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 specialised software and the spread of specialised 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 do 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[144]. 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 that the unstarred exercises.

This book is designed around a particular software system: the survey package for the R statistical environment, and one of its 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 of the survey package: `http://faculty.washington.edu/tlumley/survey`. If you find mistakes or infelicities in the book or the package I would welcome an email: `tlumley@u.washington.edu`.