HZ BOOKS
华章教育

CAMBRIDGE

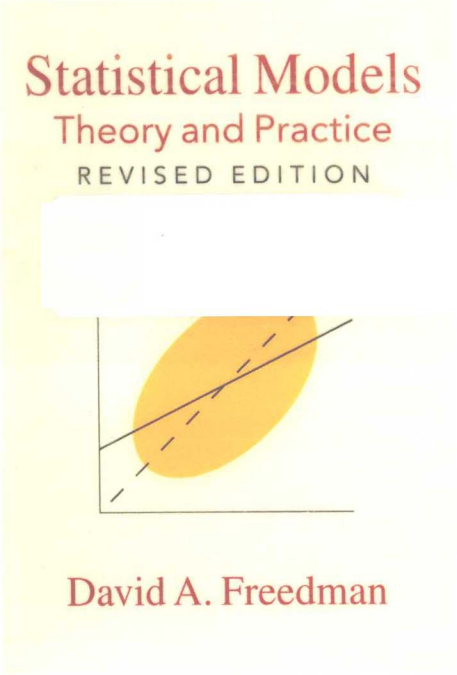
华章数学·统计学原版精品系列

统计模型

理论和实践

Statistical Models Theory and Practice (Revised Edition)

（英文版·第2版）



（美）David A. Freedman
加州大学伯克利分校 著

机械工业出版社
China Machine Press

华章数学·统计学原版精品系列

统计模型
理论和实践

Statistical Models Theory and Practice (Revised Edition)

（英文版·第2版）

机械工业出版社
China Machine Press

本书版权登记号：图字：01-2010-5147

图书在版编目（CIP）数据

统计模型：理论和实践（英文版·第2版）/（美）弗里德曼（Freedman, D. A.）著. —北京：机械工业出版社，2010.9

（华章数学·统计学原版精品系列）

书名原文：Statistical Models: Theory and Practice, Revised Edition

ISBN 978-7-111-31797-5

Ⅰ. 统… Ⅱ. 弗… Ⅲ. 统计模型－英文 Ⅳ. C81

中国版本图书馆 CIP 数据核字（2010）第 175153 号

机械工业出版社（北京市西城区百万庄大街 22 号　邮政编码 100037）

责任编辑：李俊竹

北京京师印务有限公司印刷

2010 年 9 月第 1 版第 1 次印刷

150 mm × 214 mm · 14.25 印张

标准书号：ISBN 978-7-111-31797-5

定　　价：38.00 元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换

客服热线：（010）88378991；88361066

购书热线：（010）68326294；88379649；68995259

投稿热线：（010）88379604

读者信箱：hzjsj@ hzbook. com

Foreword to the Revised Edition

Some books are correct. Some are clear. Some are useful. Some are entertaining. Few are even two of these. This book is all four. *Statistical Models: Theory and Practice* is lucid, candid and insightful, a joy to read. We are fortunate that David Freedman finished this new edition before his death in late 2008. We are deeply saddened by his passing, and we greatly admire the energy and cheer he brought to this volume—and many other projects—during his final months.

This book focuses on half a dozen of the most common tools in applied statistics, presenting them crisply, without jargon or hyperbole. It dissects real applications: a quarter of the book reprints articles from the social and life sciences that hinge on statistical models. It articulates the assumptions necessary for the tools to behave well and identifies the work that the assumptions do. This clarity makes it easier for students and practitioners to see where the methods will be reliable; where they are likely to fail, and how badly; where a different method might work; and where no inference is possible—no matter what tool somebody tries to sell them.

Many texts at this level are little more than bestiaries of methods, presenting dozens of tools with scant explication or insight, a cookbook, numbers-are-numbers approach. "If the left hand side is continuous, use a linear model; fit by least-squares. If the left hand side is discrete, use a logit or probit model; fit by maximum likelihood." Presenting statistics this way invites students to believe that the resulting parameter estimates, standard errors, and tests of significance are meaningful—perhaps even untangling complex causal relationships. They teach students to think scientific inference is purely algorithmic. Plug in the numbers; out comes science. This undervalues both substantive and statistical knowledge.

To select an appropriate statistical method actually requires careful thought about how the data were collected and what they measure. Data are not "just numbers." Using statistical methods in situations where the underlying assumptions are false can yield gold or dross—but more often dross.

Statistical Models brings this message home by showing both good and questionable applications of statistical tools in landmark research: a study of political intolerance during the McCarthy period, the effect of Catholic schooling on completion of high school and entry into college, the relationship between fertility and education, and the role of government institutions in shaping social capital. Other examples are drawn from medicine and

epidemiology, including John Snow's classic work on the cause of cholera—a shining example of the success of simple statistical tools when paired with substantive knowledge and plenty of shoe leather. These real applications bring the theory to life and motivate the exercises.

The text is accessible to upper-division undergraduates and beginning graduate students. Advanced graduate students and established researchers will also find new insights. Indeed, the three of us have learned much by reading it and teaching from it.

And those who read this textbook have not exhausted Freedman's approachable work on these topics. Many of his related research articles are collected in *Statistical Models and Causal Inference: A Dialogue with the Social Sciences* (Cambridge University Press, 2009), a useful companion to this text. The collection goes further into some applications mentioned in the textbook, such as the etiology of cholera and the health effects of Hormone Replacement Therapy. Other applications range from adjusting the census for undercount to quantifying earthquake risk. Several articles address theoretical issues raised in the textbook. For instance, randomized assignment in an experiment is not enough to justify regression: without further assumptions, multiple regression estimates of treatment effects are biased. The collection also covers the philosophical foundations of statistics and methods the textbook does not, such as survival analysis.

Statistical Models: Theory and Practice presents serious applications and the underlying theory without sacrificing clarity or accessibility. Freedman shows with wit and clarity how statistical analysis can inform and how it can deceive. This book is unlike any other, a treasure: an introductory book that conveys some of the wisdom required to make reliable statistical inferences. It is an important part of Freedman's legacy.

David Collier, Jasjeet Singh Sekhon, and Philip B. Stark
University of California, Berkeley

Preface

This book is primarily intended for advanced undergraduates or beginning graduate students in statistics. It should also be of interest to many students and professionals in the social and health sciences. Although written as a textbook, it can be read on its own. The focus is on applications of linear models, including generalized least squares, two-stage least squares, probits and logits. The bootstrap is explained as a technique for estimating bias and computing standard errors.

The contents of the book can fairly be described as what you have to know in order to start reading empirical papers that use statistical models. The emphasis throughout is on the connection—or lack of connection—between the models and the real phenomena. Much of the discussion is organized around published studies; the key papers are reprinted for ease of reference. Some observers may find the tone of the discussion too skeptical. If you are among them, I would make an unusual request: suspend belief until you finish reading the book. (Suspension of disbelief is all too easily obtained, but that is a topic for another day.)

The first chapter contrasts observational studies with experiments, and introduces regression as a technique that may help to adjust for confounding in observational studies. There is a chapter that explains the regression line, and another chapter with a quick review of matrix algebra. (At Berkeley, half the statistics majors need these chapters.) The going would be much easier with students who know such material. Another big plus would be a solid upper-division course introducing the basics of probability and statistics.

Technique is developed by practice. At Berkeley, we have lab sessions where students use the computer to analyze data. There is a baker's dozen of these labs at the back of the book, with outlines for several more, and there are sample computer programs. Data are available to instructors from the publisher, along with source files for the labs and computer code: send email to solutions@cambridge.org.

A textbook is only as good as its exercises, and there are plenty of them in the pages that follow. Some are mathematical and some are hypothetical, providing the analogs of lemmas and counter-examples in a more conventional treatment. On the other hand, many of the exercises are based on actual studies. Here is a summary of the data and the analysis; here is a

specific issue: where do you come down? Answers to most of the exercises are at the back of the book. Beyond exercises and labs, students at Berkeley write papers during the semester. Instructions for projects are also available from the publisher.

A text is defined in part by what it chooses to discuss, and in part by what it chooses to ignore; the topics of interest are not to be covered in one book, no matter how thick. My objective was to explain how practitioners infer causation from association, with the bootstrap as a counterpoint to the usual asymptotics. Examining the logic of the enterprise is crucial, and that takes time. If a favorite technique has been slighted, perhaps this reasoning will make amends.

There is enough material in the book for 15–20 weeks of lectures and discussion at the undergraduate level, or 10–15 weeks at the graduate level. With undergraduates on the semester system, I cover chapters 1–7, and introduce simultaneity (sections 9.1–4). This usually takes 13 weeks. If things go quickly, I do the bootstrap (chapter 8), and the examples in chapter 9. On a quarter system with ten-week terms, I would skip the student presentations and chapters 8–9; the bivariate probit model in chapter 7 could also be dispensed with.

During the last two weeks of a semester, students present their projects, or discuss them with me in office hours. I often have a review period on the last day of class. For a graduate course, I supplement the material with additional case studies and discussion of technique.

The revised text organizes the chapters somewhat differently, which makes the teaching much easier. The exposition has been improved in a number of other ways, without (I hope) introducing new difficulties. There are many new examples and exercises.

Acknowledgements

I've taught graduate and undergraduate courses based on this material for many years at Berkeley, and on occasion at Stanford and Athens. The students in those courses were helpful and supportive. I would also like to thank Dick Berk, Máire Ní Bhrolcháin, Taylor Boas, Derek Briggs, David Collier, Persi Diaconis, Thad Dunning, Mike Finkelstein, Paul Humphreys, Jon McAuliffe, Doug Rivers, Mike Roberts, Don Ylvisaker, and PengZhao, along with several anonymous reviewers, for many useful comments. Russ Lyons and Roger Purves were virtual coauthors; David Tranah was an outstanding editor.

Table of Contents

1

Observational Studies and Experiments

1.1 Introduction

This book is about regression models and variants like path models, simultaneous-equation models, logits and probits. Regression models can be used for different purposes:

(i) to summarize data,
(ii) to predict the future,
(iii) to predict the results of interventions.

The third—causal inference—is the most interesting and the most slippery. It will be our focus. For background, this section covers some basic principles of study design.

Causal inferences are made from *observational studies*, *natural experiments*, and *randomized controlled experiments*. When using observational (non-experimental) data to make causal inferences, the key problem is *confounding*. Sometimes this problem is handled by subdividing the study population (*stratification*, also called *cross-tabulation*), and sometimes by modeling. These strategies have various strengths and weaknesses, which need to be explored.

In medicine and social science, causal inferences are most solid when based on randomized controlled experiments, where investigators assign subjects at random—by the toss of a coin—to a *treatment group* or to a *control group*. Up to random error, the coin balances the two groups with respect to all relevant factors other than treatment. Differences between the treatment group and the control group are therefore due to treatment. That is why causation is relatively easy to infer from experimental data. However, experiments tend to be expensive, and may be impossible for ethical or practical reasons. Then statisticians turn to observational studies.

In an observational study, it is the subjects who assign themselves to the different groups. The investigators just watch what happens. Studies on the effects of smoking, for instance, are necessarily observational. However, the treatment-control terminology is still used. The investigators compare smokers (the treatment group, also called the *exposed group*) with nonsmokers (the control group) to determine the effect of smoking. The jargon is a little confusing, because the word "control" has two senses:

(i) a control is a subject who did not get the treatment;
(ii) a controlled experiment is a study where the investigators decide who will be in the treatment group.

Smokers come off badly in comparison with nonsmokers. Heart attacks, lung cancer, and many other diseases are more common among smokers. There is a strong *association* between smoking and disease. If cigarettes cause disease, that explains the association: death rates are higher for smokers because cigarettes kill. Generally, association is circumstantial evidence for causation. However, the proof is incomplete. There may be some hidden confounding factor which makes people smoke and also makes them sick. If so, there is no point in quitting: that will not change the hidden factor. Association is not the same as causation.

> Confounding means a difference between the treatment and control groups—other than the treatment—which affects the response being studied.

Typically, a confounder is a third variable which is associated with exposure and influences the risk of disease.

Statisticians like Joseph Berkson and R. A. Fisher did not believe the evidence against cigarettes, and suggested possible confounding variables. Epidemiologists (including Richard Doll and Bradford Hill in England, as well as Wynder, Graham, Hammond, Horn, and Kahn in the United States) ran careful observational studies to show these alternative explanations were

not plausible. Taken together, the studies make a powerful case that smoking causes heart attacks, lung cancer, and other diseases. If you give up smoking, you will live longer.

Epidemiological studies often make comparisons separately for smaller and more homogeneous groups, assuming that within these groups, subjects have been assigned to treatment or control as if by randomization. For example, a crude comparison of death rates among smokers and nonsmokers could be misleading if smokers are disproportionately male, because men are more likely than women to have heart disease and cancer. Gender is therefore a confounder. To control for this confounder—a third use of the word "control"—epidemiologists compared male smokers to male nonsmokers, and females to females.

Age is another confounder. Older people have different smoking habits, and are more at risk for heart disease and cancer. So the comparison between smokers and nonsmokers was made separately by gender and age: for example, male smokers age 55–59 were compared to male nonsmokers in the same age group. This controls for gender and age. Air pollution would be a confounder, if air pollution causes lung cancer and smokers live in more polluted environments. To control for this confounder, epidemiologists made comparisons separately in urban, suburban, and rural areas. In the end, explanations for health effects of smoking in terms of confounders became very, very implausible.

Of course, as we control for more and more variables this way, study groups get smaller and smaller, leaving more and more room for chance effects. This is a problem with cross-tabulation as a method for dealing with confounders, and a reason for using statistical models. Furthermore, most observational studies are less compelling than the ones on smoking. The following (slightly artificial) example illustrates the problem.

Example 1. In cross-national comparisons, there is a striking correlation between the number of telephone lines per capita in a country and the death rate from breast cancer in that country. This is not because talking on the telephone causes cancer. Richer countries have more phones and higher cancer rates. The probable explanation for the excess cancer risk is that women in richer countries have fewer children. Pregnancy—especially early first pregnancy—is protective. Differences in diet and other lifestyle factors across countries may also play some role.

Randomized controlled experiments minimize the problem of confounding. That is why causal inferences from randomized controlled experiments are stronger than those from observational stud-

ies. With observational studies of causation, you always have to worry about confounding. What were the treatment and control groups? How were they different, apart from treatment? What adjustments were made to take care of the differences? Are these adjustments sensible?

The rest of this chapter will discuss examples: the HIP trial of mammography, Snow on cholera, and the causes of poverty.

1.2 The HIP trial

Breast cancer is one of the most common malignancies among women in Canada and the United States. If the cancer is detected early enough—before it spreads—chances of successful treatment are better. "Mammography" means screening women for breast cancer by X-rays. Does mammography speed up detection by enough to matter? The first large-scale randomized controlled experiment was HIP (Health Insurance Plan) in New York, followed by the Two-County study in Sweden. There were about half a dozen other trials as well. Some were negative (screening doesn't help) but most were positive. By the late 1980s, mammography had gained general acceptance.

The HIP study was done in the early 1960s. HIP was a group medical practice which had at the time some 700,000 members. Subjects in the experiment were 62,000 women age 40–64, members of HIP, who were randomized to treatment or control. "Treatment" consisted of invitation to 4 rounds of annual screening—a clinical exam and mammography. The control group continued to receive usual health care. Results from the first 5 years of followup are shown in table 1. In the treatment group, about 2/3 of the women accepted the invitation to be screened, and 1/3 refused. Death rates (per 1000 women) are shown, so groups of different sizes can be compared.

Table 1. HIP data. Group sizes (rounded), deaths in 5 years of followup, and death rates per 1000 women randomized.

	Group size	Breast cancer		All other	
		No.	Rate	No.	Rate
Treatment					
Screened	20,200	23	1.1	428	21
Refused	10,800	16	1.5	409	38
Total	31,000	39	1.3	837	27
Control	31,000	63	2.0	879	28

Which rates show the efficacy of treatment? It seems natural to compare those who accepted screening to those who refused. However, this is an observational comparison, even though it occurs in the middle of an experiment. The investigators decided which subjects would be invited to screening, but it is the subjects themselves who decided whether or not to accept the invitation. Richer and better-educated subjects were more likely to participate than those who were poorer and less well educated. Furthermore, breast cancer (unlike most other diseases) hits the rich harder than the poor. Social status is therefore a confounder—a factor associated with the outcome and with the decision to accept screening.

The tip-off is the death rate from other causes (not breast cancer) in the last column of table 1. There is a big difference between those who accept screening and those who refuse. The refusers have almost double the risk of those who accept. There must be other differences between those who accept screening and those who refuse, in order to account for the doubling in the risk of death from other causes—because screening has no effect on the risk.

One major difference is social status. It is the richer women who come in for screening. Richer women are less vulnerable to other diseases but more vulnerable to breast cancer. So the comparison of those who accept screening with those who refuse is biased, and the bias is against screening.

Comparing the death rate from breast cancer among those who accept screening and those who refuse is *analysis by treatment received*. This analysis is seriously biased, as we have just seen. The experimental comparison is between the whole treatment group—all those invited to be screened, whether or not they accepted screening—and the whole control group. This is the *intention-to-treat analysis*.

Intention-to-treat is the recommended analysis.

HIP, which was a very well-run study, made the intention-to-treat analysis. The investigators compared the breast cancer death rate in the total treatment group to the rate in the control group, and showed that screening works.

The effect of the invitation is small in absolute terms: $63 - 39 = 24$ lives saved (table 1). Since the absolute risk from breast cancer is small, no intervention can have a large effect in absolute terms. On the other hand, in relative terms, the 5-year death rates from breast cancer are in the ratio $39/63 = 62\%$. Followup continued for 18 years, and the savings in lives persisted over that period. The Two-County study—a huge randomized controlled experiment in Sweden—confirmed the results of HIP. So did other studies in Finland, Scotland, and Sweden. That is why mammography became so widely accepted.