HUBERT M. BLALOCK, JR.

SOCIAL STATISTICS

REVISED SECOND EDITION

Social Statistics

Revised Second Edition

Hubert M. Blalock, Jr.

Professor of Sociology
University of Washington

McGraw-Hill Book Company

New York St. Louis San Francisco Auckland Bogotá
Düsseldorf Johannesburg London Madrid Mexico Montreal
New Delhi Panama Paris São Paulo Singapore Sydney
Tokyo Toronto

To Ann, Susie, Katie, and Jim

Social Statistics

1 2 3 4 5 6 7 8 9 0 DODO 7 8 3 2 1 0 9

This book was set in Mono 8A by Monotype Composition Company, Inc.
The editors were Richard R. Wright and James R. Belser;
the designer was J + M Condon;
the production supervisor was Dominick Petrellese.
The drawings were done by J & R Services, Inc.
The cover was designed by John Hite.
R. R. Donnelley & Sons Company was printer and binder.

Library of Congress Cataloging in Publication Data

Blalock, Hubert M
Social statistics.

(McGraw-Hill series in sociology)
Bibliography: p.
Includes indexes.
1. Statistics. 2. Social sciences—Statistical methods. I. Title.
HA29.B59 1979 519.5 78-17764
ISBN 0-07-005752-4

Preface

This text is written primarily for those students of sociology, both advanced undergraduates and graduate students, who actually intend to engage in social research.

In the nineteen years that have elapsed since the publication of the first edition, the level of training and the sophistication in applied statistics has undergone considerable improvement not only in sociology but also in political science, anthropology, geography, and social work. Nevertheless, the overwhelming majority of students and practitioners in these fields still lack the necessary mathematical backgrounds to take full advantage of the rapidly accumulating technical literature in mathematical statistics and econometrics. With these basic facts in mind, this text has been written so as to avoid mathematical derivations insofar as possible, and only a quick review of certain algebraic principles listed in Appendix 1 should therefore be sufficient preparation for the average student. Although it is not necessary in a first course in statistics to stress mathematical derivations, the author is convinced that certain basic and fundamental ideas underlying the principles of statistical inference must be thoroughly understood if one is to obtain more than a mere "cookbook" knowledge of statistics. Therefore, there is a relatively heavy emphasis on the underlying logic of statistical inference, including a chapter on probability, with relatively less attention being given to some of the more or less routine topics ordinarily discussed in elementary texts.

One of the most difficult problems encountered in the teaching of applied statistics is that of motivating students, both in enabling them to overcome their fears of mathematics and in learning to apply statistics to their own field of interest. It is for the latter reason that the author has not attempted to cover a wide range of applications but has selected examples of primary interest to sociologists. To some extent, examples

have also been chosen from disciplines which border on sociology: fields such as social psychology, social work, and political behavior. In most instances each new topic has been illustrated by a single example, under the assumption that most students will lose track of the basic line of thought if too many examples are used to illustrate the same point. Additional examples are therefore given in the form of exercises at the end of each chapter. In general, the author has tried to strike a reasonable compromise between the desirability of stating basic principles as clearly and concisely as possible and the necessity of repeating some of the more difficult ideas each time a new topic is discussed. Insofar as possible, new ideas have been introduced gradually, and, equally important, an effort has been made to relate each new topic to those which have preceded it. In so doing, the major goal has been to give an appreciation of the basic similarities underlying many of the most commonly used tests and measures.

Almost all the suggestions I have received from those wishing to help improve the first and second editions have implied additions to the book, rather than subtractions, and they have also implied that many of the topics originally treated should be discussed more technically. My own position is that sociologists and political scientists, in particular, need greater exposure to the more technical literature on experimental designs and on the use of simultaneous-equation procedures in connection with nonexperimental research. Yet, it became clear that if these materials were added to the original text, it would lose its appeal as an introductory text appropriate for advanced undergraduates majoring in the social sciences.

It was therefore decided to treat such topics as gaining facility with computer programming, the general linear model, experimental designs, simultaneous-equation procedures, path analysis, and the handling of measurement errors in a separate volume written with two of my former colleagues, Lewis F. Carter and N. Krishnan Namboodiri. This book, titled *Applied Multivariate Analysis and Experimental Designs*, has now been published by McGraw-Hill and may be used as a supplement to this text.

Apart from additions to the exercises and an updating of the bibliographies at the ends of the chapters in this revised second edition, virtually all of the changes and additions to the previous edition have been made for the purpose of introducing the student to a variety of more advanced or specialized topics that are becoming increasingly useful to the social scientist. With a few exceptions, these additions appear in Chapters 15 through 20, and, for the most part, involve topics in multivariate analysis.

In Chapter 15 on nominal-scale procedures there are additions of varying length dealing with the likelihood ratio chi square, the partitioning of chi square, odds ratios, prediction logic, and statistical interaction. Chapters 18 and 19 contain more extensive discussions of ordinal measures of association, cautions regarding grouping and measurement errors, Quade's matching procedure for handling partials for ordinal measures, and significance tests for ordinal measures. These expanded discussions of nominal and ordinal procedures are intended to help the student move more easily to the specialized literatures on these subjects, which have become more readily available to social scientists and which are now integral parts of standard packaged computer programs such as SPSS.

Chapter 16 now begins with a brief discussion of the general linear model, which is emphasized again later in the chapter and throughout the remaining chapters of Part 4. The discussion of experimental designs has not been expanded, since this topic constitutes fully one-third of the coverage of the companion volume, *Applied Multivariate Analysis and Experimental Designs,* written with Carter and Namboodiri. In Chapter 19 there is a somewhat expanded discussion of causal models and path analysis, as well as a brief section introducing the matrix algebra representation of the general linear model, though the major discussion of these topics has been reserved for *Applied Multivariate Analysis and Experimental Designs.* Chapter 20 has been expanded to include expository discussions of Sonquist and Morgan's Automatic Interaction Detector procedure and log-linear models.

Recognizing that the learning of statistics depends very heavily on the working of numerical exercises and in gaining facility with computer programs, it was decided to develop an instructor's manual as a supplement to the text. It has been written by James Henney.

In the instructor's manual, Henney has developed a number of exercises that are keyed not only to the chapters in *Social Statistics* but also to the SPSS manual, so that students and instructors desiring to use these two sources simultaneously may be provided added guidance. The Namboodiri, Carter, and Blalock text also contains a chapter designed to familiarize the reader with FORTRAN and to enable him or her to modify programs so as to handle nonstandard types of statistical questions that are not covered in the packaged programs such as SPSS.

Included in the text are a number of sections, paragraphs, and exercises which are either conceptually difficult or which presuppose that the student is reasonably familiar with topics ordinarily covered in courses on research methods. These portions of the text have been marked with an asterisk (*) and may be skimmed on first reading or omitted entirely. Instructors

using the text for a one-semester course may wish to indicate that students should omit these materials.

For assistance in the preparation of these editions, I would like to thank the many students at The University of Michigan, Yale University, The University of North Carolina, and The University of Washington, who have made numerous suggestions for the book's improvement. To Richard T. LaPiere, Sanford Dornbusch, Robert Ellis, Santo Camilleri, Theodore Anderson, Richard G. Ames, Erica Borden, and Louis Goodman my appreciation for reading and criticizing earlier drafts and editions of the volume. For proofing, typing, and checking computations, I would like to thank Ann Blalock, Diane Etzel, Ann Laux, and Doris Slesinger.

My deep appreciation and many thanks go to Daniel O. Price, who deserves the major credit for stimulating my interest in statistics.

I am indebted to Professor Sir Ronald A. Fisher, formerly of Cambridge, to Dr. Frank Yates, Rothamsted, and to Messrs. Oliver and Boyd Ltd., Edinburgh, for permission to reprint Tables III, IV, and V from their book *Statistical Tables for Biological, Agricultural and Medical Research.* I am equally grateful to those other publishers and authors, acknowledged in the appropriate places, who kindly gave their permission for use of various tables and computing forms.

Hubert M. Blalock, Jr.

Contents

Introduction

Part 1

Introduction: Purposes and Limitations of Statistics

1

The field of statistics has widespread applications as indicated by the fact that statistics courses are offered in such dissimilar subjects as dentistry and sociology, business administration and zoology, and public health and education. In spite of this fact, there are many misconceptions concerning the nature of this rapidly developing discipline. The layman's conception of statistics is apt to be very different from that of the professional statistician. A statistician is sometimes thought of as a person who manipulates numbers in order to prove a point. On the other hand, some students of sociology or other social sciences have tended to worship the statistician as someone who, with the aid of a magical computer, can make almost any study "scientific." Possibly because of the awe many persons have for anything with a hint of mathematics, students often find it difficult to approach a course in statistics with other than mixed feelings. Although they may be terrified at the prospect of working with numbers, they may also come to expect too much of a discipline that appears so formidable. Before jumping into the subject too quickly and thereby losing perspective, let us first ask ourselves just what statistics is and what it can and cannot do.

It is perhaps easiest to begin by stating what statistics is not. Statistics first of all is not a method by which one can prove almost anything one wants to prove. In fact, we shall find statisticians very carefully laying down rules of the game to ensure that interpretations do not go beyond the limits of the data. There is nothing inherent in statistical methods to prevent the careless or intellectually dishonest individual from drawing his or her own conclusions in spite of the data, however, and one of the most important functions of an introductory course in statistics is to place the student on guard against possible misuses of this tool.

Statistics is not simply a collection of facts. If it were, there would hardly be much point in studying the subject. Nor is statistics a substitute for abstract theoretical thinking or for careful examination of exceptional cases. In some of the older textbooks on research methods one used to find lengthy discussions on the relative merits of the case-study method versus the statistical method. It is now clearly recognized that statistical methods are in no sense opposed to the qualitative analysis of case studies but that the two approaches are complementary. It is not even true that statistics is applicable only when there are a large number of cases or that it cannot be used in exploratory studies. Finally, statistics is not a substitute for measurement or the careful construction of an interview schedule or other instrument of data collection. This last point will receive further attention toward the end of this chapter and in Chap. 2.

Having indicated what statistics is not, can we say definitely what it is? Unfortunately, persons who call themselves statisticians seem to disagree somewhat as to exactly what is covered under the general heading of statistics. Taking a pragmatic approach to the problem, we shall say that statistics has two very broad functions. The first of these functions is description, the summarizing of information in such a manner as to make it more usable. The second function is induction, which involves either making generalizations about some population, on the basis of a sample drawn from this population, or formulating general laws on the basis of repeated observations. Each of these functions will be discussed in turn.

1.1 Functions of Statistics

Descriptive statistics Quite frequently in social research we find ourselves in the position of having so much data that we cannot adequately absorb all this information. We may have collected 200 questionnaires and be in the embarrassing position of having to ask, "What do I do with it all?" With so much information it would be exceedingly difficult for any but the most photographic minds to grasp intuitively what is in the data. The information must somehow or other be boiled down to a point at which the researcher can see what is in it; it must be summarized. By computing measures such as percentages, means, standard deviations, and correlation coefficients it may be possible to reduce the data to manageable proportions. In summarizing data by substituting a very few measures for many numbers, certain information is inevitably lost, and, more serious, it is very possible to obtain results

that are misleading unless cautiously interpreted. Therefore, the limitations of each summarizing measure must be clearly indicated.

Descriptive statistics is especially useful in instances where the investigator finds it necessary to handle interrelationships among more than two variables. For example, suppose one needed to use eight or ten variables to help explain delinquency rates, and furthermore suppose these explanatory or *independent* variables were themselves highly intercorrelated. If we wished to isolate the effects of one or two of these variables, controlling for the remainder, how would we proceed? And what kinds of assumptions would be needed? Questions of this degree of complexity arise in a branch of statistics referred to as *multivariate analysis.* Relatively simple problems of multivariate analysis will be discussed in Chaps. 15, 16, 19, and 20, but more complex questions must be reserved for more advanced courses.

Inductive statistics An equally important function of statistics, and certainly the one which will occupy much of our attention in this text, is that of *induction* or inferring properties of a population on the basis of known sample results. *Statistical inference,* as the process is called, involves rather complex reasoning, but when properly used and understood becomes a very important tool in the development of a scientific discipline. Inductive statistics is based directly on probability theory, a branch of mathematics. We thus have a purely deductive discipline providing a rational basis for inductive reasoning. To the writer's knowledge there is no other rational basis for induction. This general point will be discussed in more detail in Chap. 8.

There are several practical reasons why it is often necessary to attempt to generalize on the basis of limited information. The most obvious is the time-cost factor. It would be wholly impractical, if not prohibitively costly, to ask every voter how he or she intended to vote in order to predict a national election. Nor can the ordinary researcher afford to tap every resident of a large city in order to study prejudice, social mobility, or any other phenomenon. We first decide upon the exact nature of the group about which we wish to generalize (the population). For example, we may select all citizens of voting age or all white males over eighteen residing within the city limits of Detroit. We then usually draw a sample consisting of a relatively small proportion of these people, being primarily interested, however, not in this particular sample but in the larger population from which the sample was drawn. For example, we may find that within a particular sample of 200 white males there is a negative relationship between education and prejudice. Recognizing

that had another set of 200 individuals been sampled the results might have been different, we would nevertheless like to make certain inferences as to the nature of the relationship had the entire population of adult white males in Detroit been studied.

Another reason for generalizing on the basis of limited information is that it may be impossible to make use of the entire population simply because the population is infinite or not easily defined. In replicating an experiment in the natural or social sciences the goal always seems to be some kind of generalization which it is hoped will apply under similar circumstances. Or a social scientist may have collected data on all of the cases available. For example, we may have used all 50 states as the units of analysis in studying internal migration. Nevertheless, one may want to generalize about migration under *similar* conditions. In each of these instances the situation calls for inductive statistics.

At this point you may ask a question of the following sort: "If statistics is so important, why is it that sciences such as physics and chemistry have been able to get along so well without the extensive use of statistical techniques? Is there anything different about these sciences?" Quite obviously there is. Some of the natural sciences have developed for centuries without the use of inductive statistics. But this seems to be primarily a matter of good fortune or, to give these scientists credit for their own efforts, a relatively satisfactory control over disturbing elements in the environment. As will become apparent in later chapters, to the degree that carefully controlled laboratory conditions prevail, there is less practical need for statistical techniques. In this sense statistics is a poor person's substitute for contrived laboratory experiments in which all important relevant variables have been controlled. It should be emphasized, however, that many of the same statistical principles apply to laboratory experiments in physics, to somewhat less precise agricultural experiments, and to social surveys. For example, if an experiment in physics has been replicated 37 times with similar results, it is nevertheless conceivable that subsequent trials will yield different outcomes. The scientist must therefore generalize on the basis of a limited number of experiments, and the inferences made are essentially statistical in nature. Also, the problem of measurement error can be conceived in statistical terms. No matter how precise the measuring instrument, the scientist never obtains exactly the same results with each replication. One may attribute these differences either to measurement error or to disturbing effects of uncontrolled variables. Statistics becomes especially necessary whenever there is so much variation from one replication to the next that differences cannot be ignored or attributed to measurement error. Basically, then, statistical inference underlies all scientific general-