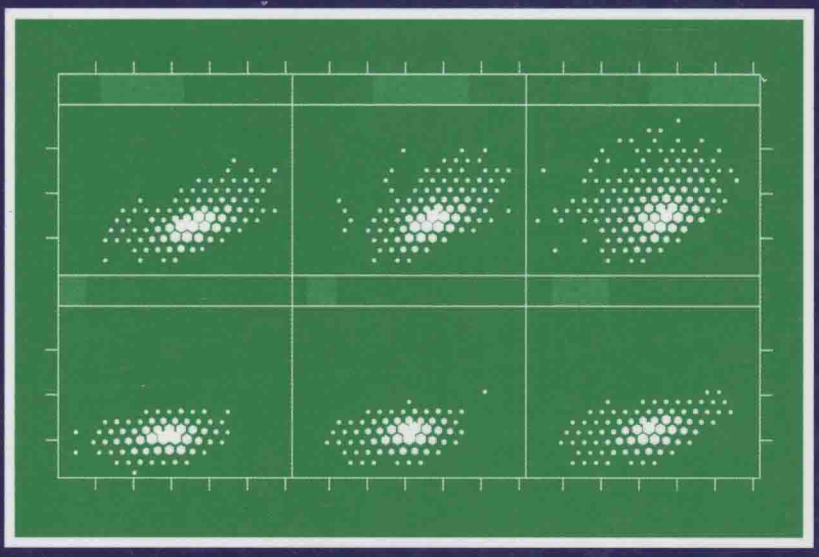
WILEY SERIES IN SURVEY METHODOLOGY

Complex Surveys

A Guide to Analysis Using R



Thomas Lumley

WILEY

WWW. LINK AVAILABLE

Complex Surveys

A Guide to Analysis Using R

Thomas Lumley
University of Washington
Department of Biostatistics
Seattle, Washington

WILEY

A John Wiley & Sons, Inc., Publication

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

Library of Congress Cataloging-in-Publication Data:

Lumley, Thomas, 1969–
Complex surveys : a guide to analysis using R / Thomas Lumley.
p. cm.
Includes bibliographical references and index.
ISBN 978-0-470-28430-8 (pbk.)
1. Mathematical statistics—Data processing. 2. R (Computer program language) I. Title.
QA276.45.R3L86 2010
515.0285—dc22 2009033999

Printed in the United States of America.

10 9 8 7 6 5 4 3

WILEY SERIES IN SURVEY METHODOLOGY

Established in Part by WALTER A. SHEWHART AND SAMUEL S. WILKS

A complete list of the titles in this series appears at the end of this volume.

Acknowledgments

Most of this book was written while I was on sabbatical at the University of Auckland and the University of Leiden. The Statistics department in Auckland and the Department of Clinical Epidemiology at Leiden University Medical Center were very hospitable and provided many interesting and productive distractions from writing.

I had useful discussions on a number of points with Alastair Scott and Chris Wild. Bruce Psaty, Stas Kolenikov, and Barbara McKnight gave detailed and helpful comments on a draft of the text. The "'s" interpretation of the $ operator came from Ken Rice. Hadley Wickham explained how to combine city and state data in a single map. Paul Murrell made some suggestions about types of graphics to include. The taxonomy of regression predictor variables is from Scott Emerson. I learned about some of the references on reification from Cosma Shalizi's web page. The students and instructors in STAT/CSSS 529 (Seattle) and STATS 740 (Auckland) tried out a draft of the book and pointed out a few problems that I hope have been corrected.

Some financial support for my visit to Auckland was provided by Alastair Scott, Chris Wild, and Alan Lee from a grant from the Marsden Fund, and my visit to Leiden was supported in part by Fondation Leducq through their funding of the LINAT collaboration. My sabbatical was also supported by the University of Washington.

The survey package has benefited greatly from comments, questions, and bug reports from its users, an attempt at a list is in the `THANKS` file in the package.

Preface

This book presents a practical guide to analyzing complex surveys using R, with occasional digressions into related areas of statistics. Complex survey analysis differs from most of statistics philosophically and in the substantive problems it faces. In the past this led to a requirement for specialized software and the spread of specialized jargon, and survey analysis became separated from the rest of statistics in many ways. In recent years there has been a convergence of ways. All major statistical packages now include at least some survey analysis features, and some of the mathematical techniques of survey analysis have been incorporated in widely-used statistical methods for missing data and for causal inference.

More importantly for this book, researchers in the social science and health sciences are increasingly interested in using data from complex surveys to conduct the same sorts of analyses that they traditionally conduct with more straightforward data. Medical researchers are also increasingly aware of the advantages of well-designed subsamples when measuring novel, expensive variables on an existing cohort.

This book is designed for readers who have some experience with applied statistics, especially in the social sciences or health sciences, and are interested in learning about survey analysis. As a result, we will spend more time on graphics, regression modelling, and two-phase designs than is typical for a survey analysis text. I have presented most of the material in this book in a one-quarter course for graduate students who are not specialist statisticians but have had a graduate-level introductory

course in applied statistics, including linear and logistic regression. Chapters 1–6 should be of general interest to anyone wishing to analyze complex surveys. Chapters 7–10 are, on average, more technical and more specialized than the earlier material, and some of the content, particularly in Chapter 8, reflects recent research.

The widespread availability of software for analyzing complex surveys means that it is no longer as important for most researchers to learn a list of computationally convenient special cases of formulas for means and standard errors. Formulas will be presented in the text only when I feel they are useful for understanding concepts; the appendices present some additional mathematical and computational descriptions that will help in comparing results from different software systems. An excellent reference for statisticians who want more detail is *Model Assisted Survey Sampling* by Särndal, Swensson, and Wretman [151]. Some of the exercises presented at the end of each chapter require more mathematical or programming background, these are indicated with a $\star$. They are not necessarily more difficult than the unstarred exercises.

This book is designed around a particular software system: the survey package for the R statistical environment, and one of its main goals is to document and explain this system. All the examples, tables, and graphs in the book are produced with R, and code and data for you to reproduce nearly all of them is available. There are three reasons for choosing to emphasize R in this way: it is open-source software, which makes it easily available; it is very widely known and used by academic statisticians, making it convenient for teaching; and because I designed the survey package it emphasizes the areas I think are most important and readily automated about design-based inference. For other software for analyzing complex surveys, see the comprehensive list maintained by Alan Zaslavsky at `http://www.hcp.med.harvard.edu/statistics/survey-soft/`.

There are important statistical issues in the design and analysis of complex surveys outside design-based inference that I give little or no attention to. Small area estimation and item response theory are based on very different areas of statistics, and I think are best addressed under spatial statistics and multivariate statistics, respectively. Statistics has relatively little positive to say about about non-sampling error, although I do discuss raking, calibration, and the analysis of multiply-imputed data. There are also interesting but specialized areas of complex sampling that are not covered in the book (or the software), mostly because I lack experience with their application. These include adaptive sampling techniques, and methods from ecology such as line and quadrat sampling.

Code for reproducing the examples in this book (when not in the book itself), errata, and other information, can be found from the web site: `http://faculty.washington.edu/tlumley/svybook`. If you find mistakes or infelicities in the book or the package I would welcome an email: `tlumley@u.washington.edu`.

Acronyms

N	Population size
N_k	Population size for stratum k
n	Sample size
n_k	Sample size for stratum k
π_i	Sampling probability for unit i
$\check{X}_i$	Weighted observation X_i/π_i
$\hat{T}$	Estimate of T
IPW	Inverse-Probability Weighted
IPTW	Inverse-Probability of Treatment Weighted
fpc	Finite-population correction (to standard errors)
$E[\,]$	Expected value
Pr[ ]	Probability
$\mathbb{I}$	An influence function

CONTENTS

CHAPTER 1

BASIC TOOLS

In which we meet the probability sample and the R language.

1.1 GOALS OF INFERENCE

1.1.1 Population or process?

The mathematical development for most of statistics is *model-based*, and relies on specifying a probability model for the random process that generates the data. This can be a simple parametric model, such as a Normal distribution, or a complicated model incorporating many variables and allowing for dependence between observations. To the extent that the model represents the process that generated the data, it is possible to draw conclusions that can be generalized to other situations where the same process operates. As the model can only ever be an approximation, it is important (but often difficult) to know what sort of departures from the model will invalidate the analysis.

Complex Surveys: A Guide to Analysis Using R. By Thomas Lumley

The analysis of complex survey samples, in contrast, is usually *design-based*. The researcher specifies a population, whose data values are unknown but are regarded as fixed, not random. The observed sample is random because it depends on the random selection of individuals from this fixed population. The random selection procedure of individuals (the *sample design*) is under the control of the researcher, so all the probabilities involved can, in principle, be known precisely. The goal of the analysis is to estimate features of the fixed population, and design-based inference does not support generalizing the findings to other populations.

In some situations there is a clear distinction between population and process inference. The Bureau of Labor Statistics can analyze data from a sample of the US population to find out the distribution of income in men and women in the US. The use of statistical estimation here is precisely to generalize from a sample to the population from which it was taken.

The University of Washington can analyze data on its faculty salaries to provide evidence in a court case alleging gender discrimination. As the university's data are complete there is no uncertainty about the distribution of salaries in men and women in this population. Statistical modelling is needed to decide whether the differences in salaries can be attributed to valid causes, in particular to differences in seniority, to changes over time in state funding, and to area of study. These are questions about the process that led to the salaries being the way they are.

In more complex analyses there can be something of a compromise between these goals of inference. A regression model fitted to blood pressure data measured on a sample from the US population will provide design-based conclusions about associations in the US population. Sometimes these design-based conclusions are exactly what is required, e.g., there is more hypertension in blacks than in whites. Often the goal is to find out why some people have high blood pressure: is the racial difference due to diet, or stress, or access to medical care, or might there be a genetic component?

1.1.2 Probability samples

The fundamental statistical concept in design-based inference is the *probability sample* or *random sample*. In everyday speech, "taking a random sample" of 1000 individuals means a sampling procedure when any subset of 1000 people from the population is equally likely to be selected. The technical term for this is a "simple random sample". The Law of Large Numbers implies that the sample of 1000 people is likely to be representative of the population, according to essentially any criteria we are interested in. If we compute the mean age, or the median income, or the proportion of registered Republican voters in the sample, the answer is likely to be close to the value for the population.

We could also end up with a sample of 1000 individuals from the US population, for example, by taking a simple random sample of 20 people from each state. On many criteria this sample is unlikely to be representative, because people from states with low populations are more likely to be sampled. Residents of these states have a similar age distribution to the country as a whole but tend to have lower incomes and

be more politically conservative. As a result the mean age of the sample will be close to the mean age for the US population, but the median income is likely to be lower, and the proportion of registered Republican voters higher than for the US population. As long as we know the population of each state, this *stratified random sample* is still a probability sample. Yet another approach would be to choose a simple random sample of 50 counties from the US and then sample 20 people from each county. This sample would over-represent counties with low populations, which tend to be in rural areas. Even so, if we know all the counties in the US, and if we can find the number of households in the counties we choose, this is also a probability sample.

It is important to remember that what makes a *probability sample* is the procedure for taking samples from a population, not just the data we happen to end up with.

The properties we need of a sampling method for design-based inference are as follows:

1. Every individual in the population must have a non-zero probability of ending up in the sample (written π_i for individual i)
2. The probability π_i must be known for every individual who does end up in the sample.
3. Every pair of individuals in the sample must have a non-zero probability of both ending up in the sample (written π_{ij} for the pair of individuals (i, j)).
4. The probability π_{ij} must be known for every pair that does end up in the sample.

The first two properties are necessary in order to get valid population estimates; the last two are necessary to work out the accuracy of the estimates. If individuals were sampled independently of each other the first two properties would guarantee the last two, since then $\pi_{ij} = \pi_i \pi_j$, but a design that sampled one random person from each US county would have $\pi_i > 0$ for everyone in the US and $\pi_{ij} = 0$ for two people in the same county. In the survey package, as in most software for analysis of complex samples, the computer will work out π_{ij} from the design description, they do not need to be specified explicitly.

The world is imperfect in many ways, and the necessary properties are present only as approximations in real surveys. A list of residences for sampling will include some that are not inhabited and miss some that have been newly constructed. Some people (me, for example) do not have a landline telephone, others may not be at home or may refuse to answer some or all of the questions. We will initially ignore these problems, but aspects of them are addressed in Chapters 7 and 9.

1.1.3 Sampling weights

If we take a simple random sample of 3500 people from California (with total population 35 million) then any person in California has a $1/10000$ chance of being sampled, so $\pi_i = 3500/3500000 = 1/10000$ for every i. Each of the people we sample represents 10000 Californians. If it turns out that 400 of our sample have high