Contents

Chapter 1

Introduction

Statistics and statistical reasoning have become important in many aspects of society and the social sciences. Before this century, life was carried on with little reference to statistics or quantitative data. But at the end of this century, statistics often appears to govern many aspect of our lives. The television programs we watch, the political positions taken by politicians and government, the products we find offered for sale, and the public policies of governments, all are heavily influenced by the analysis of quantitative data. Statistical forms of reasoning form the basis for the production and analysis of this data, and the conclusions drawn from this analysis. The administration of privately owned businesses, of various levels of government, and of non governmental organizations generates quantitative data, and all of these organizations depend heavily on the analysis of this data.

Many of us resent being quantified in this manner, and consider that quantitative data and statistics are associated with deception and injustice. Taxation that seems unjust, large companies laying off workers because projected profits do not appear adequate, or cutting unemployed people from unemployment insurance or welfare benefits, may all be justified by using administrative data. Trade and economic policies which appear to harm most Canadians have been justified on the basis of statistical studies which promised great benefits.

At the same time as some people regard statistics as mostly lies and deception, others place excessive faith in statistics and statistical conclusions. The pollster who designs the image and policies of a political party, the researcher who insists that his or her statistical analysis yields precise conclusions, the private credit agency bureaucrat who cannot admit any error

in a data base, or the instructor who insists that grades must be allocated according to the bell curve, may be using statistical methods beyond what they were intended for.

The approach adopted in this textbook is that both of the above approaches are incorrect. Statistics has proven to be an extremely useful tool in the hands of researchers in most disciplines. The discovery of sources of disease, the amassing of statistics concerning poverty, and the effects of environmental hazards have all relied heavily on the statistical approach to research. In administering society, whether this be done by private or public agencies, the ability to generate and analyze quantitative data has helped to produce great wealth for society. It is certainly true that there have been many misuses of statistics, and that quantitative data has been used to deceive and to hide ignorance. But on balance, the statistical approach to the study of society, and the administration of society, have produced great benefits.

Statistics can perhaps best be regarded as a tool or a technique, rather than as a substantive discipline. There are certainly many other tools or techniques which can be used to understand and improve society. But statistics has proven to be a very useful tool, and one that must be understood by anyone who studies the social sciences. It is useful to study statistics both because statistics is often misused, and because it is such a powerful and potentially beneficial technique. By developing a good understanding of the sources of data, how data is produced, and how it can be analyzed with statistics, it is possible to be critical of improper uses of statistics. Students who develop a good understanding of, and facility in, quantitative data and statistical analysis, can use this ability positively, to develop new knowledge, and new approaches to improving society.

Social science students have often ignored statistics, and attempted to steer clear of quantitative data. While nonquantitative approaches are essential in the social sciences, and have provided much of our knowledge, quantitative approaches have become increasingly useful. In order to read most social science journals and understand much of social science research, it is necessary to develop some ability to comprehend and work with quantitative approaches.

It is hoped that this textbook will assist in developing this understanding of statistics. This is an introductory textbook, one which does not deal heavily with the mathematical roots of statistics. Rather, the approach taken here is that it is important to develop an appreciation of the sources of data and how this data can be used. The textbook emphasizes problems,

and attempts to illustrate the common statistical approaches with examples which are relevant to Canadians today. You should not expect to do extensive statistical research work after studying only this textbook. A second or third course in statistics should be taken by those who contemplate undertaking quantitative research. However, this textbook forms the basis for these later courses, and presents the essential aspects of the statistical approach. After studying this textbook, you should be able to have a better understanding of quantitative research, and develop a positive and critical approach to social science research.

Textbook Outline. This textbook begins by examining the production of data. In Chapter 2 it is argued that data is not so much collected, as it is produced. Some data sources and the potential biases and errors which emerge in data production are discussed. Statistics Canada's Labour Force Survey is used as an example of both the contribution to and potential errors associated with data production. Chapter 3 discusses measurement in the social sciences. Social scientists use many different types of scales when measuring characteristics of human populations. These vary from sex or ethnicity, through attitudes and opinions, to characteristics such as income or age. Each of these characteristics has a different type of scale by which it is measured. The nature of the scale has important implications for the type of statistical analysis which can be used.

Chapter 4 shows how to begin using data, presenting data in the form of tables, graphs and charts. Characteristics of populations are presented as distributions in this chapter. For example, techniques of showing the distribution of characteristics such as political views or family income, across members of a population, are given in this chapter.

Chapter 5 takes the same types of data, but summarizes this data into various summary measures of samples and populations. Measures such as the average of a population, or the variation among members of a population are presented. This chapter uses some algebra, and requires that you develop some facility in calculation and manipulation of data. The measures presented in this chapter form the basis for the statistical analysis of much of the rest of the textbook.

Chapters 2 through 5 constitute **descriptive statistics**. Through the end of Chapter 5, the aim is to obtain data, present this data, and develop tables, charts or summary measures which provide a relatively accurate picture of people. Chapters 7 through 10 are usually referred to as **inferential**

statistics, discussing the methods of making inferences concerning whole populations on the basis of data from samples or experiments.

Chapter 6 provides some additional building blocks for the inferential statistics of the later chapters. Chapter 6 is concerned with probability, and the uses of probability in statistical research. The aim of Chapter 6 is not so much to show how to obtain probabilities of poker or bridge hands, but to develop an understanding of the principles of probability. This chapter presents the two most commonly used distributions in statistics, the binomial and the normal.

Inferential statistics is composed of two main parts, estimation and hypothesis testing. Chapter 7 brings together the descriptive statistics of the earlier chapters with probabilities, to show how estimates of population characteristics can be obtained. Most quantitative data that is obtained by researchers is obtained from samples. Principles of probability are involved in these samples because the samples are ordinarily random samples, or samples drawn on some probabilistic basis. Chapter 7 shows how the data from these samples can be used to provide estimates of the nature of whole populations. Some of the limitations of these estimates are discussed in this chapter.

The other form of inferential statistics is hypothesis testing. This forms the basis for Chapter 8-10. The idea behind hypothesis testing is that the researcher develops an hypothesis concerning a population, and then uses data from a sample or experiment to either support or refute this hypothesis.

For example, a researcher may hypothesize that higher income people tend to vote Conservative, and working class people are more likely to vote NDP. Such an hypothesis may be based on results obtained by previous researchers, or may be developed on the basis of hunches or observations made by the researcher. Data are obtained from a sample, either developed by the researcher, or carried out by others. This data from a sample will then show either that the voting patterns are as hypothesized, or some other pattern of voting may be observed.

In making estimates and tests of hypotheses, conclusions are never absolutely certain. Since research data are obtained from samples or experiments, and are based on the principles of probability, there is always some uncertainty regarding the conclusions. However, this uncertainty is quantified, at least in a probabilistic sense. A statistical result may be 90% certain, or 99% certain, but never 100% certain. This may seem inadequate to those who are familiar with the certainty of conclusions in deductive logic or in mathematical proofs. But this is the method of statistics, a probability is

associated with each statistical conclusion.

The last two chapters of the textbook will deal with regression, correlation and sampling. Regression and correlation are methods used to determine how closely related to each other different characteristics of a population may be. A short discussion of some of the common methods of sampling is contained in the last chapter.

Approach of this Textbook. This textbook takes the approach that statistics is best learned by doing statistics. While there are many mathematical assumptions and proofs on which statistical methods are based, some of these may be best learned in a second course on statistics. The most important aspect of statistics is to develop an understanding of how to work with numbers, how to manipulate them legitimately, and how to use them to describe populations. As a result, this textbook presents many examples and problems, and shows how to analyze these. There are a few mathematical proofs, but these are kept to a minimum. Some use of algebra, and unfamiliar forms of mathematical notation, are required. Each of these is explained as it is introduced. If you can follow the examples, and understand the steps in each of these examples, then you should be able to do the statistical work required in an introductory course in statistics.

In this textbook, some of the examples are presented in *slanted type like this*. This slanted type is intended to set the examples off from the main lines of argument in the textbook. If you have trouble with the explanation in ordinary type, study the examples, and then go back to the explanation. Often the explanations make more sense after you have seen an example.

Conclusion. Examples from the social sciences in Canada are used in this textbook wherever possible. Hopefully these examples will lead to an appreciation of how statistics can assist in developing a critical understanding Canadian society. Do not expect to become an expert in statistics based on this one textbook. But if you are able to understand and be critical of the research of others, this book will have accomplished its purpose.