Statistical Methods for Comparative Studies

Techniques for Bias Reduction

SHARON ANDERSON
ARIANE AUQUIER
WALTER W. HAUCK
DAVID OAKES
WALTER VANDAELE
HERBERT I. WEISBERG

with contributions from

ANTHONY S. BRYK
JOEL KLEINMAN

JOHN WILEY & SONS

New York • Chichester • Brisbane • Toronto • Singapore

Library of Congress Cataloging in Publication Data

Main entry under title:
Statistical methods for comparative studies.
(Wiley series in probability and mathematical statistics)
Includes bibliographical references and index.
1. Mathematical statistics. I. Anderson, Sharon, 1948-
II. Title: Bias reduction.
QA276.S783 001.4'22 79-27220
ISBN 0-471-04838-0

Printed in the United States of America

10 9 8

Preface

Investigators in many fields need methods for evaluating the effectiveness of new programs or practices involving human populations. To determine whether a program is more effective than the status quo or another alternative, we must perform comparative studies. An ideal study would apply the different programs to identical groups of subjects. Randomized experiments are often advocated as approximating this ideal. Often, however, randomization is not feasible, resulting in difficult problems of design and analysis. To address these problems, a variety of statistical methods have been developed. Many of these methods are quite recent, and to date have appeared only in technical journals. Although they are potentially very useful to researchers in many fields, these techniques are presently not readily accessible.

In this book we bring together for the first time the various techniques for the design and analysis of comparative studies. The book includes, at a relatively nontechnical level, both familiar techniques and more recent developments. Although we present theoretical results concerning the performance of the various techniques, we emphasize primarily practical implications for the applied researcher. Throughout the book we develop for the applied research worker a basic understanding of the problems and techniques and avoid highly mathematical presentations in the main body of the text.

Overview of the Book

The first five chapters discuss the main conceptual issues in the design and analysis of comparative studies. We carefully motivate the need for standards of comparison and show how biases can distort estimates of treatment effects. The relative advantages of randomized and nonrandomized studies are also presented.

Chapters 6 to 10 present the various methods: matching (including multivariate matching); standardization and stratification; analysis of covariance; and two relatively new multivariate methods, logit analysis and log-linear analysis. We emphasize the assumptions under which the techniques were developed and, whenever possible, assess quantitatively their effectiveness in reducing bias. Although we emphasize estimation as opposed to hypothesis testing, we do indicate the appropriate tests and provide references.

Chapter 11, on survival analysis, deals with the special problem of subject losses during the course of a study and discusses how to form estimates which are not biased by these losses. Chapter 12 discusses repeated measures designs, where subjects are assessed on the same variable both before and after treatment intervention, and presents new methods to handle these problems.

An important feature of the book is Chapter 13. In this chapter we describe the comparative effectiveness of the techniques in reducing bias. In addition, we discuss methods that combine features of two or more techniques. Chapter 14 deals with many of the practical issues that must be faced before drawing causal inferences from comparative studies.

Use of the Book

The book is intended for students, researchers, and administrators who have had a course in statistics or the equivalent experience. We assume that the reader has a basic familiarity with such techniques as regression and analysis of variance, in addition to the basic principles of estimation and hypothesis testing. Depending on the reader's background, some of the relatively more technical sections may be too difficult. The book is written, however, so that the more technical sections can be skipped without loss of understanding of the essentials.

We view this book as serving two different functions. First, the book can be used in a course in research and evaluation methods for students in fields such as public health, education, social welfare, public safety, psychology, medicine, and business. Second, the book serves as a reference for applied researchers wishing to determine which techniques are appropriate for their particular type of study.

However this book is used, we encourage the reader to begin with the first five chapters, because these chapters provide the definitions and lay the foundation for a clear understanding of the problems. A knowledge of the terminology is particularly important, because the fields of application and the statistical literature tend to lack a common terminology. The reader could then refer to Chapter 13, which serves to identify the most appropriate technique(s) (see especially Table 13.1).

Acknowledgments

Work on this book began in the fall of 1974 with discussions in the Observational Studies Group of the Faculty Seminar on the Analysis of Health and Medical Practices at Harvard University. This seminar was supported in part by grants from the Robert Wood Johnsons Foundation; the Edna McConnell Clark Foundation; and the Commonwealth Fund through the Center for the Analysis of Health Practices, Harvard School of Public Health. We wish to thank the organizers and sponsors of this seminar for providing us with a stimulating environment in which to begin our project.

We are indebted to Richard Light for suggesting the idea for this book, starting us in the right direction, and contributing early drafts. Joel Kleinman deserves thanks for early drafts of the first five chapters and helpful comments. We also wish to acknowledge the contribution of Anthony Bryk, who assumed major responsibility for Chapter 12.

Among colleagues who have read drafts and helped with comments, special thanks go to William Cochran, Ted Colton, Allan Donner, William Haenszel, and John Pratt. We are especially grateful to Frederick Mosteller, who provided us with very helpful comments on several drafts and facilitated the production of this book through National Science Foundation Grant SOC-75-15702.

The preparation of this work was partly facilitated by National Science Foundation Grants SOC-75-15702 and SOC-76-15546 (W.V.), and National Institute of Education Contract C-74-0125 (H.W.) and Grant G-76-0090 (H.W.).

SHARON ANDERSON
ARIANE AUQUIER
WALTER W. HAUCK
DAVID OAKES
WALTER VANDAELE
HERBERT I. WEISBERG

Chicago, Illinois
Paris, France
Chicago, Illinois
London, England
Cambridge, Massachusetts
Boston, Massachusetts
July, 1980

Contents

CHAPTER 1

Introduction

This book is concerned with the design and analysis of research studies assessing the effect on human beings of a particular *treatment*. We shall assume that the researchers know what kinds of effects they are looking for and, more precisely, that there is a definite *outcome* of interest. Examples of such treatments and the corresponding outcomes include the administration of a drug (treatment) claimed to reduce blood pressure (outcome), the use of seat belts (treatment) to reduce fatalities (outcome) among those involved in automobile accidents, and a program (treatment) to improve the reading level (outcome) of first graders. As is seen from these examples, the word "treatment" is used in a very general sense.

1.1. PROBLEMS OF COMPARATIVE STUDIES: AN OVERVIEW

It is useful to begin with what might at first sight appear to be an obvious question: What do we mean by the effect of a treatment? We would like to ascertain the differences between the results of two studies. In the first study we determine what happens when the treatment is applied to some group, in the second we determine what would have happened to the same group if it had not been given the treatment of interest. Whatever differences there may be between

the outcomes measured by the two studies would then be direct consequences of the treatment and would thus be measures of its effect.

This ideal experiment is, of course, impossible. Instead of doing the second study, we establish a standard of comparison to assess the effect of the treatment. To be effective, this standard of comparison should be an adequate proxy for the performance of those receiving the treatment—the *treatment group*—if they had not received the treatment. One of the objectives of this book is to discuss how to establish such standards of comparison to estimate the effect of a treatment.

Standards of comparison usually involve a *control* or *comparison group* of people who do not receive the treatment. For example, to measure the effect of wearing seat belts on the chance of surviving an automobile accident, we could look at drivers involved in auto accidents and compare the accident mortality of those who wore seat belts at the time of the accident with the accident mortality of those who did not. Drivers who were wearing seat belts at the time of the accident would constitute the treatment group, those who were not would constitute the control group. Ideally, the accident mortality of the control group is close to what the accident mortality of the treatment group would have been had they not worn seat belts. If so, we could use the accident mortality of the control group as a standard of comparison for the accident mortality of the treatment group.

Unfortunately, the use of a control group does not in itself ensure an adequate standard of comparison, since the groups may differ in factors other than the treatment, factors that may also affect outcomes. These factors may introduce a bias into the estimation of the treatment effect. To see how this can happen, consider the seat belt example in more detail.

Example 1.1 Effect of seat belts on auto accident fatality: Consider a hypothetical study attempting to determine whether drivers involved in auto accidents are less likely to be killed if they wear seat belts. Accident records for a particular stretch of highway are examined, and the fatality rate for drivers wearing seat belts compared with that for drivers not wearing seat belts. Suppose that the numbers of accidents in each category was as given in Table 1.1.

From Table 1.1, the fatality rate among drivers who wore seat belts was $10/50 = 0.2$ and the rate among those not wearing seat belts was $20/50 = 0.4$. The difference of $0.4 - 0.2 = 0.2$ between the two rates can be shown by the usual chi-square test to be statistically significant at the .05 level. At first sight the study appears to demonstrate that seat belts help to reduce auto accident fatalities.

Table 1.1 Hypothetical Auto Accident Data

	Seat Belts		
	Worn	Not Worn	Total
Driver killed	10	20	30
Driver not killed	40	30	70
Total	50	50	100
Fatality rate	0.2	0.4	

Table 1.2 Auto Accident Data Classified by Speed at Impact

	Low Impact Speed			High Impact Speed		
	Seat Belts Worn	Seat Belts Not Worn	Total	Seat Belts Worn	Seat Belts Not Worn	Total
Driver killed	4	2	6	6	18	24
Driver not killed	36	18	54	4	12	16
Total	40	20	60	10	30	40
Fatality rate	0.1	0.1		0.6	0.6	

A major problem with this study, however, is that it takes no account of differences in severity among auto accidents, as measured, for example, by the speed of the vehicle at impact. Suppose that the fatalities among accidents at low speed and at high speed were as given in Table 1.2.

Notice that adding across the cells of Table 1.2 gives Table 1.1. Thus $10 = 6 + 4$, $20 = 2 + 18$, $40 = 36 + 4$, and $30 = 18 + 12$. However, Table 1.2 tells a very different story from Table 1.1. At low impact speed, the fatality rate for drivers wearing seat belts is the same as that for drivers not wearing seat belts, namely 0.1. The fatality rate at high impact speed is much greater, namely 0.6, but is still the same for belted and unbelted drivers. These fatality rates suggest that seat belts have no effect in reducing auto accident fatalities.

The data of Example 1.1 are hypothetical. The point of the example is not to impugn the utility of seat belts (or of well-conducted studies of the utility of seat belts) but to illustrate how consideration of an extra variable (speed at impact) can completely change the conclusions drawn.

A skeptical reader might ask if there is a plausible explanation for the data of Table 1.2 (other than that the authors invented it). The crux of the example is that drivers involved in accidents at low speed are more likely to be wearing seat belts than those involved in accidents at high speed. The proportions, calculated from the third line of Table 1.2, are $40/60$ and $10/40$, respectively. Perhaps slow drivers are generally more cautious than are fast drivers, and so are also more likely to wear seat belts.

We say that speed at impact is a *confounding factor* because it confounds or obscures the effect, if any, of the risk factor (seat belts, or the lack of them) on outcome (death or survival). In other words, the confounding factor results in a *biased* estimate of the effect.

Fortunately, if (as in Example 1.1) the confounding factor or factors can be identified and measured, the bias they cause may be substantially reduced or even eliminated. Our purpose in this book is to present enough detail on the

various statistical techniques that have been developed to achieve this bias reduction to allow researchers to understand when each technique is appropriate and how it may be applied.

1.2 PLAN OF THE BOOK

In Chapters 2 and 3 we discuss the concepts of bias and confounding. In Chapter 3 we also consider the choice of the summary measure used to describe the effect of the treatment. In Example 1.1 we used the difference between the fatality rates of the belted and unbelted drivers to summarize the apparent effect of the treatment, but other choices of measure are possible, for example the ratio of these rates.

The construction of standards of comparison is the subject of Chapter 4. As we have said, these usually involve a control or comparison group that does not receive the treatment. When the investigator can choose which subjects enter the treatment group and which enter the control group, randomized assignment of subjects to the two groups is the preferred method. Since randomization is often not feasible in studies of human populations, we discuss both randomized and nonrandomized studies. In nonrandomized studies statistical techniques are needed to derive valid standards of comparison from the control group, which, as we have seen in Example 1.1, may otherwise give misleading results. Although randomized studies are less likely to mislead, their precision can often be improved by the same statistical techniques.

Chapter 5 discusses the choice of variables to be used in the analysis, a choice that must be related to the context and aims of the study. We also show how the specification of a mathematical model relating the chosen variables is crucial to the choice of an appropriate method of analysis and consider the effects of inadequacies in the model specification.

Chapters 6 to 10 each consider one statistical technique for controlling bias due to confounding factors. These techniques fall into two major categories, *matching* and *adjustment*.

In matching (Chapter 6), the members of the comparison group are selected to resemble members of the treatment group as closely as possible. Matching can be used either to assemble similar treatment and control groups in the planning of the study before the outcomes are determined, or to select comparable subjects from the two groups after a treatment has been given and outcomes measured. Unlike randomization, which requires control over the composition of both groups, matching can be used to construct a comparison group similar to a preselected or self-selected treatment group.

The other major category, adjustment techniques, consists of methods of analysis which attempt to estimate what would have happened if the treatment and comparison groups had been comparable when in fact they were not. In other words, the estimate of the effect of the treatment is adjusted to compensate for the differences between the groups. These adjustment methods include standardization and stratification (Chapter 7), analysis of covariance (Chapter 8), logit analysis (Chapter 9), and log-linear analysis (Chapter 10).

A common problem with longitudinal studies is that subjects may be lost to follow-up at the end of or during the course of the study. Chapter 11, on survival analysis, discusses the analysis of such studies, including the control of confounding factors. Chapter 12 discusses repeated measures designs, where the same subjects are assessed on the outcome variable before and after the intervention of a treatment.

Two summary chapters conclude the book. Chapter 13 discusses the choice of statistical technique and shows how two techniques can sometimes be used together. Finally, Chapter, 14 presents criteria to consider in drawing causal inferences from a comparative study.

The methodological Chapters (6 to 12) may be read in any order, but they all use material from Chapters 1 to 5. Chapter 13 refers in detail to Chapters 6 to 10. Chapter 14 may be read at any point.

The book presents the general rationale for each method, including the circumstances when its use is appropriate. The focus throughout is on unbiased, or nearly unbiased estimation of the effect of the treatment. Tests of significance are given when these can be performed easily. Although we give many examples to illustrate the techniques, we do not dwell on computational details, especially when these can best be performed by computer. We shall assume throughout the book that the researchers have chosen a single outcome factor for study. For simplicity of presentation we often also restrict attention to the estimation of the effect of a single treatment in the presence of a single confounding factor, although extensions to multiple confounding factors are indicated. Some special issues that arise with multiple confounding factors are discussed in Chapter 5.

Throughout the book the main concern will be *internal validity*—attaining a true description of the effect of the treatment on the individuals in the study. The question of *external validity*—whether the findings apply also to a wider group or population—is not discussed in depth as it is primarily determined by the subject matter rather than by statistical considerations.

1.3 NOTES ON TERMINOLOGY

Throughout, we shall refer to the effect of interest as the *outcome factor*. A common synonym is *response factor*. The agent whose effect on the outcome factor is being studied will be called the *treatment, treatment factor,* or *risk*

factor. The word "treatment" is generally used to describe an agent applied specifically to affect the outcome factor under consideration (as was true for all the examples in the first paragraph of this chapter). The term "risk factor," borrowed from epidemiology, is used when exposure to the agent is accidental or uncontrollable, or when the agent is applied for some purpose other than to affect the specific outcome factor under consideration. An example would be the study of the effect of smoking on the incidence of lung cancer. The use of the term "risk factor" does not in itself imply that the agent is "risky" or in fact, that risk enters the discussion at all. We use whichever term ("treatment" or "risk factor") appears more natural in context.

In later chapters we talk about quantities or labels that measure the presence, absence, level or amount of a risk factor, treatment, outcome factor, or confounding factor. Such quantities or labels will be termed *variables.* In studying the effect of seat belts on accident mortality (Example 1.1) we may define a risk variable taking the value 1 or 0, depending on whether or not the driver was wearing a seat belt at the time of the accident. The logical distinction between a factor and a variable which measures that factor is not always made in the literature, but it can be useful.

The term "comparison group" is used interchangeably with the more familiar "control group." When the important comparison is between a proposed new treatment and the present standard treatment, the standard treatment (rather than no treatment) should be given to the comparison group. In dealing with risk factors it is natural to speak of "risk groups" or of "exposed" and "nonexposed" groups. We may have several different "exposed" or "treatment" groups, corresponding to different levels of the risk factor or treatment.

CHAPTER 2

Confounding Factors

In the discussion of Example 1.1 (effect of wearing seat belts on auto accident fatality) we saw that a background factor (speed at impact) could seriously distort the estimate of the effect of the risk factor on the outcome. The distortion will arise whenever two conditions hold:

1. The risk groups differ on the background factor.
2. The background factor itself influences the outcome.

Background factors which satisfy conditions 1 and 2 are called confounding factors. If ignored in the design and analysis of a study, they may affect its conclusions, for part of the effect of the confounding factor on the outcome may appear to be due to the risk factor. Table 1.1 is misleading because the effect on accident fatality apparently due to wearing seat belts (the risk factor) is actually due to speed at impact (the confounding factor).

In Section 2.1 we show by another example how the effect of a risk factor can