Contents

Chapter 2

Production of Data

2.1 Introduction

Statistics is primarily concerned with how data are used. Data concerning human or other populations may be used to describe samples or populations, analyze relationships among variables, or draw inferences concerning populations from samples. This textbook shows how each of these can be carried out using data which have already been produced. While the manner in which data are produced is not analyzed in detail in this textbook, the production of data is an integral part of the research process. This introductory section comments on the production of data so that those who use data can be aware of some of the issues involved in the production of data.

Users of data often assume that the data are accurate, and that the work of the statistician is to analyze these data. This section is intended to show that such an assumption is often mistaken. The various methods involved in producing data may have as much or more impact on the conclusions than does the method of statistical analysis.

The purpose of this section is to provide an overview of the issues related to data production so that those who produce or use data are aware of some of the limitations of the data. The points raised in this section are likely to be discussed in much greater detail in courses or books on Social Science Methodology. However, these comments provide a brief introduction to the problems of data production, so that those who intend to produce their own data have some idea of the various stages and difficulties involved in doing this. In addition, for those who work with data that has already been produced by others, it is worthwhile to spend some time examining

the sources and limitations of this data. These provide some idea of what the data means and how it can be interpreted. In many ways, the issues and problems discussed in this section are more basic than the statistical issues raised later. If a researcher does not have well founded data to begin with, statistical analysis of the data may be misleading or useless.

In this chapter, several of the issues involved in the production of data are examined, with various examples being provided. Following that, these issues are illustrated in greater detail in a specific example, that of the production of labour force statistics. As an introduction, some general comments concerning the meaning of the title of this section **the production of data** follow.

2.2 Data as Social Product

Data refer to the facts or information obtained from observation of people, objects or phenomena. These data may be quantitative in form, if they refer to concepts such as height, income, or age, each of which can be measured on a well known and well understood scale. Data could instead refer to characteristics which are qualitative in form, but which are either difficult or impossible to quantify. For example, much historical data is a record of people, events and dates, none of which can be considered easily quantifiable. Concepts like ethnicity are inherently nonquantifiable, referring to differences in ancestry and experiences. In the case of these latter types of qualitative data, statistics may be used because one is concerned with counting the number of people, events, objects, etc. Much of the use of statistics in the social sciences is concerned with counting, determining the number of people or objects with a particular characteristic, for example, the number of Metis, the number of people in poverty, etc.

For each type of data which is used for statistical analysis, the data are not merely collected, but the data are **produced**. This means that at all stages of obtaining data, researchers who develop and use data must make decisions concerning how these data are to be defined and collected. Data are not freely available as some product of nature, one that can be gathered with little effort. Instead, theoretical, practical, and perhaps political, decisions must often be made concerning what data to develop and how to produce these data.

Data can be considered to be a **social product**, produced by individuals or groups in society. Researchers or organizations having particular

purposes or interests produce data for many different reasons. The aim of each researcher and organization in obtaining and organizing data may determine the particular form that the data take. Even when looking at a specific social issue, quite different sets of data concerning this issue may be produced by those having different reasons for analyzing this issue.

When producing data, researchers must carefully consider what they want to achieve in tackling a particular issue and how they will go about producing data that answer the questions they are concerned with in dealing with that issue.

This chapter examines a number of the issues involved in the production of data, the various stages of organizing the production of these data, and the potential problems or errors that can emerge in data production. The following discussion is not an exhaustive examination of all the issues involved in the production of data but raises some of the main practical issues and problems that should be considered by researchers and students who produce or work with data.

2.3 The Availability of Data

Once researchers have decided to examine or study a particular social issue, they will likely first be concerned with the types of data that are available concerning this particular issue. A research project should always begin by attempting to determine whether some individual or group has already produced data concerning the issue of interest. There are two main reasons for this. First, social analysis should not start from the beginning in each case, assuming that no research has been done in the area being examined. Rather, the contributions and types of approaches used by other researchers will be important guides to any new research that is carried on. Even if a researcher strongly disagrees with previous approaches, these earlier research efforts will be useful in providing data which can be questioned or reanalyzed with current research.

The second reason for using data which have already been produced, is that it is very costly and time consuming to produce data. Given the budget and time constraints most researchers face, there is often little choice but to use data which is already available. If researchers are to produce meaningful results, it may often be a more preferable and efficient approach to use data that are already available, rather than attempting to produce completely new data. This may be the case, even if earlier data have considerable

shortcomings.

For example, suppose that an organization is interested in studying some issue related to poverty in Canada. Since there are considerable data already available concerning this issue, it would be foolish to ignore these data. The Census of Canada, various Statistics Canada surveys, and data provided by non governmental organizations, all provide considerable information concerning the demographic, income, labour force, and housing situation of individuals and families, both those in and not in poverty. While there may be considerable inaccuracies in these data, it would be best to begin by examining these data. Based on these data, the researcher will probably be able to answer some questions of interest to the organization carrying out the study.

At the same time, there may be little information concerning some aspects of poverty. For example, the state of mind, the political views, or even the health, of people who suffer from inadequate incomes, are not likely to have been examined in any of the Statistics Canada data sources. If these latter issues are a major concern of the organization's study, the researcher will be faced with the choice of using only fragmentary sets of data on these issues, or deciding to produce some new data concerning these issues. If the researcher decides to do the latter, then this may entail considerable time and cost. If these issues are really are of concern to the organization, then the researcher will have to embark on data production.

If data are not available at all concerning some issue or problem, then the researcher is faced with the problem of how to begin defining and collecting data. Some of the issues relating to this are discussed in later sections.

2.3.1 Uneven Availability of Data

As has been noted, the availability of data related to different social issues or problems is very uneven. In some areas of study, researchers and organizations have developed many data sets, while in other areas there is almost no data available. Since data is a social product, one must consider how knowledge is developed in society. In most areas of investigation, this is an uneven process, developing rapidly in some areas, but developing much more slowly in other areas.

One key to understanding which types of data are developed and which types are not developed, or only poorly developed, is the interests, needs and power of the individuals and organizations which have produced data. For

those who have financial and organizational resources, the production of data is an important aspect of maintaining and furthering their control of these resources. Those individuals or groups who have little power, resources, knowledge or interest in a particular topic, may have little data concerning issues which affect them.

In the economy, where production of goods and services forms the basis for the economic or financial wealth of individuals and groups, the availability of data is generally fairly good. Most countries have considerable data concerning the quantities and value of the production and distribution of most agricultural and industrial products. Statistics Canada produces many volumes of data concerning the production and sale of products such as diverse as wheat, steel, and restaurant meals. Detailed statistics are available concerning exports and imports of many of these products, and concerning the distribution and consumption of these products.

At the same time, data concerning the environmental and health effects of the production of these goods and services, or the manner in which the workers in these industries, or the consumers of the products of these industries, are affected is generally not available. While agricultural and industrial production may have serious negative effects on workers, communities and the environment, these are generally less well understood or ignored. Only in the last few years have many people become aware of the effects of acid rain, insecticides, PCBs, etc. In the case of many of the products that involve these potentially harmful substances, the potential profits and incomes to producers have been so great that they have either ignored or covered up these problems. Data concerning these problem areas are unlikely to be made available by those whose profits or incomes depend on producing the products that create these problems.

Government statistical agencies have not made the production of data concerning these problem areas and issues to be a priority. Those people adversely affected by such production are unlikely to have the resources to obtain or produce the required data. The consequence of this is that data concerning these problems are usually not available.

One of the first steps in solving some potential health and environmental problems associated with various types of production would be to develop data concerning the adverse effects of these types of production. For example, workers who are working with products or processes which may be unsafe, may need to develop data concerning the prolems associated with these processes.

Another example of the scarcity of data is the very limited amount of

information concerning the situation faced by people of aboriginal origin in the cities of Canada. Until very recently, almost no data were available at all on this issue. In contrast, the Census made available to market researchers very considerable details concerning the incomes of individuals and families, especially better off ones, in Canadian cities. As an illustration of the difficulty that native groups and researchers have had when dealing with the urban problems of people of aboriginal origin, the question of how may people of aboriginal origin there are in cities has not been adequately answered. Without being able to even identify the number of people and families involved, it seems unlikely that much progress will be made in improving the the economic and social conditions faced by many of these people. If the economic and social condition of people of aboriginal origin is to be improved, one step that needs to be taken is to improve the data concerning the situation of these people. This can help provide some of the raw material for beginning to solve some of the problems faced by urban aboriginal people.

2.3.2 Control of Data

One of the problems encountered in working with data is that data may be available, but researchers are denied access to them. This may be because the data are either confidential or is the property of some group or individual. These two situations are somewhat different in nature, with the latter point being discussed first.

Data as Property

Data that are produced and collected are often the **property** of the person or group that produces the data or has paid for the production of the data. If the groups owning the data are unwilling to share with others access to the data, then researchers have little choice but to look for other data sources.

Ownership of data, and denial of access by researchers to the data, may be quite legitimate in some instances. In other cases, the data may be used by those having access to these data to further only their own interests, or to use it at the expense of others.

Data collected by corporations or government administrators may be necessary for these organizations to carry on their activities. Most often, these data are closely guarded and become the private property of the company or agency which has paid for the production of these data. Where these data give information on a situation which affects other individuals or

groups, it can be argued that these data should become much more widely available. This would be especially true of cases where the individuals or groups affected, are harmed as a result of processes or policies to which the data refer.

An example of the debate over the control of data comes from Western Canada, where rail line abandonment has been of great concern for rural communities. There, data concerning the practices of railroads and the profitability of rail lines are available to the railroad companies themselves. One of the debates concerning the availability of data has centred around the question of how much access farmers, rural communities and regulatory agencies will have to these data, data which are regarded by the railroad companies as their property. Small towns and rural communities have argued that these data should be public because of the manner in which the data concern and affect the communities. The availability of data concerning potential occupational or environmental health problems presents similar problems.

Closely guarded data such as military data may sometimes legitimately regarded as confidential. However, if what is classified as confidential involves data concerning issues which are not really related to national security, then wider access to these data should be allowed. Related to this is the use of opinion polls and surveys of public opinion. Where this involves polling during an election campaign, or at other times, and is paid for by political parties, these data are quite legitimately regarded as the property and confidential data of the party. However, where public money is used for polls, or where the results of polls are used by politicians to hang on to political power or sway public opinion, such confidentiality is quite illegitimate. Any poll paid for with taxpayers' money should become public information.

Confidential Personal Data

The issues related to **confidential personal information** are rather different. These data are not ordinarily used for personal or corporate profit, but are confidential in order to protect the privacy and rights of individuals. For example, when the Census of Canada collects data from a family or individual, these data are aggregated with data from many other families or individuals. The publication or release of these data is organized in such a way that no individual or family can be identified. Data are released only if they refer to enough people so that information concerning any individual

cannot be determined.

While the Census takes great care to protect the confidentiality of individuals, other organizations may be less careful. With the growth of vast data banks, considerable information concerning individuals is available to anyone who can obtain access to the data bank. This may involve financial, health and other types of data concerning individuals. In recent years, with the growth of computers and the use of Social Insurance Numbers (SIN), unique to each individual, concern has been expressed by many people that data in different data banks can be easily linked together on the basis of the SIN. If this were done, a profile of any individual could be provided, perhaps with the individual not being aware of its existence.

Many organizations do place safeguards on the use of the data they produce. For example, in Saskatchewan, the Medical Care Insurance Commission (MCIC), the Hospital Services Plan (SHSP) and the Drug Plan all have large data banks with considerable information concerning individuals and doctors. However, data concerning individuals from these data banks are not available to those outside each of these organizations. For research purposes, data from these data banks may be made available to researchers. These data are aggregated, so data concerning individuals cannot be identified, or researchers are sworn to confidentiality.

2.3.3 Types of Data Sources

Much of the data we encounter in daily life, such as labour force statistics, consumer price data, and public opinion polls, is obtained from government or private surveys. Some of the statistical issues related to survey data are discussed in more detail later in the next section. The Census of Canada, conducted every five years by Statistics Canada, is another major data source.

Another major source of data for research work is based on administrative records of public or private organizations. These are data records which were not collected primarily for purposes of data analysis, but were collected for administrative reasons. A high quality administrative data set in Canada is the immigration records of Employment and Immigration. Apart from illegal immigrants, upon entry to Canada, each immigrant is required to provide a considerable amount of personal information to Employment and Immigration. These records are obtained by Employment and Immigration as a means of monitoring and controlling the amount and type

of immigration. But because these immigration records provide a considerable amount of information on individual immigrants, they can be used to develop a good profile of immigrants. In contrast, emigration from most countries is not controlled or monitored. Data on emigrants are usually quite incomplete and most countries, including Canada, have only very indirect means of estimating the amount of emigration from the country.

As another example, in Saskatchewan, MCIC and SHSP have relatively complete information on the number of people in the province, as well as their age, sex and place of residence. This data set is a byproduct of bureaucratic needs involved in administering these programs in the province. Since almost all residents of the province register with these programs, the data sets produced from these administrative records are fairly complete and accurate. This data set would be one means of providing estimates, albeit indirect, of the amount of emigration from the province.

2.4 The Problem of Definition

When approaching a specific research problem or issue in the social sciences, it is necessary for the researcher or analyst to become familiar with this area or problem and identify a particular approach to research. In doing this, several definitional issues are likely to emerge. The reseacher must define the exact problem, issue or area to be investigated, and must decide which individuals or objects are to be studied. In addition, the researcher will have to make decisions concerning which characteristics of these individuals or objects are relevant, and how data concerning these is to be obtained.

In producing the data which will assist in answering some of the research issues, the particular approach used by a researcher may be quite different than that adopted by other researchers. The researcher comes to the research problem with his or her own individual and academic background and experiences. This is likely to lead to a specific set of ideas and approaches. The researcher will base these approaches on a review of the work of others, but this is likely to be supplemented by the researcher's own experiences and ideas. There may also be various suspicions, statements or hypotheses which the reseacher wishes to either verify or negate. If the problem involved is one involving the use of some quantitative methods, then several definitional decisions must be made by the researcher. Some of these definitional issues are discussed in this section.

Even if the researcher or analyst does not produce his or her own data, but uses data produced by others, or merely reviews or analyzes the work done by other researchers, some of the same problems emerge. While the approach of the earlier studies cannot be changed after the fact, those who use these earlier studies should make themselves very clear concerning which research approach these other studies adopted. Where data production was involved in this earlier research, the definitional decisions made by these other researchers should be well understood. When using the data and research of others, the first stage of research usually involves a very close examination of the definitions and approaches used by these other researchers.

When a decision has been made to use quantitative or statistical methods, the main definitional issues revolve around the definition of the **population** and the **variables**. That is, the researcher must decide which specific set of numerical quantities he or she wishes or needs to investigate. In doing this, the researcher will need to decide which particular individuals, objects or measurements are required, and exactly where to draw the limits in defining these. Work done by previous researchers will guide this. However, it may be that previous approaches were not completely adequate or appropriate to the current research problem. In this case, a decision will need to be made to change the approach, perhaps adopting some new definitions. If the data have already been produced, those using this data should become very familiar with the definitions used by those who originally produced the data.

In the following sections, the two major definitional issues are (1) how the population is defined and (2) which variables concerning that population are examined, and how these variables are defined.

2.4.1 The Population

The population, or universe, is what the researcher wishes to examine. This may be an actual population of people such as the population of a province of Canada, the population of Political Science majors at a university, or the population of women who are married. Each of these is a fairly clearly identifiable group, and should involve relatively few definitional problems. Alternatively, populations could be objects, for example, the population of all maple trees in Canada. A population could instead constitute a set of measurements on some phenomenon such as temperatures, for example, the population of all temperatures at Inuvik over the course of a year. The

notion of a population can be defined more clearly as follows.

The **population** is the set of all people or objects having a common characteristic (or some common characteristics).

In this definition, a population can be thought of as composed of members of the population. These could be individual people or individual objects; collectively they make up a population. In defining a population, the researcher must decide, or be aware of, exactly which individuals or objects are the concern of the study or investigation. These individuals or objects must have at least one common characteristic, or they may have several characteristics in common.

This is a fairly straightforward definition in that it includes each person who is registered in the class. The common characteristic is that each member of the population must satisfy the criterion that he or she is registered in the class. Even here though, the researcher would need to be careful to ensure that all those registered in the class are included, and that none of those being considered as a member of this population has dropped the class. Also, if some people were auditing the class or attending the class without having being registered, the researcher would have to make a decision concerning whether or not to include these.

Example 2.4.1 A Study of Trees in Saskatchewan

As in the last example, it is likely to be relatively clear what is and what is not a member of this population. The individual members of the population are objects which are trees, and they must be trees growing within the geographical boundaries of the province of Saskatchewan. However, it may sometimes not be clear what is a tree and what is only a shrub or bush, so that the researcher would have to make clear exactly what is meant by **tree**.

In the above examples, the members of the population are individual people or objects. However, it may be that the members of the population are themselves groups. For example, the members of the population could be families, households, towns, etc. In the case of the latter, the population would then be the set of all towns in some area.

Example 2.4.2 The Distribution of Incomes in Canada

Much data have been produced concerning the distribution of incomes of Canadians. Some of this research involves investigation of the income levels of individuals Canadians. In this case, the population is likely to be defined as the **set of all adults living in Canada**. *Children are usually excluded from studies of this sort, because children ordinarily would have little or no income independent of their parents. Those individuals who are in Canada but not residents of Canada are also excluded.*

For purposes of investigating how well off people are in terms of the resources at their disposal, **family income** *may be a more useful concept. The majority of people in Canada live in some family, or household, setting. In this case, the members of the population are the families of Canada, and this population is the* **set of all Canadian families**. *The researcher would have to define exactly what is meant by a family and which groups of people living together would be considered families and which would not be considered families. There may be considerable problems in doing this, given the variety of family forms and living arrangements that exist in Canada today. For example, if a researcher were to limit analysis to families having both a husband and wife present in the household, a large portion of both adults and children would be left out of consideration. In light of these changing living arrangements, it may make more sense to measure* **household income**. *In this case, the relevant population would be* **the set of all households in Canada**. *This would then create the definitional problem of defining what constitutes a household.*

An Alternative Definition of Population or Universe. Statisticians may prefer to define a population without reference to people or objects. Instead, they may view the population as a set of measurements so that the population is defined as **the set of all measurements that concern the researcher**. Another variant is to define the population as **the set of potential units for observation**. In this approach, it may be more common to refer to the population as **the universe** of observations.

The advantage of this definition is that it focuses attention on the characteristic or characteristics being investigated. The researcher may not be particularly concerned with the individuals being studied as a set of people. Rather the researcher may be more concerned with some characteristic concerning these people, such as their attitudes or opinion on some issues, their income, their ethnicity or some other characteristic. That is, the focus is

on the measurements and the nature of these, rather than on the people or objects. For example, in the example of temperatures measured in Inuvik, it is difficult to associate these temperatures with any individuals or objects, rather the temperatures are measurements of a phenomenon of nature. It would then make sense to refer to all the temperatures as the universe which the researcher is investigating. Similarly, if a researcher were examining the distribution of incomes in Canada, the population, or universe, could be considered to be the set of all incomes in Canada. The distribution of these could then be examined to see, for example, how may high, middle and low incomes there were.

While this alternative approach may be useful when only one characteristic is being examined, it is more common to think of the population as being defined by the original definition as the set of individuals or objects having some common characteristic with which the researcher is concerned. This original definition also corresponds more closely to the sampling or experimental procedures which are used to obtain most social science data. In surveys and experiments, a set of individuals or objects is clearly defined as the set with which the researcher is concerned. Following this, various characteristics of the members of this population are obtained and examined. That is, once the population has been defined in this manner, then the researcher can think of which set of individuals or objects are to be investigated. Then, the various characteristics of these people or objects can be identified and measured.

2.4.2 The Variables

The various characteristics of the members of the population are the variables that describe the population. Researchers usually are attempting to describe various characteristics of a population, and then examine the relationship among these variables. For example, researchers may be attempting to measure a variable such as political attitudes of the population of Saskatchewan and the relationship of these political attitudes to other variables such as area of residence, income, occupation, and education. In this study, it would be necessary for the researcher to define which people in Saskatchewan are to be surveyed, and then obtain data on variables such as area of residence, income, occupation, education and political attitudes.

Definition 2.4.1 A **variable** is a characteristic of the members of the population.

Once the population has been defined, the researcher can begin to define some of the characteristics of the members of the population. If these values of any characteristic change across the members of the population, then this characteristic can be considered to be a variable. In the special case where the characteristic is identical for all members of the population, then this characteristic could more properly considered to be a **constant**.

Example 2.4.3 The Distribution of Income in Canada

In this example, the major variable to be considered is the income of Canadian individuals or families. In obtaining data on incomes, the researcher would have to make clear exactly what constitutes income. Wages and salaries obviously should be included, but what about gambling wins or losses, capital gains or losses from stocks and real estate, and inheritances? Clear definitions of exactly which items the researcher wishes to include and which to exclude, would have to be made. The definition the researcher adopts may differ considerably from that Revenue Canada uses on income tax returns. In fact, the latter contain several different definitions of income.

In researching this issue in Canada, the definitions most commonly used are those developed and adopted by Statistics Canada. These are summarized as follows. As can be seen there, some of the items which would add to or subtract to the amount of money a person has in any given year are not included. In general, the definition adopted appears to be that those items which count as **regular** *income sources for an individual or family are included as income by Statistics Canada. Those receipts of individuals or families which are unusual, or are one time gains or losses, are not included as part of income.*

Statistics Canada's Definition of Income and Its Components

Included as income sources:

1. Wages and salaries.
2. Net income from self-employment.
3. Investment income
4. Government transfer payments such as Family Allowance, Old Age Security, Unemployment Insurance.

5. Pensions.
6. Miscellaneous income such as scholarships or alimony.

Some items not included as income sources:

1. Gambling gains and losses.
2. Money or property inherited.
3. Capital gains or losses.
4. Lump sum insurance settlements.

Source: Statistics Canada, **Income Distributions by Size in Canada 1989**, catalogue number 13-207, pages 36-38.

It is likely that the researcher is also interested in other variables associated with the incomes of Canadians. Researchers would likely want to know, for example, how incomes change with age, education, sex, area of residence, family size, and so on. These other characteristics such as age, education, and sex are also variables in that they are characteristics of the population which vary across the members of the population. Also note that, in this study, one characteristic that does not vary across members of the population is **residence in Canada**. By definition of the population, all those being investigated are resident in Canada, and this characteristic can be considered to be a constant here.

Example 2.4.4 Attitudes of Moose Jaw Residents Toward Abortion

Here the population could be the set of all residents of Moose Jaw. However, children would usually be excluded from such an attitude survey and a researcher would ordinarily restrict the definition of the population to Moose Jaw adults. There may be many different types of attitudes concerning abortion, such as who should decide concerning whether or not an abortion should be performed, under what circumstances it should be available, how it should be regulated, should abortion be available only in hospitals, etc. In investigating this issue, the researcher will have to decide which attitudes to investigate, and how to design questions which will give an accurate idea of the true attitudes of the members of the population.

2.4.3 Theoretical and Operational Decisions.

The last two examples show that there are many decisions which the researcher must make in deciding how variables will be defined. In doing this and in obtaining data concerning the variables describing the members of the population, the researcher will be guided by various theories and theoretical approaches. The **theoretical decision** which must be made by the researcher is to decide which theoretical concept he or she will attempt to measure or determine. The particular theoretical approach adopted by the researcher can have a major impact on the manner in which the variables are defined. For example, if the researcher wishes to work with the theoretical concept of social class, a Marxist approach to theory will define this quite differently than will conventional stratification theory. (See Example 2.4.5 below). Different researchers coming from different theoretical approaches may be looking at the world in such different ways that the variables of concern are themselves different, as well as being defined differently.

Once the researcher identifies and adopts a particular theoretical approach, this theory must be made **operational** by deciding how information will be obtained concerning each theoretical concept. A variable may be quite clearly defined in theory, but in the real world, where actual conditions may differ rather considerably from the theoretical world, the exact approach to be used may not be nearly so clear cut. Operationalizing a theoretical concept may involve features such as designing questions for a survey, designing an experiment, constructing a scale of measurement, and then measuring the variable. Each step of this set of procedures may involve several judgments that the researcher must make.

The above example concerning income of Canadians showed some of the particular operational decisions taken by Statistics Canada concerning which monetary receipts would be considered to be income, and which monetary receipts were not regarded as income. The particular operation decision there appeared to be to include only those monetary receipts which occur fairly regularly, and exclude those which are one time receipts. However, a certain judgment concerning the regularity of each has been made by Statistics Canada, one which other researchers might have operationalized somewhat differently. In the case of the measurement of attitudes and opinions, the researcher must design particular questions, along with possible responses, which the researcher feels will measure the concept which he or she has in mind. This is often a very difficult and hazardous process, since small changes in the wording of questions may considerably change the reported

results.

In many cases, especially when measuring attitudes, there may be no absolutely right or absolutely wrong way to design questions. Rather, the researcher must attempt to avoid obvious biases, and attempt to avoid influencing the respondent. It is impossible to remove all of bias on the part of the researcher when invesigating attitudes. Since it is important to measure such attitudes and opinions, the researcher should always clearly report the research approach used, the definitions adopted, the questions used, the range of possible responses as well as the method of sampling. In doing this, the nature of the definitional and research decisions can be seen by others, and anyone analyzing the results will be better informed concerning potential biases in the data.

Example 2.4.5 Social Class

Different theoretical approaches taken by sociologists and other social scientists give quite different views of the social class composition of a society. The variable or theoretical concept to be investigated here is **social class**. That is, research in this area concerns the general question of which people in the population are members of each social class. The concept of social class in theory, though, is defined quite differently by different theorists. In addition, the operationalizing of this concept will be different in each theoretical approach so that different researchers produce quite different values for this variable.

1. **Social Stratification Approach**. In North America, it has been common to approach the study of social class by classifying the population into strata from high to low. This results in a view of society as being divided into strata such as lower class, middle class, or upper class. Researchers adopting this approach consider it possible to measure social class on a continuum from the lowest to the highest social class, and they would consider it possible and meaningful to rank or place people or families at some position on this scale.

   The researcher who has decided to use this approach will still have to operationalize this definition by considering other variables such as income, occupation or prestige which can be used to rank people on this scale. These other variables will then need to be combined into a single scale which can be used to rank people. In Canada, the

Blishen and the Pineo-Porter scale are examples of this. In the United States, the Duncan scale of socioeconomic status is an example of this approach. Following this, the researcher might also have to decide the points along the scale which would distinguish the lower from the middle class, and the middle from the upper class.

2. **Marxian Approach**. A quite different approach is taken by Marxists. In this tradition, social class is not considered as a continuum along a scale, and Marxists might consider it impossible or meaningless to attempt to construct such a scale. Rather, for the Marxist, people are members of quite different classes, classes which are usually opposed to or in conflict with each other. In fact, the class may be at least partly defined by its being in opposition to other classes or in opposition to the interests of other classes. The two major classes in this approach are workers, or proletarians, and capitalists, or the bourgeoisie. There are intermediate, but less important, classes such as the petit bourgeoisie (small businesspeople), farmers and landlords.

   Marxists would also have to operationalize their definitions. This would involve deciding exactly who is a worker and who is a capitalist. Should everyone who works for a wage be considered a worker, including highly paid managers and professionals? How much property can a person own and still be considered a worker - a house, automobile, some personal possessions only - or some savings bonds and a few stocks, and if so, how many? How much would a business have to be worth, and how many employees would there need to be in the business before the owner could be considered to be a capitalist? All these are operational issues which Marxist social researchers have had to struggle with.

3. **Weberian Approach**. A third major approach is that based on the approach taken by the German social theorist, Max Weber. This approach accepts the Marxist approach of owners and workers as one basis for the formation of social class. In addition, the Weberian approach provides for many more types of social classes, built on the relationship that individuals have with markets. For Weberians, there may be other classes such as creditors, debtors, professional groups, landlords, renters, etc., a multiplicity of classes. In addition, a Weberian considers status honour to be an important aspect of class formation. This constitutes an additional theoretical difficulty, because

> the researcher must define not only market situation, but also status honour. In addition to defining each of these concepts theoretically, the Weberian researcher would then have to decide how each of these concepts could be measured and, once measured, how a small set of social classes could be constructed so that members of any population could be classified as being in one or other of these classes.

Based on these considerations, it is easy to see why there is considerable disagreement concerning how many people are in each social class. Researchers cannot even agree on what categories are most appropriate, let alone how many members of each category there is in any society. As a result, anyone who uses the concept of social class in their research should make very clear which approach and set of measurements they are using.

2.5 Methods of Obtaining Data.

After defining the population to be investigated, and which variables concerning this population need to be obtained and analyzed, the researcher must then decide how to obtain the data. If the population is relatively small, or if the researcher has a large research budget, it may be possible to obtain information concerning the desired variables from all members of the population. More likely though, the population is too large or the time and budget constraints are too restrictive to allow the researcher to obtain information from each member of the population.

In most cases, the researcher will select a sample of members of the population, or conduct an experiment using some members of the population as subjects for the experiment. As will be seen in this textbook, if carefully carried out, such an approach can provide adequate and precise information concerning the population and the variables which describe this population. Much of the statistical approach to producing and analyzing data is built on the notion of sampling from a population, in order to obtain the required data.

This section contains a short discussion of some principles concerning selection of the members of the population which the researcher is studying. The section following this contains a discussion of some of the limitations of such methods, and the possibility of errors when producing data. To begin this discussion, a distinction between a census and a sample is useful.

Definition 2.5.1 A **census** is a complete enumeration of all members of a population.

In a census of a population, information on the relevant variable or variables is obtained from each member of the population. There are a variety of methods of obtaining the data from each member of the population. The data might be obtained on the basis of questionnaires, interviews, administrative records, or experiments. Regardless of the method, the researcher will have to obtain information concerning the characteristics of each member of the population.

If the population is small and the information required is not too detailed, then a census of the population is feasible. For example, in an undergraduate class in Statistics, the professor would find it easy to conduct a census of the students in the class. The class may not be too large, and most students enrolled in the class are present most days, so that the process of surveying the members of such a class is likely to be relatively easy for the professor. However, if the population is quite large or if considerable information is required, it may not really be feasible to undertake a complete census of the population. For example, a census of all university students is likely to be a much more difficult task. However, it a researcher has access to the records of the university registrar, and the information the researcher desires is contained on these records, then a census of university students could be conducted by using the administrative records.

For Canada as a whole, a census is conducted by Statistics Canada only once every 5 years. Censuses of Canada are conducted in years ending in 1 or 6, in 1981, 1986, 1991 and 1996. The Censuses which are conducted in years ending in 1 have generally obtained information on more variables than have the Censuses in years ending in 6. However, in all of these Censuses, the aim is make sure that every resident of Canada is counted in the Census and the desired information obtained about all Canadian residents.

The information obtained in a census is often considered superior to that obtained in a sample that surveys only some members of the population. There are several limitations, however, on the quality and availability of data that can be produced from a census.

Possible Limitations of Census Data.

1. **Incomplete Coverage.**

   If a census really does survey all members of the population, then this is the only way of obtaining an absolutely accurate count of the number

of members of the population. In most cases, though, some members of the population are missed in a census enumeration, or the variables which are being investigated are missing for some members of the population who have been surveyed in the census. Refusal to answer certain questions, or lack of knowledge of some questions may mean information obtained from a census is incomplete. (In Section 2.6.1 these are referred to as nonsampling errors).

2. **Cost and Time.** The cost of conducting a census may be prohibitive, especially if the population is large. In addition, it can take considerable time to conduct a proper census of a population. For a fixed and limited budget, and a short time schedule, it may be possible to obtain information of greater quality more quickly through a sampling method.

3. **Nuisance Factor.** Frequent censuses of a population could become tiring or a nuisance for respondents. A small sample of a population might be conducted with most members of a population being unaware of it.

4. **Nointerference.** Related to the last point, the very process of obtaining information from population members may cause the nature of the population to change. If the researcher is attempting only to observe a population, and not to alter the nature of this population, then a sample of population members may mean there is less interference with the population.

   For example, suppose workers at a workplace are each asked to identify what is wrong with the manner in which the workplace is organized. If everyone is asked this same question, it is likely that workers will begin to discuss the problems associated with the workplace and begin to demand some changes in the manner in which work is organized in the workplace. If the intent of the researcher is to provoke such a response on the part of workers, then it may be most appropriate to survey each worker. Most employers would not allow such a research project in the workplace, so that if the researcher wishes to obtain the permission of the employer, information could be gathered on the basis of a small survey of workers. Alternatively, the researcher could survey workers outside the workplace.

5. **Quality of Responses.** A census may rely on people reporting information on themselves, and sending results in by mail. Such is the case in the Census of Canada. This can result in a considerable number of inaccuracies in the data, where people do not understand the questions, or give improper responses. In contrast, a survey sampling procedure may be able to use trained interviewers, people who can obtain reasonably accurate responses to questions. This may mean that the data obtained from a sample survey are actually more accurate than data produced from a complete enumeration of a population.

Because of the time and cost factors involved in conducting a census of a population, most data are obtained only from some members of the population. The set of those members of the population from which the data is obtained is referred to as a sample of the population.

Definition 2.5.2 A **sample** is any subset of the population.

This is the simplest and most general definition of a sample, saying only that any group of members of a population can be considered to be a sample of that population. Based on this definition the sample could be large, or it could be as few as 1 or 2 members of the population. In the latter case, one would ordinarily doubt the accuracy of such a small sample, at least in terms of providing an idea of the nature of the population as a whole. A sample of only one, two or a very small number of members of a population is very unlikely to represent a cross section of the population.

In most cases, the researcher will try to select a sample which can be selected and surveyed relatively quickly, at not too great a cost. The sample should also be obtained in a manner such that the population being investigated is not altered in the sampling process. As well, the researcher wants this sample to represent the characteristics of the population being investigated. That is, the sample should survey, and obtain data from, a reasonable cross section of the population, representative of the whole population at least in the main characteristics of the population which the researcher wishes to investigate.

Definition 2.5.3 A **representative sample** is a sample which has characteristics that closely match the characteristics of the population from which the sample is drawn.

That is, to be representative, the variables provided by the sample should have fairly similar values or distributions as do these same variables or characteristics in the population as a whole. For example, suppose research on the distribution of opinions concerning voting preferences is being conducted in a population. If a sample of this population is taken, this sample could be said to be representative of the population if the voting preferences expressed in the sample fairly closely match the way people actually do vote. In this case, the sample need not be representative of the population in characteristics such as hair colour or height, characteristics that have nothing to do with voting patterns. But if a more detailed analysis of voting patterns were to be conducted, where the researcher wanted to determine socioeconomic patterns associated with voting preferences, then the sample should be reasonably representative of the population in characteristics such as education, region of residence, age, ethnicity, etc. These are all characteristics of the population that are likely to have some relationship with voting patterns.

In the above definition, the notion of how closely the sample may match the population in various characteristics, is not clearly stated. While researchers would like the sample to be an exact cross section of a population, obtaining such a representative sample is almost impossible. Much of the art of sampling is attempting to balance the cost of sampling with representativeness.

Obtaining a representative sample. Some methods of sampling yield samples which are quite representative of a population, other methods provide very nonrepresentative samples. If a researcher relies on volunteers and uses his or her own judgment to select a sample, the sample may not be representative of the population. These methods may legitimately be used though because they provide data quickly and at low cost. This can provide valuable information, but the researcher must always be aware that these types of samples may not be representative of the population as a whole.

While there are many methods that sampling agencies use to obtain a representative sample, the simplest and perhaps most important method is that of random sampling.

Definition 2.5.4 A **random sample** is a method of sampling whereby each member of the population has an equal chance of being selected in the sample.

According to this definition, no individual or group in a population is systematically left out of the selection process. In random sampling, no member of the population has a greater probability of being selected in the sample than does any other member. Of course, not everyone in the population can actually be selected for the sample. Since there is a randomness to this method of sampling, the random sample may not be exactly representative of the population. But if a relatively large random sample is selected by a researcher, this sample will produce a fairly representative cross section of the population. If the sample is smaller, or if another method of selecting members of the population for the sample is used, the sample may be less representative of the population.

A further advantage of random sample is that this method of sampling is based on the principles of probability. The results derived from probability theory can be used to describe many of the characteristics of random samples. A discussion of probability and how it can be applied to sampling is contained in Chapter 6.

One characteristic of sampling, sampling error, is discussed briefly in the next section. A fuller discussion of sampling, and sampling distributions, is contained in Chapter 7. One of the major aims of sampling is to obtain an idea of the characteristics of the population. With the method of random sampling, the researcher can make estimates of the characteristics of the population, and test hypotheses concerning the population. This is the subject of most of the latter half of the textbook.

2.6 Potential Errors in Data

All data produced by researchers will have some errors associated with the data. While this may seem surprising to those who are impressed by the seeming precision of data and the power of many statistical techniques, it is best to begin by assuming that errors will be built into the process of data production. High quality data based on well constructed research designs will contain relatively few errors, while poorly developed research projects are likely to have a considerable number of errors that emerge in the process of data production. Researchers hope to construct data which have relatively few errors, so that the data reasonably accurately measure the concepts and characteristics of the population which the researcher wishes to investigate.

When using the term errors, researchers and statisticians do not usually

mean mistakes in the ordinary sense of the term, although mistakes may emerge in the research process. Rather, errors refer to systematic problems which emerge in the research and data production process as a result of the research design. For example, errors may refer to poorly constructed questions, nonresponse to questions, unrepresentative samples, etc. Errors in data are of several types. If sampling is involved, errors are ordinarily classified into sampling errors and nonsampling errors. If experiments are involved, then various forms of experimental error may also emerge.

Since most social science data come from samples, it is useful to distiguish sampling from nonsampling error. **Sampling error** refers to the potential error involved because the researcher uses a sample rather than surveying all members of the population. When estimating a characteristic of a population, the sampling error is the differece between the value of this characteristic that one finds in the sample and the true value of this characteristic if all members of the population could be surveyed.

Nonsampling error is a catch all term used to describe other types of possible error that emerge in the production of data, errors other than sampling error. The latter is discussed first, and this is followed by a discussion of sampling error.

2.6.1 Nonsampling Error

Nonsampling error refer to all the errors that can emerge in data production other than sampling error. Nonsampling errors are the type of errors that would emerge even from a census, or a complete survey, of all members of a population. There are many possible such errors, and these are discussed as follows.

1. **Incomplete Coverage**. One type of error common to almost all forms of data production is the inability of the researcher to obtain data concerning all members of the population which the researcher wishes to investigate. In a census, some respondents may not be easy to find, and some may refuse to be interviewed or respond to questions. This may not be such a problem in a sample survey or in an experimental research design, because missing respondents may be able to be replaced by other respondents. For a survey of the whole population, this need not be a serious problem if those who are not covered in a census represent a cross section of the whole population. In the

more likely case where those members of the population who cannot be found and enumerated, or who refuse to respond, are not typical of the population as a whole, then the data obtained are not likely to be a complete record of the population.

In a national census, this problem of incomplete coverage is usually referred to as undercounting. In the United States Census, this has become a serious problem, with the African-American population, and some other minority groups, being seriously underrepresented in the Census. Since the undercount tends to be concentrated in certain urban areas of the United States, this has led to underestimation of the reported population totals for some of these urban areas. This in turn has meant that some of these cities have lost government grants - grants which are based on reported population totals from the Census. In Canada, the undercount is not so great as in the United States, but there are undoubtedly some areas of the country, and some socioeconomic groups which are inadequately counted.

When conducting sample surveys, there are likely to be a considerable number of refusals, people who do not wish to respond to a telephone or personal interview. While these people can be replaced by others in a survey, these refusals create several problems. First, even if they are replaced by others, this increases the time and cost that must be devoted to obtaining the replacements. Second, people who refuse to respond to interviews, or provide data to researchers, are likely to differ somewhat in a number of characteristics from those who do not refuse such interviews. Third, the sampling procedures, and the reliability of these procedures is likely to be adversely affected.

Considerable efforts may have to be made to reduce the rate of refusals, and increase the response rate. In a survey, it is always useful to obtain data on a number of socioeconomic and demographic characteristics of respondents. Then the researcher can compare the profile of survey respondents with the more complete profile of the population from the Census or other sources.

2. **Incomplete Response**. A related problem is incomplete response by those who do respond. Respondents to a survey may respond to some questions, and fail to, or refuse to, respond to other questions. For example, researchers have often found it difficult to obtain information from respondents concerning individual or family income. Even

though the researcher may assure the respondent of confidentiality and anonymity, some respondents may be suspicious concerning the possible uses of the survey results. In addition, if the question concerns family or household income, the respondent in the family or household being surveyed may not be familiar with the level of household income. Again, such lack of information may bias the survey result, resulting in either over or under estimation of the value of the characteristic being investigated.

3. **Improper Construction of Questions.** For some characteristics such as age and sex, constructing a question which will produce accurate data is relatively straightforward. Even here though, some people may misreport age. Or the researcher may ask only the age, not the birth date of the person, with the result that some people round age to the nearest year, others to age at last birthday, and others to the nearest integer ending in 0 or 5.

   For variables such as attitudes and opinions, there is considerably greater difficulty in constructing a question which provides good data for the researcher. While attitudes and opinions may seem relatively straightforward in theory, in practice the situation is much more difficult. Some respondents to an opinion questionnaire may have little knowledge of the issue being investigated, some may not have a well formulated opinion on the issue, some may have contradictory opinions on the issue, and some may not wish to respond to the questions. In addition, others may have formulated the issue in their minds in quite a different manner from the way in which the researcher wishes to obtain responses. This may make the question meaningless for those being surveyed, with the result that the responses are also meaningless for the researcher.

   Many other issues emerge with the construction of questions. The order in which questions are placed in a survey, the set of possible options for response which the researcher provides, the nature of the preamble to the question, the exact wording of the question, all can have considerable effect on the responses. In some cases, it will be obvious to the researcher that the responses are meaningless. In other cases, it is quite possible that the researcher considers the questions well structured, but in fact the responses are quite meaningless, and the researcher never becomes very aware of this. When construct-

ing questions, or using information from previously constructed data sources, the researcher needs to be extremely careful, consulting others who have conducted similar surveys, and carrying out tests on the reliability of the data production techniques.

4. **Improper Tabulation and Coding.** Once the data have been obtained from a questionnaire or form, the researcher will usually want to enter the data obtained on a computer so that they can be analyzed. This involves attaching numerical codes to the responses of those who have been surveyed, and entering these codes on the computer. While this stage is relatively straightforward, there is also great opportunity for mistakes to creep into the coding and entry of the data on the computer.

   Some of these errors are just ordinary mistakes, entering a code of 2 when a code of 1 should have been entered, etc. Once mistakes of this sort have occurred, it may be extremely difficult to locate these mistakes. Other coding and data entry problems may relate more to judgments the researcher must make. For example, when obtaining information concerning the occupations of respondents, there may be a wide variety of occupations, some of which are not all that precisely reported in the survey. Once the responses have been completed, the researcher must make a judgment concerning what the occupation really is, and what code is to be given to the occupation. Again, if the wrong decision is made, then the analysis of data concerning occupation of respondents may have considerable errors.

   Researchers can reduce the number of such errors by having two separate people code and enter such data, and then compare results. Where there is a discrepancy between the values entered by the two people, the original responses can be checked again, and the proper codes entered. However, this is likely to double the costs of coding the data and entering the data on the computer.

5. **Conclusion.** The possibility of all of these errors existing in data should not deter researchers from producing data. Each of the problems outlined in this section can be kept under control and the number of errors reduced, if proper procedures of research design are followed. Survey sampling agencies, and organizations such as Statistics Canada, have developed a solid set of procedures which, if carefully followed, can produce quite accurate results. If the researcher can use these

agencies, or follow these procedures, then the nonsampling errors can be minimized, and data of high quality can be obtained.

2.6.2 Sampling Errors.

Sampling error refers to the specific type of error that emerges because a surveyor draws a sample of the population, rather than surveying all members of the population. When investigating a population, a researcher gathers data concerning a number of variables describing the population. Unless the sample is exactly representative of the population in all characteristics, the values of these variables based on sample data will differ somewhat from the values that would be obtained if all members of the population were surveyed. The difference in these values is the sampling error associated with the sample. When constructing a research design, a major aim of the researcher is to minimize the sampling error as much as possible, given the time and cost constraints.

For example, suppose the true average age of the population of all students at a particular university is 22.3 years, but that the researcher does not know this, or does not have the necessary information on student ages to be able to determine this true average age. Researcher A conducts a sample of university students and finds that the average age of the students in the sample is 21.1 years. In this case, the sampling error would be $22.3 - 21.1 = 1.2$ years. If researcher B had a different sample where the average age was reported as 20.3 years, then this sample would have a sampling error of 2 years. With respect to determining the correct average age of students, the latter sample would be regarded as being inferior to the sample of researcher A, because it has greater sampling error when estimating the average age of students.

While the above example is clear, the problem the researcher faces is that the true value of the variable is usually unkown, even after the sample has been obtained. That is, after conducting the survey of students, researcher A knows that the average age of students surveyed was 21.1 years and researcher B knows that the average age of the students he or she surveyed was 20.3 years. But neither researcher knows that that the true average age of all students is 22.3 years. As a result, it is not clear from the samples themselves which of the two samples provides a superior estimate of the mean age of all students.

The true values of characteristics of a population are usually unkown. If they were known, it is likely that the researcher would not have had to

conduct the survey in the first place. As a result, anyone evaluating the respective merits of different samples, may not be able to decide which is the more accurate survey. In examples like the one above, the true average age of all students would be unknown, so that there is no clear criterion for deciding which of the two samples has lower sampling error.

Rather than giving up at this point, and deciding that no conclusion can be made, there are several statements that can be made concerning sampling error. The type of statement that can be made concerning sampling error depends on how the sample is selected. If the sample is selected using the principles of probability, then some probability statements can be made concerning the level of sampling error. In the case of samples drawn on bases other than probability, it is more difficult to make statements concerning the level of sampling error. This is briefly outlined in the following paragraphs, with addition discussion in Section 2.7.6, where the interpretation of sampling error in labour force data is discussed.

Sampling Error in Probability Samples

When principles of probability are used to draw a sample, as in the case of a simple random sample, then the probability of any given level of sampling error can be determined. That is, the exact amount by which the characteristic from a sample differs from the true value of the characteristic if the whole population were to be surveyed cannot be determined. But probabilities can be attached to the various values of the difference between these characteristics. In general, the sampling error is said to exceed no more than a certain value, with some probability. This is most easily illustrated with an example.

Example 2.6.1 Gallup Opinion Polls

Each month, Gallup, Inc., conducts surveys of public opinion across Canada. These surveys cover a wide range of social and political issues. The most commonly reported example of this is the monthly Gallup reports concerning the percentage of the Canadian population which supports each political party. Gallup asks a question such as:

> If a federal election were to be held today, which party's candidate do you think you would favor?

While Gallup does not use a simple random sample of the Canadian population, the method used is based on probabilities, so that Gallup can estimate

the level of sampling error associated with its estimates. The Gallup reports usually say something like:

> A national sample of this size is accurate within a four percentage point margin, 19 in 20 times.

The characteristic of the population being estimated in this case is P, the percentage of Canadian adults who would voted for each particular party. The true value of p is unknown, but what is known is the percentage of those in the sample who say they would vote for the particular party, p. The sampling error is $p - P$, and Gallup says that this exceeds no more than 4 percentage points, 19 in 20 times, or equivalently, 95 times out of 100. That is, in 95 out of 100 samples, the difference between p and P would be no more than 4 percentage points. Another way of stating this is that the probability is 0.95 that there is no more than a 4 percentage point difference in the sample percentage of voters who favour a particular party, p, and the true value, P, if the whole population were to be surveyed. This also means that in 5 samples out of 100, or with probability 0.05, the sampling error could exceed 4 percentage points. As with all of statistics, statements are never made with absolute certainty. Rather, each statistical statement holds with only a certain level of probability.

Sampling error can be reduced to a lower level if the sample size is increased. In Chapter 8, formulas for determining sample size are given. The general thrust of these formulas is that in probability samples that survey more people, sampling error is reduced, for any given probability. Alternatively stated, for any given level of sampling error, a larger sample size is associated with a higher probability of this level not being exceeded, as compared with a smaller sample size.

Nonprobability Samples.

In the case of nonprobability samples, it is much more difficult to estimate sampling error. In this case, the sample may have been chosen in a more or less arbitrary manner, for example, a sample of one's friends, or a sample of people in the street who agree to give interviews. Since there is nothing systematic in the manner in which the sample is being selected, there is no real way to determine the sampling error. About all that can be done in this case is to obtain data on a number of known characteristics when conducting the survey, and then compare the survey results with the known characteristics of the whole population.

As an example, when carrying out person on the street interviews, a researcher should obtain the age and sex of respondents, and then check the distribution of these characteristics among those surveyed with the distribution of age and sex from the Census, or from some other source. In this way, the researcher can at least determine how representative of the population the survey is in terms of known characteristics. If it is representative in these characteristics, then perhaps the sample is also representative of the population in the charactertistics the researcher is investigating. However, the researcher has no way of attaching a probability or a level of sampling error to these results.

While nonprobability based samples are deficient in that they do not provide a way of estimating sampling error, or representativeness of the sample, often they are the only way in which data on a particular topic can be determined at all, or obtained quickly. When using this method, a researcher should attempt to select a cross section of people in the sample. If this is done carefully, the nonprobability sample may provide much useful data.

2.7 The Labour Force Survey

2.7.1 Introduction

Statistics Canada's Labour Force Survey and the data relating to the labour force which result from this Survey provide an example of what is involved in producing data. Even though this Survey provides what is generally considered to be very high quality data, the Survey also illustrates many of the problems and controversies that are associated with the production of data.

2.7.2 Labour Force Data as a Social Product

Monthly labour force data have become an important social indicator in Canada. Each month reports of the results of Statistics Canada's Labour Force Survey are reported in the media. In times of economic prosperity, declining rates of unemployment are taken as a sign of improvement in economic conditions. At other times, when jobs are scarce, high or rising rates of unemployment are taken as a sign of how Canada is experiencing

tough economic times. Commentators may make much of either a decline or an increase in the monthly unemployment rate.

The unemployment rate is taken as a sign of the health or weakness of the Canadian economy. In the financial and business world, the unemployment rate and other labour force statistics may be used as a signal by government and business people as to how future plans should be organized. While most ordinary Canadians are less likely to pay attention to these statistics, the state of labour markets, the types of jobs available, and the extent of unemployment, are all factors which can have a considerable effect on people's livelihood and income levels. As a result, the products of the labour statistics produced by Statistics Canada are important social indicators for many Canadians.

Labour force data are a social product in terms of the manner in which they are constructed and obtained. Before 1945, there was little such information available, and yet the Canadian economy grew during many of these earlier years. About the only source of information on the labour force in these early years were the numbers of "gainfully employed" that were measured each ten years in the Census of Canada. As the Canadian economy grew and became more complex, for business and government order to carry on their activities, it appeared necessary to obtain more information concerning various aspects of the economy. In particular, the Dominion Bureau of Statistics (as Statistics Canada was called until 1971), decided that more information should be made available on the labour force. Since 1960, the Labour Force Survey has provided the official measure of unemployment in Canada.

The particular manner in which the federal government decided to obtain information on the labour force was through a survey sampling method. While the survey methodology has changed over the years, the basic method used each month has been to attempt to find a representative sample of Canadian adults, and study the labour force activity of these individuals. It should be noted that this is not the only manner in which such information could be obtained. Similar information could be derived by surveying employers across the country, or by using administrative records of government employment agencies. While the survey method has resulted in high quality information, the individual survey approach taken by Statistics Canada does represent a particular approach to obtaining data on the state of the country's labour force and employment.

A further way in which labour force data are a social product is in the manner in which the variables are defined and the labour statistics organized

and presented. These are discussed in detail below.

2.7.3 Availability of Labour Force Data

The Labour Force Survey is a Canada wide sample survey conducted each month by Statistics Canada. This Survey aims to provide reliable and timely information on the Canadian work force. The data and statistics produced from the Survey are contained in the monthly Statistics Canada publication, **The Labour Force**, catalogue no. 71-001. More analytical studies, along with historical and comparative data are provided in **Perspectives on Labour and Income**, a quarterly Statistics Canada publication, catalogue no. 75-001E. A discussion of the types of data that can be found in the Survey is contained in Statistics Canada, **Guide to Labour Force Survey Data**, catalogue no. 71-528. The methods used to obtain labour force data are discussed in detail in Statistics Canada, **Methodology of the Canadian Labour Force Survey 1984-1990**, catalogue no. 71-526. Most of the discussion in this section is drawn from these sources.

The Census of Canada also contains detailed information on the occupations, industry and education level of those Canadians who are employed. These Census reports are not nearly as timely as the Labour Force Survey, with new data being made available only once every five years. But since the Census surveys almost all Canadians, considerably greater information on the detailed characteristics of the Canadian labour force is available from the Census.

Many of the statistical agencies in the individual provinces of Canada also publish data on the labour force. For example, in Saskatchewan, **Labour Force Statistics** provides a summary of Saskatchewan labour force data. This is published monthly by the Saskatchewan Bureau of Statistics. Further data on the Saskatchewan labour market is provided in the quarterly publication **Saskatchewan Labour Report**, published by Saskatchewan Human Resources, Labour and Employment. All of these publications can be obtained in the Government Documents section of the University of Regina Library.

2.7.4 Definitional Issues in Labour Force Survey Data.

The Population.

The aim of the Survey is to determine the size and nature of the Canadian labour force. Potentially, this could include everyone in Canada, apart from small children. Statistics Canada begins its analysis from this point, attempting to obtain some information on a sample of all Canadians except for children. The particular cut off age chosen by Statistics Canada is age 15, so that the base population for the Survey is all Canadian adults, aged 15 and over.

For various reasons, Statistics Canada limits the population somewhat further. In the discussion of its research methodology, Statistics Canada notes in the monthly issues of **The Labour Force**, page D-2:

> The sample used in the Labour Force Survey has been designed to represent all persons in the population 15 years of age and over residing in the provinces of Canada, with the exception of the following: persons living on Indian reserves, full-time members of the armed forces and people living in institutions (for example, inmates of penal institutions and patients in hospitals or nursing homes who have resided in the institution for more than 6 months.

Statistics Canada has decided to exclude residents of Indian reserves and of the Yukon and North West Territories from the Survey because of

> ... operational and statistical considerations, namely, the difficulties involved in carrying out monthly surveys in such areas and the general inapplicability of the survey concepts and definitions to the measurement of labour market conditions on reserves and northern communities. (**Guide to Labour Force Survey Data**, p. 7).

However, this can create some problems for measurement of the labour force and unemployment in areas of the country which has a large Status Indian population, areas such as northern Manitoba or Saskatchewan. Since unemployment has been very high on Indian reserves in these areas, by excluding residents of Indian reserves from the Survey, Statistics Canada very likely is underestimating the extent of unemployment in these areas. In **Guide to Labour Force Survey Data**, Statistics Canada also notes:

> Inmates of institutions and full-time members of the Canadian Armed Forces are excluded because they are considered to exist outside the labour market to which this survey applies.

As will be noted in Statistics Canada's supplementary measures of unemployment on page 51, by including members of the Armed Forces in the labour force count, there is little change in the estimate of the overall rate of unemployment in Canada.

The above noted exclusions amount to ignoring only 2-3 per cent of the Canadian population, although in some parts of the country, a larger percentage of the population may be excluded.

The Variables.

The next stage involved in examining the nature of the labour market in Canada is to define **the labour force**. One aim of the Labour Force Survey is to provide estimates of the numbers of employed and unemployed Canadian adults. Together these two define the **labour force**. The **theoretical** concept behind the definition of the labour force is the **supply of labour**, as economists define it. The supply of labour is the amount of labour that is supplied, or available to employers, given the prevailing rates of pay and working conditions. This definition excludes household work from the definition of the labour force as well as many unpaid workers in family farms or family businesses. It also excludes from the labour force all persons without employment who are not actively looking for paid employment. This theoretical approach is **operationalized** in the Survey with the following definitions.

1. **Labour Force**. The labour force is composed of that portion of the civilian non-institutional population 15 years of age and over who, during the reference week, were employed or unemployed.

2. **Employed**. The employed includes all persons who, during the reference week:

   (a) did any work at all for pay or profit.

   (b) did unpaid work where that work contributed directly to the operation of a farm, business or professional practice owned or operated by a related member of the household.

(c) had a job but were not at work due to own illness or disability, personal or family responsibilities, bad weather, labour dispute, vacation, or for various other reasons. Excluded from the employed are persons on layoff and persons whose job attachment was to a job to start at a definite date in the future.

(Note that this definition of employed includes all workers who did any work at all for pay or profit. This is further operationalized as all those who did such work for as little as one hour per week. The implication of this definition is that all part time workers are included with full time workers as being employed. When there are considerable numbers of part time workers who would prefer to work full time, this definition would seem to overstate the number of members of the labour force who are fully employed in the sense of having full time work).

3. **Unemployed**. The unemployed are those persons who during the reference week:

   (a) were without work, had actively looked for work in the past four weeks (ending with reference week), and were available for work;

   (b) had not actively looked for work in the past four weeks but had been on layoff and were available for work;

   (c) had not actively looked for work in the past four weeks but had a new job to start in four weeks or less from reference week, and were available for work.

   (Note that full time students without jobs or those who did not actively look for work in the four weeks before the survey are excluded. In places and times of high unemployment, considerable numbers of people without jobs may either be enrolled in school or may not actively look for work, with the result that the unemployment rate may be understated. All those adults without jobs but not in the category of unemployed are left out of the labour force and are in the following category).

4. **Not in the Labour Force**. Those persons in the civilian non-institutional population 15 years of age and over who, during the reference week were neither employed nor unemployed.

(These non labour force respondents may be individuals who are attending school, are retired, have personal or family responsibilities outside the labour force, have an extended illness or disability, etc. For whatever reason, they neither have jobs, nor are they actively looking for employment at the time of the Survey. As a result, they are neither technically employed nor unemployed and are defined by Statistics Canada as outside the labour force).

Many other concepts are used in the Survey and information is collected from respondents concerning their hours of work, type of work, occupation, etc. This discussion concentrates on the measures of employment and unemployment. Once the data have been collected using the definitions just given, various other statistics can be calculated. The most common of these are the unemployment rate and the labour force participation rate. These are defined as follows.

1. **Unemployment Rate**. The unemployment (UR) rate is really a percentage because it is defined as the percentage of the labour force that is unemployed. It is defined as:

$$\text{UR} = \frac{\text{number of workers unemployed}}{\text{number of workers in the labour force}} \times 100\%$$

Statistics Canada reported that in Saskatchewan, 432,000 people aged 15 and over were employed in January 1992, with 39,000 unemployed workers for a total labour force size of 471,000.

The Saskatchewan unemployment rate in January, 1992 was reported as:

$$\frac{39,000}{471,000} \times 100\% = 8.3\%$$

2. **Labour Force Participation Rate**. This is another measure of employment conditions. It provides a measure of the extent to which adults participate in the labour force, as being either employed or unemployed. Like the unemployment rate, the labour force participation rate (or simply the participation rate, PR) is really a percentage and measures the percentage of the adult population which is in the labour force (either employed or unemployed). The labour force participation rate is defined as:

$$\text{PR} = \frac{\text{number of labour force participants}}{\text{number of adults aged 15 and over}} \times 100\%$$

For example, in January 1992 in Saskatchewan, there were 720,000 adults aged 15 and over and, of these, 471,000 were in the labour force, either employed or unemployed. The labour force participation rate for Saskatchewan in July was:

$$\frac{471,000}{720,000} \times 100\% = 65.4\%.$$

The labour force participation rate is generally higher in good economic times or regions, and lower in poor economic times or in areas where there are few jobs available. In January 1992, the labour force participation rate was 71.7% in Alberta, where jobs have been relatively plentiful. In contrast, the labour force participation rate in January 1992 was only 50.5% in Newfoundland, a province known for its limited number of jobs. The rate also varies rather considerably over each year, as seasonal work picks up and declines and as students move into and out of the labour force.

3. **Employment Rate**. A third measure of labour market conditions is the employment rate. Again this is a percentage, the percentage of adults who are employed. In some cases, this rate may be superior to the unemployment rate in providing a measure of the availabity of jobs, or changes in the availability of jobs. The employment rate (ER) is defined as:

$$\text{ER} = \frac{\text{number of employed}}{\text{number of adults aged 15 and over}} \times 100\%$$

For example, in January 1992 in Saskatchewan, there were 720,000 adults aged 15 and over and, of these, 432,000 were employed. The employment rate for Saskatchewan in January was:

$$\frac{432,000}{720,000} \times 100\% = 60.0\%.$$

The employment rate generally rises as the number of jobs increases and declines as the number of jobs declines. While the unemployment

rate generally moves in the opposite direction to the employment rate, it is possible that the extent of unemployment is understated when the number of jobs declines. This can occur if considerable numbers of the unemployed drop out of the labour force. When this is the case, the employment rate may give a better idea of employment conditions than does the unemployment rate.

In describing changes in the labour market over time, all of these measures should be considered. The annual average of these various rates for Saskatchewan for the years 1986-1990 are shown in Table 2.1, with the monthly fluctuations shown in Table 2.4.

Year	1986	1987	1988	1989	1990	1991
Unemployment Rate	7.7%	7.3%	7.5%	7.4%	7.0%	7.4%
Participation Rate	66.9%	66.7%	66.4%	66.2%	66.8%	67.1%
Employment Rate	61.5%	61.5%	61.4%	61.1%	62.1%	62.2%
Employed ('000s)	452	453	451	446	449	449
Unemployed ('000s)	37	36	37	36	34	36
Labour Force ('000s)	489	489	488	482	483	484
Population ('000s)	734	736	734	728	723	722

Table 2.1: Labour Force Annual Averages, Saskatchewan, 1986-1991

Sources for Table 2.1: 1986-1990 data from Statistics Canada, **Historical labour force statistics**, catalogue 71-201; 1991 data from **The Labour Force**, catalogue 71-001.

Alternative Measures of the Unemployment Rate

The last section showed how the theoretical conception of economists and the practical considerations of Statistics Canada's survey have led to some particular definitions of the labour force, employed and unemployed. As defined by Statistics Canada, these definitions may not correspond to the conceptions most people may have in mind. The most severe criticism of Statistics Canada has been that it has often undercounted the number of those without jobs who would like to find jobs. As noted earlier, this is likely to be most serious in years of high unemployment, and in areas of the country where limited numbers of jobs are available. In these circumstances, many workers without jobs may not very actively look for work because they know that none are available in the area where they live.

In order to obtain some estimate of the number of such people, and because of the debate concerning the appropriate concept of unemployment and employment, various researchers have attempted to calculate alternative estimates of the level of unemployment and employment in Canada.

The first of these estimates, contained in Table 2.2 was made by two researchers at the University of Regina, Bob Stirling and Denise Kouri. (See J.A. Fry, **Economy, Class and Social Reality**, pp.186-192). While these figures are now somewhat dated, the data show how different the reported unemployment rate would have been if somewhat different definitions of employment and unemployment had been used.

Statistics Canada's Supplementary Measures of Unemployment. The data in Table 2.3 are based on Statistics Canada's attempts to show how the unemployment rate would have been altered under various definitions of employment and unemployment. The rate R5 represents the official unemployment rate. Rates R1 to R4 are alternatives that attempt to indicate longer term or more 'hard core' unemployment. Rates R6 through R9 represent alternative definitions of unemployment based on underutilization of potential supplies of labour. It might be noted that these different definitions lead to quite different levels of the unemployment rate. However, the general trend in the rates over time is fairly similar, rising from the late 1970s through the recession of 1982-3, and then declining again as the economic recovery has proceeded.

The data in Table 2.3 for 1977-1985 are taken from Statistics Canada's publication **The Labour Force**, February 1987. The 1987-1990 figures come from Statistics Canada, **Perspectives on Labour and Income**,

Unemployment Rate, Canada, 1977

Official Rate	8.1
Adjust for part-time work	9.1
Include as unemployed those who looked for work in last 6 months	10.1
Include as unemployed those who are without work and want work	16.9
Raise labour force participation rate of each province to the Alberta average	15.0

Table 2.2: Alternative Measures of the Unemployment Rate

catalogue 75-001E, Summer 1990, page 78, and Spring 1992, page 64.

R1: Persons unemployed 14 or more weeks as a percentage of the labour force.

R2: Unemployment rate of persons heading families with children under sixteen years of age.

R3: Unemployment rate excluding full-time students.

R4: Unemployment rate including full-time members of Canadian Armed Forces.

R5: Official unemployment rate.

R6: Unemployment rate of full-time labour force.

R7: Unemployment rate including persons not in the labour force who sought work in the past six months but are not now looking for labour market related reasons.

R8: Underutilization rate based on hours worked and hours "lost".

R9: Unemployment rate of part-time labour force.

Supplementary Measures of Unemployment,
Canada, 1977-89, Per cent

	1977	1979	1981	1983	1985	1987	1989	1990
R1	2.8	2.6	2.6	5.8	4.8	3.8	2.9	3.1
R2	6.3	6.0	6.2	10.0	9.1	7.5	6.8	7.3
R3	7.9	7.2	7.3	11.7	10.3	8.6	7.4	8.0
R4	8.0	7.4	7.5	11.8	10.4	8.8	7.5	8.1
R5	8.1	7.4	7.5	11.9	10.5	8.8	7.5	8.1
R6	8.8	8.2	8.5	13.9	12.5	10.6	9.0	9.6
R7	9.0	8.4	8.6	13.3	11.6	9.7	8.2	8.7
R8	9.2	8.7	9.0	14.5	13.1	11.1	9.5	10.2
R9	9.6	10.2	9.8	13.0	12.8	11.5	9.7	10.1

Table 2.3: Supplementary Measures of the Unemployment Rate

2.7.5 Methods Used in the Labour Force Survey

As noted earlier, Statistics Canada uses a survey method to obtain Labour Force Survey Data. Some of the main features of the Survey are as follows. These points are all discussed in greater detail in each issue of **The Labour Force**.

1. The Survey is a monthly survey conducted by Statistics Canada. Each month approximately 62,000 representative households (approximately 115,000 individuals) across Canada are surveyed concerning their labour force activity. In Saskatchewan, approximately 5,000 representative households (10-11,000 individuals) are surveyed. This makes the Labour Force Survey the largest regular survey conducted in the province.

2. The Survey has been conducted regularly since November 1945, first quarterly, and since November 1952 on a monthly basis.

3. Estimates for each month normally refer to the week containing the 15th day of the month.

4. Interviewers working for Statistics Canada conduct personal interviews in all dwellings where the interviews are being conducted for the first

time. Subsequent interviews are conducted by telephone. Respondents normally are surveyed for six consecutive months and then are dropped from the survey. The questionnaire used in the Survey is reproduced following page 59.

2.7.6 Potential Errors in Labour Force Data

Nonsampling Error

As is possible in any survey of a population, there are many possible sources of **nonsampling error** - improper questions, incorrect answers, non-response, etc. Some of these errors emerge because surveyors use improper concepts or they may design questions poorly. It is also possible that some groups in the population may be missed because of the way in which surveyors design the survey. On the other side, respondents may refuse to participate in the survey, or they may respond only to certain questions. Additionally, respondents may give misleading or incorrect answers, either deliberately or inadvertently. All of these possible sampling and nonsampling error can distort the results. However, according to Statistics Canada,

> The non-response rate for the Labour Force Survey is among the lowest in the world for a survey of this type. In 1990, the rate averaged less than 6%. (**The Labour Force**, January 1992, page D-7).

Sampling Error

The true rate of unemployment for the adult population in any month is unknown. While the rates in the table provide good estimates of the true unemployment rate, they are based on a sample of adults. As a result, they are subject to possible sampling error. This sampling error arises because not all the population is surveyed each month, but only a sample of adults is surveyed. Later in the semester the method of estimation, and the size of the potential sampling error are discussed. In the next paragraph, an estimate of potential sampling error, based on information provided by Statistics Canada is provided.

With a probability of 0.95, or in 19 samples out of 20, the possible degree of sampling error for the Saskatchewan unemployment rate is likely to be no greater than ± 0.9 percentage points. In November 1992, the reported

unemployment rate, based on the sample, was 9.0%. The true rate of unemployment for Saskatchewan in November was thus very likely to have been in the range 9.0 ± 0.9 or between 8.1% and 9.9% of the Saskatchewan labour force. As a result of this, reported changes in the unemployment rate from one month to another month may not rmean that there was a real shift in unemployment unless these changes are considerable. Based on the Saskatchewan sample, it may be that month to month changes need to be half a percentage point or more in order to reflect a real shift. In addition to sampling error, it should be remembered that there are likely to be various forms of nonsampling error in the Survey. While it is difficult to make a numerical estimate of their effect, these nonsampling errors could add to the above noted sampling error.

2.7.7 Labour Force Survey Trends in Saskatchewan

The month to month fluctuations in the unemployment, participation and employment rates have several causes. **First**, the rates change from month to month as economic and labour force conditions in the province change. **Second**, the rates reported in the table may fluctuate from month to month because they are based on samples. Because not all the population is surveyed, with the data being based on a sample of the population, there may be considerable **sampling error** involved in these monthly data. That is, some of the month to month fluctuation may represent the fact that only a small fraction of Saskatchewan adults are being sampled each month. The people in the sample are different people each month, with some months having more employed respondents, and other months having more unemployed respondents. While each month's sample is likely to be fairly representative of the whole Saskatchewan population, no month is exactly representative. A **third** reason the rates differ from month to month is the **seasonal** nature of employment and unemployment. Each of these is discussed briefly here.

General Labour Force Trends, 1986-1990

The basic trends in the Saskatchewan labour force can be seen by looking at the annual data in Table 2.1. These annual data take out the effect of month to month changes. Based on these data, it can be seen that the **unemployment rate** over the last four years has been relatively stable, between 7% and 8%. The **number of unemployed** has also remained

	Unemployment Rates		Participation Rates	Employment Rates
	Actual	Adjusted	Actual	Adjusted
1989				
Jan.	8.7	7.3	65.5	62.2
Feb.	9.1	7.9	64.7	60.8
Mar.	9.0	7.7	65.0	61.2
Apr.	8.9	8.1	65.1	60.9
May	7.4	7.5	66.9	61.0
June	6.8	7.7	67.6	61.0
July	7.1	7.5	68.6	61.4
Aug.	6.5	7.2	68.7	61.5
Sept.	5.9	7.2	67.0	61.8
Oct.	6.0	7.1	66.2	61.6
Nov.	7.1	7.3	65.1	61.1
Dec.	7.0	6.9	64.3	61.1
1990				
Jan.	8.9	7.5	64.6	61.3
Feb.	8.1	6.9	64.9	61.7
Mar.	7.6	6.3	65.3	62.3
Apr.	7.6	6.6	65.5	62.2
May	6.8	6.8	67.8	62.2
June	6.4	7.4	68.7	62.4
July	6.6	7.0	69.2	62.1
Aug.	6.5	7.2	69.8	62.4
Sept.	5.7	7.0	67.8	62.6
Oct.	6.2	7.3	66.8	62.9
Nov.	6.7	6.9	66.0	62.1
Dec.	7.0	6.9	65.4	62.1
1991				
Jan.	8.4	7.0	65.9	62.7
Feb.	8.9	7.6	65.8	62.1
Mar.	8.6	7.2	66.1	62.4
Apr.	8.2	7.4	66.8	62.8
May	6.9	7.2	68.3	62.8

Table 2.4: Saskatchewan Monthly Labour Force Statistics, 1989-1991

around an annual average of 34 to 38 thousand unemployed. In contrast, the **number of employed** workers, has fallen from the all time peak of 453,000 employed in 1987.

The decline in the number of jobs in the province has also been paralleled by a considerable decline in the population aged 15 and over. Over the last two or three years, there has been considerable outmigration from the province. The Saskatchewan population aged 15 and over reached an all time peak of 736,000 in 1987, but dropped to 723,000 in 1990. Because the number of employed and the number of people have declined by about the same amount, the employment rate and the participation rate have not changed all that dramatically.

As a result of these considerations, perhaps not too much should be made of the month to month fluctuations, especially in the unemployment rate. The most notable Saskatchewan labour market feature of the last 3 or 4 years has been the decline in jobs and population, although in recent months, these appear to have levelled off, or perhaps even made modest gains.

2.7.8 Seasonal Adjustment

When one examines the month to month changes in the actual or unadjusted unemployment or participation rates, one notices a fairly regular pattern each year. The rate of unemployment is usually higher in the winter and lower in the summer, rising in the fall and declining in the spring. In contrast, the labour force participation rate shows the opposite pattern, falling in the winter months and rising in the summer months. In Canada, and especially in Saskatchewan, these patterns are fairly regular on a year to year basis.

This seasonal fluctuation in employment is important for workers, especially those looking for jobs, and those working in industries such as construction, where the seasonal pattern is even more marked than in the labour force as a whole. However, in order to obtain an idea of the **trend** in the rate of unemployment or labour force participation, statisticians and labour force analysts find it useful to **seasonally adjust** the rates. The seasonal adjustment is a **moving average** or smoothing of the monthly figures. This can be clearly seen in the diagram following page 59. There, one could plot the actual or unadjusted figures and then draw the solid line with a pencil, smoothing out the regular month to month fluctuations over the course of

a year. Statistics Canada uses a complex computer program to smooth the line, but the principle involved is the same.

The advantage of using the seasonally adjusted figures is that one can obtain an idea of the trend in the unemployment or participation rate by comparing the seasonally adjusted rates for any two months. Referring to Table 4, one can see that in 1989, the actual unemployment rate varied considerably, from a high of 9.1% in February, to a low of 5.9% in September. However, most of this reflects seasonal shifts, rather than major changes in the basic structure of the provincial labour merket. Looking at the comparable seasonally adjusted unemployment rates over the months of 1989, one can see that these rates fluctuate by only a little over 1 percentage point.

Again in 1990, the actual reported unemployment rate seems to have fallen by about 3.2 percentage points, from 8.9% in January, to 5.7% reported in September. However, most of this was the usual seasonal decline. The adjusted unemployment rate also fluctuated over this same period, falling from 7.5% in January to 6.3% in March, but then increasing again so it was at 7.4% in June. After that it fell again, and averaged close to 7% for the rest of the year. Most of these changes, and the decline of about half a percentage point from the beginning of 1990 to mid 1990 may represent nothing other than sampling error.

2.7.9 Monthly Shifts, 1989-1991

As noted earlier, there were no dramatic shifts in the unemployment rates over this period. However, some tentative statements could be made concerning changes in the labour market over the period shown in Table 2.4. For the latter half of 1989, the adjusted unemployment rate appears somewhat lower than in the first half of 1989, averaging about half a percentage point lower in the second than in the first half of that year. It appears that this decline in unemployment continued into the first part of 1990. The seasonally adjusted unemployment rate was below 7% between February and May of 1990. In addition, as evidence of improved employment conditions, the seasonally adjusted employment rate continued to increase from around 61% in mid 1989 to around 62% by the spring of 1990. As Table 2.1 shows, for 1990 as a whole, employment did increase by 3,000 over its 1989 level.

After mid 1990, it appears that the unemployment rate rose a little again, and in the first few months of this year, the seasonally adjusted rate has stayed in the 7.0% to 7.6% range. This increase over early 1990 may just

represent sampling error, but the rate has consistently stayed above 7% this year. One can also note that the employment rate has not increased much since mid 1990, whereas it did increase between early 1989 and early 1990. In contrast, the monthly labour force participation rates in 1990 generally do appear to have increased over their level in the corresponding month of 1989.

What appears to have happened in the past year is that there has been a very small increase in the number of jobs, but more people are participating in the labour force. As a result, the number of people looking for work has grown more than has the number of jobs, contributing to a small rise in the unemployment rate. From 1987 through 1989 the unemployment rate did not increase all that much as jobs were lost because many people left the province. This exodus of people reduced both the number of unemployed workers left in the province and the number of workers in the labour force. Based on the more recent figures, it appears that this loss of jobs, and of people, from the province, may have stopped in 1991, at least for the time being.

2.8 Example Problem Involving Data Production

Question.

Using some of the ideas concerning the production of data, what questions might you ask or how might you go about trying to reconcile the statements in the following quotes. In the Toronto **Globe and Mail** of January 6, 1988, columnist Geoffrey York states:

> About 20 per cent of Regina's population is Indian or Métis, the highest percentage of any major North American city.

In **Native Economic Conditions in Regina and Saskatoon**, published in 1983, Stewart J. Clatworthy and Jeremy Hull state that

> The present native population of Regina is estimated to be 11,700 ... .

(Note that in 1981, the total population of Regina was 162,984).

Answer.

Some comments on the two quotes could be as follows. First, note that Clatworthy and Hull's estimate of the percentage of native people in Regina is $(11,700/162,984) \times 100\% = 7.2\%$. This is considerably less than the estimate used by York.

The first issue that might be raised is the definition of the population involved. York refers to the **Indian** or **Métis** population and Clatworthy and Hull refer to the **native** population. While these might be considered the same, some investigation of the meaning of each of these should take place. Following this, the exact questions asked of respondents to determine if they were of native, Indian or Métis origin should be determined. There may be some major difference in the concept behind, or wording of, the question which might explain some of this difference.

Another issue is what it means to be considered a member of the population of Regina. Exactly what criteria need to be satisfied in order that one be considered a member of the Regina population. It may be that some Status Indian people essentially live in Regina, but are not considered members of the Regina population because they have an attachment to a band or reserve outside Regina.

The next stage of investigation would begin to look at how the reported figures were derived. The sampling or survey methods used to determine the number of native people in the city could have a major effect on the results. Whether the data were based on a complete survey or census of the whole population, or on some other method, could affect the results. Regardless of which method was used, one should take account of possible underenumeration of the native population, and the reasons why there could be more underenumeration of people of native origin, than of the population as a whole.

Another consideration would be the difference in time periods. While it seems unlikely that a growth from 7 or 8% to 20 per cent could occur in 5 years, a considerable growth in the numbers of native people in Regina probably has taken place in recent years.

The two quotes do not provide sufficient information to determine the percentage of the Regina population that is of aboriginal origin. Further investigation of some of the issues raised here would be required.