.22	1.26	1.33
Other province	1.03	0.66	1.04	1.13	1.22	1.31	1.37	1.56
Canada	0.56	…	…	…	…	…	…	…
Mother tongue of family reference person – Other than English or French	1.09	0.00	0.94	1.08	1.21	1.35	1.45	1.71
All other economic family characteristics	1.00	…	…	…	…	…	…	…

… not applicable

Note:

1. Effective October 20, 2008, the names 'Yukon Territory' in English and 'Territoire du Yukon' in French become 'Yukon' in English and in French. There is no change to the abbreviations as well as to the numeric and alpha codes for Yukon. The abbreviations remain 'Y.T.' in English and 'Yn' in French, and '60' for the numeric code 'YT' for the alpha code.

Source: Statistics Canada, 2006 Census.

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 2006 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 enumerators. Calibrating sample estimates to known population counts as part of the census weighting procedures helped to reduce the impact of these biases.

Appendix A Glossary of terms

The definitions of census terms, variables and concepts are presented here as they appear in the *2006 Census Dictionary* (Catalogue no. 92-566-XWE). Users should refer to the *2006 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): Census division (CD) is the 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/territory level and the municipality (census subdivision).

Census subdivision (CSD): Census subdivision (CSD) is the general term for municipalities (as determined by provincial/territorial legislation) or areas treated as municipal equivalents for statistical purposes (e.g., Indian reserves, Indian settlements and unorganized territories).

Census tract (CT): Census tracts (CTs) are small, 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.

A committee of local specialists (for example, planners, health and social workers, and educators) initially delineates census tracts in conjunction with Statistics Canada. Once a census metropolitan area (CMA) or census agglomeration (CA) has been subdivided into census tracts, the census tracts are maintained even if the urban core population subsequently declines below 50,000.

Collection unit (CU): Small geographic units used for the collection of census data. Collection units cover all the territory of Canada.

Collective dwelling: Refers to a dwelling of a commercial, institutional or communal nature. It may be identified by a sign on the premises or by a census representative speaking with the person in charge, a resident, a neighbour, etc. Included are lodging or rooming houses, hotels, motels, tourist homes, nursing homes, hospitals, staff residences, communal quarters (military bases), work camps, jails, missions, group homes, and so on. Collective dwellings may be occupied by usual residents or solely by foreign and/or temporary residents.

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

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 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.

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 2006 and 2001 census weights

Person constraints at the WA and DA level	**Household constraints at the WA and DA level**
-Total persons	-Households of size 1
-Total persons aged 15+	-Households of size 2
	-Households of size 3
-Males	-Households of size 4
-Males aged 15+	-Households of size 5
	-Households of size 6+
-Persons aged 0 to 4	-Total households
-Persons aged 5 to 9	
-Persons aged 10 to 14	-Dwelling type (new in 2006)
-Persons aged 15 to 19	-Apartments less than 5 storeys (new in 2006)
-Persons aged 20 to 24	
-Persons aged 25 to 29	
-Persons aged 30 to 34	**Other DA-level constraints**
-Persons aged 35 to 39	-Total households in DA
-Persons aged 40 to 44	-Total persons in DA
-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	

Appendix C Statistics used in sampling bias study

In Chapter 6, it is stated that under random sampling,
should follow an approximately normal (0,1) distribution. A justification for this is given here. Sampling was done independently in each CU. Therefore $\hat{X}^{(0)}$ is the sum of *H* independent random variables, where *H* is the number of CUs in Canada. There are 46,510 CUs in Canada that contain sampled households, so *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 CU 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

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.

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.

Dick, Peter. 2007. 'The 2006 Dwelling Classification Survey'. Report for the Federal-Provincial-Territorial Meeting on Demography September 21, 2007.

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.

———. 2006. *2006 Census Dictionary*. Catalogue no.92-566-XWE. Version updated February 14, 2008. Ottawa, Ontario, http://www12.statcan.ca/english/census06/reference/dictionary/index.cfm (accessed November 5, 2008).

Statistics Canada National Contact Centre

E-mail: infostats@statcan.gc.ca

8:30 a.m. to 4:30 p.m., Monday to Friday

Toll-free telephone (Canada and the United States):

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

Local or international calls:

- 1-613-951-8116 – Inquiries line
- 1-613-951-0581 – Fax line