INTRODUCTION TO THE MATHEMATICS *of*

DEMOGRAPHY

THIRD EDITION

ROBERT L. BROWN, Ph.D., FSA, FCIA, ACAS

INTRODUCTION to the MATHEMATICS of DEMOGRAPHY

(Third Edition)

Robert L. Brown, FSA, FCIA, ACAS

ACTEX Publications
Winsted, Connecticut

Requests for permission should be addressed to
ACTEX Publications, Inc.
P.O. Box 974
Winsted, CT 06098

Manufactured in the United States of America

10 9 8 7 6 5 4 3

Cover design by MUF

Library of Congress Cataloging-in-Publication Data

Brown, Robert L., 1949-
Introduction to the mathematics of demography / Robert L. Brown
p. cm.
Includes bibliographical references and index.

1. Demography--Mathematical models. I. Title.
HB849.51.B76 1991
304.6'01'5118--dc20 91-13446
CIP

TABLE OF CONTENTS

PREFACE

In spite of an increased interest and publication activity in the general subject area of demography, a major void still remains: a suitable textbook to introduce undergraduate students to the mathematics of demography.

Textbooks and journal papers on the topic of demography will typically have one of two basic foundational sources. Many find their basis in sociology, and are almost entirely descriptive and qualitative in presentation. Then there is a wide variety of texts and papers written from a mathematical perspective, but these are often either devoid of the descriptive material necessary for a sound pedagogical presentation, or they are advanced works more suitable for graduate courses.

This text, appropriately called *Introduction to the Mathematics of Demography*, tries to strike a happy compromise between a qualitative presentation and introductory, undergraduate quantitative analysis. There is much material that is purely descriptive, but there is also an extensive survey of the foundational mathematics required in much of the work done by demographers. The mathematical content can be understood by undergraduate students with a background of algebra, introductory calculus, and introductory probability.

Examples are drawn from North American applications and studies. Each chapter includes a number of worked examples and a set of exercises for the reader to solve. Many of the text exercises

appeared on recent Course 161 examinations given by the Society of Actuaries. Their permission to include these questions in the text is greatly appreciated.

The author wishes to acknowledge the heavy influence on this text of predecessor works by Keyfitz [21], Pollard, Yusuf, and Pollard [33], and Keyfitz and Beekman [22]. The last was especially important, as many of this text's examples and problems arose from similar problems presented earlier in that text.

The author wishes to thank a number of professionals who reviewed several drafts of the text and made valuable contributions to the finished product. Notable among these persons are Jeffrey A. Beckley, FSA, Beckley & Associates, Inc.; John A. Beekman, ASA, Ph.D., Ball State University; Richard Bilisoly, FSA, Society of Actuaries; James C. Hickman, FSA, Ph.D., University of Wisconsin; Robert Hupf, FSA, Mutual of Omaha Insurance Company; Bertram M. Kestenbaum, ASA, Social Security Administration; Richard F. Lambert, FSA, Prudential Insurance Company of America; and Elias S. Shiu, ASA, Ph.D., University of Manitoba.

Special thanks to Alice H. Wade, ASA, Social Security Administration, who, as chairperson of the Society of Actuaries Course 161 Examination Committee, was very supportive of the project and coordinated the manuscript review efforts of the committee members. Thanks also to Ms. Wade for supplying the U.S. Life Tables which appear in Appendix A.

Finally the author expresses his appreciation to the people at ACTEX Publications, who turned a rough manuscript into a finished textbook. The principal players on that team included Marilyn J. Baleshiski, format and layout editor, Sandi Lynn Fratini, style editor, Dick London, FSA, mathematics editor, and Marlene F. Lundbeck, cover and graphic arts.

While the contributions of the above-named persons are sincerely acknowledged, the author takes full responsibility for any remaining errors or deficiencies in the text.

Robert L. Brown, FSA, FCIA, ACAS
University of Waterloo
Waterloo, Ontario
April 1991

PREFACE TO THE SECOND EDITION

In the two years that have passed since the publication of *Introduction to the Mathematics of Demography*, classroom use of the text by the author and others has revealed the need for some important revisions. These revisions are in the nature of improvements in presentation rather than in content. Thus the Second Edition covers substantially the same material that was covered in the original text, but has clarified and improved the presentation in many areas.

In Chapter 1, the presentation of Myers' blending technique has been improved through an example, and an appendix has been added showing the derivation of the formula for the standard error of an estimate in a one-in-six sample. In Chapter 4, an example has been added to clarify the higher-age extrapolation procedure used in constructing the U.S. Life Tables, an issue that has perplexed students in this area for years.

The introductory presentation of stationary and stable population theory, in Chapters 5 and 6, respectively, has been completely rewritten to provide a clearer understanding of those models. In particular, the nature of the stationary model as a special case of the stable model has been more clearly shown.

Throughout the text an effort has been made to attain consistency in notation for similar concepts, and, conversely, to use different notation for sufficiently different concepts. For example, the new edition

now uses b_c for the crude birth rate of a population, which is a statistic of actual data, and b_i for the intrinsic birth rate of a stable population, which is a parameter of a mathematical model. The original text unwisely used the symbol b for both concepts. Other examples of notational improvement can be found by readers familiar with both editions of the text.

We believe the improvements contained in the new edition will further the ability of the text to attain its goal of providing a clear introduction to the mathematics of demography.

RLB
April 1993

PREFACE TO THE THIRD EDITION

It has now been four years since the publication of the Second Edition of this text. The Preface to that edition points out that it covers substantially the same material as in the original text, but with significant improvements in its pedagogic presentation. The new Third Edition of *Introduction to the Mathematics of Demography* similarly makes some presentation improvements, but includes several important changes in content as well.

In Chapter 1, the description of census-taking in Canada and the United States now refers to the 1991 Canadian census and the 1990 U.S. census, updating the Second Edition which described the 1986 and 1980 censuses in Canada and the United States, respectively. This updating echoes throughout the text, in that later numerical examples and exercises are evaluated using the most recent census data.

The material on Myers' blended population method, a technique for detecting and correcting digit preference in the reporting of ages, has been deleted from the text. Modern censuses now ask for year of birth, rather than age, and there is very little evidence of age misreporting in a well-educated population.

Consistent with the updating of Chapter 1 for the most recent national censuses, Chapter 4 now presents a description of the construction of the most recent national life tables, the 1989-91 U.S. Life Tables (replacing the 1979-81 version) and the 1990-92 Canadian

Life Tables (replacing the 1985-87 version). Also in Chapter 4, improvements have been made in Section 4.4 dealing with abridged life tables.

In Chapter 6, some confusing material in Section 6.3 has been rewritten to clarify the various methods of estimating a population's intrinsic growth rate from census data, and the most recent census data has been used for these estimates. The Section 6.5 description of quasi-stable population theory, always a difficult topic to understand, has been improved.

A major addition to the new edition is Section 7.4, describing *dynamic* methodologies for population projection in lieu of the *static* approaches presented in Section 7.3. The fertility theories of Easterlin and Ermisch, which relate to the topic of dynamic population projection, have been moved to this chapter. (In the prior edition of the text, these theories were described in Chapter 2.)

Throughout the text a considerable number of new exercises has been added.

Finally, in Appendix D, a mathematical proof of the Sharpe-Lotka Theorem, a crucial part of the stable population theory described in Chapter 6, is presented. The author would like to express his appreciation to Professor Gordon E. Willmot, Ph.D., FSA, of the University of Waterloo, for contributing this proof to the new edition.

The author would also like to thank Claire Bilodeau, ASA, a graduate student at the University of Waterloo, for her invaluable help in updating many parts of the text.

We hope the reader familiar with prior editions of this text will agree that significant improvements have been made in the new edition, both in the area of important new material and improved pedagogic presentation.

RLB
April 1997

INTRODUCTION to the MATHEMATICS of DEMOGRAPHY

(Third Edition)

CHAPTER 1

DATA: SOURCES AND ERRORS

1.1 INTRODUCTION

Demography is a term derived from two Greek words, which, if translated literally, means "*to draw or write about people.*" According to the United Nations Multilingual Demographic Dictionary, "*Demography is the scientific study of human populations, primarily with respect to their size, their structure, and their development.*" A more precise definition, given by Bogue, is as follows:

> "*Demography is the statistical and mathematical study of the size, composition, and spatial distribution of human populations, and of changes over time in these aspects through the operation of the five processes of fertility, mortality, marriage, migration, and social mobility*" ([6], p. 1).

Mathematical demography had its beginnings in the development of procedures for the formation of life tables (see Chapters 3 and 4). The 1662 work by John Graunt, *Natural and Political Observations Upon the Bills of Mortality* (see Keyfitz and Smith [25], pp. 11-20), is cited as the first substantive demographic work to be published. Graunt was able to derive an impressive array of demographic information using only lists of christenings and deaths during the time of the

London plague. A portion of Graunt's report is illustrated in Table 1.1 below.

A second work of importance in mathematical demography was the Breslau Table of 1693, a life table developed by Edmund Halley, after whom Halley's comet is named. (See Keyfitz and Smith [25], pp. 21-26.)

TABLE 1.1

Table of Notorious Diseases		Table of Causalities	
Apoplexy	1306	Bleeding	9
Cut of the Stone	38	Burnt and Scalded	135
Falling Sickness	74	Drowned	829
Dead in the Streets	243	Excessive drinking	2
Gowt	134	Frighted	22
Headache	51	Grief	279
Jaundice	998	Hanged themselves	222
Lethargy	07	Killed by several	
Leprosy	6	accidents	1021
Lunatick	158	Murdered	86
Overlaid and Starved	529	Poisoned	14
Palsy	423	Smothered	265
Rupture	201	Shot	17
Stone and Strangury	863	Staved	51
Sciatica	5	Vomiting	136
Sodainly	454		

Another important step in the development of mathematical demography was the publication in 1798 of *An Essay on the Principle of Population* by Reverend Thomas Robert Malthus. He postulated that human populations would naturally increase faster than the food supply needed to sustain them. Malthus believed that what kept the population at a sustainable level were checks that he categorized as vice, misery, and moral restraint. Vices which controlled population growth were wars and excesses of all kinds. The misery category included diseases, epidemics, famine, plague, and so on.

1.2 THE COLLECTION OF DEMOGRAPHIC STATISTICS

Virtually all basic demographic data come from censuses, surveys, or vital statistics registration systems.

A ***census*** (from the Latin *censere*, to assess) has been defined as "*the total process of collecting, compiling and publishing demographic, economic, and social data pertaining, at a specified time or times, to all persons of a defined territory.*" An in-depth discussion of census-taking follows by looking at how the United States and Canada take a census and what is produced.

A census may be taken on either a ***de facto*** or a ***de jure*** basis. Under the *de facto* method, persons are counted wherever they happen to be at the time of the census. Under the *de jure* method, persons are counted according to their usual place of residence, so that those temporarily absent would be counted as if at home.

Both Canada and the United States use the *de jure* method, except in the case of transients where the *de facto* method applies. In the rest of the world, the *de facto* method is more common and is recommended by the United Nations Population Commission.

The advantage of the *de jure* method is that it gives a picture of the permanent population of the communities enumerated. This information could be used, for example, to determine how many political representatives that jurisdiction should have.

The disadvantage of the *de jure* method is that some persons may be omitted from the count, since they are absent from their usual residence, or could be counted twice, once at their temporary residence and again at their usual residence. Further, information secured secondhand about persons who are temporarily absent may be incomplete or incorrect.

Conversely, the main advantage of the *de facto* method is that it offers less chance of double-counting or omission of persons. The disadvantages of the *de facto* method are threefold: it is difficult to obtain information about persons in transit, it provides an incorrect picture of the usual population of a community, and vital statistic rates may be distorted because the population base is not relative to the vital statistics (e.g., persons normally return to their usual residence for the birth of a child).

In Canada and the United States, most census forms are mailed out, one per household, and mailed back. The head of the household will usually answer the questions for all members of that household. In countries with lower literacy levels, each individual is interviewed by a trained enumerator or canvasser.

An advantage of the mail-out approach is that the census can be taken on one particular day. The use of enumerators usually means that the census must be spread out over several days or weeks. This may lead to problems with respect to births, deaths, and migration during that period of time. If the time lapse is prolonged, it may also lead to inaccuracies because of poor memory.

Surveys are used to discover certain errors in the census (see Section 1.4), to provide information between censuses at more frequent intervals, and to provide information on topics not included in the census. For example, the monthly sample surveys conducted by the Census Bureau in the United States and by Statistics Canada provide data on monthly labor force statistics, including unemployment rates.

Sample surveys contain errors of coverage, classification, and sampling error, which will be discussed in detail later (see Section 1.4). While these surveys are usually very accurate on a national basis, they should be used with care on a sub-national basis. Each survey publication should indicate the accuracy of the data. Users of the survey are advised to review this information carefully in order to understand the limits of the presentation. All definitions should be carefully checked so that the terms used are clearly understood. This is especially true for data analysis done by agencies other than the Census Bureau or Statistics Canada. Terms in the data published by other agencies are commonly used incorrectly, such as the use of the term rate in place of ratio (see Chapter 2).

The term ***vital statistics*** generally refers to data regarding vital events such as birth, adoption, death, marriage, divorce, legal separation, and annulment. By legal requirement the data are usually recorded at the time of the occurrence of the event. The registration of these events is a provincial responsibility in Canada. In the United States, each state has the responsibility for the registration of its own vital statistics. The cities of Baltimore, New Orleans, and New York maintain systems independent of the states in which they are located.

In some instances, vital statistics are combined with census data for certain reports, such as the Life Tables of the United States and Canada. For example, the 1989-91 U.S. National Life Tables combine mortality data from state registries for the period of 1989 to 1991 and

compare it to the census population of April 1, 1990. This combination derives a central death rate at each age. A similar combination of data is used in the 1990-92 Canadian Life Tables.

A ***population register*** is a system of continuous registration with an entry for each individual. Population registers are used by a number of countries, including Netherlands, Belgium, Finland, Sweden, Norway, Denmark, Iceland, Italy, Gibraltar, Germany, Israel, Japan, Taiwan, Bulgaria, and Czechoslovakia.

Population registers are useful in providing a permanent up-to-date data base for legal identification of persons, elections, military service, and so on. Special studies can be done by choosing appropriate samples of the data. On the negative side, population registers are expensive to maintain and may become defective. They are usually used only in countries with a high literacy rate and low migration.

Finally, data on immigrants and emigrants are usually collected at the point of their entry or exit, such as an airport. This presumes legal migration. The number of illegal immigrants into the United States makes statistics on migration there less credible.

1.3 TAKING THE CENSUS

There are indications of population enumerations being made as early as 3800 B.C. in Babylonia and 2275 B.C. in China. The first census in North America was taken in New France (Quebec) in 1666 by Jean Talon. Decennial censuses have been taken in the United States since 1790 and in Canada since 1851. Canada has also taken a smaller quinquennial census since 1956. The Canadian census takes place on June 1, and the U.S. census takes place on April 1, or as close to these dates as possible; for example, the census would not be taken on a Sunday. These dates are chosen to maximize the number of people at home at their normal residence (the *de jure* method) while being close to midyear.

In Canada, the census, and several other demographic surveys, are the responsibility of Statistics Canada. In the United States, the census is the responsibility of the Census Bureau, which is a part of the Department of Commerce. In addition to taking the decennial Census of Population and Housing, the Census Bureau conducts the Economic Censuses, a Census of Agriculture, and a Census of Governments

every five years. The Bureau also conducts hundreds of surveys, some as often as once a month.

The U.S. Census is required by law, under Article 1, Section 2 of the Constitution, to ensure representation in the House of Representatives by population size. However, the census is designed to satisfy the needs of a broad range of possible users. While keeping the questions the same from census to census enhances continuity and comparative statistics, the questions posed change slightly in every census. The number of questions posed is a compromise between the needs of users and the cost of collecting and processing the information. Both the Census Bureau and Statistics Canada expend considerable energy trying to optimize the questions to be asked so as to satisfy as many users as possible within budgetary constraints.

Both Canada and the United States widely pretest the census questions and census forms before the actual census, and provide extensive publicity at the time of the census. Both countries also provide telephone backup to the census, whereby individuals can phone in toll-free to seek guidance with the questionnaire, usually in their language of choice. Complete confidentiality of individual information is guaranteed. No one can get data from the census bureau that would allow identification of any individual information.

To optimize the collection of a wide database, while limiting the overall cost, both Canada and the United States use ***sampling techniques*** whereby not every household answers every question.

In the 1991 Canadian census, 80% of the households in urban areas were given a short form that asked questions as to name, date of birth, sex, marital status, mother tongue, type of dwelling, and dwelling ownership. The other 20% of urban households, and all rural and northern households, were given the long form which repeated all of the questions from the short form, but also included questions on labor force activity, income, education, disability, citizenship, housing (dwelling characteristics and shelter costs), ethnicity, and language. The census forms were available in 32 languages in addition to English and French. Braille versions were also available.

The 1990 United States census included a short form questionnaire that approximately five out of every six urban households in the nation received. The short form asked 14 questions, of which 7 applied to each person, 6 applied to housing conditions, and the final was a control question to assure that all household members were

counted and no visitors were counted. A random sample of approximately 1 in 6 urban households received a long form of the questionnaire that asked these questions, and an additional 19 questions about housing conditions and 26 additional questions about each individual. Several of these questions had multiple parts, but it was not necessary for everyone to answer every question. Some questions applied only to households or persons with certain characteristics.

In smaller towns and counties (less than 2500 in population) one out of every two households received the long form. The 50% sample rate was used in areas that constituted approximately one-tenth of the nation's population. In total, about 81% of the population completed the short form, and 19% completed the long form.

Both the United States and Canada have used a combination of ***mail-out/mail-back*** and ***enumerators***. For example, in the 1990 United States census, the Postal Service delivered census questionnaires to about 83% of all addresses in the country, primarily in metropolitan areas. For another 11% of the nation's housing units, mostly in rural and seasonal-housing areas, enumerators visited every housing unit before census day, and left a census questionnaire to be completed and returned by mail. The overall mail return rate was 74.1%. In sparsely populated parts of the country, where it is often difficult to determine mailing addresses and not cost effective for enumerators to drop off questionnaires, the Postal Service delivered unaddressed questionnaires to all known housing units. Members of these households were to complete the forms and hold them for collection by enumerators. Enumerators recorded the addresses when they picked up the questionnaires. This technique applied to only about 6% of all households but covered 50% of the nation's land area.

Special procedures were used with people who lived in special places such as group homes, prisons, hospitals, nursing homes, convents, military installations (both at home and abroad), colleges and universities, and so on. The 1990 U.S. census was the first to have an in-depth enumeration of selected components of the homeless population at shelters, subsidized hotels and motels, and city-designated street locations, as well as bus, subway, and train stations. March 31, 1990, between 4 p.m. and 10 p.m. was designated as T-night, when as many transients as possible were enumerated. T-night enumerators visited and interviewed transients at YWCA's, YMCA's, hostels, campgrounds, marinas, carnivals, and so on.

In the 1991 Canadian census, less than 2% of households were enumerated by canvassers. Canvassers were used to enumerate each household in remote or northern areas and on Indian reserves where there is irregular mail service. Some remote areas were enumerated during March 1991 (as opposed to June) if the community became migrant after the spring breakup. For the first time, an attempt to enumerate homeless people, through soup kitchens, was done on an experimental basis by interviewers. A special short form was used to enumerate individuals in hospitals and jails. The enumeration was actually done from the institution's administrative records, and only basic information was collected for each resident.

In the 1991 Canadian census, the population count included both permanent and non-permanent residents. This was only the second time that non-permanent residents were enumerated (the other time being 1941 to reflect the situation imposed by World War II). Non-permanent residents are persons who hold student or employment authorizations, special permits, or who are refugee claimants. They were included in the 1991 census because they now make up a growing segment of the population. Their presence can affect the demand for services such as health care, schooling, employment programs, and language training. Also, vital statistics (e.g., births and deaths) include non-permanent residents. Finally, this census definition is now closer to the United Nations recommendation that all long-term residents be enumerated. Because of this change, population counts, and all statistics derived from population counts (e.g., birth and death rates), will be affected. As a result, users should be especially careful when comparing data from 1991 and previous censuses.

1.4 SOURCES OF ERRORS AND THEIR CORRECTIONS

The significance of data error to the users of that data depends on the nature of the error, the intended use of the data, and the level of detail involved. Some errors occur more or less at random and tend to cancel out when individual responses are aggregated for a sufficiently large group. For example, some people may overestimate their incomes, while others underestimate them. If there is no systemic tendency to err in either direction, then the errors will offset each other

in any large aggregation. On the other hand, if there is a systemic tendency for people to err in a particular direction (e.g., if incomes are generally understated), then the average reported will be different from the true average. Such ***systemic errors*** are far more serious a problem for users of the data than are ***random errors***. The bias they create in the data persists no matter how large the group, and it is very difficult to measure.

Errors can arise form many sources, but can be grouped into a few broad categories which will now be reviewed in some detail.

1.4.1 Coverage Errors

Every effort is made to minimize coverage errors, but such errors do occur. Examples include an enumerator missing a dwelling entirely, or the householder not listing all the usual residents of the dwelling. Sometimes errors lead to double counting or ***overcoverage***, although this is usually much less of a problem than ***undercoverage***, which occurs when individuals or households are missed.

In the United States, the primary program for measuring undercoverage is the ***Post Enumeration Survey***. This program involves several post enumeration surveys that attempt to measure the quality of the census data. In addition to measuring undercoverage, the program tries to measure the number of persons erroneously included in the census (such as babies born after April 1), persons counted more than once, and persons coded to the wrong geographic area.

The Census Bureau combines results from these surveys to form an estimate of the net undercount or net overcount. The combined samples from the surveys are large enough to permit publication of relatively reliable data at the state level. The 1990 census was adjusted from the actual count of 248.7 million to 252.7 million based on such surveys. Thus the post enumeration surveys indicated an undercount of 4 million, or 1.58%.

In Canada, the ***Coverage Error Measurement Program*** serves a similar purpose. On June 4, 1991, the actual census count was 27,296,859. However, to account for an estimated net population undercoverage of 2.87% (807,254 persons), the census count was adjusted to 28,104,113.

Another method used to quantify the census undercount or overcount is to estimate the expected result of the census count by building

up the population from the last census count (unadjusted) with births, deaths, and immigration. This method can be formulated as

$$P(t) = P(0) + B - D + I - E, \tag{1.1}$$

where $P(t)$ is the predicted population at time t, B is the births in the period, D is the deaths in the period, I is the immigrants (usually legal only) in the period, and E is the emigrants in the period. The excess of the actual census count over the predicted census count, divided by the actual count, is called the ***error of closure***.

According to this method, the U.S. census missed 3.3% of the population in 1950, 2.6% in 1960, and 2.3% in 1970. In 1980, the component approach given by Equation (1.1) yielded an estimate of about 221.7 million for the true population of legal residents. The final 1980 census count was roughly 226.5 million, showing a significant improvement in coverage (mostly of illegal immigrants). For the 1990 census, the projection of the 1980 census gave an estimated count of 250.2 million versus the actual count of 248.7 million, for an error of closure of -0.6% (an undercount).

Undercoverage might not be of great concern if it were random among population groups or geographic areas. Previous censuses, however, have indicated that undercoverage varies in some specific regions and groups. These errors can seriously affect the federal and state funding provided to major U.S. cities, such as New York.

1.4.2 Content or Response Errors

Sometimes it proves impossible to obtain a complete response from a household, even if the dwelling was identified as occupied and a questionnaire was delivered. The household members may be away during the entire census period, or the members may refuse to complete the form even though required by law. More often, the questionnaire is returned but information is missing for some questions or some individuals.

Census enumerators in the United States and census representatives in Canada are responsible for seeing that questionnaires are completed for every address in their area. Where questionnaires arrive with missing information, enumerators phone or visit the households until all questions have been answered.

In the 1990 U.S. census, enumerators were required to attempt to contact the household up to six times: the initial visit, three telephone attempts, and two additional personal visits. If the enumerator could not make contact during these attempts, or if the household refused to provide any information, the enumerator would try to obtain last-resort information from another knowledgeable source, such as a neighbor or apartment manager. Enumerators also delisted nonexistent units and sorted out duplicate listings.

Some nonresponse is inevitable, and although certain adjustments for missing data can be made during processing, some loss of accuracy must follow.

After detecting nonresponse errors, values for missing or incomplete entries are imputed. ***Imputation*** can be by ***allocation*** or ***substitution***. When omissions are completed by inferring the correct value from other questionnaire answers, it is referred to as an allocation. An example would be a missing marital status for a 19-year-old son that is filled in as single because a 22-year-old son was also reported as single. This is also called the ***deterministic*** approach. A substitution, on the other hand, selects a record that has a number of characteristics in common with the record that is missing or in error, and imputes the missing information from this "donor" record. For example, suppose a housing unit is reported as occupied, but no other information is given. The full set of information for that unit would be taken from a similar "donor" unit. The method of substitution is also referred to as the ***probabilistic*** approach.

In general, the higher the allocation rate the more variance one can expect in the data. Allocations may also introduce bias in the data, if, on average, characteristics of nonrespondents differ from those of respondents (e.g., income level).

Response errors also occur. The respondent may have misinterpreted the question or may not know the answer, especially if given for an absent household member. Even an enumerator can introduce responses that are in error.

The Census Bureau gets a measure of response errors by means of content evaluation studies, including a reinterview of about 12,000 long-form housing units. The Bureau compares the results with the responses recorded in the original records of the census, and adjusts the original census accordingly.

1.4.3 Misstatement of Age

In both Canada and the United States today, nearly all census data analysis starts with quinquennially grouped data. Furthermore, the census forms now ask for year of birth, rather than age. Along with the higher levels of education existing today, these factors have eliminated the problem of digit preference, so no adjustments for it need be made. Historically, a preference for reporting certain ages, such as those ending in 0 or 5, could be detected in the data, and special adjustments were required to offset these errors.

1.4.4 Processing Errors

After census day the questionnaires are sent to regional processing sites. Nonwritten responses are usually in machine readable form. Written responses, such as name, must be separately coded. The coded information is then computerized either by keypunching or by electronic transferral. Mistakes can occur in coding or transmission, despite rigorous quality checks.

1.4.5 Sampling Errors

As previously noted, in both Canada and the United States all households receive the short-form questionnaire, but only a subset of the population receives the long-form questionnaire. The information collected from these households is weighted to produce estimates for the entire population. The simplest weighting procedure would be to multiply the results for the sample households by the sample ratio (five for many Canadian households and six for many U.S. households). This procedure is not used, however.

Sampling error can be reduced by using a complicated technique called ***ratio estimation***. First, weights are derived from the ratio of complete-count short-form questionnaires to long-form samples within particular areas and population subgroups. Then instead of multiplying the sample responses by five (in Canada) or six (in the U.S.), the weights are adjusted so as to reproduce the total demographic characteristics for the area or subgroup revealed by the short-form questionnaires. For example, suppose in a particular area we know from the short form that there are 200 heads of household who have university degrees, but a one-in-five sampling reports only 36 rather

than the expected 40. For this variable, all responses from the sample would be multiplied by $\frac{200}{36}$ rather than by 5 to estimate the total population response.

As described, this method assumes that a primary variable, such as university degree, is used to extrapolate from the sample to the 100% survey. In 1991 Statistics Canada estimated weights (all close to 5) for each of the variables age, sex, and marital status, and then chose a single weight that produced the least squared error.

It is possible to present a mathematical statement of the variance associated with these sampling techniques. For example, the ***standard error*** of an estimate based on a sample of one in six households is

$$SE_{\hat{X}} = \sqrt{5\hat{X}\left(1 - \frac{\hat{X}}{N}\right)}, \tag{1.2}$$

where $\hat{X}$ is the estimated number of units with some characteristic and N is the total number of units (households) in the area. (See Appendix A for a derivation and fuller explanation of Equation (1.2).) For example, if a one-in-six sample estimates that 247 persons ($\hat{X} = 247$) in an area with 5021 inhabitants ($N = 5021$) have a certain characteristic, then the standard error is

$$SE_{247} = \sqrt{5 \times 247 \times \left(1 - \frac{247}{5021}\right)} = \sqrt{1174.25} = 34.3.$$

The corresponding formula for the standard error for estimated percentages is

$$SE_{\hat{p}} = \sqrt{\frac{5\hat{p}(100 - \hat{p})}{N}}, \tag{1.3}$$

where $\hat{p}$ is the estimated percentage, given by $.01\hat{p} = \frac{\hat{X}}{N}$. For example, the standard error of an estimate of 15 percent ($\hat{p} = 15$) with a total of 1243 units ($N = 1243$) can be computed as

$$SE_{.15} = \sqrt{\frac{5 \times 15 \times (100 - 15)}{1243}} = \sqrt{5.1} = 2.3,$$

measured in percentage points. (See Appendix A for the derivation of Equation (1.3).)

1.4.6 Gross Error Ratio / Net Error Ratio

Demographic data are subject to several sources of error, including errors of coverage, errors of content (or response errors), misstatement of age, processing errors, and sampling errors.

If data are published in grouped form, which is the norm, then it is possible for some errors to cancel. For example, some people may report their age incorrectly, but remain in the correct age group. Or a number of people may overstate a value while an equal number understate the same value.

This results in two measures of error, the ***gross error ratio*** and the ***net error ratio***. The gross error ratio would measure the proportion of persons who are missclassified because of an error of content. The net error ratio is the net difference between the theoretically correct answer and the answer provided by the count, with the allowance of some errors being offset as described above. Net error ratios are usually much smaller than gross error ratios. It is also the case that both measures may only be estimates since it may not be possible to arrive at the absolutely correct answer.

To formulate this presentation, consider the following table.

TABLE 1.2

Perfect Count	Reported Count		Total
	Number in Class	Number Not in Class	
Number in Class	a	b	$a+b$
Number Not in Class	c	d	$c+d$
Total	$a+c$	$b+d$	$n=a+b+c+d$

The reported count showed $a + c$ in the class, whereas the true or correct census should have been $a + b$. The net error ratio is given by

$\frac{(a+b)-(a+c)}{n} = \frac{b-c}{n}$, usually expressed as a percentage. The total gross error, on the other hand, is represented by the sum $b+c$, so the gross error ratio is given by $\frac{b+c}{n}$, and is usually given as a percentage. The following example illustrates these concepts.

Suppose in a community of 1000 people, it is known that 20% are university graduates. Then $n = 1000$ and the number of people actually in this class is $a+b = 200$. Suppose the reported responses resulted in a net error ratio of -6% and a gross error ratio of $+14\%$. This tells us that $b+c = 140$ people responded incorrectly, either by saying they were graduates (although they actually were not) or by saying they were not (although they actually were). Similarly we see that $b-c = -60$ from the net error ratio, so we can now find $b = 40$, $c = 100$, $a = 160$, and $d = 700$. Thus we see that of the 200 university graduates, 160 confessed to it and 40 tried to hide it!

For census and vital registrations it is impossible to achieve a perfect result. However, there are several techniques in use that attempt to discover and correct errors in the data. We have already discussed methods used to reduce errors of coverage and content in the U.S. Census.

1.5 CONCLUSION

Census data quality information is disseminated in two ways. All census publications include a section on data quality that examines sources of errors and provides cautionary notes for users. In some cases, estimates of the magnitude of errors are given, such as estimates of sampling error. Information is also available in reports that summarize the results of data quality studies.

Other methods to determine and correct errors are limited only by the ingenuity of the demographer. For example, the update formula method, given by Equation (1.1), can be used to predict not only the total population, but any subcategory as well. For example, one group that is prone to census count errors is young males. Because this is a highly mobile group, they may be missed or double-counted in the census. By updating the age groups [5-10) and [10-15) from a previous census, the demographer can get an early estimate of the number to be expected among young males in the current census.

Similarly the demographer can look at sex ratios from adjacent age groups. Obviously, there should be a smooth and natural progression of sex ratios. The number of male births normally exceeds the number of female births. For example, in Canada in 1991, 51.3% of births were male and 48.7% were female. This is normally stated as 105 male births for every 100 female births. At higher age groups, because of the higher mortality for males, this ratio will decline to 50/50 and beyond. The presence of age groups where the ratio produced by the census count is at odds with the projected ratio will indicate a possible source of error and will also provide the demographer with a possible correction factor.

Users of census and vital registration statistics must take the responsibility to determine the errors inherent in the data. Further, users must be sure to understand all the terms used in the presentation of the data. For example, does the term "income" refer to family income or individual income? What does the word "unemployed" actually mean? Users should take the care to review all definitions and source-of-error information carefully and to be aware of data limitations in any subsequent analysis.

1.6 EXERCISES

1.1 Introduction; 1.2 The Collection of Demographic Statistics

1-1. With respect to the taking of a census, (a) define and differentiate the *de facto* and *de jure* approaches, and (b) list their advantages and disadvantages.

1-2. Define and give examples of vital statistics. At which political level are vital statistics collected in Canada and the United States?

1-3. Define population register and list its advantages and disadvantages.

1.3 Taking the Census

1-4. Outline the sampling techniques used in taking the 1991 Canadian census and the 1990 United States census.

1.4 Sources of Errors and Their Corrections

1-5. With respect to census-taking, define and differentiate the following items.

(a) Coverage errors
(b) Content errors
(c) Processing errors
(d) Sampling errors

1-6. How are coverage and content errors addressed?

1-7. In a region with 50,000 inhabitants, one-in-six sampling was used to estimate the number of inhabitants owning BMW's. If the unadjusted standard error of the estimate was 17.31011, how many reported owning a BMW?

1-8. Define and differentiate between allocations and substitutions.

1-9. A small country has three provinces. A *defacto* census shows population counts of 45,000 for Province X, 25,500 for Province Y, and 64,500 for Province Z. A post-census survey indicated that on census day, (i) 5% of Province X residents were in Province Y and 10% were in Province Z, (ii) 10% of Province Y residents were in Province X and 10% were in Province Z, (iii) 0% of Province Z residents were in Province X and 5% were in Province Y. What would a *de jure* census find?

1-10. A dog breeder had to go away on a one-month trip. Before he left, he counted how many dogs he had. He also directed his assistant to record the births, deaths, sales, and purchases while he was away. Upon his return, he performed another count of the dogs in the kennel. From the following data, determine the error of closure.

First count, before departure	53
Assistant's records:	
Births	12
Deaths	1
Sales	10
Purchases	3
Second count, after return	58

1-11. Consider the following data for Canada from 1986 to 1991:

Resident population June 1, 1986 (from census)	26,010,310
Non-permanent residents as of June 1, 1986	166,987
Activity during 1986-1991:	
Births	1,930,140
Deaths	944,936
Immigrants	985,127
Emigrants	212,465
Net influx of non-permanent residents, exclusive of births and deaths	207,849
Total population June 1, 1991	27,293,910

Modify Equation (1.1) to incorporate non-permanent residents, and determine the error of closure.

1-12. In a certain community of 1000 persons, 460 were reported as having a certain characteristic. Upon re-enumeration it was discovered that really only 400 persons had this characteristic. Find the net error ratio. Can the gross error ratio be calculated?

1-13. Consider the following information:
Actual population on January 1, 1980: 100,000

	Number of Occurrences Recorded Between	
	1/1/80 and 1/1/81	1/1/80 and 1/1/90
Births	1,500	18,000
Deaths	850	10,000
Net Immigration	350	6,500

The recorded occurrences between 1/1/80 and 1/1/81 were used to model the expected population change from 1/1/80 to 1/1/90. Only then did the true statistics become available. Determine (a) the linearly projected 1/1/90 population, (b) the actual 1/1/90 population, and (c) the error of closure.