Richard Valliant
Jill A. Dever
Frauke Kreuter

Practical Tools for Designing and Weighting Survey Samples

Springer

Statistics for Social and Behavioral Sciences

Advisors:

S.E. Fienberg
W.J. van der Linden

Richard Valliant • Jill A. Dever • Frauke Kreuter

Practical Tools for Designing and Weighting Survey Samples

Springer

Richard Valliant
University of Michigan
Ann Arbor, MI, USA

Jill A. Dever
RTI International
Washington, DC, USA

Frauke Kreuter
University of Maryland
College Park, MD, USA

Printed on acid-free paper

To Carla and Joanna
  Vince, Mark, and Steph
  Gerit and Konrad

Preface

Survey sampling is fundamentally an applied field. Even though there have been many theoretical advances in sampling in the last 40 or so years, the theory would be pointless in isolation. The reason to develop the theory was to solve real-world problems. Although the mathematics behind the procedures may seem, to many, to be impenetrable, you do not have to be a professional mathematician to successfully use the techniques that have been developed. Our goal in this book is to put an array of tools at the fingertips of practitioners by explaining approaches long used by survey statisticians, illustrating how existing software can be used to solve survey problems and developing some specialized software where needed. We hope this book serves at least three audiences:

(1) *Students* seeking a more in-depth understanding of applied sampling either through a second semester-long course or by way of a supplementary reference
(2) *Survey statisticians* searching for practical guidance on how to apply concepts learned in theoretical or applied sampling courses
(3) *Social scientists* and *other survey practitioners* who desire insight into the statistical thinking and steps taken to design, select, and weight random survey samples

Some basic knowledge of random sampling methods (e.g., single- and multistage random sampling, the difference between with- and without-replacement sampling, base weights calculated as the inverse of the sample inclusion probabilities, concepts behind sampling error, and hypothesis testing) is required. The more familiar these terms and techniques are, the easier it will be for the reader to follow. We first address the student perspective.

A familiar complaint that students have after finishing a class in applied sampling or in sampling theory is: "I still don't really understand how to design a sample." Students learn a lot of isolated tools or techniques but do not have the ability to put them all together to design a sample from start to

finish. One of the main goals of this book is to give students (and practitioners) a taste of what is involved in designing single- and multistage samples in the real world. This includes devising a sampling plan from sometimes incomplete information, deciding on a sample size given a specified budget and estimated response rates, creating strata from a choice of variables, allocating the sample to the strata given a set of constraints and requirements for detectable differences, and determining sample sizes to use at different stages in a multistage sample. When appropriate, general rules of thumb will be given to assist in completing the task.

Students will find that a course taught from this book will be a combination of hands-on applications and general review of the theory and methods behind different approaches to sampling and weighting. Detailed examples will enable the completion of exercises at the end of the chapters. Several small, but realistic projects are included in several chapters. We recommend that students complete these by working together in teams to give a taste of how projects are carried out in survey organizations.

For survey statisticians, the book is meant to give some practical experience in applying the theoretical ideas learned in previous courses in balance with the experience already gained by working in the field. Consequently, the emphasis here is on learning how to employ the methods rather than on learning all the details of the theory behind them. Nonetheless, we do not view this as just a high-level cookbook. Enough of the theoretical assumptions are reviewed so that a reader can apply the methods intelligently. Additional references are provided for those wishing more detail or those needing a refresher. Several survey data sets are used to illustrate how to design samples, to make estimates from complex surveys for use in optimizing the sample allocation, and to calculate weights. These data sets are available through a host web site discussed below and in the R package `PracTools` so that the reader may replicate the examples or perform further analyses.

This book will also serve as a useful reference for other professionals engaged in the conduct of sample surveys. The book is organized into four parts. The first three parts—*Designing Single-Stage Sample Surveys*, *Multistage Designs*, and *Survey Weights and Analyses*—begin with a description of a realistic survey project. General tools and some specific examples in the intermediate chapters of the part help to address the interim tasks required to complete the project. With these chapters, it will become apparent that the process toward a solution to a sample design, a weighting methodology, or an analysis plan takes time and input from all members of the project team. Each part of the book concludes with a chapter containing a solution to the project. Note that we say "a solution" instead of "the solution" since survey sampling can be approached in many artful but correct ways.

The book contains a discussion of many standard themes covered in other sources but from a slightly different perspective as noted above. We also cover several interesting topics that either are not included or are dealt with in a limited way in other texts. These areas include:

- Sample size computations for multistage designs
- Power calculations as related to surveys
- Mathematical programming for sample allocation in a multi-criteria optimization setting
- Nuts and bolts of area probability sampling
- Multiphase designs
- Quality control of survey operations
- Statistical software for survey sampling and estimation

Multiphase designs and quality control procedures comprise the final part of the book—*Other Topics*. Unlike the other areas listed above, aspects related to statistical software are used throughout the chapters to demonstrate various techniques.

Experience with a variety of statistical software packages is essential these days to being a good statistician. The systems that we emphasize are:

- R® (R Core Team 2012; Crawley 2007)
- SAS®[1]
- Microsoft Excel®[2] and its add-on Solver®[3]
- Stata®[4]
- SUDAAN®[5]

There are many other options currently available, but we must limit our scope. Other software is likely to be developed in the near term, so we encourage survey practitioners to keep their eyes open.

R, a free implementation of the S language, receives by far the most attention in this book. We assume some knowledge of R and have included basic information plus references in Appendix C for those less familiar. The book and the associated R package, `PracTools`, contain a number of specialized functions for sample size and other calculations and provide a nice complement to the base package downloaded from the main R web site, www.r-project.org. The package `PracTools` also includes data sets used in the book. In addition to PracTools, the data sets and the R functions developed for the book are available individually through the book's web site hosted by the Joint Program in Survey Methodology (JPSM) located at www.jpsm.org, from the Faculty page. Unless otherwise specified, any R function referred to in the text is located in the `PracTools` package.

Despite the length of this book, we have not covered everything that a practitioner should know. An obvious omission is what to do about missing data. There are whole books on that subject that some readers may find

---

[1] www.sas.com.

[2] office.microsoft.com.

[3] www.solver.com.

[4] stata.com.

[5] www.rti.org/sudaan.

useful. Another topic is dual or multiple frame sampling. Dual frames can be especially useful when sampling rare populations if a list of units likely to be in the rare group can be found. The list can supplement a frame that gives more nearly complete coverage of the group but requires extensive screening to reach member of the rare group.

At this writing, we have collectively been in survey research for more years than we care to count (or divulge). This field has provided interesting puzzles to solve, new perspectives on the substantive research within various studies, and an ever growing network of enthusiastic collaborators of all flavors. Regardless from which of the three perspectives you approach this book, we hope that you find the material presented here to be enlightening or even empowering as your career advances. Now let the fun begin . . .

<table>
<tr><td>Ann Arbor, MI</td><td>Richard Valliant</td></tr>
<tr><td>Washington, DC</td><td>Jill A. Dever</td></tr>
<tr><td>College Park, MD</td><td>Frauke Kreuter</td></tr>
<tr><td></td><td>October 2012</td></tr>
</table>

Acknowledgments

We are also extremely grateful to Robert Pietsch who created the TeX files, Florian Winkler who programmed the `PracTools` package in R, Valerie Tutz who helped put together the bibliography, Melissa Stringfellow who checked many of the exercises, and Barbara Felderer who helped check the R package. There were also many students and colleagues (unnamed here) who contributed to improving the presentation with their many questions and criticisms.

Contents

List of Figures

Chapter 1
An Overview of Sample Design and Weighting

This is a practical book. Many techniques used by survey practitioners are not covered by standard textbooks but are necessary to do a professional job when designing samples and preparing data for analyses. In this book, we present a collection of methods that we have found most useful in our own practical work. Since computer software is essential in applying the techniques, example code is given throughout.

We assume that most readers will be familiar with the various factors that affect basic survey design decisions. For those we recommend skipping the next section and reading through the chapter guide (Sect. 1.2) instead. For all others, Sect. 1.1 will provide a very brief background on where sample design and weighting fits into the large task of designing a survey. Some terminology and associated notation is defined here that will come in handy throughout the book. The glossary in Appendix A is a more complete list of the notation used throughout the book. Some topics, like multistage sampling, require fairly elaborate (and difficult to remember) notation. The Notation Glossary will be a useful reference for most readers.

1.1 Background and Terminology

Choosing a sample design for a survey requires consideration of a number of factors. Among them are (1) specifying the objective(s) of the study; (2) translating a subject-matter problem into a survey problem; (3) specifying the target population, units of analysis, key study variables, auxiliary variables (i.e., covariates related to study variables and for which population statistics may be available), and population parameters to be estimated; (4) determining what sampling frame(s) are available for selecting units; and (5) selecting an appropriate method of data collection. Based on these considerations,

(6) a sample can be designed and selected. Across all these steps trade-off decisions are to be made as a function of budget and time constraints for completing the work.

Introductory books such as *Survey Methodology* (Groves et al. 2004) or *Introduction to Survey Quality* (Biemer and Lyberg 2003) cover these issues nicely and are strongly recommended as supplements to the material presented in this book. A primary focus here is the sixth step and thus we will only briefly comment on the other five to the extent that they are necessary to understand our discussion of sample design, selection, and weighting.

(1) The *study objectives* may be stated very generally, in which case it is the responsibility of the survey researcher to help the sponsor (i.e., client) to specify some measurable goals. Although it seems obvious that no one would undertake data collection without some well-planned intentions, this is often not the case. Part of the craft of survey design is to translate a subject-matter problem into a survey problem. This may entail turning vague ideas like the following into specific measurements: "measure the attitudes of the employees of a company"; "determine how healthy a particular demographic group is, such as persons with a household income below the poverty line"; "decide how well a local school system is serving its students." Some objectives are very broad and very difficult to operationalize. For example, measuring price changes in all sectors of a nation's economy is a goal of most Western governments. Consumer, producer, and import/export price indices are usually the vehicles for doing this. Economic theory for a cost of living index (COLI) is formulated at the level of a single consumer. On the other hand, a price index is meant to apply to a large group of consumers. The translation of the subject-matter problem into a survey problem requires deciding which of some alternative price indices best approximates the COLI. The study objective will affect all other aspects of the survey design.

(2) No matter if one faces a simple or complex objective, to determine what type of sample and associated sample size are adequate to achieve the objective, the theoretical concepts under study must be translated into *constructs* that can be measured through a survey, and the goals themselves must be quantified in some way.

An example of an economic objective is to estimate the unemployment rate. This is often done via a household survey like the Current Population Survey (CPS)[1] in the US or the Labour Force Survey (LFS)[2] in Canada. Measuring the unemployment rate requires defining constructs such as what it means to be in the labor force, i.e., have a job or want a job, and what it means to be employed, to be looking for a job if you do not already have one, and whether doing unpaid work in a family business constitutes having a

---

[1] http://www.census.gov/cps/.

[2] http://www.statcan.gc.ca/.

job. Often compromises need to be made between concepts and specific items that can be collected. For example, the following question is taken from the US National Health Interview Survey (NHIS)[3] survey instrument:

> Have you EVER been told by a doctor or other health professional that you had coronary heart disease?

Since a respondent's understanding of his/her own health problems can be faulty, the more valid method might be to ask a doctor directly whether the respondent has heart disease. But, asking the respondent seems to be a compromise intended to reduce costs.

Once the key measurements have been identified, statistical goals can be set. The goals are usually stated in terms of measures of precision. Precision estimates include standard errors (SEs) or relative standard errors, defined as the SE of an estimator divided by the population parameter that is being estimated. A relative standard error of an estimator is also called a coefficient of variation (CV). We use the term CV throughout the book. A precision target might be to estimate the proportion of adults with coronary heart disease with a CV of 0.05, i.e., the standard error of the estimated proportion is 5 % of the proportion itself. These targets may be set for many different variables.

(3) Specifying a *target population* also requires some thought. A target population is the set of units for which measurements can be obtained and may differ from the (inferential) population for which scientific inferences are actually desired. For instance, in doing a survey to measure the relationship between smoking and health problems, health researchers are interested in relationships that exist generally and not just in the particular year of data collection. The *analytic units* (or units of observation) are the members of the target population that are subjected to the survey measurements. Additionally, the study may specify the analysis of units that have particular characteristics, known as the *eligibility criteria*. For example, a survey of prenatal care methods may include only females of age 18 to 50, and a study to estimate rates of sickle-cell anemia in the US may only include African Americans.

(4) Rarely is there a one-to-one match between target populations and *sampling frames* available to researchers. If a frame exists with contact information such as home or email addresses, then it may be relatively quick and cheap to select a sample and distribute hard-copy or electronic surveys. Such frames usually exist for members of a professional association, employees of a company, military personnel, and inhabitants of those Scandinavian countries with total population registries. Depending on the survey sponsor, these frames may or may not be available for sampling. In the absence of readily available sampling frames, area probability samples are often used. Those take some time to design and select (unless an existing sample or address list frame can be used).

---

[3] http://www.cdc.gov/nchs/nhis.htm.

At this writing, a fairly new sampling frame exists in the USA that is based on the US Postal Service (USPS) Delivery Sequence File (DSF) (Iannacchione et al. 2003; Iannacchione 2011; Link et al. 2008). The DSF is a computerized file that contains nearly all delivery point addresses serviced by the USPS. Some researchers use the DSF as a replacement for random digit dialed (RDD) telephone surveys or as an adjunct to field listings collected in area samples (see below). Commercial vendors of survey samples sell "enhanced" versions of the DSF that, for many addresses, may include a landline telephone number, a name associated with the address, Spanish surname indicator, estimated age of the head of household, as well as some geocoded (i.e., latitude and longitude) and census tract information. If accurate, these items can improve the efficiency of a sample by allowing the targeting of different groups.

(5) One of the critical decisions that must be made and has a direct bearing on sample design is the *method of data collection*. The method of data collection is chosen by weighing factors such as budget, time schedule, type of data collected, frame availability, feasibility of using the method with members of the target population, and expected outcome rates (e.g., contact and response rates) for different methods. Collection of blood specimens in addition to questionnaire responses might suggest an in-person interview with a field interviewer accompanied by or also trained as a phlebotomist. A study of high school students may, for example, include data collection through the Web in a classroom setting. Collecting data through a self-administered (hard-copy) questionnaire, however, would not be practical for an illiterate population. Today many surveys consider the use of multiple modes to find the right balance between cost, timeliness, and data quality.

If personal interviews are required or when no nationwide sampling frames are available, clustered area sampling may be necessary. Clustering allows interviewers to be recruited for a limited number of areas and helps control the amount of travel required to do address listing or interviewing. Clustering of a sample, as in multistage sampling, typically will lead to larger variances for a given sample size compared to an unclustered sample. Two measures that are simple, but extremely useful to express the effect of clustering on survey estimates, are the *design effect* and the *effective sample size* introduced by Kish (1965). We define them here and will use them repeatedly in the coming chapters:

- *Design effect* (*deff*)—the ratio of the variance of an estimator under a complex design to the variance that would have been obtained from a simple random sample (*srs*) of the same number of units. Symbolically, $deff\left(\hat{\theta}\right) = \frac{V(\hat{\theta})}{V_{srs}(\hat{\theta})}$ where $\hat{\theta}$ is an estimator of some parameter, V denotes variance under whatever sample design is used (stratified simple random sample, two-stage cluster sample, etc.), and V_{srs} is the *srs* variance of the *srs* estimator of the same parameter. Generally an *srs* selected with replacement (*srswr*) is used for the denominator calculation. The sample

size for V_{srs} is the same as the sample size of units used in the numerator
estimate.
- *Effective sample size* (n_{eff})—the number of units in the sample divided by
  the *deff*. This is the sample size for an *srswr* that yields the same variance
  for an estimate as the variance obtained from the sample design used to
  collect the data.

As apparent from the definition, the *deff* is specific to a particular esti-
mator, like a mean, total, quantile, or something else. People often have
averages in mind when they use *deff*'s, but the idea can be applied more
generally. Usually, the variance in the denominator of a *deff* is for simple ran-
dom sampling *with replacement*, although without replacement could be used.
Which to use is mostly a matter of personal preference. However, since the
values of the with- and without-replacement variances can be quite different
when the sampling fraction is large, it is important to know which is used
in the denominator of any *deff* that you are supplied. The *deff* and n_{eff} are
especially handy when computing total sample sizes for clustered samples.
However, often good estimates of *deff* and n_{eff} can be hard to come by and
are likely to vary by survey item.

(6) With a method of data collection in mind and knowledge of the avail-
able sampling frames, the survey researcher next determines the appropriate
type of *random sampling (mechanism) design*. The general designs that we
consider in our text can be categorized as one of these three:

- *Stratified single-stage designs*—units of observation are selected directly
  from a sampling frame, sometimes referred to as a list frame, containing
  data such as contact or location information and stratification variables.
- *Stratified multistage designs*—units are selected from lists constructed
  "on-site" for aggregate units from a previous design stage (e.g., actively
  enrolled students within schools).
- *Stratified multiphase designs*—a primary sample of units is selected from
  the designated frame (phase one), and samples of phase-one units are
  selected in subsequent phases using information obtained on the units in
  phase one (e.g., a study where a subsample of nonrespondents is recon-
  tacted using a different mode of data collection, or a study of individuals
  who are classified as having some condition based on tests administered in
  a previous phase of the design).

Each of the three general designs above usually involves probability sam-
pling. Särndal et al. (1992, Sect. 1.3) give a formal definition of a probability
sample, which we paraphrase here. A probability sample from a particular
finite population is one that satisfies four requirements:

- (*i*) A set of samples can be defined that are possible to obtain with the
  sampling procedure.
- (*ii*) Each possible sample s has a known probability of selection, $p(s)$.

(*iii*) Every element in the target population has a nonzero probability of selection.

(*iv*) One set of sample elements is selected with the probability associated with the set. Weights for sample elements can be computed that are intended to project the sample to the target population.

However, a survey designer often loses control over which set of elements actually provide data because of nonresponse and other sample losses. There are also samples that are not probability samples, even initially. For example, persons who volunteer to be part of an Internet survey panel do not constitute a sample selected with known probabilities. Inferences from such samples may be possible if the non-probability sample can be linked to the nonsample part of the population via a model.

The decision on whether to use a single or multistage design is in part a function of the available sampling frame. Two general types of sampling frames are available for unit selection—*direct* and *indirect*. Sampling frames containing a list of *the* units of observation are referred to as a direct list frame. Single-stage designs are facilitated by these frames. Indirect frames, however, allow initial access only to groups of units. With a multistage design, units are selected from within the groups, often referred to as clusters. For example, in a survey of households, a common practice is to first select a sample of geographic areas, called primary sampling units (PSUs). Within the sample PSUs, households may be selected from (*i*) lists compiled by research personnel (called listers) who canvass the area (in a process known as *counting and listing*) or (*ii*) lists maintained by organizations such as the USPS.

If no list of eligible units is available for a target population, some type of screening process is necessary. Screening for households with children under the age of three could be done by calling a sample of landline telephone numbers and administering screening questions to determine if the household is eligible (i.e., contains at least one child less than three years of age). This method is often used but suffers from several problems. One is the fact not all eligible households have landline telephones and would thus be missed through the screening process. Until recently, cell phones were usually not included in most US telephone surveys. Another problem is associated with the large number of phone numbers required to screen for a rare subpopulation. An example of how onerous the screening process can be is provided by the National Immunization Survey (NIS).[4] The goal of the NIS is to estimate the proportions of children 19–35 months old in the USA who have had the recommended vaccinations for childhood diseases like diphtheria, pertussis, poliovirus, measles, and hepatitis. In 2002, 2.06 million telephone numbers were called. Of those, 1.02 million were successfully screened to determine whether they had an age-eligible child. About 34,000 households were identified as having one or more in-scope children—an eligibility rate of 3.4 % among those households successfully screened (Smith et al. 2005).

---

[4] http://www.cdc.gov/nis/.

Ideally, the sample frame covers the entire target population. A telephone sample that only covers landlines clearly falls short of that goal, but there are other more subtle reasons for coverage errors too. In principle, an area sample that uses all of the land area in-scope of the survey should have 100 % coverage. However, this does not pan out in practice. Kostanich and Dippo (2002, Chap.16) give some estimates of proportions of different demographic groups that are covered by the CPS. In the 2002 CPS, young Black and Hispanic males had coverage rates of 70–80 %, using demographic projections from the 2000 Decennial Census as reference points (U.S. Census Bureau 2002). The reasons for this undercoverage are speculative but may include the possibility that some of these young people do not have permanent addresses or that other household members do not want to divulge who lives at the sample address (Tourangeau et al. 2012). In urban areas, it may also be difficult to identify all the households due to peculiar apartment building configurations, inability to gain entry to buildings with security protection, or other reasons.

In the case of a commercial buildings survey, there is some ambiguity about what constitutes a business, especially in small family-owned businesses, leading to uncertainty about whether a building is "commercial" or not. As a result, listers may skip some buildings that should be in-scope based on the survey definitions (Eckman and Kreuter 2011).

As is evident from the preceding discussion, many frames and the samples selected from them will imperfectly cover their target populations. A frame may contain ineligible units, and eligible units may not be reliably covered by the frame or the sample. In some applications, the best sample design practices will not correct these problems, but there are weighting techniques that will reduce them. All of these issues are covered in later chapters, as described in the next section.

1.2 Chapter Guide

The book is divided into four parts: I: Designing Single-Stage Sample Surveys (Chaps. 2–7), II: Multistage Designs (Chaps. 8–11), III: Survey Weights and Analyses (Chaps. 12–16), IV: Other Topics. Parts I–III begin with descriptions of example projects similar to ones encountered in practice. After introducing each project, we present the tools in the succeeding chapters for accomplishing the work. The last chapter in Parts I–III (Chaps. 7, 11, and 16) provides one way of meeting the goals of the example project. Something that any reader should appreciate after working through these projects is that solutions are not unique. There are likely to be many ways of designing a sample and creating weights that will, at least approximately, achieve the stated goals. This lack of uniqueness is one of many things that separate the lifeless homework problems in a math book from real-world applications.

Practitioners need to be comfortable with the solutions they propose. They need to be able to defend decisions made along the way and to understand the consequences that alternative design decisions would have. This book will prepare you for such tasks.

Part I addresses techniques that are valuable in designing single-stage samples. Chapter 2 presents a straightforward project to design a personnel survey. The subsequent chapters concentrate on methods for determining the sample size and allocating it among different groups in the population. Chapter 3 presents a variety of ways of calculating a sample size to meet stated *precision goals* for estimates for the full population. Chapter 4 covers various methods of computing sample sizes based on *power requirements*. Using power as a criterion for sample size calculation is more common in epidemiological applications. Here the goal is to find a sample size that will detect with high probability some prespecified difference in means, proportions, etc., between some subgroups or between groups at two different time periods.

Chapters 3 and 4 focus on sample size decisions made based on optimizing precision or power for *one single* variable at a time. For surveys with a very specific purpose, considering a single variable is realistic. However, many surveys are multipurpose. Not one, but several key variables are collected across a variety of subgroups in the population. For example, in health surveys, questions are asked on a variety of diseases and differences between racial or socioeconomic groups are of substantive interest. In such surveys analysts may use data in ways that were not anticipated by the survey designers. In fact, many large government-sponsored surveys amass an array of variables to give analysts the freedom to explore relationships and build models. To meet multiple goals and respect cost constraints, the methods in Chaps. 3 and 4 could be applied by trial and error in the hopes of finding an acceptable solution. A better approach is to use mathematical programming techniques that allow optimization across *multiple* variables.

Chapter 5 therefore presents some *multicriteria programming methods* that can be used to solve these more complicated problems. Operations researchers and management scientists have long used these algorithms, but they appear to be less well known among survey designers. These algorithms allow more realistic treatment of complicated allocation problems involving multiple response variables and constraints on costs, precision, and sample sizes for subgroups. Without these methods, sample allocation is a hit-or-miss proposition that may be suboptimal in a number of ways. In decades past, specialized, expensive software had to be purchased to solve optimization problems. However, software is now readily available to solve quite complicated allocation problems. Even under the best circumstance not every person, business, or other unit sampled in a survey will respond in the end. As discussed in Chap. 6, adjustments need to be made to the initial sample size to account for these losses.

Some samples need to be clustered in order to efficiently collect data and therefore require sample design decisions in *multiple stages*. This is the

concern of Part II, which begins with a moderately complex project in Chap. 8 to design an area sample and allocate units to geographic clusters in such a way that the size of the samples of persons is controlled for some important demographic groups. Chapters 9 and 10 cover the design of samples of those geographic clusters. The US National Health and Nutrition Examination Survey (NHANES; Center for Disease Control and Prevention 2009) is a good example of a survey that could not be afforded unless the interviews were clustered. Elaborate medical examinations are conducted on participants from whom a series of measurements are taken: body measurements like height and weight; bone density measured via body scans; dental health lung function using spirometric tests to name just a few. The equipment for performing the tests is housed in trailers called Mobile Examination Centers, which are trucked from one sample area to another. Moving the trailers around the country and situating them with proper utility hookups in each location is expensive. Consequently, a limited number of PSUs have to be sampled first. Other surveys require sampling in multiple stages for a different reason, for example, if target sample sizes are required for certain subgroups. These subgroups often have to be sampled at rates other than their proportion in the population as a whole.

Part III discusses the computation of survey weights and their use in some analyses. We begin with a project in Chap. 12 on calculating weights for a personnel survey, like the one designed in Project 1 (see Chap. 2). Chapters 13 and 14 describe the steps for calculating base weights, making adjustments for ineligible units, nonresponse, and other sample losses, and for using auxiliary data to adjust for deficient frame coverage and to reduce variances. Some of the important techniques for using auxiliary data are the general regression estimator and calibration estimation. Since software is now available to do the computations, these are within the reach of any practitioner.

Intelligent use of these weight calculation tools requires at least a general understanding of when and why they work based on what they assume. Chapter 13 sketches the rationale behind the nonresponse *weight adjustment methods.* In particular, we cover the motivation behind cell adjustments and response propensity adjustments. Adjustment cells can be formed based on estimated propensities or regression trees. Understanding the methods requires thinking about models for response. The chapter also describes how use of auxiliary data can correct for frames that omit some units and how structural models should be considered when deciding how to use auxiliary data. We cover applications of calibration estimation, including poststratification, raking, and general regression estimation in Chap. 14. Methods of weight trimming using quadratic programming and other more *ad hoc* methods are also dealt with in that chapter.

Chapter 15 covers the major approaches to *variance estimation* in surveys—exact methods, linearization, and replication. Thinking about variance estimation in advance is important to be sure that data files are prepared in a way that permits variances to be legitimately estimated. To use

linearization or exact estimators, for example, fields that identify strata and PSUs must be included in the data file. The weighting procedures used in many surveys are fairly elaborate and generate complex estimators. Understanding whether a given method reflects the complexity of weight creation and what it omits, if anything, is important for analysts. There are a number of software packages available that will estimate variances and standard errors of survey estimates. We cover a few of these in Chap. 15.

Part IV covers two specialized topics—multiphase sampling and quality control. If subgroups are to be sampled at different rates to yield target sample sizes and a reliable list of the units in these subgroups is not available in advance of sampling, the technique of *multiphase sampling* can be used as described in Chap. 17. A large initial sample is selected and group identity determined for each unit through a screening process. Subsamples are then selected from the groups at rates designed to yield the desired sample sizes. Multiphase sampling can be combined with multistage sampling as a way of controlling costs while achieving target sample sizes. Another commonly used multiphase survey design involves the subsampling of phase-one nonrespondents for a phase-two contact, typically with a different mode of data collection than used initially.

An essential part of good survey practice is controlling the quality of everything that is done. Mistakes are inevitable, but procedures need to be developed to try and avoid them. Chapter 18 discusses some general *quality control measures* that can be used at the planning and data processing stages of a survey. These things are done by every professional survey organization but are seldom addressed in books on sampling. Quality control (QC) of statistical operations goes beyond merely checking work to make sure it is done correctly. It includes advance planning to ensure that all tasks needed to complete a project are identified, that the order of tasks is listed and respected, and that a proposed time schedule is feasible. Tracking the progress of data collection over time is another important step. Chapter 18 summarizes various rates that can be used, including contact, response, and balance on auxiliaries.

Documenting all tasks is important to record exactly what was done and to be able to backtrack and redo some tasks if necessary. In small projects the documentation may be brief, but in larger projects, detailed written specifications are needed to describe the steps in sampling, weighting, and other statistical tasks. Having standard software routines to use for sampling and weighting has huge QC advantages. The software may be written by the organization doing the surveys or it may be commercial off-the-shelf software. In either case, the goal is to use debugged routines that include standard quality checks.

Most of the code examples are written in the R language (R Core Team 2012), which is available for free. Additional materials are provided in the Appendices. Appendix C contains a primer on the R programming language and functions developed for chapter examples. Note that unless otherwise specified any R function referred to in the text is located in the `PracTools`

package. Data sets used in many of the examples are described in Appendix B; small data sets are provided within these pages while larger files are available through the book's Web address. Appendix A is the glossary of the notation used throughout the book. We recommend that you keep the glossary in mind as you read individual chapters since the notation needed for some topics is elaborate.

   With that brief overview, you are ready to see what a real sample design project looks like. The next chapter describes the requirements of a business organization for a survey of its employees.

Part I
Designing Single-Stage Sample Surveys

Chapter 2
Project 1: Design a Single-Stage Personnel Survey

Our primary goal is to equip survey researchers with the tools needed to design and weight survey samples. This chapter gives the first of several projects that mirror some of the complexities found in applied work. The three goals of this project are:

- Determine the allocation of a single-stage sample to strata in a multipurpose survey, accounting for specified precision targets for different estimates and differing eligibility and response rates for subgroups.
- Examine how sensitive the precision of estimates is to incorrect assumptions about response rates.
- Write a technical report describing the sample design.

As you proceed through the following chapters in Part I of the book, we suggest that you return to this chapter periodically, refresh your memory about the aims of Project 1, and think about how the methods in Chaps. 3–6 can be used in the development of the sampling design. In this chapter we outline the task that you should be able to solve after reading Part I.

2.1 Specifications for the Study

The Verkeer NetUltraValid (VNUV) International Corporation is preparing to conduct Cycle 5 of its yearly work climate survey of employees in their Survey Division. The climate survey assesses employee satisfaction in various areas such as day-to-day work life, performance evaluations, and benefits. In the first three cycles of the survey, the VNUV Senior Council attempted to do a census of all employees, but many employees considered the survey to be burdensome and a nuisance (despite their being in the survey business themselves). The response rates progressively declined over the first three cycles. In the fourth cycle, the Senior Council decided to administer an intranet survey only to a random sample of employees within the Survey Division. The aim

was to control the sampling so that continuing employees would not be asked to respond to every survey. In Cycle 5, a more efficient sample is desired that will improve estimates for certain groups of employees. The Senior Council requires a report from your design team that specifies the total number of employees to be selected, as well as their distribution by a set of characteristics noted below. They wish the quality and precision of the estimates to be better than the Cycle 4 survey. Note that this is the first survey in which the Senior Council has sought direction from sampling statisticians on the allocation of the sample.

Three business units are contained in the Survey Division: (*i*) the Survey Research Unit (SR) houses both survey statisticians and survey methodologists; (*ii*) the Computing Research Unit (CR) contains programmers who support analytic and data collection tasks; and (*iii*) Field Operations (FO) is populated by data collection specialists. The Senior Council would like to assess the climate within and across the units, as well as estimates by the three major salary grades (A1–A3, R1–R5, and M1–M3) and by tenure (i.e., number of months employed) within the units. However, the climate survey will only be administered to full- and part-time employees within these units. Temporary employees and contractors are excluded from the survey.

The Senior Council has identified three questions from the survey instrument that are most important to assessing the employee climate at VNUV. They are interested in the percentages of employees answering either "strongly agree" or "agree" to the following questions:

Q5.
    Overall, I am satisfied with VNUV as an employer at the present time.
Q12.
    There is a clear link between my job performance and my pay at VNUV.
Q15.
    Overall, I think I am paid fairly compared with people in other organizations who hold jobs similar to mine.

Note that the response options will remain the same as in previous years, namely, a five-level Likert scale: strongly agree, agree, neutral, disagree, and strongly disagree. A sixth response option, don't know/not applicable, is also available.

Additionally, the Senior Council would like estimates of the average number of training classes attended by the employees in the past 12 months. Relevant classes include lunchtime presentations, formal instructional classes taught at VNUV, and semester-long courses taught at the local universities.

2.2 Questions Posed by the Design Team

After receiving the study specifications document from the Senior Council, a design team is convened to discuss the steps required to complete the assigned

task. At this initial meeting, the following information was determined from the specifications:

- Data will be collected from employees through a self-administered intranet (i.e., web site internal to the corporation) questionnaire.
- All full- and part-time employees in the three business units within the Survey Division are eligible for the survey. Employees in other units within VNUV, as well as temporary employees and contractors, are ineligible and will be excluded from the sampling frame.
- The sample of participants will be randomly selected from a personnel list of all study-eligible employees provided by the head of VNUV's Human Resources (HR) Department.
- A single-stage stratified sampling design is proposed for the survey because (*i*) study participants can be selected directly from the complete HR (list) sampling frame and (*ii*) estimates are required for certain groups of employees within VNUV.
- The stratifying variables will include *business unit* (SR, CR, and FO), *salary grade* (A1–A3, R1–R5, and M1–M3), and potentially a categorized version of *tenure*.
- The analysis variables used to determine the allocation include three proportions, corresponding to each of the identified survey questions, and one quantitative variable. Estimates from the previous climate survey will be calculated by the design team from the analysis data file maintained by HR.

As is often the case when reviewing a sponsor's specifications for a project, there were a number of issues that needed clarification. Based on the initial discussion, the design team submitted the following clarifying questions to the Senior Council and received the responses noted below each:

1. Currently, HR defines tenure as the number of months of employment at VNUV. Is there a grouping of tenure years that would be informative to the analysis? For example, analysis of the previous climate survey suggests that responses differ among employees with less than 5 years of employment at VNUV in comparison to those with a longer tenure. Response: *Yes. Dichotomize tenure by less than 5 years and 5 years or greater.*
2. What is the budget for the climate survey and should we consider the budget when deciding on the total sample size? Response: *The budget permits two staff members to be assigned part-time to process and analyze the data over a three-month period. This does not affect the sample size. However, the council has decided that individual employees should not be surveyed every cycle to reduce burden and attempt to get better cooperation. Selecting a sample large enough to obtain 600 respondents will permit the annual samples to be rotated among employees.*

3. We are interested in classifying a difference between two estimates as being substantively meaningful to VNUV. Could you provide us with a meaningful difference? Response: *At least a five percentage point difference between any two sets of employee climate estimates is a meaningful difference. A difference of 2 to 3 in the average number of training classes is also of interest.*

4. Should the proportion answering "strongly agree" or "agree" to the three questions include or exclude the "don't know/not applicable" response category? Response: *Exclude.*

5. How precise should individual estimates be for this round of the survey? The quality of the data from prior versions of the climate survey has been measured in terms of estimated coefficients of variation (CV). Response: *The target CVs of overall estimates by business unit, by tenure within business unit, and by salary grade within business unit are listed in Table 2.1 below.*

6. Are there additional requirements for the design, such as estimates by gender and by number of dependents in addition to estimates by business unit, business unit by salary grade, and business unit by tenure? Response: *No.*

7. The VNUV Climate Survey Cycle 4 report does not detail the previous sampling design. The design team assumes that the Cycle 4 sample was randomly drawn from an updated list of employees within certain employee subgroups (i.e., a stratified simple random sample design). Is this correct? If so, where might we locate the stratifying information? Response: *No strata were used in the last design. The previous employee file was sorted by a random number and an equal probability, systematic sample was selected.*

8. Are the eligibility and response rates expected to be the same in Cycle 5 as they were in Cycle 4? Response: *The eligibility rates should be about the same, but we are not sure about the response rates. We would like to understand how sensitive the CVs will be if the response rates turn out to be lower than the ones in Cycle 4.*

2.3 Preliminary Analyses

HR provided the team with two data files. The first file contained information on all current VNUV employees such as employee ID, division, business unit, tenure in months, part-time/full-time status, and temporary/permanent employee status. The team eliminated all records for employees currently known to be ineligible for the survey, created a dichotomized version of tenure, and calculated population counts for the 18 design strata (Table 2.2).

The second file contained one record for each employee selected for the previous climate survey. In addition to the survey status codes (ineligible, eligible respondent, and eligible nonrespondent) and the survey responses,

Table 2.1: Target coefficient of variation by reporting domain: VNUV climate survey cycle 5, survey division.

Reporting domain	Target CV [a]
Business unit	0.06
Unit × Salary grade	0.10
Unit × Tenure	0.10

[a] Coefficient of variation

Table 2.2: Current distribution of eligible employees by business unit, salary grade, and tenure: VNUV climate survey cycle 5, survey division.

		Business unit			
Salary grade	Tenure	SR	CR	FO	Total
A1–A3	Less than 5 years	30	118	230	378
	5+ years	44	89	115	248
R1–R5	Less than 5 years	106	86	322	514
	5+ years	253	73	136	462
M1–M3	Less than 5 years	77	12	48	137
	5+ years	44	40	46	130
A1–A3	*Total*	74	207	345	626
R1–R5	*Total*	359	159	458	976
M1–M3	*Total*	121	52	94	267
Total	Less than 5 years	213	216	600	1,029
	5+ years	341	202	297	840
Total	*Total*	554	418	897	1,869

this file included the characteristics that should be used to define sampling strata in the new survey. This file, however, did not contain employee names or other identifying information to maintain the confidentiality promised to all survey participants. Sample members were classified as ineligible if, for example, they had transferred to another business unit within VNUV or retired after the sample was selected but before the survey was administered. The team isolated the eligible Survey Division records, created the sampling strata defined for the current climate survey design, and created the binary analysis variables for Q5, Q12, and Q15 from the original five-category questions (Table 2.3).

The information in Tables 2.4–2.6 was tabulated from the Survey Division responses to the Cycle 4 survey. No survey weights were used because the Cycle 4 sample employees were selected with equal probability and no weight adjustments, e.g., for nonresponse, were made.

Table 2.3: Documentation for recode of question responses to binary analysis variable: VNUV climate survey cycle 4, survey division.

Question responses	Binary analysis variable
1 = Strongly agree	1 = Strongly agrees or agrees
2 = Agree	1 = Strongly agrees or agrees
3 = Neutral	0 = Does not (strongly) agree
4 = Disagree	0 = Does not (strongly) agree
5 = Strongly disagree	0 = Does not (strongly) agree
6 = Don't know/not applicable < missing category >	

Table 2.4: Distribution of response status by business unit, salary grade, and tenure: VNUV climate survey cycle 4, survey division.

Business unit	Salary grade	Tenure	Total sample n	Ineligible[a] n	pct[b]	Total n	Resp. n	pct[c]	Nonresp. n	pct
SR	A1–A3	Less than 5 years	10	0	0.0	10	9	88.9	1	11.1
		5+ years	11	0	0.0	11	9	84.6	2	15.4
	R1–R5	Less than 5 years	34	3	9.7	31	16	51.6	15	48.4
		5+ years	71	1	1.3	70	55	78.7	15	21.3
	M1–M3	Less than 5 years	23	0	0.0	23	21	91.3	2	8.7
		5+ years	13	2	15.4	11	9	84.6	2	15.4
CR	A1–A3	Less than 5 years	41	3	7.1	38	22	58.6	16	41.4
		5+ years	20	0	0.0	20	10	50.0	10	50.0
	R1–R5	Less than 5 years	28	0	0.0	28	14	50.0	14	50.0
		5+ years	19	0	0.0	19	10	53.8	9	46.2
	M1–M3	Less than 5 years	6	0	0.0	6	6	100.0	0	0.0
		5+ years	9	1	11.1	8	7	88.9	1	11.1
FO	A1–A3	Less than 5 years	85	26	30.3	59	23	39.4	36	60.6
		5+ years	16	0	0.0	16	6	39.4	10	60.6
	R1–R5	Less than 5 years	101	2	2.2	99	65	65.2	34	34.8
		5+ years	34	1	2.6	33	24	71.8	9	28.2
	M1–M3	Less than 5 years	14	0	0.0	14	14	100.0	0	0.0
		5+ years	14	2	15.4	12	10	84.6	2	15.4
Total			549	41	7.5	508	330	65.0	178	35.0

[a] Ineligible sample members were those employees selected for the cycle 4 survey who retired or left the company prior to data collection

[b] Unweighted percent of total sample within each design stratum (row)

[c] Unweighted percent of total eligible sample within each design stratum (row)

Table 2.5: Estimates for four key questions by business unit, salary grade, and tenure: VNUV climate survey cycle 4, survey division.

Business unit	Salary grade	Tenure	Proportion (Strongly) agree			Avg number of training classes	
			Q5	Q12	Q15	Mean	SE[a]
SR	A1–A3	Less than 5 years	0.93	0.88	0.77	8.2	0.82
		5+ years	0.75	0.71	0.62	12.4	1.24
	R1–R5	Less than 5 years	0.84	0.80	0.69	22.3	2.23
		5+ years	0.80	0.76	0.66	24.0	1.92
	M1–M3	Less than 5 years	0.91	0.86	0.75	8.3	0.83
		5+ years	0.95	0.90	0.79	3.6	0.36
CR	A1–A3	Less than 5 years	0.99	0.94	0.92	7.2	0.87
		5+ years	0.80	0.76	0.74	10.9	1.09
	R1–R5	Less than 5 years	0.82	0.78	0.76	19.6	3.92
		5+ years	0.90	0.86	0.84	21.1	2.11
	M1–M3	Less than 5 years	0.97	0.92	0.90	7.3	0.73
		5+ years	0.97	0.92	0.90	3.2	0.32
FO	A1–A3	Less than 5 years	0.50	0.48	0.45	4.6	0.69
		5+ years	0.52	0.49	0.47	6.9	1.04
	R1–R5	Less than 5 years	0.75	0.71	0.68	12.5	1.87
		5+ years	0.70	0.67	0.63	13.4	2.02
	M1–M3	Less than 5 years	0.93	0.88	0.84	4.6	0.70
		5+ years	0.94	0.89	0.85	2.0	0.30

[a] Standard error

2.4 Documentation

With the preliminary analysis complete, the design team began to draft the sampling report to the Senior Council using the annotated outline below. This outline will be used to write a formal report when the project is completed:

Title = *VNUV Climate Survey Cycle 5 Sample Design Report*

1. Executive summary

   - Provide a brief overview of the survey including information related to general study goals and year when annual survey was first implemented.
   - Describe the purpose of this Cycle 5 document.
   - Provide a table of the sample size to be selected per business unit (i.e., respondent sample size inflated for ineligibility and nonresponse).

Table 2.6: Estimates by business unit, salary grade, and tenure: VNUV climate survey cycle 4, survey division.

Business unit	Salary grade	Tenure	Proportion (Strongly) agree			Avg number of training classes	
			Q5	Q12	Q15	Mean	SE
SR			0.84	0.80	0.69	18.1	0.98
CR			0.90	0.85	0.83	12.6	0.90
FO			0.67	0.63	0.60	8.9	0.60
SR	A1–A3		0.82	0.78	0.68	10.7	0.65
	R1–R5		0.81	0.77	0.67	23.5	2.26
	M1–M3		0.92	0.88	0.76	6.6	0.30
CR	A1–A3		0.91	0.86	0.85	8.8	0.46
	R1–R5		0.86	0.81	0.80	20.3	5.45
	M1–M3		0.97	0.92	0.90	4.1	0.09
FO	A1–A3		0.51	0.48	0.46	5.4	0.33
	R1–R5		0.74	0.70	0.66	12.8	2.09
	M1–M3		0.93	0.89	0.84	3.4	0.15
SR		Less than 5 years	0.88	0.83	0.73	15.3	1.33
		5+ years	0.81	0.77	0.67	19.9	2.06
CR		Less than 5 years	0.92	0.88	0.86	12.2	2.67
		5+ years	0.87	0.83	0.81	13.1	0.82
FO		Less than 5 years	0.67	0.64	0.60	8.8	1.08
		5+ years	0.67	0.63	0.60	9.2	1.02

- Discuss the contents of the remaining section of the report.

2. Sample design

- Describe the target population for Cycle 5.
- Describe the sampling frame including the date and source database.
- Describe the type of sample and method of sample selection to be used.

3. Sample size and allocation

- Optimization requirements
  - Optimization details including constraints and budget.
  - Detail the minimum domain sizes and mechanics used to determine the sizes.

- Optimization results
  – Results: minimum respondent sample size per stratum
  – Marginal sample sizes for key reporting domains
  – Estimated precision achieved by optimization results
- Inflation adjustments to allocation solution
  – Nonresponse adjustments
  – Adjustments for ineligible sample members
- Final sample allocation
  – Marginal sample sizes for key reporting domains
- Sensitivity analysis
  – Results from comparing deviations to allocation after introducing changes to the optimization system

4. Appendix

- Sample size per strata (table), full sample and expected number of respondents
- Other relevant detailed tables including preliminary analysis

2.5 Next Steps

The optimization problem and a proposed solution to the sampling design task discussed in this chapter will be revealed in Chap. 7. The methods discussed in the interim chapters will provide you with the tools to solve the allocation problem yourself. We will periodically revisit the VNUV design team discussions prior to Chap. 7 to provide insight into the design team's decisions and procedures.

Chapter 3
Sample Design and Sample Size for Single-Stage Surveys

Chapter 3 covers the problem of determining a sample size for single-stage surveys with imposed constraints such as a desired level of precision. To determine a sample size, a particular type of statistic must be considered. Means, totals, and proportions are emphasized in this chapter. We concentrate on simple random samples selected without replacement in Sect. 3.1. Precision targets can be set in terms of coefficients of variation or margins of error for unstratified designs as discussed in Sect. 3.1.1. We cover stratified simple random sampling in Sect. 3.1.2. Determining a sample size when sampling with varying probabilities is somewhat more complicated because the without-replacement variance formula is complex. A useful device for determining a sample size when sampling with probability proportional to size (*pps*) is to employ the design-based variance formula for with-replacement sampling, as covered in Sect. 3.2.1. Although we mainly cover calculations based on design-based variances, models are also especially useful when analyzing *pps* sampling as discussed in Sect. 3.2.2.

The remainder of this chapter covers some more specialized topics, including systematic, Poisson, and some other sampling methods in Sect. 3.3. Population parameters are needed in sample size formulas; methods for estimating them are covered in Sect. 3.4. Other important special cases are rare characteristics and domain (subpopulation and subgroup) estimates discussed in Sect. 3.5. The chapter concludes with some discussion of design effects and software for sample selection in Sects. 3.6 and 3.7.

The methods discussed here are limited to analyses for estimates based on a single y variable. Granted, this is extremely restrictive because most surveys measure a number of variables and make many estimates for domains such as the design strata. The more applicable problem of determining sample sizes and allocations for a multipurpose survey will be studied in Chap. 5.

3.1 Determining a Sample Size for a Single-Stage Design

One of the most basic questions that a survey designer must face is, how many? This is not easy to answer in a survey with multiple goals and estimates. A sample size that is adequate to estimate the proportion of persons who visited a doctor at least once last year may be much different from the sample size needed to estimate the proportion of persons with some extremely rare disorder like Addison's disease. Neither of these sample sizes is likely to be the same as that required to estimate the average salary per person.

This section discusses methods for estimating sample sizes for single-stage designs with one goal specified on the level of precision for a key analysis variable. Within the text, we consider several commonly used probability sampling plans. Methods applied with this simple survey design are the basis for understanding their application in more complex settings such as the project included in Chap. 2. Later in Chap. 5 we cover mathematical programming, which is the best tool for sample size calculation for complicated multi-goal surveys. Sample size determination for area samples requires a sample size calculation for each stage of the design and is discussed in Chap. 9.

Before getting into the details of the sample size calculations, a word about terminology is needed:

- Mathematicians like to distinguish between an *estimator*, which is a random quantity, and an *estimate*, its value in a particular sample. This distinction is of no importance for our purposes and we will use the terms interchangeably.
- We will use the phrase *population standard deviation* to mean the square root of a finite population variance. For example, the standard deviation of an analysis variable Y is $S = \sqrt{S^2}$ where the population variance, or *unit variance*, is $S^2 = \sum_{i=1}^{N} (y_i - \bar{y}_U)^2 /(N-1)$ where $\bar{y}_U = \sum_{i=1}^{N} y_i /N$ is the finite population mean and N is the number of elements in the population. U denotes the universe (i.e., the population) of N units.
- The *population (or unit) coefficient of variation* of Y is $CV_U = S /\bar{y}_U$. The square of the population CV, $CV_U^2 = S^2 /\bar{y}_U^2$, is called the *population (or unit) relvariance*.
- The term *standard error of an estimate*, abbreviated as SE, means the square root of the variance of the estimate. If $\hat{\theta}$ is an estimate of some population value, θ, then its standard error is $SE\left(\hat{\theta}\right) = \sqrt{V\left(\hat{\theta}\right)}$, where V is the variance computed with respect to a particular sample design. Common usage is to say *standard error* as shorthand for *standard error of an estimate*, although the former can be ambiguous unless everyone is clear about which estimate is being discussed. The standard error, $\sqrt{V\left(\hat{\theta}\right)}$, is a theoretical quantity that must be estimated from a sample. If we estimate $V\left(\hat{\theta}\right)$ by $v\left(\hat{\theta}\right)$, then the *estimated standard error of the estimate*

$\hat{\theta}$ is $se\left(\hat{\theta}\right) = \sqrt{v\left(\hat{\theta}\right)}$. Shorthand for this is to call $\sqrt{v\left(\hat{\theta}\right)}$ the *estimated standard error*.

- The *coefficient of variation (CV)* of an estimate $\hat{\theta}$ is defined as $CV\left(\hat{\theta}\right) = \sqrt{V\left(\hat{\theta}\right)}/\theta$, where $\theta = E\left(\hat{\theta}\right)$, the design-based expected value of the estimate $\hat{\theta}$, assuming that $\hat{\theta}$ is unbiased. This, too, must be estimated from a sample by $cv\left(\hat{\theta}\right) = \sqrt{v\left(\hat{\theta}\right)}/\hat{\theta}$, which is referred to as the *estimated coefficient of variation of the estimate* $\hat{\theta}$ or sometimes as the *estimated relative standard error*. Note that practitioners will often say "standard error" when they mean "estimated standard error" and *CV* when they mean "estimated *CV*."

The *CV* is usually expressed as a percentage, i.e., $100 \times CV\left(\hat{\theta}\right)$ and is a quantity that has a more intuitive interpretation than either the variance or SE. The *CV* has no unit of measure. For example, if we estimate the number of employees, both the SE and $\bar{y}_U$ are in units of employees, which cancel out in the *CV*. Because the *CV* is unitless, it can be used to compare the relative precision of estimates for entirely different kinds of quantities, e.g., dollars of revenue and proportion of businesses having health plans that pay for eyeglasses.

- An *auxiliary variable* is a covariate that is related to one or more of the variables to be collected in the study. An auxiliary variable may be available for every unit in a sampling frame, in which case, it can be used in designing an efficient sample. If the population total of an auxiliary is available from some source outside the survey, the auxiliary variable can be used in estimation. For estimation, having the value of one or more auxiliaries only for the sample cases is usually sufficient as long as population totals are available.

Other terms will be defined in later chapters as needed.

Regardless of the naming convention, in this book, theoretical quantities that are a function of population parameters are capitalized, e.g., $\sqrt{V\left(\hat{\theta}\right)}$, and the corresponding sample estimators are represented in lowercase, e.g., $\sqrt{v\left(\hat{\theta}\right)}$. A sample estimate of a population parameter θ is denoted with "hat," i.e., $\hat{\theta}$.

As long as all participants on a project understand the shorthand phrases in the same way, there will be no confusion. But, you may find it useful to occasionally verify that your understanding is the same as that of your colleagues. In the remainder of this section, we will calculate sample sizes using theoretical quantities like $CV\left(\hat{\theta}\right)$. However, bear in mind that precise

sample estimates of various ingredients (like S and CV_U) typically will be needed to evaluate the sample size formulas.

Criteria for Determining Sample Sizes

To determine a sample size, some criterion must be adopted for deciding how big is big enough. This is a question of how precise you want an estimate to be. We discuss several precision criteria that may be used in the sections that follow:

- Standard error of an estimate—Setting a target SE requires a judgment to be made about an acceptable level of SE. This can be difficult because an SE has the same units as the analysis variable (e.g., persons, dollars, milligrams of mercury).
- Coefficient of variation—CVs are more useful than SE's because they have no units of measure. Target values can be set without regard to the scale of an analysis variable.
- Margin of error (MOE)—This is related to the width of a confidence interval. MOE's are useful because survey sponsors or analysts are often comfortable making statements like "I want to be able to say that the population value is within 3% of the sample estimate." For later use, we denote the MOE as e.

Deciding which of these is the best criterion for a given survey is, to some extent, arbitrary. A practitioner should develop the knack of explaining the options to survey sponsors and guiding the sponsors toward choices that they both understand and accept. As we will emphasize, a key consideration is the budget. The sample size must be affordable; otherwise the survey cannot be done.

3.1.1 Simple Random Sampling

First, take the simple case of a single variable y and a simple random sample of units selected without replacement (*srswor*). Suppose we would like to estimate the (population) mean of y using the estimated (sample) mean based on a simple random sample of n units:

$$\bar{y}_s = \frac{1}{n} \sum_{i=1}^{n} y_i.$$

The theoretical population variance of the sample mean from an *srswor* (design) is

$$V(\bar{y}_s) = \left(1 - \frac{n}{N}\right)\frac{S^2}{n} \tag{3.1}$$

$$= \left(\frac{1}{n} - \frac{1}{N}\right)S^2$$

where N is the number of units in the target population on the sampling frame and $S^2 = \sum_{i=1}^{N}(y_i - \bar{y}_U)^2/(N-1)$ is the population unit variance, with $\bar{y}_U = \sum_{i=1}^{N} y_i/N$ the mean of all units in the target population. The term $1 - n/N$ is called the *finite population correction (fpc)* factor. The variance in expression (3.1) is called a *design variance* or *repeated sampling variance*, meaning that it measures the variability in $\bar{y}_s$ calculated from different possible samples of size n selected from the frame. In advance of sampling, the design variance is generally considered to be the one to use in computing a sample size. After a particular sample has been selected and data collected, the variance computed under a reasonable model may be more appropriate for inference from that particular sample (e.g., see Valliant et al. 2000). Since we are concerned about the design at the planning stage, we will usually consider design variances—in this case, ones calculated with respect to repeated simple random sampling.

Sometimes it will be handy to write a sum over the set of sample units as $\sum_{i \in s}$ with s denoting the set of sample units and a sum over the whole population as $\sum_{i \in U}$ where U denotes the population, or universe, of all units. To estimate the total of y from an *srswor*, use

$$\hat{t} = N\bar{y}_s \ , \tag{3.2}$$

whose (design) variance is

$$V(\hat{t}) = N^2\left(1 - \frac{n}{N}\right)\frac{S^2}{n}$$

$$= N\left(\frac{N}{n} - 1\right)S^2 \ .$$

To determine a sample size for an *srswor*, it does not matter whether we think about estimating a mean or a total—the result will be the same. There are situations, like domain estimation, to be covered later in this chapter where the estimated total is not just the estimated mean times a constant. In those cases, the variances of the two estimators are not as closely related and computed sample sizes may be different.

The square of the coefficient of variation for $\bar{y}_s$ and $\hat{T}$ is

$$CV^2(\bar{y}_s) = \left(\frac{1}{n} - \frac{1}{N}\right)\frac{S^2}{\bar{y}_U^2}. \tag{3.3}$$

We can set the squared CV or relvariance in Eq. (3.3) to some desired value, say CV_0^2 (like 0.05), and solve for the required sample n:

$$n = \frac{\frac{S^2}{\bar{y}_U^2}}{CV_0^2 + \frac{S^2}{N\bar{y}_U^2}}. \tag{3.4}$$

The sample size is a function of the (population) unit relvariance. When the population is large enough that the second term in the denominator is negligible compared to the first, the sample size formula is approximately

$$n \doteq \frac{S^2/\bar{y}_U^2}{CV_0^2} . \tag{3.5}$$

The more variable y is, the larger the sample size must be to achieve a specified CV target. Naturally, if the calculated n is more than the budget can bear, the survey will have to be scaled back or abandoned if the results would be unacceptably imprecise. Another way of writing Eq. (3.4) is

$$n = \frac{n_0}{1 + \frac{n_0}{N}}, \tag{3.6}$$

where $n_0 = \frac{S^2/\bar{y}_U^2}{CV_0^2}$, as in expression (3.5). The term n_0 is also the required sample size if a simple random sampling *with-replacement (srswr)* design was used. Thus, n_0/N in Eq. (3.6) accounts for the proportion of the population that is sampled. In two populations of different size but with the same variance S^2, Eq. (3.6) reflects the fact that the smaller size population will require a smaller sample to achieve a given CV.

Notice that setting the CV to CV_0 is equivalent to setting the desired variance to $V_0 = CV_0^2 \times \bar{y}_U^2$. Multiplying the numerator and denominator of expression (3.3) by $\bar{y}_U^2$ gives the equivalent sample size formula,

$$n = \frac{S^2}{V_0 + \frac{S^2}{N}} \doteq \frac{S^2}{V_0} . \tag{3.7}$$

As noted earlier, expression (3.4) is likely to be the easier formula to use than expression (3.7) because CV's are easier to understand than variances.

The R function, nCont, will compute a sample size using either CV_0 or V_0 as input (see Appendix C for an R code introduction). The parameters used by the function are shown below:

```
nCont(CV0=NULL, V0=NULL, S2=NULL, ybarU=NULL, N=Inf,
      CVpop=NULL)
```

If CV_0 is the desired target, then the unit CV, $S/\bar{y}_U$, or the population mean and variance, $\bar{y}_U$ and S^2, must also be provided. If V_0 is the constrained value, then S^2 must also be included in the function call. The default value of N is infinity, but a user-specified value can also be used. This and all subsequent functions discussed in the book are listed in Appendix C. The functions can be used after loading the PracTools package.

Example 3.1 (Sample size for a target CV). Suppose that we estimate from a previous survey that the population CV of some variable is 2.0. If the population is extremely large and CV_0 (the target CV) is set to 0.05, then the call to the R function is `nCont(CV0=0.05, CVpop=2)`. The resulting sample size is 1,600. If the population size is $N = 500$, then `nCont(CV0=0.05, CVpop=2, N=500)` results in a (rounded) sample size of 381. The fpc factor has a substantial effect in the latter case. ∎

Setting CV_0

To put the method described above into practice, a value for the target coefficient of variation, CV_0, must be set. To some extent, the value is arbitrary although rules of thumb have been developed over the years. A CV of an estimate of 50 % would imply that a normal-approximation confidence interval formed by adding and subtracting two standard errors of an estimate would cover zero. Such an estimate obviously is highly imprecise. The US National Center for Health Statistics flags any estimate it publishes that has a CV of 30 % or more and labels it as "unreliable."[1] Often, an estimate with a CV of 10 % or less is considered "reliable," but the purposes to which the estimate will be put must be considered.

Another way of setting precision would be to meet or beat the CV achieved in a previous round of a survey, assuming that level of precision was satisfactory. In that case, the same sample design and allocation could be used again. Some values of CVs from government-sponsored surveys in the USA are listed in Table 3.1. These obviously have quite a large range. CVs for published estimates from a given survey will also vary considerably because survey sponsors are usually anxious to publish estimates for many different domains whose sample sizes can vary. Some of the estimates will be very precise while others will not be.

In some instances, a precision target may be set by an administrative group. For example, the Council of the European Union (1998) specifies that certain types of labor force estimates have a CV of 8 % or less. The EU also recommends that member nations achieve certain *effective sample sizes* (Council of the European Union 2003) for income and living conditions estimates. An effective sample size, n_{eff}, was defined in Chap. 1 and is equal to the number of analytic sample units divided by the design effect, *deff*, for an estimator. The use of a *deff* or n_{eff} is a handy way of approximating required sample sizes in multistage surveys, as we will see in Chaps. 9 and 10.

Example 3.2 (Finding a sample size for tax returns). The US Internal Revenue Service (IRS) allows businesses, in some circumstances, to use sample estimates on their tax returns instead of dollar values from a 100 % enumeration of all accounts. For example, a business may estimate the total value

---

[1] www.cdc.gov/nchs/data/statnt/statnt24.pdf.

Table 3.1: Coefficients of variation or standard errors of some published estimates in US government-sponsored surveys.

Survey	Estimate	CV or standard error (SE)
Current Population Survey[a]	National unemployment rate of 6 %	1.9 % CV
Consumer Price Index[b]	National 1-month percentage price change	0.04 SE in percentage points
National Health & Nutrition Examination Survey III (1988–1994)[c]	Estimated median blood lead concentration (μg/dL) in US women, 17–45 years of age	1.24 % CV
2000 Survey of Reserve Component Personnel[d]	Percentage of Marine personnel saying that serving the country had a very great influence on their decision to participate in the National Guard/Reserve	3.22 % CV
National Hospital Discharge Survey 2005[e]	Total days of hospital care for heart disease	21.3 % CV

[a] Bureau of Labor Statistics (2006)
[b] Bureau of Labor Statistics (2009)
[c] Thayer and Diamond (2002)
[d] Deak et al. (2002, Table 28a.1)
[e] Center for Disease Control and Prevention (2005, Tables I, II)

of all capital assets that can be depreciated on a five-year schedule. The estimate may come from a sample of stores, buildings, or other appropriate units. In order to be allowed to use the point estimate from such a sample, the taxpayer must demonstrate that the increment used to compute a one-sided 95 % confidence interval is no more than 10 % of the point estimate. That is, if a total is estimated and a normal-approximation confidence interval is used, the requirement is that the MOE be $e = 1.645 \times CV\left(\hat{T}\right) \le 0.10$. If this condition is met, $\hat{T}$ can be used on the tax return; if not, either $\hat{T} - 1.645 \times SE\left(\hat{T}\right)$ or $\hat{T} + 1.645 \times SE\left(\hat{T}\right)$ must be used, whichever is the most disadvantageous to the taxpayer (Internal Revenue Service 2004, 2007). Since $CV(\bar{y}_s) = CV\left(\hat{T}\right)$ under simple random sampling, the IRS bound is equivalent to $CV\left(\hat{T}\right) \le 0.10 / 1.645$. If the population CV is 1, the sample size that would meet the IRS requirement is 271, which is obtained via `nCont(CV0=0.10/1.645, CVpop=1)`. ∎

Example 3.3 (VNUV sample sizes). Revisiting the data gathered for the VNUV Climate Survey (Project 1 in Chap. 2), the design team uses the previous survey data to estimate the population CV's for the average number of classes per year taken by a VNUV employee in the Survey Research

(SR) business unit. Since $CV^2(\bar{y}_s) = (n^{-1} - N^{-1})CV_U^2$ where $CV^2(\bar{y}_s)$ is from the previous survey, the population (unit) CV within each stratum can be computed as $CV_U^2 = CV^2(\bar{y}_s)/(n^{-1} - N^{-1})$. Information for the SR business unit, key to calculating the sample sizes, includes the following:

Business unit	Salary grade	Eligible employees	Previous sample size	Estimated average number of classes		
				Mean[a]	SE[b]	CV
SR	all	554	149	18.1	0.98	0.054
	A1–A3	74	20	10.7	0.65	0.061
	R1–R5	359	96	23.5	2.26	0.096
	M1–M3	121	33	6.6	0.30	0.045

[a] Counts of employees shown in Table 2.2
[b] Estimated means and standard errors were obtained from a prior survey and are shown in Table 2.6

The unit CV's estimated from the formula above for the three salary grades are 0.319, 1.099, and 0.303 and is 0.771 for all grades combined. To improve on the precision obtained from the prior round of the survey, the design team evaluates the target CV for each of the four estimates above at $CV_0 = 0.05$. The code to determine the new sample sizes is shown below. R comments (code that is not executed) are given after the pound sign (#) to help in understanding each section of the program:

```
Nh <- c(74, 359, 121)
Npop <- sum(Nh)
nh.old <- c(20, 96, 33)
n.old <- sum(nh.old)
cv.old <- c(0.061, 0.096, 0.045)
cv.SR <- 0.054
        # estimate unit CV from last survey
CVpoph <- cv.old/sqrt((1/nh.old - 1/Nh))
CVpop_ <- cv.SR/sqrt(1/n.old - 1/Npop)

        # salary grade samples
nCont(CV0=0.05, CVpop = CVpoph, N=Nh)
        # SR business unit sample
nCont(CV0=0.05, CVpop = CVpop_, N=Npop)
```

The results follow. Note that the decision to constrain the estimates within salary grade, in addition to across all the salary grades within this business unit, has cost implications. A total sample of 167 will meet the 0.05 CV target for the full business unit. However, the sum of the required sample sizes across the salary grades is approximately 261, indicating that over half of the maximum (respondent) sample size set (n=500) would need to be

allocated to these three strata (a likely problem toward finding a feasible solution):

Business unit	Salary grade	Sample size
SR	*all*	166.3
	A1–A3	26.3
	R1–R5	205.9
	M1–M3	<u>28.2</u>
	Sum	260.4

∎

Estimating Proportions

Many surveys estimate the proportion of units that have some characteristic. Coding y_i to one if unit i has the characteristic and zero if not (i.e., zero-one indicator variable), the estimated proportion is also the sample mean,

$$p_s = \sum_{i \in s} y_i \, / n \ .$$

In Project 1 (Chap. 2), the design team defined indicators for "agree" or "disagree" responses to three survey questions. The unit relvariance is then defined as

$$\frac{S^2}{\bar{y}_U^2} = \frac{N}{N-1} \frac{q_U}{p_U} \doteq \frac{q_U}{p_U},$$

where $p_U = \sum_{i \in U} y_i \, / N$ and $q_U = 1 - p_U$. The relvariance of p_s is

$$CV^2 (p_s) = \left(\frac{1}{n} - \frac{1}{N} \right) \frac{N}{N-1} \frac{q_U}{p_U} \ ,$$

which is a special case of Eq. (3.3). The sample size that will achieve a target CV of CV_0 comes from specializing the expression in Eq. (3.4):

$$n = \frac{\frac{N}{N-1} \frac{q_U}{p_U}}{CV_0^2 + \frac{1}{N-1} \frac{q_U}{p_U}} \doteq \frac{\frac{q_U}{p_U}}{CV_0^2} \tag{3.8}$$

The last approximation comes from again assuming that N, the size of the target population, is large.

Based on Eq. (3.8), the sample size will be larger for rare characteristics than for more prevalent ones. This coincides with the unit relvariance, $q_U \, / p_U$,

being larger for rare characteristics. Note that this, at first, seems to contradict the counsel that, when computing a sample size for estimating a proportion, you should assume that $p_U = 0.5$ because this will lead to the most conservative, i.e., largest, sample size (Cochran 1977, Sect. 4.4). However, that advice is based on the assumption that a target value of $V(p_s)$ is set. In that case, we can use the fact that $V(p_s) = \left(\frac{1}{n} - \frac{1}{N}\right) \frac{N}{N-1} p_U q_U$ to find that the sample size that will achieve a specified variance of V_0 is

$$
\begin{aligned}
n &= \frac{\frac{N}{N-1} p_U q_U}{V_0 + \frac{p_U q_U}{N-1}} \\
&\doteq \frac{p_U q_U}{V_0} \ .
\end{aligned}
\tag{3.9}
$$

Since $p_U q_U$ is maximized at $p_U = 0.5$, the largest sample size occurs when $p_U = 0.5$. You will explore the difference in setting a sample size based on a CV and based on a standard error target in Exercises 3.1 and 3.2.

Whether the sample size should be computed via the formula given in Eqs. (3.8) or (3.9) depends on the context. The same expression is not always desirable. A CV target of, say, 0.05 is far harder to hit for a rare characteristic than for a more prevalent one because the unit relvariance, q_U/p_U, depends inversely on the mean, p_U—the smaller the value of p_U, the bigger the relvariance. Figure 3.1 graphs the approximate sample sizes from Eq. (3.8) needed for CVs of 0.05 and 0.10 for p_U ranging from 0.10 to 0.90. If $p_U = 0.10$ and we want a CV of 0.05, the required sample size is 3,600. In contrast, if $p_U = 0.50$, the sample size is 400.

The R function, nProp, will compute the sample size using Eq. (3.8), assuming that a target CV_0 is set, or using Eq. (3.9), assuming a target variance, V_0. In either case, a value of p_U must be supplied. The parameters used by the function are nProp(CV0=NULL, V0=NULL, pU=NULL, N=Inf).

Example 3.4 (Sample size for rare characteristic). Consider the case of a rare characteristic in the population with $p_U = 0.01$. If we require a CV of 0.05, this means that the standard error of the proportion would be 0.0005. The sample size needed for this level of precision is 39,600, which is far larger than the budgets for many surveys could support (and larger than some populations!). The call to the R function to calculate this sample size is either nProp(V0=0.0005^2, N=Inf, pU=0.01) or nProp(CV0=0.05, N=Inf, pU=0.01).

On the other hand, it may be substantively interesting if we were able to estimate the proportion plus or minus 1/2 of 1%. This would, at least, confirm any suspicion that the proportion is quite small. If 1/2 of the 1% goal is translated to mean that a 95% confidence interval should have a half-width of 0.005, this means that
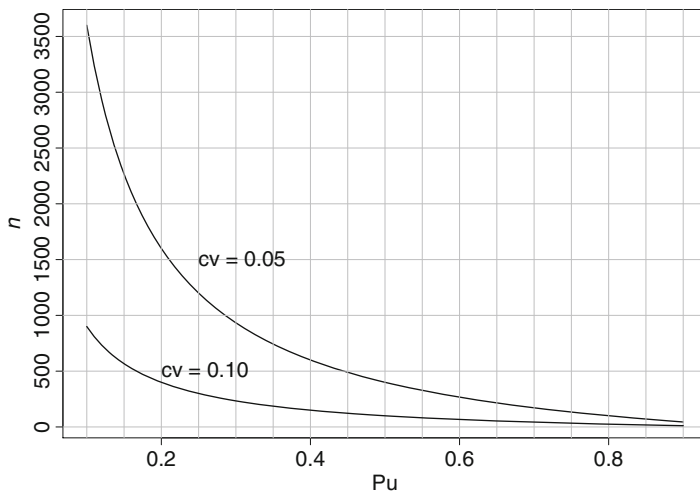
Fig. 3.1: Approximate sample sizes from Eq. (3.8) required to achieve CVs of 0.05 and 0.10 for population proportions ranging from 0.10 to 0.90. The population size is assumed to be large so that the finite population correction is one.

$$1.96\sqrt{\frac{p_U\,(1-p_U)}{n}} = 0.005\ ,$$

i.e., the standard error is about 0.0026. This, in turn, implies that the sample size needed to meet this goal is $n = 1,522$—far less than 39,600. The call to `nProp` to compute this is `nProp(V0=(0.005/1.96)^2, N=Inf, pU=0.01)`.

The function `nProp` will also take a vector `pU` as input. For example, if we want the sample sizes for p_U in `(0.01, 0.05, 0.10)`, the command is `nProp(CV0=0.05, N=Inf, pU=c(0.01, 0.05, 0.10))` with results, $n = $ 39,600, 7,600, and 3,600. ∎

Example 3.5 (Effect of the fpc). Returning again to Project 1 in Chap. 2, the following estimated "strongly agree" proportions were calculated from the previous climate survey for question 5 (*Q5. Overall, I am satisfied with VNUV as an employer at the present time*) for employees in the Survey Research unit:

The design team decides to initially constrain all the estimated proportions with $CV_0 = 0.06$. However, one member of the team recommends the use of `N=Inf` with the `nProp` function citing from statistics class that any population size greater than 30 is large. Others on the team disagree but concede to run the sample size calculations both ways for comparison, e.g., `nProp(CV0=0.06, N=Inf, pU=0.82)` for salary grades A1–A3, which gives $n = 61$, compared with `nProp(CV0=0.06, N=68, pU=0.82)`,

Business unit	Salary grade	Eligible employees[a]	Q5[b]	Sample size	
				$N = Inf$	$N = N_h$
SR	A1–A3	74	0.82	61.0	33.7
	R1–R5	359	0.81	65.2	55.3
	M1–M3	121	0.92	24.2	20.3
Total		554		150.4	109.3

[a] Counts of employees shown in Table 2.2

[b] Estimated proportion of employees who strongly agree with the statement in question 5

which yields $n = 33$. The results shown above highlight the need for specifying the population size (if known) when calculating sample sizes unless the population is extremely large. ∎

Setting a Margin of Error

The method just described is also equivalent to setting a tolerance for how close an investigator would like the estimate to be to the population value. In fact, many investigators prefer to think of setting tolerances rather than CVs. If the tolerance (sometimes called the MOE) is e and the goal is to be within e of the population mean with probability $1 - \alpha$, this translates to

$$\Pr\left(|\bar{y}_s - \bar{y}_U| \le e\right) = 1 - \alpha . \tag{3.10}$$

This is equivalent to setting the half-width of a $100\,(1 - \alpha)\,\%$ two-sided confidence interval (CI) to $e = z_{1-\alpha/2}\sqrt{V\left(\bar{y}_s\right)}$, assuming that $\bar{y}_s$ can be treated as being normally distributed. The term $z_{1-\alpha/2}$ is the $100\,(1 - \alpha\,/2)$ percentile of the standard normal distribution, i.e., the point with $1 - \alpha\,/2$ of the area to its left. If we require

$$\Pr\left(\left|\frac{\bar{y}_s - \bar{y}_U}{\bar{y}_U}\right| \le e\right) = 1 - \alpha , \tag{3.11}$$

this corresponds to setting $e = z_{1-\alpha/2}\,CV(\bar{y}_s)$. (See Exercise 3.4.) If we set the MOE to e_0, then Eq. (3.10) can be manipulated to give the required sample size as

$$n = \frac{z_{1-\alpha/2}^2 S^2}{e_0^2 + z_{1-\alpha/2}^2 S^2 / N} \ . \tag{3.12}$$

Similarly, if the MOE in Eq. (3.11) is set to e_0, we obtain

$$n = \frac{z_{1-\alpha/2}^2 S^2 / \bar{y}_U^2}{e_0^2 + z_{1-\alpha/2}^2 S^2 / (N \bar{y}_U^2)} \ . \tag{3.13}$$

In the particular case of estimating a proportion, we set $S^2 = N p_U q_U / (N - 1)$ in Eq. (3.12). Solving for n gives

$$
\begin{aligned}
n &= \frac{\frac{N}{N-1} z_{1-\alpha/2}^2 p_U q_U}{e^2 + z_{1-\alpha/2}^2 \frac{p_U q_U}{N-1}} \\
&\doteq z_{1-\alpha/2}^2 \frac{p_U q_U}{e^2},
\end{aligned}
\tag{3.14}
$$

which is the same as Eq. (3.9) once we note that $V_0 = e^2 \big/ z_{1-\alpha/2}^2$. Either Eqs. (3.9) or (3.14) may be convenient, depending on how one phrases the goal for estimation.

If we require that the half-width of a CI be a specified proportion of p_U, then set $S^2 \big/ \bar{y}_U^2 = N q_U / [(N-1) p_U]$ in Eq. (3.13). The solution for the sample size is then

$$
\begin{aligned}
n &= \frac{\frac{N}{N-1} z_{1-\alpha/2}^2 \frac{q_U}{p_U}}{e^2 + z_{1-\alpha/2}^2 \frac{q_U}{p_U (N-1)}} \\
&\doteq \frac{z_{1-\alpha/2}^2}{e^2} \frac{q_U}{p_U}.
\end{aligned}
\tag{3.15}
$$

Because $CV_0^2 = e^2 \big/ z_{1-\alpha/2}^2$, expression (3.15) is the same as Eq. (3.8). The R function, `nPropMoe`, will calculate sample sizes using Eqs. (3.14) or (3.15), corresponding to whether we set the MOE in terms of Eqs. (3.10) or (3.11). The type of MOE is selected by the parameter `moe.sw` where `moe.sw=1` invokes Eq. (3.14), i.e., $e = z_{1-\alpha/2} \sqrt{V(p_s)}$, and `moe.sw=2` invokes Eq. (3.15), i.e., $e = z_{1-\alpha/2} \sqrt{V(p_s)} / p_U$. The full set of parameters is shown in the function call below:

```
nPropMoe(moe.sw, e, alpha=0.05, pU, N=Inf)
```

Example 3.6 (Sample size based on MOE). Suppose that we want to estimate a proportion for a characteristic where $p_U = 0.5$ with a MOE of e when $\alpha = 0.05$. In other words, the sample should be large enough that a normal-approximation 95 % confidence interval should be $0.50 \pm e$. For example, if $e = 0.03$ and p_s were actually 0.5, we want the confidence interval to be $0.50 \pm 0.03 = [0.47, 0.53]$. The sample size is highly dependent on the width of the confidence interval as seen in the following table. Sample sizes were

evaluated using the formula given in Eq. (3.14) with $p_U = 0.5$ and $z_{0.975} = 1.96$. The command to generate the sample sizes listed in the table below is

```
nPropMoe(moe.sw=1, e=seq(0.01,0.08,0.01), alpha=0.05,
         pU=0.5)
```

e	n	e	n
0.01	9,604	0.05	384
0.02	2,401	0.06	267
0.03	1,067	0.07	196
0.04	600	0.08	150

Notice that the terminology in this example may seem a little loose. When a sample is selected and the proportion is estimated, p_s will almost certainly not equal p_U. The computed CI will be $p_s \pm e$, not $p_U \pm e$. Consequently, it is best to think of p_U in Example 3.6, and in the subsequent discussion, as a value, hypothesized in advance of sampling. ∎

Wilson Method for Proportions

A problem with normal-approximation confidence intervals (CIs) for proportions, computed as $p_s \pm z_{1-\alpha/2} \sqrt{V(p_s)}$, is that the interval may not be confined to $[0, 1]$ when the proportion is extreme (i.e., extremely rare or highly prevalent). One method that will produce endpoints in the allowable range is due to Wilson (1927). Brown et al. (2001) and Newcombe (1998) showed that the Wilson method has better coverage properties than several alternative methods, including the standard normal-theory intervals. The general idea is to treat $t = (p_s - p_U) \big/ \sqrt{p_U q_U / n}$ as having a standard normal distribution. Then, rearranging the inequality $|t| \leq z_{1-\alpha/2}$ gives a quadratic in p_U. The roots of the quadratic are the endpoints of the Wilson confidence interval:

$$\frac{\left(2p_s n + z^2\right) \pm z\sqrt{z^2 + 4p_s q_s n}}{2\left(z^2 + n\right)} .$$

This interval is not symmetric, but to parallel the earlier methods, we will consider half the width of the interval as the MOE. The half-width of this confidence interval is

$$\frac{1}{2} \frac{z\sqrt{z^2 + 4p_s q_s n}}{z^2 + n},$$

where $z \equiv z_{1-\alpha/2}$. If we set the half-width to some desired value e, substitute an advance estimate of p_U for p_s, and solve for n, this leads to another quadratic in n whose largest root is

$$n = \frac{1}{2}\left(\frac{z}{e}\right)^2 \left[\left(p_U q_U - 2e^2\right) + \sqrt{e^2 - p_U q_U \left(4e^2 - p_U q_U\right)}\right] . \qquad (3.16)$$

If a complex sample were selected, then similar steps apply after treating $t = (\hat{p} - p_U) \big/ \sqrt{p_U q_U / n_{eff}}$ as being standard normal.

The R function nWilson will calculate a sample size using inputs for p_U and e. As in nPropMoe, the desired MOE can be specified as the CI half-width on the proportion (moe.sw=1) or as the CI half-width on a proportion of the population value p_U (moe.sw=2). The function does not include an *fpc* although the reader could modify the code to include one if the associated sampling rate (n/N) is sizeable. The full set of parameters is nWilson(moe.sw,alpha=0.05,pU,e).

The function returns a list containing the sample size, the anticipated endpoints of the CI, and the length of the CI. The last value, 'length of CI', simply verifies that the anticipated length of the CI equals twice the input value e when moe.sw=1 and equals $2e\,p_U$ when moe.sw=2.

Example 3.7 (Wilson sample size). Suppose that $p_U = 0.04$ and the desired half-width of the CI is 0.01. The function call and output are

```
nWilson(moe.sw =1, pU=0.04, e=0.01)

$n.sam
[1] 1492.151
$'CI lower limit'
[1] 0.0311812
$'CI upper limit'
[1] 0.0511812
$'length of CI'
[1] 0.02
```

Thus, a sample of about 1,492 is needed. Notice that the anticipated CI is not symmetric around $p_U = 0.04$. The corresponding MOE computation using nPropMoe is

```
nPropMoe(moe.sw=1, e=0.01, alpha=0.05, pU=0.04, N=Inf)

[1] 1475.120
```

where the value after the "[1]" is the output from the R function. In other words, the estimated sample size is about the same with either function. The usefulness of the Wilson method in practice is more in the actual computation of the confidence interval itself rather than in estimating a sample size. ∎

Log-Odds Method for Proportions

Another method of CI construction for proportions is to transform the proportion to the log-odds scale, put a confidence interval on the log-odds, and then back-transform the endpoints of the CI to the proportion scale. Like the Wilson method, this approach produces a CI on the proportion that is confined to $[0, 1]$. Based on the empirical results in Brown et al. (2001), the Wilson method performs somewhat better in small to moderate size samples. However, the use of the log-odds is better known among practitioners, and the

sample sizes calculated with the two methods will be similar. The log-odds of the sample estimate is $\log\left(p_s/q_s\right)$ with $q_s = 1 - p_s$. A linear approximation to the log-odds is

$$\log\left(p_s/q_s\right) \doteq \log\left(p_U/q_U\right) + \left(p_s - p_U\right)/\left(p_U q_U\right) \ .$$

The approximate variance of $\log\left(p_s/q_s\right)$ is then

$$v\left[\log\left(p_s/q_s\right)\right] = \frac{1}{p_U q_U}\left(\frac{1}{n} - \frac{1}{N}\right)\frac{N}{N-1} \ .$$

A normal-approximation CI on $\log\left(p_U/q_U\right)$ is $\log\left(p_s/q_s\right) \pm z_{1-\alpha/2}\sqrt{v\left[\log\left(p_s/q_s\right)\right]}$. Defining (L, U) as the endpoints of this confidence interval, the back-transformed endpoints of a CI on p_U is $\left[\left(1 + \exp\left(-L\right)\right)^{-1}\right.$, $\left.\left(1 + \exp\left(-U\right)\right)^{-1}\right]$. Computing the half-width of this CI and setting this to a MOE e give

$$e = \frac{1}{2}\frac{\exp\left(-L\right) - \exp\left(-U\right)}{\left[1 + \exp\left(-L\right)\right]\left[1 + \exp\left(-U\right)\right]} \ .$$

With some algebra this equation leads to a quadratic equation in

$$\exp\left[\frac{z}{\sqrt{p_U q_U}}\sqrt{\left(\frac{1}{n} - \frac{1}{N}\right)\frac{N}{N-1}}\right],$$

which can be solved to give

$$n = \left\{\frac{N}{N-1}\left[\frac{\sqrt{p_U q_U}}{z_{1-\alpha/2}}\log\left(x\right)\right]^2 + \frac{1}{N}\right\}^{-1}, \tag{3.17}$$

where

$$x = \frac{1}{k\left(1 - 2e\right)}\left[e\left(k^2 + 1\right) + \sqrt{e^2\left(k^2 + 1\right)^2 - k^2\left(1 - 2e\right)\left(1 + 2e\right)}\right]$$

and $k = q_U/p_U$. The R function, nLogOdds, will evaluate the sample size in Eq. (3.17). The function accepts the same five parameters as nPropMoe. The desired MOE can be specified as the CI half-width on the proportion (moe.sw=1) or as the CI half-width on a proportion of the population proportion p_U (moe.sw=2). The full set of parameters accepted by the function is shown in the call below:

```
nLogOdds(moe.sw, e, alpha=0.05, pU, N=Inf)
```

Another transformation that is sometimes used when calculating a CI for a proportion is the $\arcsin\left(\sqrt{p_s}\right)$. This transformation is not included here because it does not appear amenable to sample size calculation when setting a MOE.

Example 3.8 (Comparison of three methods). As in Example 3.7, suppose that $p_U = 0.04$, the desired half-width of the CI is 0.01, and the population is large. The function call and output listed after the "[1]" from our three functions for computing samples sizes are

```
nLogOdds(moe.sw=1, e=0.01, alpha=0.05, pU=0.04, N=Inf)
[1] 1500.460
nWilson(moe.sw=1, pU=0.04, e=0.01)\$n.sam
[1] 1492.151
nPropMoe(moe.sw=1, e=0.01, alpha=0.05, pU=0.04, N=Inf)
[1] 1475.120
```

The sample sizes are within about 2 % of each other although the Wilson and log-odds methods do suggest a larger sample size than the standard approach. ∎

Obtaining Population Values

As a last word before we leave simple random sampling, note that all of the sample size formulas above are written in terms of population quantities that are likely unknown during the sample design phase of the study. For example, S^2, $\bar{y}_U$, and p_U are all population values. If the same survey has been done before on an earlier rendition of the population, then the sample data can be used to estimate the parameters. If no previous data are available on the target population, it may be possible to get data on a similar population. In some cases, published summary estimates may be accessible. This is especially true of proportions. For example, the US Bureau of Labor Statistics[2] publishes estimated percentages of workers that receive different benefits from their employers, the National Center for Health Statistics[3] produces statistics on the nation's health, the National Center for Education Statistics (NCES) tabulates statistics on public and private education at all levels, and the Census Bureau[4] provides statistics on the population and many other topics. Other countries have similar statistical agencies that publish economic, epidemiological, and other statistics.

In some cases, a secondary data source for the entire population or microdata sets for earlier samples will be available. For instance, the Common Core of Data (CCD)[5] from NCES contains population data files of elementary and secondary schools that can be used to tabulate means, variances, proportions or other statistics. If the microdata are provided for individual records for a sample of units from the target population, you can estimate population parameters. We will discuss how to estimate some population parameters from samples in Sect. 3.4. Note that the design team for Project 1 in Chap. 2

---

[2] http://stats.bls.gov/.

[3] http://www.cdc.gov/nchs/.

[4] http://www.census.gov/.

[5] http://nces.ed.gov/ccd/.

had direct access to the relevant data sources and could therefore produce the estimates provided in Tables 2.2–2.6.

3.1.2 Stratified Simple Random Sampling

Simple random samples are rare in practice for several reasons. Most surveys have multiple variables and domains for which estimates are desired. Selecting a simple random sample runs the risk that one or more important domains will be poorly represented or omitted entirely. In addition, variances of survey estimates can often be reduced by using a design that is not *srswor*.

A design that remedies the problems noted for an *srswor* is referred to as stratified simple random sampling (without replacement) or *stsrswor*. As the name indicates, an *srswor* design is administered within each design stratum. Strata are defined with one or more variables known for *all* units and partition the entire population into mutually exclusive groups of units. We might, for example, divide a population of business establishments into retail trade, wholesale trade, services, manufacturing, and other sectors. A household population could be divided into geographic regions—north, south, east, and west. For an *stsrswor*, we define the following terms:

- N_h = the known number of units in the population in stratum h ($h = 1, 2, \ldots, H$)
- n_h = the size of the *srswor* selected in stratum h
- y_{hi} = the value of the y variable for unit i in stratum h
- $S_h^2 = \sum_{i=1}^{N_h} (y_{hi} - \bar{y}_{U_h})^2 / (N_h - 1)$, the population variance in stratum h
- U_h = set of all units in the population from stratum h
- s_h = set of n_h sample units from stratum h

Note that the total sample size is calculated as $n = \sum_{h=1}^{H} n_h$. The population mean of y is

$$\bar{y}_U = \sum_{h=1}^{H} W_h \bar{y}_{U_h},$$

where $W_h = N_h / N$ and $\bar{y}_{U_h}$ is the population mean in stratum h. The sample estimator of $\bar{y}_U$ based on an *stsrswor* is

$$\bar{y}_{st} = \sum_{h=1}^{H} W_h \bar{y}_{s_h}, \qquad (3.18)$$

where $\bar{y}_{s_h} = \sum_{i \in s_h} y_{hi} / n_h$. When estimating a population proportion, the estimator is similar:

$$p_{st} = \sum_{h=1}^{H} W_h p_{s_h} \qquad (3.19)$$

with p_{s_h} defined in the same way as $\bar{y}_{s_h}$ using a zero-one (indicator) y variable. The population sampling variance of the stratified estimator is

$$V(\bar{y}_{st}) = \sum_{h=1}^{H} W_h^2 \frac{1 - f_h}{n_h} S_h^2, \tag{3.20}$$

where $f_h = n_h / N_h$.

Strata are especially useful if they correspond to domains for which separate estimates are needed. In that case, the sample assigned to each stratum can be determined using the formulas in Sect. 3.1.1 and is guaranteed to result in selecting sample cases for each domain (i.e., stratum). However, the overall sample size, $n = \sum_{h=1}^{H} n_h$, may become excessively large. To remedy this problem, the overall sample size can be allocated to the strata using various techniques as discussed in the next section. An efficient allocation can lead to the variance of an overall estimator, $\bar{y}_{st}$ or p_{st}, being smaller than with an (unstratified) *srswor*.

Choosing Stratification Variables

Stratifiers can be selected on at least five grounds (see, e.g., Lohr 1999, Chap. 4):

1. To avoid selecting a sample that is poorly distributed across the population, as could occur with *srswor*
2. As a way of guaranteeing certain sample sizes in groups that will be studied separately (i.e., domains)
3. As an administrative convenience (e.g., a mail survey might be used for units in some strata but personal interviews for the remaining strata)
4. To manage cost (e.g., data collection in some strata might be more expensive than in other strata)
5. As a way of improving sample efficiency for full population estimates by grouping units together that have similar mean and variance properties

An example of the second use would be a business survey in which establishments are grouped by type of business (manufacturing, retail, service, etc.). The sample could be allocated in such a way that each sector receives a large enough sample size to meet precision targets for some important estimates. In a survey of schools, strata might be defined based on the level and ownership of a school (e.g., elementary, middle, and high school crossed with public or private ownership). Typically, an allocation to these strata designed to meet a CV target for each stratum would not be the best allocation for making an efficient estimate for the full population. However, in such cases, the domain estimates are usually more important than full population estimates. In addition, when the domain estimates have acceptable precision, then the full population estimates will also.

Stratification by size with an efficient allocation is an example of 5 above. This method uses a size variable that is correlated with whatever is to be measured in the survey. In an establishment survey, the number of employees at a previous time period should be a predictor of current employment and possibly of other variables, like revenue. To determine a good measure of size (MOS), regression modeling should be done assuming that some relevant data are available. This method of stratification is closely related to *pps* sampling described in Sect. 3.2.1 (also, see Valliant et al. 2000, Chap. 6).

Types of Allocations

There are several types of allocation methods that can be considered for a stratified sample. The first three allocations below assume that the total sample size n is fixed and corresponds to a fixed study budget (assuming the cost of collecting and processing data for each unit is the same). In the last two, the total sample size is determined to be consistent with either cost or variance constraints:

1. *Proportional allocation in which $n_h = nW_h$*

   This allocation is efficient for estimating the mean of y if the stratum standard deviations, S_h, are all equal or, at least, are very close to each other. This method may allocate very few units to some small strata and, thus, can be poor when separate stratum estimates are desired.
2. *Equal allocation with $n_h = n/H \equiv \bar{n}$*

   Equal allocation is useful if an estimate is needed for each stratum individually and if the stratum standard deviations are about the same.
3. *Neyman allocation where $n_h = n\frac{W_h S_h}{\sum_{h=1}^{H} W_h S_h}$*

   Neyman allocation minimizes the variance, $V(\bar{y}_{st})$, of the estimator of the population mean. Neyman may give high variance estimates for some individual strata. Plus, it ignores any differential costs of data collection and processing among strata (as do proportional and equal allocations).
4. *Cost-constrained optimal allocation*

   This allocation minimizes $V(\bar{y}_{st})$ subject to a fixed total budget and is discussed in detail below.
5. *Precision-constrained optimal allocation*

   This allocation minimizes total cost subject to a fixed constraint on $V(\bar{y}_{st})$ or $CV(\bar{y}_{st})$ and is also discussed more below.

We sketch the results for allocations 4 and 5 below. In both, the proportion of the sample allocated to a stratum is the same and is given in Eq. (3.23). The two methods lead to different total sample sizes as shown in Eqs. (3.22) and

(3.25). You can read more of the mathematical details in a text like Särndal et al. (1992).

The cost-constrained optimal allocation 4 uses this simple linear cost function, $C = c_0 + \sum_{h=1}^{H} n_h c_h$, where C is total cost, c_0 is the sum of cost values that *do not vary* with the number of sample cases, and c_h is the cost per sample case in stratum h. The term c_0 is usually called "fixed cost" and can include components such as salaries for a project manager, programmers, and editing supervisors. The term c_h is the cost of data collection, e.g., interviewing and mailing, and other unit costs that increase as the sample size increases. Minimizing $V(\bar{y}_{st})$ in expression (3.20) subject to a specified total budget leads to

$$n_h = (C - c_0) \frac{W_h S_h \big/ \sqrt{c_h}}{\sum_{h=1}^{H} \left(W_h S_h \sqrt{c_h} \right)} \ . \tag{3.21}$$

The total sample size is the sum of the n_h across the sampling strata:

$$n = (C - c_0) \frac{\sum_{h=1}^{H} W_h S_h \big/ \sqrt{c_h}}{\sum_{h=1}^{H} \left(W_h S_h \sqrt{c_h} \right)} \ . \tag{3.22}$$

The proportion of the sample allocated to stratum h is

$$\frac{n_h}{n} = \frac{W_h S_h \big/ \sqrt{c_h}}{\sum_{h=1}^{H} \left(W_h S_h \big/ \sqrt{c_h} \right)} \ . \tag{3.23}$$

As is apparent from the expression given in Eq. (3.23), strata that account for a larger share of the population, as measured by W_h or have larger standard deviations, S_h, get a larger portion of the overall sample size. Strata where the unit cost, c_h, is larger get less.

If the variance, $V(\bar{y}_{st})$, is fixed at V_0 and we minimize the total cost, as with the precision-constrained optimal allocation 5, the allocation to stratum h is

$$n_h = (W_h S_h \big/ \sqrt{c_h}) \frac{\sum_{h=1}^{H} W_h S_h \sqrt{c_h}}{V_0 + N^{-1} \sum_{h=1}^{H} W_h S_h^2} \ . \tag{3.24}$$

If the CV of $\bar{y}_{st}$ is fixed at CV_0, this implies that V_0 in Eq. (3.24) should be set to $V_0 = CV_0^2 \times \bar{y}_U^2$. In this case, the proportion of the sample allocated to stratum h is also given by Eq. (3.23), but the total sample size is

$$n = \sum_{h=1}^{H} (W_h S_h \big/ \sqrt{c_h}) \frac{\sum_{h=1}^{H} W_h S_h \sqrt{c_h}}{V_0 + N^{-1} \sum_{h=1}^{H} W_h S_h^2} \ . \tag{3.25}$$

When computing a cost-constrained or precision-constrained allocation, sample sizes are usually rounded up to the next integers. This is not usually something to be overly concerned about since the constraints will still be met approximately. In addition, if there is any chance of nonresponse or

other sample losses, the survey designer inevitably loses some control over the allocation. In any case, useful quality control checks after making the calculations in Eqs. (3.21) and (3.24) are:

(i) Verify that $\sum_h n_h c_h$ approximately respects the cost constraint.
(ii) Check that $V(\bar{y}_{st})$ is about equal to V_0.

These are simple ways of detecting computational mistakes.

Of the two allocations 4 and 5, the cost-constrained method in Eq. (3.21) is probably the one used more often. The usual situation is that an investigator has a predetermined amount of money to spend. Any sample that is selected must fit within that budget. Another standard occurrence is that partway through a study the budget is changed—usually cut—or that the unit costs c_h are higher than expected. Consequently, midcourse adjustments to the sample size are necessary. If the total budget is cut, the optimal allocation of the reduced sample can be computed by reducing the sample sizes in Eq. (3.21) by the same percentage in each stratum. Alternatively, some judgment can be made about whether retaining precision in some strata is more important than in others.

Regardless of the allocation chosen, formula (3.20) can be used to compute the variance of $\bar{y}_{st}$. Even though Eq. (3.20) could be specialized using the formulas for allocations 1–5, this is unnecessary and, in fact, undesirable for computer programming. When evaluating Eq. (3.20) from a sample, the population variance, S_h^2, can be estimated as described in Sect. 3.4.

The R function, strAlloc, will compute the proportional, Neyman, cost-constrained, and variance-constrained allocations defined above. The parameters accepted by the function are:

n.tot	fixed total sample size
Nh	vector of pop stratum sizes (N_h) or pop stratum proportions (W_h), required
Sh	stratum unit standard deviations (S_h), required unless alloc = "prop"
cost	total variable cost ($C - c_0$)
ch	vector of cost per unit in stratum h (c_h)
V0	fixed variance target for estimated mean
CV0	fixed CV target for estimated mean
ybarU	pop mean of y ($\bar{y}_U$)
alloc	type of allocation, must be one of "prop", "neyman", "totcost", "totvar"

The parameters can only be used in certain combinations, which are checked at the beginning of the function. Basically, given an allocation, only the parameters required for the allocation are allowed and no more. For example, the Neyman allocation requires Nh, Sh, and n.tot. The function returns

a list with three components—the allocation type, the vector of sample sizes
(n_h), and the vector of sample proportions allocated to each stratum (n_h / n).

Example 3.9 (Cost allocation). Table 3.2 gives stratum population counts and
standard deviations of total expenditures based on the 1998 Survey of Men-
tal Health Organizations (SMHO).[6] The survey data set is treated as the
population (`smho98`) for this example. (See B for details of this and other
data sets.) The y variable is the total expenditures during a calendar year
for an individual organization. With a small number of strata, as is the
case in this example, a spreadsheet is a good tool for computing different
allocations.

To illustrate the difference that cost can make in the allocation to strata,
Table 3.3 shows the proportions of the total sample that would be allocated
with the Neyman allocation and with an allocation that uses the unit costs in
the c_h column. Neyman allocates about 73 % (0.346 + 0.386) of the sample
to the psychiatric and multiservice or substance abuse hospitals. After con-
sidering cost, these two strata account for only 60 % of the sample (a 13 %
point reduction) because the cost per organization is higher than for other
strata.

Table 3.2: Statistics on total expenditures for a population of mental health organi-
zations.

Stratum h	Organization type	N_h	Mean $\bar{y}_{U_h}$	Standard deviation S_h	Population coefficient of variation $S_h / \bar{y}_{U_h}$
1	Psychiatric hospital	215	21,240,408	26,787,207	1.261
2	Residential	65	10,024,876	10,645,109	1.062
3	General hospital	252	4,913,008	6,909,676	1.406
4	Military veterans	50	11,927,573	11,085,034	0.929
5	Partial care or out-patient	149	6,118,415	9,817,762	1.605
6	Multiservice or sub-stance abuse	144	15,567,731	44,553,355	2.862
Total		875	11,664,181		

We can also compute the total sample sizes that would be implied by
different budgets or precision targets. For maximum variable-cost budgets,
$C - c_0$, of $100,000 and $200,000, the total sample sizes are 119 and 238, as
shown below. If the target $CV(\bar{y}_{st})$ is set to a value CV_0, then V_0 in Eq. (3.25)
is $(CV_0 \times \bar{y}_U)^2$. Using this to evaluate Eq. (3.25) gives sample sizes of 406 and
198 for CV targets of 0.05 and 0.10.

---

[6] Substance Abuse and Mental Health Services Administration, http://www.samhsa.
gov/.

Table 3.3: Neyman and cost-constrained allocations for the mental health organizations for estimating the mean of total expenditure.

Stratum h	Organization type	Cost c_h	Neyman $\frac{n_h}{n} = \frac{W_h S_h}{\sum_{h=1}^{H} W_h S_h}$	Cost- or precision-constrained $\frac{n_h}{n} = \frac{W_h S_h / \sqrt{c_h}}{\sum_{h=1}^{H} \left(W_h S_h / \sqrt{c_h} \right)}$
1	Psychiatric hospital	1,400	0.346	0.257
2	Residential	200	0.042	0.082
3	General hospital	300	0.105	0.168
4	Military veterans	600	0.033	0.038
5	Partial care or Outpatient	450	0.088	0.115
6	Multiservice or substance abuse	1,000	0.386	0.339
Total			1.000	1.000

Budget $(C - c_0)$	Sample size from Eq. (3.22)	CV target	Sample size from Eq. (3.25)
$100,000	119	0.05	406
$200,000	238	0.10	198

The R code for the Neyman allocation (using an arbitrary total sample size of 100) is

```
Nh <- c(215, 65, 252, 50, 149, 144)
Sh <- c(26787207, 10645109, 6909676, 11085034, 9817762, 44553355)
strAlloc(n.tot = 100, Nh = Nh, Sh = Sh, alloc = "neyman")
```

The cost-constrained allocations with variable costs of $100,000 and $200,000 are computed with

```
ch <- c(1400, 200, 300, 600, 450, 1000)
strAlloc(Nh = Nh, Sh = Sh, cost = 100000, ch = ch,
         alloc = "totcost")
strAlloc(Nh = Nh, Sh = Sh, cost = 200000, ch = ch,
         alloc = "totcost")
```

The allocations with CV targets of 0.05 and 0.10 are returned by

```
strAlloc(Nh = Nh, Sh = Sh, CV0 = 0.05, ch = ch,
         ybarU = 11664181, alloc = "totvar")
strAlloc(Nh = Nh, Sh = Sh, CV0 = 0.10, ch = ch,
         ybarU = 11664181, alloc = "totvar")
```

As for all R functions, the output can be assigned to an object for further manipulation. For instance, the components of

```
alloc1 <- strAlloc(Nh = Nh, Sh = Sh, CV0 = 0.05,
        ch = ch, ybarU = 11664181, alloc = "totvar")
```

as shown by `names(alloc1)`, are `alloc$allocation`, `alloc$nh`, and `alloc$'nh/n'`. ∎

Allocations for Comparing Stratum Means

The allocations described above were designed to be good for overall population estimates. However, individual stratum estimates or the difference in stratum estimates may be just as important. Cochran (1977, Sect. 5A.13) suggests two criteria that could be used in such cases. One is to minimize the average variance of the difference between all $H(H-1)/2$ pairs of strata. Assuming that stratum per-unit costs are equal, the optimal stratum sample sizes are

$$n_h = n \frac{S_h}{\sum_{h=1}^{H} S_h}. \tag{3.26}$$

This is similar to Neyman allocation in being proportional to the stratum standard deviations but, unlike Neyman, is unaffected by the stratum sizes W_h.

A second criterion would be to require that the variance of the estimator of the difference in any two stratum means be the same. In this case, the optimal allocation to stratum h is

$$n_h = n \frac{S_h^2}{\sum_{h=1}^{H} S_h^2}, \tag{3.27}$$

which assigns a larger fraction of the sample to the high variance strata than does Eq. (3.26).

Example 3.10 (Allocations for stratum estimates). Continuing with the previous example, the results of calculating the allocations for the mental health organizations based on the criteria in Eqs. (3.26) and (3.27) are shown in Table 3.4. These allocations are both more extreme than those in Table 3.3 in assigning more sample to stratum 6. Stratum 3 also gets only 0.015 of the total when allocating in proportion to S_h^2 due to its relatively small stratum variance. Based on other considerations, like the desire to analyze general hospitals separately, this allocation may be unsatisfactory to many analysts. ∎

Bear in mind that the examples above were developed to estimate the mean of one variable—total expenditures. Other variables may be just as important to analysts, and efficient allocations for them may be quite different from the ones we just calculated for expenditures. Chapter 5 will cover sample allocation tasks using more than one analysis variable.

Table 3.4: Allocations for the mental health organizations to optimize comparisons of stratum means of total expenditures.

		n_h / n	
Stratum h	Organization type	Allocation proportional to S_h	Allocation proportional to S_h^2
1	Psychiatric hospital	0.244	0.233
2	Residential	0.097	0.037
3	General hospital	0.063	0.015
4	Military veterans	0.101	0.040
5	Partial care or out-patient	0.089	0.031
6	Multiservice or sub-stance abuse	0.406	0.644
Total		1.000	1.000

3.2 Finding Sample Sizes When Sampling with Varying Probabilities

When samples are selected with varying probabilities, different methods are needed for sample size calculations. A useful device is to make sample size calculations based on the with-replacement variance formula as shown in Sect. 3.2.1. This formula is simpler than the without-replacement formulas, which involve joint selection probabilities. Thinking about model structure is another good way to determine sample sizes in some populations, as discussed in Sect. 3.2.2. If there are auxiliary variables on a frame that are good predictors of the variables to be collected in a survey, models for these relationships can be used in determining sample sizes. This section discusses the connection of *pps* sampling to models and the use of regression estimators of means and totals. Chapter 14 describes more extensively how to use models in estimation via calibration weighting. An interested reader can find in-depth coverage of the use of models in survey estimation in Valliant et al. (2000).

3.2.1 Probability Proportional to Size Sampling

Sampling units in proportion to some MOS can be extremely efficient in single-stage sampling for estimating totals if the MOS used for sampling is closely related to the analysis variable *y*. Texts usually distinguish between *pps* with-replacement sampling, denoted by *pps*, and without-replacement

sampling, denoted by πps. We will generally refer to either of these as *pps* but will be careful to distinguish between with-replacement and without-replacement variance formulas. Suppose that the relative size of unit i is p_i. For example, if the MOS in a hospital population is the number of beds, x_i, the relative size of hospital i is $p_i = x_i / \sum_U x_i$. If a fixed size sample of n units is selected without replacement, the selection probability is $\pi_i = np_i$. We will also refer to this method of sampling when the MOS is x as $pp(x)$ sampling or, more generally, as $pp(\mathrm{MOS})$. The π-estimator of the mean, assuming that N is known, is defined in general as $\hat{\bar{y}}_\pi = N^{-1} \sum_s y_i / \pi_i$. In the special case of $\pi_i = np_i$, the π-estimator is

$$\hat{\bar{y}}_\pi = N^{-1} \sum_s \frac{y_i}{np_i} . \tag{3.28}$$

If each y_i were exactly proportional to x, say $y_i = \beta x_i$, then the π-estimator reduces to $\hat{\bar{y}}_\pi = \beta \bar{x}_U$ in every sample. But, with $y_i = \beta x_i$, the population mean of y is $\beta \bar{x}_U$; so, $\hat{\bar{y}}_\pi$ would be perfect in every sample. Less restrictively, if y_i follows the model,

$$E_M (y_i) = \beta x_i , \tag{3.29}$$
$$V_M (y_i) = v_i,$$

where the y_i's are independent and the v_i's are positive values, then $\hat{\bar{y}}_\pi$ is model unbiased in the sense that $E_M \left(\hat{\bar{y}}_\pi - \bar{y}_U \right) = 0$. In Eq. (3.29), $E_M (y_i)$ and $V_M (y_i)$ are the theoretical expectation (or average) and variance of y_i evaluated with respect to the specified model. A good practice when constructing estimators is to do some modeling to determine whether there are any covariates that can be used as measures of size and to create estimators with lower variance than the simple π-estimator as discussed in Sect. 3.2.2.

The variance of $\hat{\bar{y}}_\pi$ is complicated because it involves joint selection probabilities of pairs of units:

$$V \left(\hat{\bar{y}}_\pi \right) = N^{-2} \sum_{i \in U} \sum_{j \in U} (\pi_{ij} - \pi_i \pi_j) \frac{y_i}{\pi_i} \frac{y_j}{\pi_j} \tag{3.30}$$

(e.g., see Särndal et al. 1992). The term π_{ij} is the probability that units i and j are simultaneously in the sample. Details on variance estimation techniques in different situations are covered in Chap. 18.

Several methods are available for selecting samples with varying probabilities; not all of these allow the joint selection probabilities, π_{ij}, to be computed. Cochran (1977) reviews several methods for selecting samples of size $n_h = 2$. Two methods for samples of size larger than two are Sampford's and sequential *pps* (Chromy 1979). Section 3.7 covers some of the software packages available for selecting samples with varying probabilities.

Determining a Measure of Size

In single-stage sampling the MOS is associated directly with the units to be sampled—beds in a hospital survey, employees in a business survey, etc. In Chaps. 9 and 10, we will discuss assigning sizes to aggregated units, like counties, in multistage sampling. In this section, some of the thinking needed to assign an MOS in the single-stage case is covered. The key finding is due to Godambe and Joshi (1965). Their result says that under model (3.29) the most efficient MOS for *pps* sampling is proportional to $\sqrt{v_i}$. This assumes that a population total is estimated and an estimator is used that is unbiased when averaging over a model and a probability sampling design. Isaki and Fuller (1982) extended this to a linear model where $E_M(y_i) = \mathbf{x}_i^T \beta$ and $V_M(y_i) = v_i$ with $\mathbf{x}_i$ defined as a vector of x's (auxiliary variables), β defined as a vector of regression slopes of the same dimension as $\mathbf{x}_i$, and $\mathbf{x}_i^T$ is the transpose of the $\mathbf{x}_i$ vector. In that case, $\sqrt{v_i}$ is still the best MOS for *pps* sampling, assuming that a regression estimator of the population total is used. We describe regression estimators in more detail in Sect. 3.2.2 and later in Chap. 14.

A model that may fit some establishment or institutional populations reasonably well has a variance with the form, $V_M(y_i) = \sigma^2 x_i^\gamma$, where x_i is an MOS and γ is a power. Typical values of γ are in the interval [0,2]. With a specification of the regression mean, $E_M(y_i)$, γ can be estimated iteratively. First, the model is fit by ordinary least squares (OLS) and the residuals calculated. The squared residual, e_i^2, is an approximate estimate of $V_M(y_i)$, regardless of its form. When $V_M(y_i) = \sigma^2 x_i^\gamma$, the slope in a regression of $\log(e_i^2)$ on $\log(x_i)$, where log is the natural logarithm, is an approximate estimate of γ. Henry and Valliant (2009) give more detail along with applications. Two R functions that will iteratively estimate γ are `gammaFit` along with `gamEst` given in Appendix C. Note that `gamEst` is set up for a regression without an intercept. If an intercept is desired, the matrix $\mathbf{X}$, which is an input to `gammaFit`, must be defined to include a column of 1's. The parameters used by `gammaFit` are:

X	matrix of predictors
x	vector of x's in V(Y)
Y	vector of response variables
maxiter	maximum no. of iterations allowed
show.iter	show values of gamma at each iteration, TRUE or FALSE
tol	relative change in gamma used to judge convergence

Example 3.11 (Estimating the power γ). Figure 3.2 is a scatterplot of an *srswor* of units 7, 17, 30, 33, 62, 111, 139, 247, 370, and 393 from the `hospital` population. A model for y that fits fairly well for the hospital population is $E_M(y_i) = \beta_1 \sqrt{x_i} + \beta_2 x_i$, $V_M(y_i) = \sigma^2 x_i^\gamma$. First, assign x and
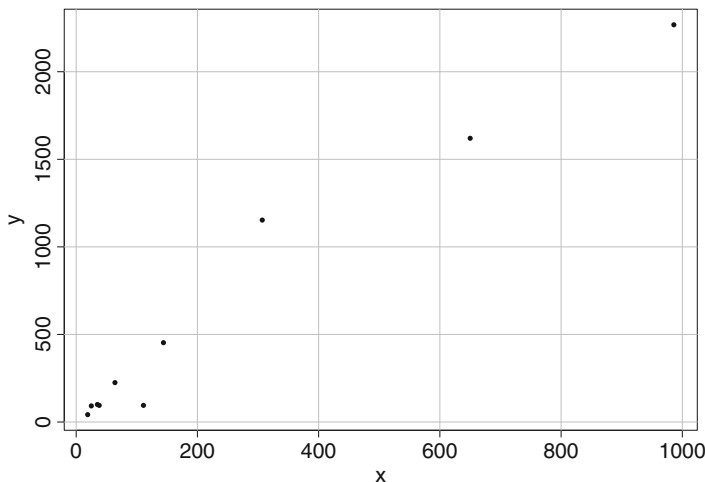
Fig. 3.2: Scatterplot of a sample of $n = 10$ sample units from the hospital population.

y to be the vectors of the ten values for these units. The matrix **X** contains columns for $\sqrt{x}$ and x. To estimate γ, the call to gammaFit and its output are

```
X <- cbind(sqrt(x), x)
gammaFit(X = X, x = x, y = y, maxiter=100, tol=0.001)

Convergence attained in 9 steps.
g.hat = 1.882531
```

In practice, the power might be rounded to 1.75 or 2 with the choice of 1.75 being selected since it would cause the MOSs to be less extreme than 2. Assuming that 1.75 is used, the MOS for *pps* would be $\sqrt{x_i^{1.75}}$. Another caution when using gammaFit is that in small samples, the algorithm may not converge. Setting the show.iter parameter to TRUE will print $\hat{\gamma}$ at each iteration, which may help in recognizing any problems. ∎

Calculations for With-Replacement Sampling

Expression (3.30) is obviously not too handy for computing a sample size. One practical approach is to use a variance formula appropriate for *pps* with replacement (*ppswr*) sampling. The simplest estimator of the mean that is usually studied with *ppswr* sampling is called "*p*-expanded with replacement" (Särndal et al., 1992, Chap. 2) and is defined as

$$\hat{\bar{y}}_{pwr} = \frac{1}{Nn} \sum_s \frac{y_i}{p_i} \ . \tag{3.31}$$

A unit is included in the sum as many times as it is sampled. Although Eq. (3.31) looks just like $\hat{\bar{y}}_\pi$ above, the selection probability of unit i is not np_i in with-replacement sampling; it is actually $1 - (1 - p_i)^n$. The variance of $\hat{\bar{y}}_{pwr}$ in *ppswr* sampling is

$$V\left(\hat{\bar{y}}_{pwr}\right) = \frac{1}{N^2 n} \sum_U p_i \left(\frac{y_i}{p_i} - T\right)^2 \equiv \frac{V_1}{N^2 n} \qquad (3.32)$$

where T is the population total of y. The obvious advantage of Eq. (3.32) when computing a sample size is that n is clearly separated from the other terms, unlike in expression (3.30).

If the desired coefficient of variation is CV_0, Eq. (3.32) can be solved to give the sample size as

$$n = \frac{V_1}{N^2} \frac{1}{\bar{y}_U^2 CV_0^2} . \qquad (3.33)$$

The difficulty with this formula is the estimation of V_1. As described in Sect. 3.4, V_1 can be estimated from a sample that was selected with the same relative measures of size, p_i, as to be used in the planned sample. Or, it can also be estimated from a *pps* sample that was selected with some other MOS.

Example 3.12 (Accounting for large units). Figure 3.3 plots total expenditures by the number of beds for the 671 organizations in the SMHO (`smho98`) population that have nonzero beds. The 204 units that reported zero beds provide only outpatient care. There is a fairly strong relationship between number of beds and expenditures with the correlation being 0.78. The gray line is a nonparametric smoother that is resistant to influence by unusual
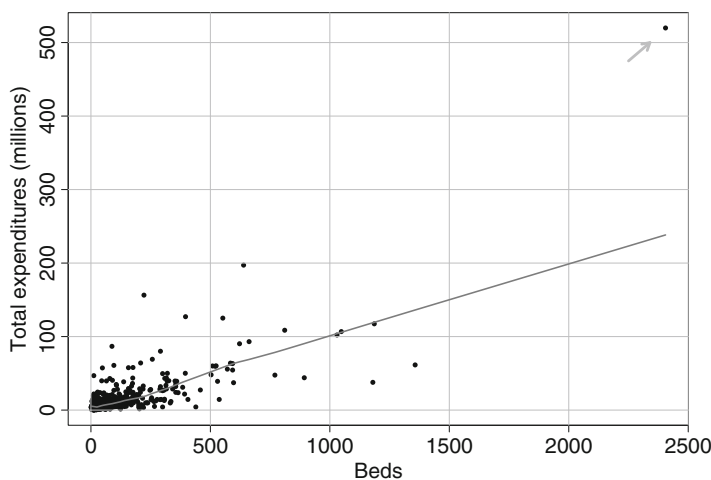
Fig. 3.3: Plot of total expenditures versus number of beds for the SMHO population. The *gray line* is a nonparametric smoother (lowess).

points. One point (marked by the arrow) with 2,405 beds is obviously much different than the others. Good practice is to select that organization for the sample with probability one. Such cases are variously called "take-alls," "certainties," or "self-representing," depending on the country of origin for the statistician. The general thinking is that a take-all is so unlike the others in the population that it should not be weighted-up to represent anything except itself. One useful rule of thumb that is often used is to compute the targeted selection probabilities for all units in the population and determine which units have values greater than or equal to one. In a *pps* sample with MOS x_i, this will occur if

$$x_i \geq \frac{N\bar{x}_U}{n} \ .$$

Sometimes this is relaxed to include all units with selection probabilities greater than some cutoff like 0.8. In that case, the take-alls would be units with $x_i \geq 0.8N\bar{x}_U/n$. Notice that these take-all cutoffs depend on how big the sample is; the larger the sample, the more units may be designated as take-alls.

If we set aside the big unit and select a *pps* sample from the remainder, the π-estimator of the mean will be

$$\hat{\bar{y}}_\pi = N^{-1}\left[(N-1)\,\hat{\bar{y}}_{\pi,nt} + y_{2405}\right] \ ,$$

where $\hat{\bar{y}}_{\pi,nt}$ is the π-estimator of the mean for the $N-1$ non-take-all units and y_{2405} is the total expenditures for the unit with 2,405 beds. More generally, if we had n_t take-alls, the estimator of the mean would be $\hat{\bar{y}}_\pi = N^{-1}\left[(N-n_t)\,\hat{\bar{y}}_{\pi,nt} + T_{yt}\right]$ where T_{yt} is the total of the Y's for the take-alls. The variance of $\hat{\bar{y}}$ is $\left(\frac{N-n_t}{N}\right)^2 V\left(\hat{\bar{y}}_{\pi,nt}\right)$ with $n_t = 1$ in this example since the big unit does not contribute to any sample-to-sample variability. But the CV of $\hat{\bar{y}}$ is still computed by dividing by $\bar{y}_U$:

$$CV\left(\hat{\bar{y}}\right) = \frac{N-n_t}{N}\sqrt{V\left(\hat{\bar{y}}_{\pi,nt}\right)}/\bar{y}_U \ .$$

To calculate a sample size, we approximate $V\left(\hat{\bar{y}}_{\pi,nt}\right)$ by the *pwr* variance in Eq. (3.32), i.e.,

$$V\left(\hat{\bar{y}}_{\pi,nt}\right) \doteq \frac{V_1}{(N-n_t)^2\,n},$$

where the V_1 in Eq. (3.32) refers only to the subuniverse of $N-n_t$ non-take-alls. The result is $V_1 = 9.53703\mathrm{e}{+}19$. The sample size formula (3.33) is then

$$n = \frac{9.53703\mathrm{e}{+}19}{671^2 \times 13,667,706^2 CV_0^2} \tag{3.34}$$

with $\bar{y}_U = 13,667,706$. For $CV_0 = 0.15$, Eq. (3.34) evaluates to $n = 51$ for the non-take-alls. ∎

Because the calculation in Example 3.12 is based on with-replacement sampling, the sample sizes may be conservatively large if a without-replacement sample with a sizeable *fpc* is actually selected. Kott (1988) gives an approximate method of inserting an *fpc* in the *pps* sampling formula that would help reduce this problem. Also, when certainties are identified as in Example 3.12 and probabilities are recomputed for the non-certainties, there may be additional units that have selection probabilities greater than 1. These should also be selected with certainty and the calculation in Eq. (3.34) recomputed for the remaining units. A few iterations may be needed to identify all of the take-alls. Alternatively, a cutoff like $x_i \geq 0.8 N \bar{x}_U / n$ could be used after the first iteration, which may eliminate some later rounds of iteration.

A final point on *pps* sampling is that it may be inefficient in single-stage sampling for estimating the proportion of units that have some characteristic. As noted earlier, $pp(x)$ sampling combined with the π-estimator or a regression estimator is efficient if y follows a linear model and the MOS is proportional to the model standard deviation. An appropriate model for a binary characteristic is typically nonlinear, e.g., logistic or complementary log–log, and not a straight-line like $E_M(y_i) = \beta x_i$. If the probability of having the characteristic does increase as the MOS increases, then $pp(\text{MOS})$ sampling may not be too bad. However, if a better model is that all units have a common probability or that different groups of units have different probabilities, $pp(\text{MOS})$ sampling will produce estimators with higher variances than *srswor* or *stsrswor*.

This is one of many illustrations that a given sampling plan cannot be ideal for all quantities that may be estimated in a survey. Finding compromises that are reasonably efficient for many different estimates is part of the art of good sample design. As we have said more than once, the mathematical programming tools in Chap. 5 will be extremely helpful in transforming the art into more of a science.

Relationship of *pps* Sampling to Stratification

Although *pps* sampling can be very efficient in some circumstances, it can have some practical disadvantages when some units do not respond. In establishment surveys, like those of businesses, schools, or hospitals, a target sample size of responders may be desired. Almost every survey faces some degree of nonresponse. Chapter 13 describes some of the mathematical ways of adjusting survey weights to attempt to correct the problem. Another method of dealing with nonresponse is to substitute another unit for any one that does not respond. This is especially common in surveys of schools. When *pps* sampling is used to select the initial units, a substitute may not have the same MOS as the original selection. This can lead to some ambiguity in assigning survey weights. Should the substitute receive the weight associated with the original selection? Or, should its weight be the one it would receive had it been an original selection itself? Another question is how to select the

substitutes themselves? Some of this uncertainty can be avoided by using stratified sampling in a way that approximates *pps* sampling.

Strata can be formed based on size as follows. Sort the frame from low to high based on the MOS. Determine the total sample size using Eqs. (3.33) or (3.37) to be described below. Divide the frame into $H = n/2$ strata such that the total of the MOS is about the same in each stratum. Then, select an *srswor* of size 2 in each stratum. If z_{hi} is the MOS for unit i in stratum h and the MOSs do not vary much within a stratum, the selection probability in stratum h will be

$$\pi_{hi} = \frac{2}{N_h} \doteq \frac{2z_{hi}}{N_h \bar{z}_h},$$

where $\bar{z}_h$ is the average MOS in stratum h, and we assume that $z_{hi} \doteq \bar{z}_h$. That is, the *stsrswor* selection probabilities are approximately the same as those in *pps* sampling. Using $n_h = 2$ is not essential but the more strata are created, the less the values of z_{hi} will vary within a stratum and the more likely it is that $z_{hi} \doteq \bar{z}_h$.

Since the sample is *stsrswor*, the sampling weight, π_{hi}^{-1}, is the same for each unit in stratum h. This means that substitutes can be selected by simple random sampling from the units that were not among the original sample and assigned the same weight as the originals. Of course, substitution is a form of imputation that affects variances in ways that may be difficult to reflect when making inferences. Consequently, devising a straightforward method of substitution does not solve all problems.

In addition, there are practical limits to the closeness of the *stsrs* selection probabilities to those from *pps*. If the population is fairly small, say, less than 500, and the measures of size used for *pps* sampling have a large range, the range of *pps* selection probabilities may itself be large in some strata. In such cases, the common *srs* selection probability of units within a stratum may differ considerably from the *pps* probabilities for some units.

Example 3.13 (Creating strata with equal total MOS). In Example 3.11 the power γ in the model $E_M(y_i) = \beta_1 \sqrt{x_i} + \beta_2 x_i$, $V_M(y_i) = \sigma^2 x_i^\gamma$ was estimated to be 1.88 for the hospitals population. We round this down to 1.75 for this example. The code below will create $H = 10$ strata in the hospitals population by cumulating the sorted list of $\sqrt{x^{1.75}}$ and forming strata that have about the same total value of $\sqrt{x^{1.75}}$. Two units are to be selected from each stratum:

```
x <- hospital$x
g <- 1.75
H <- 10; nh <- 2
hosp.pop <- hospital[order(x), ]

xg <- sqrt(x^g)
N <- nrow(hosp.pop)

    # create H strata using cume sqrt(x^g) rule
cumxg <- cumsum(xg)
size <- cumxg[N]/H
brks <- (0:H)*size
```

```
strata <- cut(cumxg, breaks = brks, labels = 1:H)
Nh <- table(strata)

str.selprobs <- rep(nh,H) / Nh

        # selection probabilities for pp(sqrt(x^g))
pps.selprobs <- H*nh*xg / sum(xg)
round(cbind(Nh = Nh, stsrs = str.selprobs, pps.means =
    by(pps.selprobs,strata,mean)),4)

    Nh  stsrs pps.means

1  129 0.0155    0.0155
2   57 0.0351    0.0345
3   42 0.0476    0.0483
4   35 0.0571    0.0574
5   30 0.0667    0.0668
6   25 0.0800    0.0771
7   23 0.0870    0.0889
8   20 0.1000    0.0979
9   18 0.1111    0.1134
10  14 0.1429    0.1451
```

The last statement above lists the numbers of hospitals in each stratum, the selection probabilities when 2 units are selected via *srs* in each stratum, and the average stratum values of the probabilities if the sample were selected using $pp\left(\sqrt{x^{1.75}}\right)$. The average *pps* selection probabilities are very close to the *srs* probabilities in each stratum. Some efficiency will be lost with this *stsrs* plan compared to the optimal probabilities, but the loss may be small. Plus, an *stsrs* plan is attractive because of its simplicity. ∎

3.2.2 Regression Estimates of Totals

Models can also be used to construct estimates of means and totals that are more efficient than π-estimators. Thinking about a model that may describe the dependence of y on an x can also be a useful way of computing a sample size. Details of this approach, given in Särndal et al. (1992, Chap. 12) and Valliant et al. (2000, Sect. 4.4), are sketched here. There is also a particularly useful connection between the model calculations that follow and *pps* sampling, as we will see. Suppose that the following linear regression model holds:

$$E_M\left(y_i\right) = \sum_{j=1}^{p} \beta_j x_{ji} , \qquad (3.35)$$

$$V_M\left(y_i\right) = \sigma^2 v_i,$$

where the subscript M means that the calculation is made with respect to a model, the β_j's are slope parameters, x_{ji} is the jth auxiliary variable associated with unit i, and v_i is a positive value. A design-based estimator of the population mean of y that is unbiased under this model is the *general regression estimator* (GREG), defined by

$$\hat{\bar{y}}_r = \hat{\bar{y}}_\pi + \sum_{j=1}^{p} b_j \left(\bar{x}_{Uj} - \hat{\bar{x}}_{\pi j} \right),$$

where b_j is the estimate of β_j using survey-weighted least squares, $\bar{x}_{Uj}$ is the population mean of x_j, and $\hat{\bar{x}}_{\pi j}$ is the π-estimator of the mean of x_j. (We will cover calculation of survey, or design, weights in Part III. For this discussion, you can think of b_j as simply a type of weighted least-squares estimator.) The "anticipated variance" (see Isaki and Fuller, 1982) is a variance computed over both the sample design and model. In the case of the GREG with *pps* without-replacement (*ppswor*) sampling and under model (3.35), the optimal selection probabilities, i.e., the ones that minimize the anticipated variance, are

$$\pi_i = \frac{n v_i^{1/2}}{N \bar{v}_U^{1/2}}$$

with $\bar{v}_U^{1/2} = \sum_U \sqrt{v_i} / N$. With these optimal probabilities, the approximate anticipated variance itself is

$$AV\left(\hat{\bar{y}}_r\right) \doteq \left[\frac{1}{n} \left(N \bar{v}_U^{(1/2)} \right)^2 - N \bar{v}_U \right] \sigma^2, \tag{3.36}$$

where $\bar{v}_U = \sum_U v_i / N$. Dividing by $[E_M(\bar{y}_U)]^2$, we get a kind of relvariance. Setting the result equal to CV_0^2 and solving for n leads to

$$n = \frac{\left[\bar{v}_U^{(1/2)} \right]^2}{CV_0^2 \frac{[E_M(\bar{y}_U)]^2}{\sigma^2} + \frac{\bar{v}_U}{N}}. \tag{3.37}$$

Exactly the same sample size formula can be derived using purely model-based arguments. In model (3.35), v_i and $\sqrt{v_i}$ must both be linear combinations of some or all of the x's to get the result. First, we look at a simple example to illustrate the model structure that is needed. If the model is

$$E_M(y_i) = \beta_1 \sqrt{x_i} + \beta_2 x_i, \tag{3.38}$$
$$V_M(y_i) = \sigma^2 x_i,$$

this fits the required structure since $v_i \propto x_i$, $\sqrt{v_i} \propto \sqrt{x_i}$, and both x_i and $\sqrt{x_i}$ are part of $E_M(y_i)$. This model allows a curved relationship between y and a single x with the amount of curvature depending on the slope coefficients. Models like this one often fit relationships in establishment populations well.

Under model (3.35), the best model-based estimator of the mean has the form $\hat{\bar{y}}_M = N^{-1} \left(\sum_{i \in s} y_i + \sum_{i \notin s} \hat{y}_i \right)$ with $\hat{y}_i = \sum_{j=1}^{p} \hat{\beta}_j x_{ji}$ and $\hat{\beta}_j$ being a weighted least-squares estimator of β_j. The ideal weights are inversely proportional to $v_i = x_i$, unlike the survey-weighted least-squares estimator which is a function of the design weights. The estimator of the mean uses the sum of y for the sample units ($i \in s$), which is observed, and predicts the y's for the nonsample units ($i \notin s$). The best sample for this estimator is one that is "balanced" on v_i and $\sqrt{v_i}$ in a certain way (Valliant et al., 2000, Theorem 4.2.1). In particular, the sample means of v_i and $\sqrt{v_i}$ should be the same as the ones obtained on average in $pp\left(\sqrt{v_i}\right)$ sampling. With the particular form of the model variance where v_i and $\sqrt{v_i}$ are linear combinations of the x's and with a balanced sample, the sample size needed to achieve a coefficient of variation of CV_0 is given by Eq. (3.37). The next example illustrates the calculation with the smho98 population.

Example 3.14 (Sample size calculation using a model). As an illustration, we regress total expenditures (EXPTOTAL) from the smho98 population on number of beds (BEDS) and the square root of number of beds with the variance specification in Eq. (3.38). The one large organization and all organizations with 0 beds are removed, leaving 670. The R code for doing this is listed below:

```
          # Isolate certainty selections (i.e., size > 2000)
cert <- smho98[,"BEDS"] > 2000

          # Remove certainties and size=0
tmp   <- smho98[!cert, ]
tmp   <- tmp[tmp[, "BEDS"] > 0, ]

          # Create model variables
x <- tmp[ , "BEDS"]
y <- tmp[ , "EXPTOTAL"]

          # Object containing model results
m <- glm(y ~ 0 + sqrt(x) + x, weights = 1/x)

          # Model results
summary(m)
```

Part of the output is

```
Call:
glm(formula = y ~ 0 + sqrt(x) + x, weights = 1/x)

Coefficients:
        Estimate Std. Error t value Pr(>|t|)
sqrt(x)  1044992      98955  10.560  < 2e-16 ***
x          34677       9612   3.607 0.000332 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(Dispersion parameter for gaussian family taken to be
    1.723118e+12)

    Null deviance: 2.3973e+15  on 670  degrees of freedom
Residual deviance: 1.1510e+15  on 668  degrees of freedom
```

The coefficients for both $\sqrt{x_i}$ and x_i are highly significant. The estimate of σ^2 in Eq. (3.38) is the residual deviance divided by its degrees of freedom or $(1.1510e+15)/668 = 1.723054e+12$. Using the same set of 670 units, the means of x, $\sqrt{x}$, and y are 105.97, 8.84, and 12,912,191. If we want a CV of 0.15 as in Example 3.11, then

$$n = \frac{8.84^2}{0.15^2 \frac{12,912,191^2}{1.723118 \times 10^{12}} + \frac{105.97}{670}} \doteq 34 .$$

An alternative to using $\bar{y}_U$, the mean of y, would be the average of the model predictions. However, in model (3.35), the special variance structure means that the two alternatives are equal. Continuing with the program above, the simple R code to compute the sample size is

```
N <- nrow(tmp)
mean(x)
mean(sqrt(x))

          # Estimate of sigma squared
sig2 <- m$deviance/m$df.residual

          # Sample size n for CV = 0.15
n <- mean(sqrt(x))^2 / (0.15^2 * mean(y)^2 / sig2 + mean(x)/N)
```

The sample size of 34 is less than the $n = 51$ found in Example 3.11. The reason for this is that the GREG and the prediction estimator are more efficient than the π-estimator since both take more advantage of the ability to predict y based on the value of x. ∎

One of the simplest estimators that flows out of a model is the ratio estimator. The ratio estimator of a mean in an *srswor* is

$$\bar{y}_R = \bar{y}_s \bar{x}_U / \bar{x}_s .$$

This estimator is a special case of the GREG when the model is $E_M(y_i) = \beta x_i$, $V_M(y_i) = \sigma^2 x_i$. Its approximate relvariance in *srswor* is

$$[CV(\bar{y}_R)]^2 = N^2 \left(\frac{1}{n} - \frac{1}{N} \right) \frac{S_R^2}{\bar{y}_U^2},$$

where $S_R^2 = (N-1)^{-1} \sum_U r_i^2$ with $r_i = y_i - x_i(\bar{y}_U / \bar{x}_U)$. Setting the CV to a target value, CV_0, and solving for n yields

$$n = \left[CV_0 \frac{\bar{y}_U^2}{S_R^2} + \frac{1}{N} \right]^{-1} , \qquad (3.39)$$

which is the same as Eq. (3.4) with S^2 replaced with S_R^2. Thus, the function nCont can be used.

Example 3.15 (Sample size for a ratio model). As in Example 3.14, suppose the mean of total expenditures (y) in the `smho98` population is to be estimated using the number of beds (x) and assume the model is a straight-line through the origin with variance proportional to x. As in the previous example, we remove the one large organization and all organizations with 0 beds. The R code for computing the sample size is:

```
m <- glm(y ~ 0 + x, weights = 1/x)
ybarU <- mean(y)
S2R <- sum(m$residuals^2/(length(x)-1))
nCont(CV0=0.15, S2=S2R, ybarU=ybarU, N=670)
[1] 51.16394
```

A sample of $n = 51$ is larger than $n = 34$ calculated in Example 3.13 because the ratio estimator is less efficient than the regression estimator used in that example. ∎

3.3 Other Methods of Sampling

Systematic sampling is often used in practice because it is fairly easy to implement and it can be used to control the distribution of a sample across a combination of auxiliary variables. For example, a field data collector might have to select a systematic sample from a list of addresses compiled by walking around a neighborhood. Selecting systematically in the field could speed the process of both sampling and data collection. Carrying it out in the field may also be less error-prone than more complicated selection methods. In other cases, it is used even though other methods could easily be implemented whose statistical properties are more well defined.

The method requires a list of units sorted in some order. Systematic sampling can be used to select equal probability samples or *pps* samples. The sampler starts somewhere on the list and skips down the list picking every kth($k = 10$ or 12 or 20, etc.) unit depending on the method. Various ways of selecting samples systematically are given in many books. As Cochran (1977, Chap. 8) notes, systematic sampling can have the characteristics of simple random sampling, stratified sampling, or cluster sampling depending on how the list is sorted. One of the most common uses of systematic sampling is to sort by some set of covariates in order to implicitly stratify units by the sorting variables. The sorting variables are implicit stratification variables in contrast to the design strata that have sample sizes explicitly defined by the sample design. For example, a frame of schools might be explicitly stratified by grade level (elementary, middle, high school). Within grade level, the schools might be sorted by urbanicity (urban/suburban/rural location), and by number of students within urbanicity. If an equal probability of selection method is used, the resulting systematic sample will contain an approximate

proportional representation of the units within the domains formed by the cross of the implicit stratification variables. Thus, the sample is controlled for more than the design strata without forming a large number of small strata that can inflate the variation in the weights (see discussions in Chap. 14).

The mathematical problem with systematic sampling is that no design-unbiased variance estimator can be constructed (see Särndal et al., 1992, Chap. 3). The general reason for this is that $\pi_{ij} = 0$ for some pairs of units. If the sorting is used to create implicit strata, the intuitive reason that an unbiased variance estimator does not exist is that only one unit is selected from a systematic selection interval. Regardless of the reasons for its use, statisticians usually collapse the selection intervals into one or more analytic strata and pretend the method of selection was something else, like *stsrswor*, *stsrswor*, or *ppswr*, in order to estimate a variance and to calculate a sample size. Thus, special sample size formulas are not needed for systematic sampling.

Poisson sampling is another technique in which units can be assigned different selection probabilities. Suppose that π_i is the probability assigned to unit $i \in U$. Each unit in the population is given an independent chance of selection. The sample size is random, which is one drawback of the method. However, it is especially useful in selecting a sample from a population where the frame must be compiled over an extended period of time. For example, in 2004, the US IRS received over 130 million tax returns for individuals and selected a sample of about 200,000 returns using Poisson sampling (Henry et al. 2008). Because people file returns for a particular tax year over an entire calendar year (and often beyond), the Poisson method allows the sampling to be done on a flow basis throughout the year rather than waiting until all returns are filed.

A typical implementation of Poisson sampling is to divide the population into groups. All units in a group are assigned the same selection probability. In this case, the sampling method in each group is called Bernoulli sampling. As shown in Särndal et al. (1992), conditional on the sample size in each group, the sample can be treated as if it were selected using *stsrswor*. Consequently, the sample size analyses for stratified simple random sampling can be used. The sample size found for each stratum would be set equal to the expected size under Bernoulli sampling. This would, in turn, determine the probability to be used for each unit in a group because $E(n_h) = N_h \pi_h$ where N_h is the frame count in stratum h and π_h is the common selection probability for units in the stratum.

3.4 Estimating Population Parameters from a Sample

The sample size formulae in Sects. 3.1 and 3.2 all involve population parameters. These must be estimated from a previous sample or from a secondary data set. If the previous sample was selected in the same way as the planned

sample, estimation is straightforward. If a different type of sample is planned from the earlier one, things are more complicated.

First, suppose that the earlier sample, s_0, was an *srswor* of size n_0. The unbiased estimators of $\bar{y}_U = \sum_{i=1}^{N} y_i / N$ and $S^2 = \sum_{i=1}^{N} (y_i - \bar{y}_U)^2 / (N-1)$ are then defined as

$$\bar{y}_{s_0} = \sum_{i \in s_0} y_i / n_0 \text{ and}$$

$$\hat{S}^2 = \sum_{s_0} (y_i - \bar{y}_{s_0})^2 / (n_0 - 1) .$$

In the special case of a binary variable, $\bar{y}_{s_0}$ reduces to the sample proportion p_0 and $\hat{S}^2 = n_0 p_0 (1 - p_0) / (n_0 - 1)$. If the planned sample is to be stratified, stratum variances must be estimated. Since s_0 is an *srswor*, the set of sample units in any domain (e.g., a stratum) is an equal probability sample from the domain. The number of sample cases in the domain is random, but there is an inferential argument that allows us to condition on the number units actually observed in each domain. As long as the achieved sample size is greater than 1, we estimate the mean and variance in a stratum h as

$$\bar{y}_{s_{0h}} = \sum_{i \in s_{0h}} y_i / n_{0h} \text{ and}$$

$$\hat{S}_h^2 = \sum_{i \in s_{0h}} (y_i - \bar{y}_{s_{0h}})^2 / (n_{0h} - 1) ,$$

where s_{0h} is the set of n_{0h} sample units in stratum h from the earlier study. If y is binary, we have similar reductions to those above: $\bar{y}_{s_{0h}} = p_{s_{0h}}$ and $s_h^2 = n_{0h} p_{0h} (1 - p_{0h}) / (n_{0h} - 1)$.

In some cases, we will have no microdata but an estimate of variance, $v(\bar{y}_{s_0})$, (or its square root) may be published. Assuming again that s_0 was an *srswor* of size n_0, the unit variance can be estimated as

$$\hat{S}^2 = \frac{n_0 v(\bar{y}_{s_0})}{1 - f_0},$$

where $f_0 = n_0 / N$. If the previous sample was more complex than *srswor* but we have a design effect for the estimated mean, $\hat{\bar{y}}$, then

$$\hat{S}^2 = \frac{n_0 v\left(\hat{\bar{y}}\right)}{1 - f_0} \frac{1}{deff\left(\hat{\bar{y}}\right)}, \tag{3.40}$$

where $deff\left(\hat{\bar{y}}\right)$ is the design effect for $\hat{\bar{y}}$. This assumes that the *deff* refers to without-replacement sampling. To approximate $f_0 = n_0 / N$, we may have to either estimate N with $\hat{N} = \sum_{i \in s_0} w_i$ or get the information from some published, secondary source. If the sampling fraction in the earlier survey is negligible or the published *deff* uses an *srswr* variance in its denominator, then just set $f_0 = 0$.

Now, consider the case where a *pps* sample of size n was selected using MOS's $\{p_i\}_{i=1}^{N}$. Even though s_0 was probably selected without replacement, the standard work-around is to treat the design as if it was *ppswr*. The estimate of the parameter V_1 in Eq. (3.32) is

$$\hat{V}_1 = \frac{1}{n-1} \sum_{s_0} \left(\frac{y_i}{p_i} - \frac{1}{n} \sum_{s_0} \frac{y_i}{p_i} \right)^2 \tag{3.41}$$

$$= \frac{n^2}{n-1} \sum_{s_0} \left(w_i y_i - \frac{1}{n} \sum_{s_0} w_i y_i \right)^2,$$

where $w_i = (np_i)^{-1}$. If the plan is to select the new sample with another set of probabilities $\{q_i\}_{i=1}^{N}$, then the new V_1 can still be estimated. The new V_1 is

$$V_1 = \sum_{U} q_i \left(\frac{y_i}{q_i} - t_U \right)^2 = \sum_{U} \frac{y_i^2}{q_i} - t_U^2. \tag{3.42}$$

The term $\sum_{U} y_i^2 / q_i$ is a population total and can be estimated by $n^{-1} \sum_{s_0} y_i^2 / (q_i p_i)$. An unbiased estimator of Eq. (3.42) is

$$\hat{V}_1 = \frac{1}{n} \sum_{s_0} \frac{y_i^2}{q_i p_i} - \left(\frac{1}{n} \sum_{s_0} \frac{y_i}{p_i} \right)^2 + v\left(\hat{t}_\pi\right) \tag{3.43}$$

where $v\left(\hat{t}_\pi\right)$ is the estimated variance of $\hat{t}_\pi = n^{-1} \sum_{s_0} y_i / p_i$. The third term on the right-hand side of Eq. (3.43) is a bias-correction term that will often be negligible compared to the other terms. The theory behind these estimators can be found in Särndal et al. (1992, Result 2.9.1). One problem with Eq. (3.43) is that it can be negative, which is, of course, impossible for a population variance. This predicament is more likely to happen in small samples than in large ones.

If s_0 is selected with varying probabilities (not necessarily *pps*) and the inverse selection probabilities are $\{w_i\}_{i \in s_0}$, the unit variance parameter can also be estimated approximately as

$$\hat{S}^2 = \frac{n}{n-1} \frac{\sum_{s_0} w_i (y_i - \bar{y}_w)^2}{\sum_{s_0} w_i - 1}, \tag{3.44}$$

where $\bar{y}_w = \sum_{s_0} w_i y_i / \sum_{s_0} w_i$. This expression also applies to estimating the stratum population variance, S_h^2, based on the sample s_{0h}. The estimator $\hat{S}^2$ does have a negative bias, although the problem will be an issue only in small samples.

Example 3.16 (Estimating unit variance for ppswr sampling). A sample of 20 from the hospital population was selected with probability proportional to the number of hospital beds (x_i), $pp(x)$, in order to estimate the average number of discharges (y_i). The data are listed in Table 3.5. We compute $\hat{V}(\hat{\bar{y}}_{pwr}) \equiv \hat{V}_1/(N^2 n)$ for the $pp(x)$ sample of size $n = 20$ from a total of $N = 393$ hospitals. The probabilities of inclusion, $\pi_i = np_i$, are calculated with $p_i = x_i/\sum_U x_i$ where $\sum_U x_i = 107,956$. The weights are calculated as the inverse of the π_i's, i.e., $w_i = (np_i)^{-1}$.

The estimate $\hat{\bar{y}}_{pwr}$ is calculated as 813.1. To estimate the sample variance of the *pwr*-estimator, we first calculate $\hat{V}_1$ in Eq. (3.41) as $\hat{V}_1 = 11,001,669,955$. Substituting this value into the $\hat{V}(\hat{\bar{y}}_{pwr})$ formula (3.32), we have

$$v(\hat{\bar{y}}_{pwr}) \equiv \frac{11,001,669,955}{393^2 \times 20} = 3561.587$$

and a CV estimate of 0.073.

Now, suppose that we plan to select a future sample with probabilities proportional to the square root of beds. Estimator (3.43) applies with $q_i = \sqrt{x_i}/\sum_U \sqrt{x_i}$ and $p_i = x_i/\sum_U x_i$:

$$\hat{V}_1 = \frac{1}{n}\left(\sum_U \sqrt{x_i}\right)\sum_{s_0}\frac{y_i^2}{\sqrt{x_i}p_i} - \left(\frac{1}{n}\sum_{s_0}\frac{y_i}{p_i}\right)^2 + v(\hat{\bar{y}}_{pwr})$$

$$= \frac{5992.3}{20}410,727,850 - 319,545^2 + 3,561.6^2$$

$$= 20,950,895,199,$$

which, in a sample of $n = 20$, would lead to an anticipated CV for either the total or the mean of $\sqrt{20,950,895,199/20}/319,545 = 0.101$.

Table 3.5: Sample data for 20 hospitals selected with probabilities proportional to the number of hospital beds.

Population ID	Discharges y_i	Beds x_i	Population ID	Discharges y_i	Beds x_i
76	244	70	320	1,239	472
155	402	160	321	1,258	474
192	732	227	329	1,657	498
200	925	235	354	2,116	562
228	632	275	360	1,326	584
243	557	300	369	1,606	635
253	1,226	310	373	1,707	670
289	896	378	376	2,089	712
297	2,190	400	378	1,283	760
315	1,948	461	381	1,239	816

∎

Example 3.17 (Estimating unit variance for srswor sampling). Continuing the previous example, suppose that we consider selecting an *srswor* from the hospital population and using the sample mean for discharges as the estimator. The sample size required to hit a specified CV is in expression (3.4). Thus, we need to estimate the unit variance S^2 using Eq. (3.44). Evaluating this with the data for the ten sample hospitals in Table 3.5, we obtain $\hat{S}^2 = \frac{20}{19} 134,350,622 / 341.478 = 414,145.8$. The anticipated CV for mean discharges in a sample of 20 is then $\sqrt{(1 - 20/393)\, 414,145.8/20}\, /813.1 = 0.172$. ∎

In these examples, either $pp(x)$ or $pp(\sqrt{x})$ sampling together with the π-estimator is more efficient than *srswor* because of the strong relationship between discharges and beds. Using a regression estimator as in Sect. 3.2.2, in conjunction with $pp(x)$, is likely to be even more efficient. A word of caution is in order, though. The estimates of the unit variances, V_1 and S^2, are themselves variable. Another sample s_0 of $n = 20$ may yield estimates that are different, and possibly quite different, from the ones above. Exercise 3.13 asks that you select several samples from the hospital population to get a feel for this.

3.5 Special Topics

Some specialized but nonetheless practical topics are sampling rare populations and making estimates for domains.

3.5.1 Rare Characteristics

Some analysts will be especially interested in estimating the occurrence of rare characteristics, like the prevalence of certain types of diseases or other unusual health conditions. Examples are the proportion of persons who have had a myocardial infarction in a given year or in their lifetimes, the proportion of the population that is blind, and the proportion of children with deficient blood iron levels. The rarer a characteristic is, the more difficult it will be to select a sample that will give reliable estimates. In fact, there may be a sizeable chance that a sample has no cases at all with the rare characteristic.

If p_U is the proportion that have a trait and selections are independent, the probability of obtaining no cases, i.e., ones that have the trait, in a sample of size n is $(1 - p_U)^n$. This calculation is appropriate for a simple random sample selected with replacement (*srswr*). If we want this probability to be no more than α, then the inequality

$$(1 - p_U)^n \leq \alpha$$

can be solved for the sample size to give

$$n \geq \frac{\log{(\alpha)}}{\log{(1 - p_U)}} \ .$$ (3.45)

(The inequality reverses since $\log{(1 - p_U)}$ is negative.) Table 3.6 below shows
that sample sizes and expected numbers of cases in the sample for $\alpha = 0.05$
and 0.01 for a range of values of the population prevalence. For extremely
rare characteristics, like $p_U = 1/100,000$ which is about the prevalence of
Addison's disease, a sample of nearly 300,000 would be needed to have only
a probability of 0.05 of not observing a case. Even with that size of sample,
the expected number of sample cases is only 3, which is not enough to be
worth analyzing.

Table 3.6: Sample sizes and expected numbers of cases with a rare characteristic.

α	p_U	n	np_U
0.05	0.10	28	2.8
	0.05	58	2.9
	0.03	98	3.0
	0.01	298	3.0
	0.005	598	3.0
	0.0001	29,956	3.0
	0.00001	299,572	3.0
0.01	0.10	44	4.4
	0.05	90	4.5
	0.03	151	4.5
	0.01	458	4.6
	0.005	919	4.6
	0.0001	46,049	4.6
	0.00001	460,515	4.6

A related problem is how to put a confidence bound on a proportion when
very few sample cases are observed to have the characteristic. Cochran (1977,
Sect. 3.6, Example 3) examines this problem using a hypergeometric distribu-
tion. In a population with N units, of which A have some rare characteristic,
e.g., an error in an audit of accounts, the probability that no units with the
characteristic are found in a sample of size n is

$$\frac{\binom{N-A}{n}}{\binom{N}{n}} = \frac{(N - A)(N - A - 1) \cdots (N - A - n + 1)}{(N - 1)(N - 2) \cdots (N - n + 1)} \doteq \left(\frac{N - A - u}{N - u}\right)^n ,$$

where $u = (n - 1)/2$. For $N = 1,000$, $n = 200$, and $A = 10$, this
approximation gives 0.107. That is, if the error rate is $A/N = 0.01$, the

probability of observing no errors in a sample of 200 is 0.107. Thus, we take $A = 10$ as the upper 90% confidence limit on the number of actual errors. Jovanovic and Levy (1997) cover an interesting method known as the "rule of three" which derives from the formula $(1 - p_U)^n \leq \alpha$ that led to Eq. (3.45). Setting this expression equal to α gives a kind of upper bound on how large p_U can be. Solving for p_U gives $p_U = 1 - \alpha^{1/n}$. A Taylor series expansion (see Sect. 15.3 for details on this type of expansion) gives $\alpha^{1/n} = 1 + \ln(\alpha)/n - [\ln(\alpha)]^2/(2n^2) + \cdots$. Retaining the first two terms gives the upper bound on p_U as

$$p_U \doteq -\ln(\alpha)/n .$$

When $\alpha = 0.05$, $-\ln(\alpha) \doteq 3$, which implies that a 95% upper confidence bound on p_U is about $3/n$. This is a handy rule of thumb for getting a quick bound on the proportion. Korn and Graubard (1998) and Kott and Liu (2009) deal with several, additional alternative methods.

For extremely rare traits, unrestricted random sampling is seldom a good idea. Large sample sizes may be needed to get acceptable precision for full population estimates. The problem is compounded if estimates for subgroups, like ones defined by age, gender, and region, are desired. Kalton (1993) gives a thorough review of the options that might be used for sampling. He distinguishes among rare characteristics, rare populations, mobile populations, population flows, and elusive populations. Stratification, use of multiple frames, multiplicity sampling, and two-phase sampling are some of the techniques available. We will touch on two-phase sampling in Chap. 17.

3.5.2 Domain Estimates

Most multipurpose surveys make separate estimates for domains or subpopulations. Kish (1987a) offered the following taxonomy of domains:

1. Design domains: subpopulations that are restricted to specific strata (e.g., Ontario in a survey in Canada where provinces are strata)
2. Cross-classes: groups that are broadly distributed across the strata and primary sampling units (PSUs) (e.g., African-Americans over the age of 50 in the USA)
3. Mixed classes: groups that are disproportionately distributed across the complex sample design (e.g., Hispanics in a sample that includes Los Angeles, an area with a large Hispanic population, as a geographical stratum)

A goal of some surveys is to sample a few domains at higher rates than they occur in the population. This is known as *oversampling*. If, for example, we want equal size samples of Whites and African-Americans in a household sur-

vey in the USA, we will have to sample the latter at a much higher rate than the former because Whites are a much larger proportion of the population.

A legitimate question is, If domains are going to be important for analysts, why not make each domain a design stratum so that the sample size in each can be controlled? There are a few reasons why this cannot always be done. First, the frame may not give domain membership for all units in advance of sampling (e.g., adults looking for work). Second, using the domains for strata may be impractical. The domains may not be disjoint. For example, we may want to analyze persons in domains defined by gender and race/ethnicity. Strata that account for both factors would have to be defined by the cross-classification of gender × race/ethnicity. When many domains are of analytic interest, the complete cross of all of them could be too cumbersome to use as individual strata.

In cases where the domain identifiers are available on the frame but explicit strata using all domains are not formed, practitioners often try to ensure representation of each by using systematic sampling. In our simple example, the frame might be sorted by gender and then by race/ethnicity within gender. A systematic, equal probability sample would be distributed by gender and race/ethnicity much like the population. This method would usually eliminate samples that are poorly distributed among the domains but would not oversample any domain.

Any time an analyst does a cross-tabulation, the cells in the table hold domain estimates. Thus, making domain estimates is a standard step in analyzing survey data. In a military personnel survey, for example, design strata might be branch of the service crossed with pay grade, while a domain could be the set of personnel who were stationed overseas at any time during the last 5 years. In a telephone survey of households, domains might be the groups of persons who report that they have a college degree or have had their homes burglarized in the last year. There can also be unintended reasons for an estimate to be treated as one for a domain. If a frame contains ineligible units, e.g., a business frame that has out-of-business listings, then the eligible units are a domain.

A key feature of the domain estimation problem is that domain membership for cross-classes and mixed classes is often not determined until data are collected. In such cases, the number of sample units in a domain is random and the total number of domain members in the population is typically unknown. This results in estimated domain means being constructed as the ratio of an estimated total divided by an estimate of the number of domain units in the population. Such ratio estimators require approximate methods for variance estimation described below.

In designing a sample to adequately cover the domains that are to be analyzed, there are two options. One is to calculate the expected numbers of units that will occur in the sample in each domain for a particular total sample size. The total sample size is then made large enough so that, in expectation, the key domains of interest will be adequately represented. For example,

according to the 2006 NHIS, about 14.8% of persons did not have any type of health insurance at the time of the interview.[7] If an equal probability sample of persons were selected and 1,000 persons were desired in the uninsured domain, we would need a sample of about 6,760 ($= 1,000/0.148$) to get 1,000 in expectation. There will, of course, be some sample variation in the number actually obtained. So, it would be prudent to select more than 6,760 to be safe.

The second option would be to select a two-phase sample, which we cover in Chap. 17. In the first phase, screening questions are administered to determine domain membership. At the second phase, units are subsampled at rates designed to obtain specified domain sample sizes. The subsampling rates will vary among the domains. Ideally, using a second phase allows the counts from the first phase to be tabulated before setting the second-phase rates. Having this flexibility allows much better control over the achieved sample sizes than does single-phase selection. In some surveys with tight time schedules, this advantage is diluted a bit because second-phase rates have to be set based on partial data from the first phase. Even in this case, two-phase sampling can be effective in controlling sample sizes for domains.

Suppose that a simple random sample is selected without replacement and that domain membership is unknown before sampling is done. The estimate of a domain total for a variable y is $\hat{t}_d = (N/n) \sum_s y_{di}$ where y_{di} is the value of the variable for a unit if it is in domain d and is 0 if the unit is not in the domain. This can also be written as $y_{di} = y_i \delta_i$ with $\delta_i = 1$ if unit i is in the domain and 0 if not. The variance of $\hat{t}_d$ is

$$V(\hat{t}_d) = \frac{N^2}{n}\left(1 - \frac{n}{N}\right)S^2,$$

where the unit variance is calculated including the zeros for non-domain units. The unit variance can be rewritten as $S^2 \doteq P_d\left(S_d^2 + Q_d\bar{y}_{Ud}^2\right)$ where S_d^2 is the variance among units that are in the domain, $\bar{y}_{Ud}$ is the population mean for those units, $P_d = N_d/N$ is the proportion of units in the population that are in the domain, and $Q_d = 1 - P_d$ (see Hansen et al. 1953a, Sect. 4.10; Cochran 1977, Sect. 2.11). Using this version of S^2, the relvariance of $\hat{t}_d$ is

$$CV^2\left(\hat{t}_d\right) \doteq \frac{1}{n}\left(1 - \frac{n}{N}\right)\frac{CV_d^2 + Q_d}{P_d}, \tag{3.46}$$

where $CV_d^2 = S_d^2/\bar{y}_{Ud}^2$ is the unit relvariance among the domain units. Setting Eq. (3.46) equal to a target value CV_0^2 and solving for n gives

$$n = \frac{CV_d^2 + Q_d}{P_d CV_0^2 + \frac{CV_d^2 + Q_d}{N}} \doteq \frac{CV_d^2 + Q_d}{P_d CV_0^2}. \tag{3.47}$$

---

[7] http://www.cdc.gov/nchs/data/nhis/earlyrelease/200706_01.pdf.

The approximation comes from assuming that the population size N is large. Notice that Eq. (3.47) reduces to the earlier formula (3.4) for a full population estimate when $P_d = 1$.

If the mean per domain unit is estimated, the required sample size is similar, but an approximate variance is needed. Suppose that the mean is estimated by $\hat{\bar{y}}_d = \hat{t}_d / \hat{N}_d$ where $\hat{N}_d = N n_d / n$. Linearly approximating $\hat{\bar{y}}_d$ leads to

$$\hat{\bar{y}}_d - \bar{y}_{Ud} \doteq \frac{1}{N_d} N \bar{e}_s,$$

where $\bar{e}_s = \sum_s e_i / n$ with $e_i = \delta_i (y_i - \bar{y}_{Ud})$. The approximate variance is then

$$V(\hat{\bar{y}}_d) \doteq \frac{1}{N_d^2} \frac{N^2}{n} \left(1 - \frac{n}{N}\right) S_e^2$$

with $S_e^2 = (N-1)^{-1} \sum_U e_i^2$. Since $e_i = y_i - \bar{y}_{Ud}$ for units in the domain $e_i = 0$ for non-domain units, $S_e^2 = (N-1)^{-1} \sum_{U_d} (y_i - \bar{y}_{Ud})^2 \doteq P_d S_d^2$. The relvariance of $\hat{\bar{y}}_d$ is then

$$CV^2(\hat{\bar{y}}_d) \doteq \frac{1}{nP_d} \left(1 - \frac{n}{N}\right) CV_d^2.$$

Setting this equal to CV_0^2 and solving for n gives

$$n = \frac{CV_d^2}{P_d CV_0^2 + \frac{CV_d^2}{N}} \doteq \frac{CV_d^2}{P_d CV_0^2}. \tag{3.48}$$

This sample size for estimating a mean can be substantially smaller than the one in Eq. (3.47) for estimating the domain total as illustrated in Table 3.7. For a small domain with unit relvariance of 1 ($CV_d^2 = 1$) an *srs* of 15,600 is required to obtain a CV for the estimated total of 0.05. However, a sample of 8,000 is needed to estimate the mean with a CV of 0.05. As the domain becomes more prevalent, i.e., P_d becomes larger, the sample sizes for totals and means become closer together.

Table 3.7: Sizes of simple random samples required to achieve a CV of 0.05 for estimated domain totals and means for different sizes of domains.

P_d	n for total	n for mean
0.05	15,600	8,000
0.25	2,800	1,600
0.50	1,200	800
0.75	667	533
1.00	400	400

The population size is assumed to be large; domain relvariance is $CV_d^2 = 1$.

Next, consider an *stsrswor* sample. The estimated mean for a domain is again defined as the estimated total for the domain $(\hat{t}_d)$ divided by an estimate of the number of units in the domain $(\hat{N}_d)$, i.e.,

$$\hat{\bar{y}}_d = \frac{\sum_h \sum_{i \in s_{dh}} w_{hi} y_{hi}}{\sum_h \sum_{i \in s_{dh}} w_{hi}} \equiv \frac{\hat{T}_d}{\hat{N}_d} ,$$

where w_{hi} is the sampling weight for unit hi and s_{dh} is the set of sample units in stratum h that are also members of domain d. In *stsrswor* the weight for a unit in stratum h is $w_{hi} = N_h / n_h$. Consequently, the domain mean can be specialized to

$$\hat{\bar{y}}_d = \frac{\sum_h W_h p_{dh} \bar{y}_{d,sh}}{\sum_h W_h p_{dh}},$$

where $p_{dh} = n_{dh} / n_h$ and $\bar{y}_{d,sh} = \sum_{s_{dh}} y_{hi} / n_{dh}$ with n_{dh} reflecting the number in the set s_{dh} of sample units in domain d within stratum h. The approximate variance of $\hat{\bar{y}}_d$ (see Cochran, 1977, Sect. 5A.14) is

$$AV(\hat{\bar{y}}_d) = \frac{1}{P_d^2} \sum_h \frac{W_h^2}{n_h} \left(1 - \frac{n_h}{N_h}\right) \left[\frac{N_{dh} - 1}{N_h - 1} S_{dh}^2 \right. \tag{3.49}$$
$$\left. + \frac{N_{dh}}{N_h - 1} \left(1 - \frac{N_{dh}}{N_h}\right) (\bar{y}_{U_{dh}} - \bar{y}_{U_d})^2 \right],$$

where $P_d = N_d / N$ is the proportion of units in the domain in the whole population, $P_{dh} = N_{dh} / N_h$ is the proportion in stratum h, $Q_{dh} = 1 - P_{dh}$, $\bar{y}_{U_{dh}} = \sum_{i \in U_{dh}} y_{hi} / N_{dh}$, U_{dh} is the population of domain units in stratum h, $\bar{y}_{U_d} = \sum_{h, U_{dh}} y_{hi} / N_d$, and $S_{dh}^2 = \sum_{i \in U_{dh}} (y_{hi} - \bar{y}_{U_{dh}})^2 / (N_{dh} - 1)$ is the variance among units in stratum h that are in the domain.

If the sample proportion of units in the domain, n_{dh} / n_h, is about the same as the population proportion, P_{dh}, then the approximate variance can be written more suggestively as

$$AV(\hat{\bar{y}}_d) \doteq \sum_h \left(\frac{P_{dh}}{P_d}\right)^2 \frac{W_h^2}{n_{dh}} \left(1 - \frac{n_h}{N_h}\right) \left[S_{dh}^2 + Q_{dh}(\bar{y}_{U_{dh}} - \bar{y}_{U_d})^2\right] . \tag{3.50}$$

When the domain is spread evenly over the strata so that $P_{dh} \doteq P_d$ (i.e., a uniformly distributed cross-class), this formula can be roughly interpreted as the sum of *(i)* the variance that would be obtained if we knew domain membership in advance and sampled a fixed number of domain units directly and *(ii)* a contribution due to the difference in the domain means among the strata. For the purpose of determining sample size, Eq. (3.50) is difficult to use. If $n_{dh} \doteq n_h P_{dh}$, this can be substituted in Eq. (3.50) to obtain an expression that depends only on the n_h's. The methods of allocating samples to strata covered in Sect. 3.1.2 can then be used by replacing S_h^2 with $S_h^{*2} = \frac{P_{dh}}{P_d^2} \left[S_{dh}^2 + Q_{dh}(\bar{y}_{U_{dh}} - \bar{y}_{U_d})^2\right]$. To use this substitution, quite a bit

of information is needed—the proportion of units in each stratum that is in the domain, the stratum variance among the domain units, and the mean per domain unit in each stratum. Thus, estimates of many population values are needed in advance of sampling. Alternatively, two-phase methods are a sound way of approximately controlling the sample sizes in domains. These methods do require special variance estimation methods to be covered later.

The formulas above do simplify if a domain consists of one or more design strata in their entirety, i.e., a design domain listed at the beginning of this section. In that case, $p_{dh} = P_{dh} = 1$ for strata in the domain and 0 otherwise. The domain mean in *stsrswor* specializes to

$$\hat{\bar{y}}_d = \frac{\sum_{h \in S_d} W_h \bar{y}_h}{\sum_{h \in S_d} W_h},$$

where S_d is the set of strata that are in the domain and $N_d = \sum_{h \in S_d} N_h$. Since $P_{dh} = 1$, the variance in Eq. (3.50) becomes

$$V\left(\hat{\bar{y}}_d\right) = \frac{1}{N_d^2} \sum_{h \in S_d} \frac{N_h^2}{n_h} \left(1 - \frac{n_h}{N_h}\right) S_h^2 . \tag{3.51}$$

In other words, the variance depends only on the contributions from the strata that are in the domain. In this case, a sample estimate of the variance in Eq. (3.51) is easily constructed by substituting s_h^2 for S_h^2 as long as N_d is known. The allocation to individual strata can be directly controlled so that desired levels of precision can be achieved in different strata.

3.6 More Discussion of Design Effects

Design effects can be used to adjust a sample size computed for a single-stage sample to, at least, approximate the size needed in a more complicated sample. The *deff* for some estimator $\hat{\theta}$ is defined as

$$deff\left(\hat{\theta}\right) = \frac{V\left(\hat{\theta}\right)}{V_{srs}\left(\hat{\theta}\right)},$$

where V denotes variance under whatever sample design is used (stratified, clustered, etc.) and V_{srs} is the *srs* variance of the *srs* estimator of the same population parameter. This notation is a bit imprecise because the estimate $\hat{\theta}$ is probably not computed in the same way in a simple random sample and in a more complex sample. If n is calculated using a simple random sampling formula, then $n \times deff$ is the sample size needed in the more complex design to achieve the same variance as the simple random sample.

In some designs this is a fairly crude calculation. For example, in a two-stage design in which clusters and elements within clusters are sampled, $n \times$ *deff* tells you nothing about how many clusters and elements per cluster should be sampled for an efficient allocation. In fact, the *deff* will not apply unless the new sample has the same numbers of clusters and elements per cluster as the one used to compute the *deff*.

If a *deff* is obtained from a software package, it is important to understand how the *deff* is computed. For example, SUDAAN (RTI International 2012) provides four different design effects that account for some or all of the effects of stratification, clustering, unequal weighting, and oversampling of subgroups. These may be informative after a sample has been selected to gauge the contribution to variance of the different factors. One of the most basic things to understand is whether the *srs* variance in the denominator of the *deff* is computed using a with-replacement or without-replacement formula. When the sampling fraction is large, these can be quite different. Often the sample that can be afforded is a small part of the population, so that *srswr* is the appropriate choice for the denominator.

However, *deff*'s from a previous survey may not be that useful when planning a new survey. You may not be repeating the same type of design for which the software computed *deff*'s. The strata and cluster definitions may be different. The desired sample sizes for subgroups may be different. The method of weighting (e.g., nonresponse adjustments and use of auxiliary data) that you will use may be different. If a new design will depart substantially from an old, the sample size methods in the following chapters that explicitly consider the effects of strata, precision goals for subgroups, variance components for multistage designs, and other design features should give more useful answers than simple *deff* adjustments.

3.7 Software for Sample Selection

In the past, a survey organization had to rely on computer programs developed by its own staff to draw the random samples. Thankfully, software is now available for this purpose, thus allowing statisticians more time for the design phase of the study. We review several functions for two of the software packages in the subsequent sections—R and SAS.

3.7.1 R Packages

The following is a list of some of the currently available R sampling functions grouped by package:

Package	Function	Description
base	sample	Select *srswr* or *srswor* samples
pps	ppss	Systematic *ppswor* sampling
	ppssstrat	Stratified *ppswor* systematic sampling
	ppswr	*pps* sampling with replacement
	stratsrs	*stsrswor*
sampling	cluster	Single-stage cluster sampling
	srswor	Select *srswor* samples
	srswr	Select *srswr* samples
	strata	Select *stsrswor*, *stsrswr*, Poisson, and systematic samples
	UPrandomsystematic	Systematic *ppswor* sampling after randomizing the order of the list
	UPsampford	Sampford's method of *ppswor*

For example, the function srswor(n,N) returns a sequence of zeros and ones where a one indicates the n units randomly selected without replacement from an ordered list of N units. pps (Gambino 2005), sampling (Tillé and Matei 2012), and samplingbook (Manitz 2012) packages offer other functions, not shown above, for selecting unequal probability samples.

Updates to the software, including new functions and new features for current functions, are made available through the R web site. User-defined functions are easily created as discussed in this and other chapters—see Appendix C for a complete list of author-defined R functions used in this text.

Example 3.18. [Select a stratified sample (stsrswor)]. We wish to select ten hospitals from each of the six strata in the smho98 data file using the R function strata from the sampling package. The following code illustrates how to import a SAS transport file (smho.xpt), create a new variable called stratum6 in the population object, and select an *stsrswor* using strata.

When reading data and doing specialized calculations, like creating the stratum6 variable, it is always wise to check your work by looking at the contents and size of the data file and tabulating summaries of derived variables. We show some of these steps in Examples 3.18 and 3.19 but will omit them from most other examples in this book. However, the reader should bear in mind that thorough checking is critical to doing high-quality work:

```
    # Load R libraries
require(foreign)
require(sampling)
    # Random seed for sample selection
set.seed(82841)
    # Load SAS transport file and examine
smho98 <- read.xport("smho98.xpt")
dim(smho98)
  [1] 875 378
```

```
smho98[1:5,1:5]
    STRATUM BEDS EXPTOTAL SEENCNT EOYCNT
  1       1   81  9066430    1791    184
  2       1   80  9853392    1870    244
  3       1   26  3906074    1273      0
  4       1   90  9853392    1781    154
  5       1   71  9853392    1839    206

    # Create 6-level stratum variable and verify
smho98$stratum6 <- 0
smho98[( 1<=smho98$STRATUM & smho98$STRATUM<=2), "stratum6"] <- 1
smho98[( 3<=smho98$STRATUM & smho98$STRATUM<=4), "stratum6"] <- 2
smho98[( 5<=smho98$STRATUM & smho98$STRATUM<=8), "stratum6"] <- 3
smho98[( 9<=smho98$STRATUM & smho98$STRATUM<=10),"stratum6"] <- 4
smho98[(11<=smho98$STRATUM & smho98$STRATUM<=13),"stratum6"] <- 5
smho98[(14<=smho98$STRATUM & smho98$STRATUM<=16),"stratum6"] <- 6

table(smho98$stratum6,smho98$STRATUM)
      1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16
  1 151  64   0   0   0   0   0   0   0   0   0   0   0   0   0   0
  2   0   0  43  22   0   0   0   0   0   0   0   0   0   0   0   0
  3   0   0   0   0 150  23  65  14   0   0   0   0   0   0   0   0
  4   0   0   0   0   0   0   0   0  38  12   0   0   0   0   0   0
  5   0   0   0   0   0   0   0   0   0   0  13  77  59   0   0   0
  6   0   0   0   0   0   0   0   0   0   0   0   0   0  86  39  19

table(smho98$stratum6)
    1   2   3   4   5   6
  215  65 252  50 149 144

    # Select 10 units by srswor per stratum
smp.IDs <- strata(data        = smho98,
                  stratanames = "stratum6",
                  size        = rep(10,6),
                  method      = "srswor")

    # Pull sampled records and verify sample counts
sample1 <- getdata(smho98,smp.IDs)
table(sample1$stratum6)
   1  2  3  4  5  6
  10 10 10 10 10 10
```

∎

Warning. A potentially confusing "feature" of R is that different packages may use the same names for functions that do different things. In Example 3.18, we used the function, strata, from the sampling package to select an *stsrswor*. The survival package (Therneau 2012) also has a strata function which does something different. Note, survival may be loaded without you realizing it because it is used by other packages like doBy (Højsgaard and Halekoh 2012), which we use in later chapters. If the

`survival` package is loaded before the `sampling` package and you try to select a stratified sample, an error is likely to occur because R will use the wrong `strata`. If so, check the order in which R searches files and packages with `search()`. If necessary, detach `survival` with the command `detach("package:survival")`.

Example 3.19 (Select a stratified pps sample). A sample of 50 hospitals is required for a study of institutions listed on the `smho98` data file. Instead of selecting an *stsrswor* as in Example 3.17, we will instead select a *pps* sample within five design strata with a size measure defined as the square root of the bed size, i.e., $pp\left(\sqrt{x}\right)$ discussed in Sect. 3.5.2. We will use the R function `ppssstrat` from the `pps` package to draw a (approximate) proportional sample within strata. The `round` function is used to eliminate the fractional sample sizes for convenience, hence the use of "approximate" in our discussion. Because outpatient facilities are not included in the target population, all hospitals with zero beds are excluded from the list frame prior to drawing the sample as shown in the code below:

```
    # Load R libraries
require(foreign)
require(pps)
    # Random seed for sample selection
set.seed(4297005)
    # Load SAS transport file
smho98 <- read.xport("smho98.xpt")
dim(smho98)
[1] 875 378

    # Eliminate outpatient facilities
smho98 <- smho98[smho98$BEDS > 0,]
dim(smho98)
[1] 671 378

    # Create 5-level stratum variable and verify
smho98$stratum5 <- 0
smho98[( 1<=smho98$STRATUM & smho98$STRATUM<=2), "stratum5"] <- 1
smho98[( 3<=smho98$STRATUM & smho98$STRATUM<=4), "stratum5"] <- 2
smho98[( 5<=smho98$STRATUM & smho98$STRATUM<=8), "stratum5"] <- 3
smho98[( 9<=smho98$STRATUM & smho98$STRATUM<=13),"stratum5"] <- 4
smho98[(14<=smho98$STRATUM & smho98$STRATUM<=16),"stratum5"] <- 5
table(smho98$stratum5)
  1   2   3   4   5
215  64 216  44 132

    # Create size measure
smho98$sqrt.Beds <- sqrt(smho98$BEDS)

    # Approx. proportional sample sizes
smp.size <- 50
(strat.cts <- as.numeric(table(smho98$stratum5)))
[1] 215  64 216  44 132
(strat.ps  <- strat.cts / sum(strat.cts))
[1] 0.32041729 0.09538003 0.32190760 0.06557377 0.19672131
```

```
    # Verify stratum proportions sum to one
sum(strat.ps)
[1] 1

    # Stratum sample sizes
smp.size.h <- round(strat.ps * smp.size,0)
[1] 16   5 16   3 10
sum(smp.size.h)
[1] 50

    # Sort data file by sampling strata and select samples
smho98 <- smho98[order(smho98$stratum5),]
smp.IDs   <- ppssstrat(sizes = smho98$sqrt.Beds,
                       strat = smho98$stratum5,
                       n     = smp.size.h)

    # Verify no duplicates in sample
length(smp.IDs)
[1] 50
length(unique(smp.IDs))
[1] 50

    # Subset to sampled records
smp.data <- smho98[smp.IDs,]
table(smp.data$stratum5)
 1   2   3   4   5
16   5 16   3 10
```

Two points to note are that `ppssstrat` selects a systematic sample from the stratum frame without doing any ordering within strata. If you want to randomize the order within strata, use the function `permuteinstrata` in the `pps` package. Also, exactly the same sample can be selected with `strata` from the `sampling` package with the code:

```
require(sampling)
    # Random seed for sample selection
set.seed(4297005)
sam <- strata(data        = smho98,
              stratanames = "stratum5",
              size        = smp.size.h,
              method      = "systematic",
              pik         = smho98$sqrt.Beds)
```

∎

3.7.2 SAS PROC SURVEYSELECT

The statistical software SAS includes a procedure called `SURVEYSELECT`[8] that selects random samples given a specified method. The general syntax for the procedure is

```
PROC SURVEYSELECT DATA=<input data file> METHOD=<method>  ...;
  STRATA <variables> / ... > ;
  CONTROL <variables>;
  SIZE <variables>;
  ID <variables>;
```

For example, `METHOD=SRS` will produce an *srswor* sample from the input data file. Including a `STRATA` variable will result in *srswor* samples within explicit strata, i.e., an *stsrswor* sample. Implicit strata (i.e., sorting variables) are identified with the `CONTROL` statement. Single-stage systematic samples can be selected with `METHOD=SYS`. *pps* samples is selected with replacement using `METHOD=PPS`. Some specialized *pps* sampling procedures (Brewer, Murthy, Sampford, and Chromy) are also included, but we will not cover them in this book. An interested reader can consult Cochran (1977) and Chromy (1979) for details of these methods.

Note that `SURVEYSELECT` selects samples only within a particular stage of a design. The code must be adapted and run for each stage of a multistage design as discussed later in Chaps. 9 and 10.

Example 3.20 (Select stsrswor with SAS). In this example, we reproduce the results for Example 3.18 using SAS `PROC SURVEYSELECT`. As with the R program in Example 3.18, the first step is to read in the SAS transport file. Here, we additionally assign a unique identification number to each hospital record:

```
*Load SAS Transport Data File;
LIBNAME inxp xport "... \smho98.xpt";
DATA SMHO98(KEEP=STRATUM HospID BEDS);
  SET inxp.SMHO98;
  HospID = _n_;
RUN;
```

After creating the stratum variable with values 1–6, the `stsrswor` is selected using

```
PROC SURVEYSELECT DATA=SMHO98 OUT=SampData
    METHOD=SRS SAMPSIZE = (10 10 10 10 10 10) SEED=82841;
  STRATA stratum6;
RUN;
```

The output data file, `SampData`, contains one record for each of the 60 randomly sampled hospitals, all variables included on the `smho98` input data file and two additional variables:

---

[8] http://support.sas.com/documentation/.

1. `SelectionProb`—the probability of selection into the sample
2. `SamplingWeight`—the sampling weight calculated as the inverse selection probability

The sampling weight is also referred to as the design weight or the base weight. Note that the R function `strata` discussed in Example 3.18 does not produce a sampling weight. Details on calculating the weights for a variety of sample designs can be found in Chaps. 13 and 14. ∎

Example 3.21 (Select a stratified pps sample with SAS). A $pp\left(\sqrt{x}\right)$ sample of 50 inpatient facilities was selected in Example 3.18 using the R function `ppssstrat` after determining an approximate proportional allocation to five design strata. The proportional allocation can be calculated with an initial call to `PROC SURVEYSELECT` as shown in the SAS code below:

```
DATA SMHO98inp DROPCASE;
  SET SMHO98;
  * Eliminate outpatient facilities;
  IF BEDS<1 THEN OUTPUT DROPCASE;
  ELSE DO;
    * Create 5-level stratum variable;
    IF       1<=STRATUM<=2  THEN stratum5=1;
    ELSE IF  3<=STRATUM<=4  THEN stratum5=2;
    ELSE IF  5<=STRATUM<=8  THEN stratum5=3;
    ELSE IF  9<=STRATUM<=13 THEN stratum5=4;
    ELSE IF 14<=STRATUM<=16 THEN stratum5=5;
    * Size measure;
    sqrtBEDS = sqrt(BEDS);
    OUTPUT SMHO98inp;
  END;
RUN;

*Approx. proportional allocation;
PROC SURVEYSELECT DATA=SMHO98inp OUT=StratSiz N=50;
  STRATA stratum5 / ALLOC=PROP NOSAMPLE;
RUN;
```

The output data file, `StratSiz`, contains the allocation for each of the five design strata. Note that the values match those calculated "by hand" with R in Example 3.18:

Stratum	Sample Size
1	16
2	5
3	16
4	3
5	10
	50

Because the `nosample` option was used, this procedure call only calculates the stratum-specific sample sizes. The following code selects the sample of 50 inpatient hospitals:

```
PROC SURVEYSELECT DATA=SMHO98inp OUT=SampDat2
     METHOD=PPS_SYS SAMPSIZE=StratSiz SEED=4297005;
     STRATA stratum5;
     SIZE sqrtBEDS;
     ID HospID;
RUN;
```

■

Exercises

3.1. According to the US Bureau of Labor Statistics, 71 % of all workers in private industry had access to employer-sponsored medical care plans, 52 % of all workers participated in medical care plans in March 2006, and 7 % of part-time workers participated in a vision care program (http://www.bls.gov/ncs/ebs/sp/ebsm0004.pdf, Tables 1 and 2). Calculate the size of a simple random sample of employees that would be needed to estimate each of these proportions using the estimation targets in (a), (b), and (c).

(a) Coefficient of variation of 10 %.
(b) Standard error of 3 % points.
(c) MOE of 3 % points.
(d) For each of the sample sizes you computed in (a), (b), and (c), what are the anticipated half-widths of 95 % confidence intervals? Use the normal approximation with a multiplier of 1.96.
(e) Comment on the differences in sample sizes that result from the three precision targets in (a), (b), and (c).

3.2. Explore the difference in setting a sample size based on a target for a coefficient of variation of an estimated proportion and setting it based on a target standard error. Assume that a simple random sample without replacement is selected but that the population size is large so that the *fpc* is negligible.

(a) Calculate $CV(p_s)$ and $\sqrt{V(p_s)}$ for a sample size of $n = 100$ for p_U in (0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.95, 0.99).
(b) Graph the values of $CV(p_s)$ versus p_U and $\sqrt{V(p_s)}$ versus p_U.
(c) Discuss the differences in the relationships.

3.3. Suppose that the population is composed of 1,000 business establishments. The mean number of full-time employees per establishment is 50. The population variance of the number of full-time employees is 150.

(a) Compute the size of a simple random sample selected without replacement that would be necessary to produce a CV of the sample mean of 5 %.
(b) What if you anticipated that only 40 % of the sample establishments would respond to a request for data? How would that affect your sample size calculation in (a)?
(c) Suppose that you conduct the survey and actually get a response rate of 35 %. Would you expect the mean for the 35 % that did respond to be a good estimate of the population mean? Why or why not?

3.4. (a) Suppose that an investigator sets a desired tolerance e such that $\Pr(|\bar{y}_s - \bar{y}_U| \leq e) = 1 - \alpha$. Assuming that $\bar{y}_s$ can be treated as normally distributed, show that this is equivalent to setting the half-width of a $100(1 - \alpha)$ % two-sided confidence interval equal to $e = z_{1-\alpha/2}\sqrt{V(\bar{y}_s)}$.

(b) If we require $\Pr\left(\left|\frac{\bar{y}_s - \bar{y}_U}{\bar{y}_U}\right| \le e\right) = 1 - \alpha$, show that this corresponds to setting the half-width of a $100\,(1-\alpha)\,\%$ two-sided confidence interval equal to $e = z_{1-\alpha/2}\,CV(\bar{y}_s)$.

3.5. Verify formula (3.12) for the sample size needed when a MOE e is set for estimating a proportion.

3.6. Verify formula (3.17) for the required sample size derived from the MOE calculation using the normal approximation for the log-odds of a proportion.

3.7. An investigator wants to estimate the prevalence of a characteristic that is speculated to be rare. The investigator's best guess is that the prevalence is $2\,\%$. She would like to estimate the prevalence with a MOE of 0.005.

(a) What sample size is required?
(b) Since the investigator seems very uncertain about the actual prevalence, what alternative calculations could you do to illustrate the effects of different sample sizes?
(c) Compare the results in (b) for the standard normal, the Wilson, and the log-odds methods of computing sample sizes.

3.8. Compute the unit relvariances of:

(a) The variables `beds` and `discharges` in the `hospital` population
(b) The variables total expenditures (`EXPTOTAL`), number of inpatient beds (`BEDS`), number of patients seen during 1998 (`SEENCNT`), the number of clients on the roles at the end of 1998 (`EOYCNT`), and number of in patient visits (`Y_IP`) in the `smho98` population

3.9. This problem uses the summary values for the population (`smho98`) of mental health organizations in Table 3.2. Assume that an *srswor* will be selected in each stratum. In all parts, round your computed sample sizes to the nearest integer.

(a) Find the Neyman allocation of a sample size $n = 115$. Round the sample sizes to the nearest integer. Calculate the total variable cost of this allocation assuming variable costs per sample unit of 1,000, 400, 200, 1,000, 200, and 1,000 in the strata.
(b) Find the allocation that minimizes the variance of the estimated population mean of total expenditures, assuming the variable costs in part (a) and a total budget for variable costs of $80,000.
(c) Compute the coefficient of variation of $\bar{y}_{st}$ for the allocations you found in (a) and (b). Compare the results. Use rounded sample sizes for these calculations.
(d) Suppose that your target for $CV(\bar{y}_{st})$ is 0.15 and that the cost structure is the same as in part (a). Calculate the optimal allocation and the total cost, $C - c_0$, for that allocation.

(e) What are the CV's for the individual estimated stratum means for your allocations in parts (a), (b), and (d)? Comment on the results.

(f) Suppose that your government client would like to publish individual stratum estimates but that the agency has an ironclad rule that an estimate must have a CV of 0.30 or less to be publishable. Do any of your allocations in (a), (b), and (d) satisfy this criterion? Find an allocation that does meet the 0.30 CV criterion for all strata; compute its cost and the CV it gives for the estimated population mean across all strata. How would you discuss the trade-offs between this new allocation and those of (a), (b), and (d) with the client?

(g) What are the design effects for $\bar{y}_{st}$ for the allocations in parts (a), (b), and (f)?

3.10. The number of inpatient visits (IPV's) during a calendar year is the variable Y_IP on the smho98 file.

(a) Use the organizations with a positive number of IPV's as the population and determine the number of sample units needed to estimate the mean IPV's per organization with a CV of 0.10. Assume that the sample will be selected with probability proportional to number of inpatient beds (BEDS) and that $\hat{\bar{y}}_{\pi}$ will be used. Determine which units should be take-alls and the breakdown of the sample size by take-alls and non-take-alls. Designate any unit with a selection probability of 0.8 or larger as a take-all.

(b) Repeat part (a) with a CV target of 0.15.

(c) Now, suppose that you decide to use a regression estimator of the mean number of discharges. Use a model with no intercept and with the square root of beds and beds itself as predictors. If this model is correct, what is the optimum MOS to use in a *pps* sample? What sample would be required to obtain an anticipated CV of 0.10 with this regression estimator and a sample selected with the optimal MOS?

(d) Explain any differences in the results for parts (a) and (c).

3.11. Show that Eq. (3.41) reduces to $\hat{V}_1 = \frac{N^2}{n-1} \sum_{s_0} (y_i - \bar{y}_{s_0})^2$ if the s_0 sample is *srswr* of size n and the planned sample is to be an *srswr*. Hence, $\hat{V}_1 / N^2 n = [n(n-1)]^{-1} \sum_{s_0} (y_i - \bar{y}_{s_0})^2$.

3.12. Researchers at a health organization are interested in estimating the number of discharges within the last 12 months from hospitals specializing in a new medical procedure ($N = 393$). The project budget was sufficient to allow data collection at ($n =$) 50 hospitals. Based on prior research, the project statistician selected a *pps* sample of size 50 using the number of hospital beds as the MOS. The total number of beds tabulated from the list sampling frame was 107,956. Data from all 50 sample hospitals is located in the text file hosp50.csv. Data for number of beds for all 393 hospitals in the frame are in the file hospital.txt or hospital.RData.

(a) Calculate the design weights for the 50 sample hospitals. How might you verify that the weights were calculated correctly? Show the verification.
(b) Estimate the average number of discharges based on the sample using the π-estimator of the mean. Assume that the population count, $N=393$, is known.
(c) Estimate the sample variance for your estimate in (b) using the formula for with-replacement sampling.
(d) Estimate the 95% confidence interval for your estimate in (b). What assumptions are you making when computing this confidence interval?
(e) Suppose you want to select a new sample with probabilities proportional to the square root of beds. Estimate the appropriate V_1 for this design. How many sample hospitals would be needed to meet the target $CV\left(\hat{\bar{y}}_\pi\right)$ $= 0.15$ with this design?

3.13. Select ten samples of size 20 from the hospital population using probability proportional to the number of beds as in Example 3.16. Calculate the estimate $\hat{V}_1$ in Eq. (3.43) for the alternate MOS $\sqrt{x_i}$ from each sample. Suppose that you set a target of $CV_0 = 0.10$ for a new sample. What is the range of anticipated sample sizes required to achieve this target? Suggest a way of attempting to reflect the variability of the estimator of the variance component V_1 when determining the size of a new sample.

3.14. In preparation for an upcoming study, you have been asked to perform sample size calculations using two separate analysis variables, y_1 and y_2. The population, from which the sample will be selected, contains 1,000 units. Data collected during a previous study using a *srswor* design are contained in the file `Domainy1y2.txt`.

(a) Determine the sample size needed to meet a target $CV = 0.05$ for the estimated mean of the two analysis variables, y_1 and y_2. Are the estimated sample sizes different? Is so, why?
(b) If the target precision level is increased to a $CV = 0.03$, how do your calculations in (a) change?
(c) Repeat your calculations in parts (a) and (b) for the proportion of units whose values for y_1 are less than or equal to 50 ($y_1 \leq 50$).
(d) Repeat your calculations in parts (a) and (b) for the proportion of units whose values for y_1 are less than or equal to 22 ($y_1 \leq 22$). Compare your results from parts (c) and (d).

3.15. Some populations can be divided into elements that have a zero value for a variable and others that have a nonzero value. For example, the US tax law allows businesses to claim a tax credit for the salaries and wages of employees engaged in research as defined in "Coordinated Issue All Industries Credit for increasing Research Activities—Qualified Research Expenses" (June 18, 2004).[9] Some employees are engaged in qualified research for some

---

[9] Available at http://www.irs.gov/businesses/article/0,,id=182094,00.html.

percentage of their time (the nonzeros); others do not do research at all (the zeros).

(a) Show that the unit variance, $S^2 = \sum_{i=1}^{N} (y_i - \bar{y}_U)^2 / (N-1)$, can be written as

$$S^2 = \frac{1}{N-1} \left[(N_1 - 1) S_1^2 + N \bar{y}_{U1}^2 P (1-P) \right]$$
$$\doteq P \left(S_1^2 + Q \bar{y}_{U1}^2 \right),$$

where N_1 is the number of elements with nonzero values, $P = N_1 / N$ is the proportion of elements with nonzero values, $Q = 1 - P$, $\bar{y}_{U1}$ is the mean for elements with nonzero values, and $S_1^2 = \sum_{i=1}^{N_1} (y_i - \bar{y}_{U1})^2 / (N_1 - 1)$ is the variance among elements with nonzero values. In the example, N_1 would be the number of employees who performed qualified research out of a total of N in a company.

(b) Suppose that an *srswor* is to be selected and N_1 and N are both large. Show that the number of sample elements required to achieve $CV\left(\hat{T}\right) = CV_0$ can be written as

$$n \doteq \frac{1}{P * CV_0^2} \left(\frac{S_1^2}{\bar{y}_{U1}^2} + Q \right)$$

(c) Graph the sample size in (b) versus P for values of the unit relvariance among nonzero elements equal to 1, 2, and 4.

3.16. Consider two different sample designs for the smho.N874 population. One is a sample of 50 units selected with probability proportional to the square root of beds, i.e., $\sqrt{x}$ where $x =$ number of inpatient beds. The other is a stratified design where 25 strata are formed by sorting the frame from low to high based on $\sqrt{x}$. The strata are then formed to each have approximately the same sum of $\sqrt{x}$. A sample of 2 units is then selected by *srswor* from each stratum.

(a) Compare the selection probabilities for these two sample designs. For example, compute the mean *pps* selection probability within each stratum and compare it to the *stsrswor* selection probabilities.
(b) Graph the *stsrswor* probabilities versus the *pps* selection probabilities.

Hint: The R functions cumsum and cut will be useful.

3.17. Use the smho.N874 population to estimate the power γ in the model $E_M(y) = \beta_1 \sqrt{x} + \beta_2 x$, $V_M(y) = \sigma^2 x^\gamma$. The Y variable is the total expenditures, which is the variable EXPTOTAL on the smho.N874 file. The x variable is number of beds (BEDS). Use the organizations with a positive number of beds as the population. Based on your estimate $\hat{\gamma}$, what type of *pps* sampling method would be efficient? What type of general regression estimator would you recommend?

3.18. Suppose that the sample of size n is to be selected with *ppswr* using a MOS x and that the *pwr*-estimator will be used to estimate the mean. There are n_t take-alls identified using some rule of thumb, say, $x_k \geq N\bar{x}_U / n$. Write down the *pwr*-estimator for this situation. Show that the size of the non-take-all sample required to achieve a coefficient of variation of CV_0 is

$$n_{nt} = \frac{V_1}{(N\bar{y}_U CV_0)^2}$$

where $V_1 = \sum_{U_{nt}} p_k \left(y_k / p_k - T_{nt}\right)^2$ with U_{nt} being the universe of non-take-alls, the p_k's being the 1-draw selection probabilities from the non-take-alls, and $T_{nt} = \sum_{U_{nt}} y_k$. Show that the CV of $\hat{\bar{y}}_{\pi}$ is

$$CV\left(\hat{\bar{y}}_{\pi}\right) = \frac{V_1}{N\bar{y}_U \sqrt{n_{nt}}} \ .$$

3.19. You plan to select a simple random sample without replacement from the population of Detroit, Michigan. The number of visits to a doctor per person is to be estimated separately for African-American and all other persons. Census data show that African-Americans are 83% of the population. You have these estimates from an earlier survey:

Group	Population variance	Mean number of visits per year
African-American	4.2	1.4
All others	3.3	2.2

(a) Determine what size of simple random sample would be needed to obtain *CV*s for the estimated mean number of visits person of 0.01, 0.05, 0.10, and 0.20. Assume that the population is so large that N can be treated as infinite.
(b) Assuming that a single sample will be selected, which group will determine the total sample size needed to hit the CV targets?

3.20. An *srswor* of size n is selected from a population of size N. The estimate of the mean per unit in domain d is $\hat{\bar{y}}_d = \hat{t}_d \big/ \hat{N}_d$ where $\hat{N}_d = N n_d / n$.

(a) Show that the linear approximation to $\hat{\bar{y}}_d$ is $\hat{\bar{y}}_d - \bar{y}_{Ud} \doteq \frac{1}{N_d} N\bar{e}_s$ where $\bar{e}_s = n^{-1} \sum_s e_i$ with $e_i = \delta_i \left(y_i - \bar{y}_{Ud}\right)$.
(b) Using this, show that the approximate variance of $\hat{\bar{y}}_d$ is $V\left(\hat{\bar{y}}_d\right) \doteq \frac{1}{N_d^2} \frac{N^2}{n} \left(1 - \frac{n}{N}\right) S_e^2$ with $S_e^2 = (N-1)^{-1} \sum_U e_i^2$.
(c) Show that the relvariance of $\hat{\bar{y}}_d$ is $CV^2\left(\hat{\bar{y}}_d\right) \doteq \frac{1}{nP_d} \left(1 - \frac{n}{N}\right) CV_d^2$.

Chapter 4
Power Calculations and Sample Size Determination

In Chap. 3 we calculated sample sizes based on targets for coefficients of variation (CVs), margins of error, and cost constraints. Another method is to determine the sample size needed to detect a particular alternative value when testing a hypothesis. For example, when comparing the means for two groups, one way of determining sample size is through a power calculation. Roughly speaking, power is a measure of how likely you are to recognize a certain size of difference in the means. A sample size is determined that will allow that difference to be detected with high probability (i.e., a detectable difference). Power can also be determined in a one-sample case where a simple hypothesis is being tested versus a simple alternative. Using power to determine sample sizes is especially useful when some important analytic comparisons can be identified in advance of selecting the sample. Although not covered in most books on sample design, most practitioners will inevitably have applications where power calculations are needed.

Suppose that a survey designer or an experimenter decides that a difference of δ ($|\delta| > 0$) between two or more true (population) means is important to recognize. If the true difference is δ, then we would like the sample size to be large enough so that there is a specified probability of showing a statistically significant difference between the domain or treatment means. Setting the detection probability (i.e., power) at 0.80 or 0.90 is common practice. Power is also often stated in percentages rather than probabilities, e.g., 0.80 is the same as 80 % power. This method of sample size determination is particularly common in medical studies. Useful references that cover sample size calculation in various types of medical studies include Armitage and Berry (1987), Lemeshow et al. (1990), Schlesselman (1982), and Woodward (1992).

The size of the budget is critical. If the power calculation leads to an unaffordable sample size, the experiment or survey will have to be scaled back. In some cases, the study may have to be abandoned entirely if meaningful differences cannot be detected with the size of sample that can be afforded.

This chapter reviews the terminology used in hypothesis testing and power analysis and describes the mechanics of power calculations for one- and

two-sample tests. The assumptions and inputs to power computations need to be understood in order to make the right sample size computations. To that end, we provide some algebraic details. We concentrate on tests for means and proportions and give some examples of how to implement the sample size calculations in R, SAS, Excel, and Stata.

4.1 Terminology and One-Sample Tests

This section discusses the ideas of Type I and II errors when performing hypothesis tests, power of a test, and *1-sided* and *2-sided tests*, along with *one-sample* and *two-sample* tests. We concentrate on tests of means, but the terms apply more generally to other population parameters. Table 4.1 summarizes the terminology used when testing hypotheses together with the decisions that can be made and errors that can occur. H_0, shown in the table, is traditionally called the null hypothesis; an alternative hypothesis is denoted by H_A.

Table 4.1: Terminology: size and power of a test.

| | | Decision | |
		Do not reject H_0	Reject H_0
State of nature	H_0 is true	Correct decision with probability $1 - \alpha$	Type I error—incorrect decision with probability α (*level* or *size* of test)
	H_0 is false (H_A is true)	Type II error—incorrect decision with probability β (at a specific alternative value)	Correct decision with probability $1 - \beta$ (*power* of test at a specific alternative value)

Analysts usually avoid saying that a null hypothesis is accepted on the grounds that a hypothesis like $H_0 : \mu = 3$ is never likely to be exactly true. If the real mean (to 3 decimal places) were 3.001, then H_0 would be false. Many people like to use the more noncommittal statement "H_0 is not rejected" rather than "H_0 is accepted" which implies that the hypothesis has been proved to be true.

Characterizing Hypotheses and Tests

Hypotheses can be simple or composite. Tests can be characterized as one-sided or two-sided. When a hypothesis contains only one value, it is called *simple* (e.g., $H_0 : \mu = 3$ is simple). A hypothesis that contains more than one value is *composite* (e.g., $H_0 : \mu \leq 3$ is composite). Whether a test is one- or

two-sided depends on the alternative. If the null hypothesis is $H_0 : \mu = 3$, a one-sided alternative is $H_A : \mu > 3$ because the alternative values of interest are only in one direction from the null value. A two-sided alternative would be $H_A : \mu \neq 3$ since the alternatives can be either greater or less than $H_0 : \mu = 3$. The alternative, $H_A : \mu \neq 3$, is also composite because it involves many values.

One-Sample Test

By *one-sample*, we mean a case where a single mean is being tested against some hypothesized value(s). For a one-sample, simple null hypothesis versus a simple alternative, we are testing:

$$H_0 : \mu = \mu_0 \text{ versus } H_A : \mu = \mu_0 + \delta$$

at level α for some δ that can be positive or negative. Usually, we think of testing the simple null hypothesis versus the composite alternative:

$$H_A : \mu \neq \mu_0.$$

The standard test statistic is

$$t = \frac{\hat{\bar{y}} - \mu_0}{\sqrt{v\left(\hat{\bar{y}}\right)}}, \tag{4.1}$$

where $\hat{\bar{y}}$ is an estimate of the mean of the variable y and $v\left(\hat{\bar{y}}\right)$ is an estimate of the variance of $\hat{\bar{y}}$. In survey sampling the finite population mean is estimated as

$$\hat{\bar{y}} = \frac{\sum_{i \in s} w_i y_i}{\sum_{i \in s} w_i}, \tag{4.2}$$

where w_i is the survey weight for unit i and s denotes the set of sample units. The sample can be selected in a complex way (e.g., stratified, multistage with varying probabilities). As long as the variance is consistently[1] estimated by $v\left(\hat{\bar{y}}\right)$, t in (4.1) is treated as having a (central) t-distribution under the null hypothesis. The t-approximation may be poor when the sample size is small and the y-variable has a very skewed distribution. But, the t is a useful starting place for the power and sample size calculations in this chapter. The

---

[1] Roughly speaking, an estimator is said to be consistent if it gets closer and closer to the value it is supposed to be estimating as the sample size increases. A variance estimator $v\left(\hat{\bar{y}}\right)$ is a consistent estimator of the true variance $V\left(\hat{\bar{y}}\right)$ if $v\left(\hat{\bar{y}}\right) / V\left(\hat{\bar{y}}\right) \xrightarrow{P} 1$ as $n \to \infty$. In survey samples, n is the number of sample units in a single-stage sample or the number of primary sampling units (PSUs) in a multistage sample. A ratio is used in this definition because both the estimator and its target approach 0 as the sample size increases.

degrees of freedom for the t are usually based on some rules of thumb. One that is often used is

$$df = \text{number of PSUs} - \text{number of strata}. \qquad (4.3)$$

For example, in a design with H strata and n_h primary sampling units (PSUs) selected from stratum h, the rule of thumb gives $\sum_{h=1}^{H} (n_h - 1) = (n_+ - H)$ df with $n_+ = \sum_h n_h$. For a household design with 50 strata and 2 sample PSUs per stratum, the rule of thumb would be $df = 50$, even though the number of sample households could be in the hundreds or thousands. These rules are not necessarily accurate, and some better approximations to df can be computed [see, e.g., (Rust 1984, 1985; Valliant and Rust 2010)].

When the sample of PSUs is large, the t-distribution will be about the same as a normal distribution. As always, it is hard to give a good answer to the question: "what is large?" Critical points of the t and normal distributions are very close to each other for $df \geq 60$. The table below shows the 97.5 percentiles of t for various df, i.e., the points $t_{0.975,df}$ such that $\Pr(t \leq t_{0.975}(df)) = 0.975$. For $df = 60$, $t_{0.975,60} = 2$—about the same as 1.96 for a standard normal distribution[2].

df	$t_{0.975,df}$
1	12.71
5	2.57
10	2.23
30	2.04
60	2.00
100	1.98
∞	1.96

Some rules of thumb are thrown around, like "(number of PSUs)—(number of strata) must be 30 or more" in order to use the normal approximation. However, the approximate df for a variance estimator is affected by how skewed the input data are in addition to the number of PSUs and number of strata. Family incomes, for example, are highly skewed while education test scores, like the Scholastic Aptitude Test (SAT), are usually constructed to have nearly normal distributions across the test takers. Skewed input data will require more sample PSUs for the t-statistic to be approximately normal than will symmetric, nearly normal input data. Extremely rare or prevalent characteristics will also have the same effect. On the other hand, getting a good fix on the approximate df is not simple, and practitioners usually are content with computing the value in (4.3) and adopting a cutoff, like 60, for using the normal approximation.

---

[2] A standard normal distribution is a normal distribution with mean $= 0$ and standard deviation $= 1$, i.e., $N(0,1)$.

Use of Finite Population Corrections in Variances

Testing the simple hypothesis that the mean is a particular value, $H_0 : \mu = \mu_0$, or, as covered later in Sect. 4.3, that the means of two groups are equal, $H_0 : \mu_X = \mu_Y$, raises an issue that may seem to be niggling but is worth a comment. Two finite population means are not likely to be exactly equal. Using the example from earlier in this section, if one mean is 3 and another 3.001, these are different. Consequently, when testing hypotheses, like $H_0 : \mu_X = \mu_Y$, that compare groups, analysts usually consider these to be tests on underlying parameters of a model that describes the population reasonably well. Thus, even if the entire finite population were enumerated, the calculated means would still have variances because they would still be estimates of the underlying, unknown model parameters. Consistent with that philosophy, variance estimates should not include finite population correction factors (like $1 - n/N$ in *srswor*).

Ignoring the *fpc* in a variance estimator has real, practical implications for the sample size calculations in later sections. If the sampling fraction is greater than about 0.05, the sample sizes computed to achieve a certain level of power can be noticeably different, depending whether an *fpc* is included or not. Incorporating a non-negligible *fpc* reduces the value of a variance estimate and, consequently, reduces the computed sample size to achieve that power. Thus, it may appear that some money can be saved simply by injecting an *fpc* into the calculations. However, the superpopulation thinking above would say this is specious reasoning. In some applications, like household surveys, sampling fractions are usually so low that fretting about an *fpc* is unnecessary. Nevertheless, you may confront the issue in other situations, like school surveys, where the population is smaller.

If your goal is really to measure how large the difference is between two finite population means, then a power calculation is probably not what you want. The appropriate sample size calculation should be done using the methods in Chap. 3 where we accounted for the *fpc*. An interesting discussion of whether an *fpc* should be used in different situations can be found at http://web.cos.gmu.edu/~wss/wss070328paper.pdf.

Before explaining how to calculate power, we give some definitions of terms that are used in hypothesis testing.

Definition 4.1 (Type I Error). A Type I error is rejecting a null hypothesis when it is actually true. The probability that H_0 is rejected in such a case is called the *level* or *size* of the test and, for a 2-sided test, is

$$\Pr\left(|t| > t_{1-\alpha/2}\left(df\right)|H_0 \text{ is true}\right) = \alpha,$$

where $t_\gamma\left(df\right)$ is the γ-quantile of the central t-distribution with df degrees of freedom, i.e., $\Pr\left(t < t_\gamma\left(df\right)\right) = \gamma$. Said another way, the level of the test is the chance that the test statistic is in the rejection region of the distribution when the null hypothesis is actually true. For a 1-sided test of $H_0 : \mu = \mu_0$

versus $H_A : \mu > \mu_0$ the Type I error rate is

$$\Pr\left(t > t_{1-\alpha}\left(df\right)|H_0 \text{ is true}\right) = \alpha.$$

Definition 4.2 (Type II Error). A Type II error is accepting that a null hypothesis is true when it is actually false. The probability that H_0 is accepted in such a case for a 2-sided test is

$$\Pr\left(|t| \le t_{1-\alpha/2}\left(df\right)|H_A \text{ is true}\right) = \beta.$$

For a 1-sided test of $H_0 : \mu = \mu_0$ versus $H_A : \mu > \mu_0$ the Type II error rate is

$$\Pr\left(t \le t_{1-\alpha}\left(df\right)|H_A \text{ is true}\right) = \beta.$$

To actually compute β, we must think of a specific value within the possibilities spanned by H_A.

Definition 4.3 (Power). Power is 1 minus the Type II error rate, i.e., the probability of rejecting H_0 when it actually is false. The power and Type II error rate vary depending on the particular value of the alternative. For a 2-sided test, the power is the chance that the test statistic is in the rejection region when $\mu = \mu_0 + \delta$ and is equal to

$$\Pr\left(|t| > t_{1-\alpha/2}\left(df\right)|\mu = \mu_0 + \delta\right) = 1 - \beta.$$

The power in a 1-sided test of $H_0 : \mu = \mu_0$ versus $H_A : \mu > \mu_0$ is

$$\Pr\left(t > t_{1-\alpha}\left(df\right)|\mu = \mu_0 + \delta\right) = 1 - \beta.$$

Notice that we could use the more elaborate notation β_δ since the power depends on the specific value of the alternative. (Examples 4.1 and 4.2 illustrate power calculations for specific values of alternatives.)

Definition 4.4 (p-value). A p-value is the smallest level of significance at which a null hypothesis would be rejected based on the observed value of the test statistic being used. Suppose that the calculated value of (4.1) is t_{obs}. Then, the p-value for a 2-sided test is

$$\Pr\left(|t| > t_{obs}|H_0 \text{ is true}\right).$$

No particular alternative hypothesis is entertained here—no decision is made to choose between H_0 and some H_A. When the analysis consists of computing a test statistic and its associated p-value, this is called *significance testing* and is probably the procedure most commonly used, especially in the social sciences.

The p-value is usually taken to be a measure of the strength of evidence for or against the null hypothesis. A small p-value is interpreted as evidence that

H_0 is false, i.e., a test statistic of size t_{obs} or more extreme is very unlikely to occur if H_0 were true. The smaller the p-value, the stronger the evidence against H_0. This interpretation is dubious since the p-value associated with a given size of effect depends on the sample size. Quoting Royall (1986):

> ...a difference between treatments that is just statistically significant at the 0.05 level may be so small that it is of no clinical significance if the study groups are enormous, whereas a difference between smaller groups yielding the same p-value corresponds to a much larger estimated treatment effect.

Because of these issues, p-values are not useful for determining sample sizes.

4.2 Power in a One-Sample Test

The power for a given sample size depends on how far away the alternative value, $\mu_0 + \delta$, is from the null value, μ_0. Alternatives that are far from the null are naturally easier to detect than the ones that are close. Three things are needed for a sample size calculation based on power:

1. Value of δ
2. Desired probability $1 - \beta$ (i.e., the power) of obtaining a significant test result when the true difference is δ
3. Significance level α of the test, which can be either 1-sided or 2-sided

1-Sided Tests

First, consider a 1-sided test of $H_0 : \mu = \mu_0$ versus $H_A : \mu > \mu_0$. The null hypothesis will be rejected if $t > t_{1-\alpha}(df)$. For example, with an $\alpha = 0.05$ level test and a large number of df, H_0 will be rejected if $t > t_{0.95}(\infty) = z_{0.95} = 1.645$. When the sample of PSUs is large, t in (4.1) can be treated as having a $N(0,1)$ distribution under H_0. If, on the other hand, the true mean is $\mu = \mu_0 + \delta$ for some $\delta > 0$, then the mean of t is

$$\frac{\delta}{\sqrt{V(\hat{\bar{y}})}},$$

where $V(\hat{\bar{y}})$ is the theoretical variance of $\hat{\bar{y}}$. Assuming that $v(\hat{\bar{y}}) \doteq V(\hat{\bar{y}})$, the probability that t is in the rejection region when $\mu = \mu_0 + \delta$ is

$$\Pr\left(t > t_{1-\alpha}\left(df\right) | \mu = \mu_0 + \delta\right) \tag{4.4}$$

$$\doteq \Pr\left(\frac{\hat{\bar{y}} - \mu_0}{\sqrt{V\left(\hat{\bar{y}}\right)}} - \frac{\delta}{\sqrt{V\left(\hat{\bar{y}}\right)}} > z_{1-\alpha} - \frac{\delta}{\sqrt{V\left(\hat{\bar{y}}\right)}} \middle| \mu = \mu_0 + \delta\right)$$

$$= \Pr\left(Z > z_{1-\alpha} - \frac{\delta}{\sqrt{V\left(\hat{\bar{y}}\right)}}\right),$$

where Z is a standard normal random variable, i.e., one with mean 0 and variance 1. Expression (4.4) is the power of the test against the alternative $\mu = \mu_0 + \delta$.

Figure 4.1 illustrates the situation. If H_0 is true and t has mean 0, the test statistic will have a standard normal distribution (on the left in the figure) given that the df is large. The rejection region is marked in light gray and has area α, 0.05 in this case. If the mean is $\mu_0 + \delta > 0$, then the mean of t is $\delta / \sqrt{V\left(\hat{\bar{y}}\right)}$ and the distribution of t is shifted to the right. The probability of being in the rejection region for the shifted distribution is the area to the right of $z_{1-\alpha} = 1.645$ (light gray plus darker gray).

Example 4.1 (Power for a previous survey). Suppose that you plan to select a sample of households from a particular Canadian province and measure the mean household income for married-couple households ($\hat{\bar{y}}$). Based on earlier surveys of the same design and size, you anticipate that the mean is about \$55,000 Canadian dollars and will be estimated with a 6 % *CV*. You would like to test the hypothesis $H_0 : \mu = \$55,000$ versus $H_A : \mu > \$55,000$ at the $\alpha = 0.05$ level. Thus, the anticipated standard error of $\hat{\bar{y}}$ is $0.06 \times 55,000 = 3,300$. You would also like to know how much power you have to detect that the mean is really \$60,000. Substituting in (4.4) and using the normal approximation, the anticipated power is

$$\Pr\left(Z > z_{1-\alpha} - \frac{\delta}{\sqrt{V\left(\hat{\bar{y}}\right)}}\right) = \Pr\left(Z > 1.645 - \frac{60,000 - 55,000}{3,300}\right)$$

$$= 0.448.$$

When the survey is actually conducted, the sample estimate of the mean turns out to be \$59,000 with a 7.5 % *CV*. The *t*-statistic for testing $H_0 : \mu = \$55,000$ is, thus,

$$t_{obs} = \frac{59,000 - 55,000}{0.075 \times 59,000} = 0.9040.$$

The *p*-value associated with this statistic is $\Pr\left(t > 0.9040 | \mu = 55,000\right) \doteq 0.183$. Consequently, whether the mean is larger than \$55,000 seems doubtful. A check on this conclusion is to calculate a confidence interval for the

population mean. In this case, a one-sided $95\,\%$ interval is $59,000 - 1.645 \times 0.075 \times 59,000 = 51,721$, which is less than the hypothesized $55,000. ∎
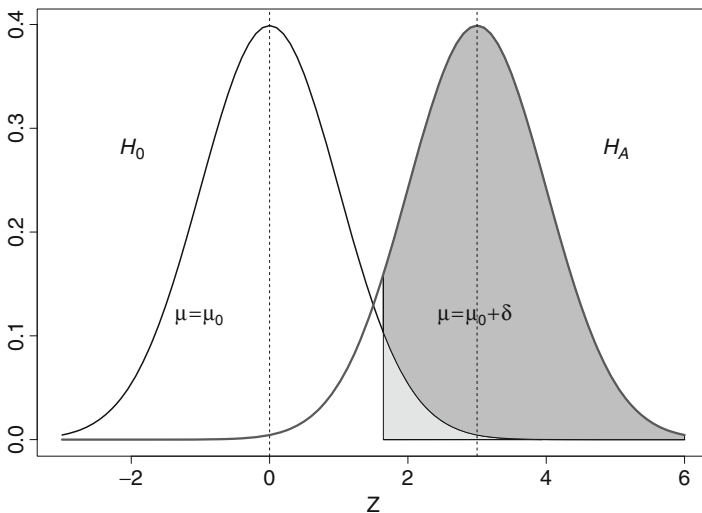
Fig. 4.1: Normal densities of test statistics under H_0 and H_A. $\delta \left/ \sqrt{V\left(\hat{\bar{y}}\right)}\right.$ is set equal to 3 in this illustration so that $E\{t\,|H_A \text{ is true}\} = 3$. A 1-sided test is conducted at the 0.05 level.

The power calculation can also be done using a t-distribution if the degrees of freedom for the variance estimator are not large, say less than 60. The statistic $\left(\hat{\bar{y}} - \mu_0\right) \left/ \sqrt{v\left(\hat{\bar{y}}\right)}\right.$ will have a noncentral t-distribution with noncentrality parameter $\delta \left/ \sqrt{V\left(\hat{\bar{y}}\right)}\right.$ when the mean is $\mu = \mu_0 + \delta$. The power of the t-test is then the probability that a noncentral t random variable with df degrees of freedom is greater than $t_{1-\alpha}\left(df\right)$. This is the method used by the R function `power.t.test` in the `stats` package (R Core Team and contributors worldwide 2012c) described in Sect. 4.4.

Suppose we want the power, i.e., the probability of being to the right of $z_{1-\alpha}$ to be $1 - \beta$ (e.g., 0.80). Let z_β be the point on the standard normal distribution with area β to its left and $1 - \beta$ to its right. Now, suppose that $V\left(\hat{\bar{y}}\right) = \sigma_y^2/n$ where n is the sample size of analytic units and σ_y^2 is the population unit variance of Y. Working from (4.4), we set $z_{1-\alpha} - \frac{\delta}{\sqrt{\sigma_y^2/n}}$ equal to $z_\beta\left(= -z_{1-\beta}\right)$ and solve for the sample size n to obtain

$$n = \left[\sigma_y \frac{(z_{1-\alpha} - z_\beta)}{\delta}\right]^2 = \left[\sigma_y \frac{(z_{1-\alpha} + z_{1-\beta})}{\delta}\right]^2. \qquad (4.5)$$

For designs other than *srswor*, $V(\hat{\bar{y}}) = \sigma_y^2/n$ does not hold. A common work-around is to set the solution to (4.5) equal to the effective sample size, defined as $n_{eff} = n/deff$ where $deff = V(\hat{\bar{y}})/V_{SRS}(\hat{\bar{y}})$, the ratio of the variance under the complex design to the variance under *srswor*. Of course, this does not fully solve the problem since a value for the *deff* is required for the particular design and analysis variable in question. Its value will depend on whether the design is stratified single-stage, clustered, or something else and on how the sample is allocated to strata and clusters.

Example 4.2 (Finding a sample size for specified power). In Example 4.1, suppose that microdata has been used to estimate the population standard deviation via one of the methods discussed in Sect. 3.4 obtaining $\hat{\sigma}_y = 74,000$. If the population mean is \$55,000, this implies that the unit relvariance is $74^2/55^2 = 1.8$. (Unit relvariances in the range 1 to 5 are typical for continuous variables.) A one-sided $\alpha = 0.05$ level test is to be conducted and a simple random sample of households can be selected. Suppose, in particular, that $H_0 : \mu = \$55,000$ and $H_A : \mu > \$55,000$. If we want power of 0.80 ($z_{1-\beta} = z_{0.80} \doteq 0.84$) to detect that the mean is \$60,000, then the sample size from (4.5) is

$$n = \left[74,000\frac{(1.645 + 0.84)}{5,000}\right]^2 \doteq 1,355 \text{ households.}$$

If a clustered design is used and we estimate *deff* to be 1.6, then the required sample size is $n = 1,355\,(1.6) \doteq 2,170$. On the other hand, if we want the same power against an alternative of \$57,500, then the *deff* adjusted sample size is

$$n = 1.6\left[74,000\frac{(1.645 + 0.84)}{2,500}\right]^2 \doteq 8,670.$$

Clearly, the goals of the analysis have a big impact on sample size. Careful thought needs to be given to the size of the alternative that is substantively important to detect. ∎

In applications like Example 4.2, σ_y must be estimated from a previous sample or guessed based on experience. The sample size of 8,670 is itself an estimate of the size actually needed for power of 0.80. Because this is done in advance, it would be better to call this the *anticipated* power. When data are collected in the new survey, we can estimate the *achieved* power based on that data. Random variation being what it is, the anticipated and achieved power are rarely the same. As a safeguard, a sample of more than 8,670 might be selected in case the $\hat{\sigma}_y$ is too small.

2-Sided Tests

Calculation of power for a 2-sided test is similar but a bit more involved. The null hypothesis is rejected if $|t| > t_{1-\alpha/2}\,(df)$. If the goal is to detect a departure from the null hypothesis value of δ in either direction, then alternatives of the form $\mu_0 \pm \delta$ are of interest. We will examine these one at a time—first $\mu = \mu_0 + \delta$ and then $\mu = \mu_0 - \delta$. Again, assuming that the normal approximation is good enough and noting that $z_{\alpha/2} = -z_{1-\alpha/2}$, the Type II error probability that the test statistic is in the acceptance region, when $\mu = \mu_0 + \delta$, is

$$\Pr\left(|t| \le t_{1-\alpha/2}\,(df)\,|\mu = \mu_0 + \delta\right)$$

$$\dot{=} \Pr\left(-z_{1-\alpha/2} \le \frac{\hat{\bar{y}} - \mu_0}{\sqrt{V(\hat{\bar{y}})}} < z_{1-\alpha/2}\,\middle|\, \mu = \mu_0 + \delta\right)$$

$$= \Pr\left(-z_{1-\alpha/2} - \frac{\delta}{\sqrt{V(\hat{\bar{y}})}} \le \frac{\hat{\bar{y}} - \mu_0}{\sqrt{V(\hat{\bar{y}})}} - \frac{\delta}{\sqrt{V(\hat{\bar{y}})}} < z_{1-\alpha/2} - \frac{\delta}{\sqrt{V(\hat{\bar{y}})}}\,\middle|\, \mu = \mu_0 + \delta\right)$$

$$= \Pr\left(-z_{1-\alpha/2} - \frac{\delta}{\sqrt{V(\hat{\bar{y}})}} \le Z < z_{1-\alpha/2} - \frac{\delta}{\sqrt{V(\hat{\bar{y}})}}\right)$$

$$= \Pr\left(Z \le z_{1-\alpha/2} - \frac{\delta}{\sqrt{V(\hat{\bar{y}})}}\right) - \Pr\left(Z \le -z_{1-\alpha/2} - \frac{\delta}{\sqrt{V(\hat{\bar{y}})}}\right).$$

The power of the test against the alternative is then

$$\Pr\left(|t| > t_{1-\alpha/2}\,(df)\,|\mu = \mu_0 + \delta\right) \tag{4.6}$$

$$\dot{=} 1 - \Pr\left(Z \le z_{1-\alpha/2} - \frac{\delta}{\sqrt{V(\hat{\bar{y}})}}\right) + \Pr\left(Z \le -z_{1-\alpha/2} - \frac{\delta}{\sqrt{V(\hat{\bar{y}})}}\right).$$

The last term on the right-hand side of (4.6) will be near 0 in many cases.

By a similar computation, the power of the test against the alternative $H_A : \mu = \mu_0 - \delta$ is

$$\Pr\left(|t| > t_{1-\alpha/2}\,(df)\,|\mu = \mu_0 - \delta\right) \tag{4.7}$$

$$\dot{=} 1 - \Pr\left(Z \le z_{1-\alpha/2} + \frac{\delta}{\sqrt{V(\hat{\bar{y}})}}\right) + \Pr\left(Z \le -z_{1-\alpha/2} + \frac{\delta}{\sqrt{V(\hat{\bar{y}})}}\right).$$

In this case, the second term on the right-hand side of (4.7) will often be near 1 and expression (4.7) will be approximately $\Pr\left(Z \le -z_{1-\alpha/2} + \delta\,\middle/\,\sqrt{V(\hat{\bar{y}})}\right)$.

Suppose we want the power against either $\mu_0 + \delta$ or $\mu_0 - \delta$ to be $1 - \beta$. We can set (4.6) or (4.7) to $1 - \beta$ and then solve for n. Using either (4.6) or (4.7) leads to the same sample size as we now show. First, approximate (4.7) by

$$1 - \Pr\left(Z \leq z_{1-\alpha/2} - \frac{\delta}{\sqrt{V(\hat{\bar{y}})}}\right) = \Pr\left(Z > z_{1-\alpha/2} - \frac{\delta}{\sqrt{V(\hat{\bar{y}})}}\right)$$

and set this equal to $1 - \beta$. This implies that $z_{1-\alpha/2} - \frac{\delta}{\sqrt{V(\hat{\bar{y}})}} = z_\beta$. Using $V(\hat{\bar{y}}) = \sigma_y^2/n$ and solving gives

$$n = \left[\sigma_y \frac{(z_{1-\alpha/2} - z_\beta)}{\delta}\right]^2. \tag{4.8}$$

Approximating (4.7) by $\Pr\left(Z \leq -z_{1-\alpha/2} + \delta \big/ \sqrt{V(\hat{\bar{y}})}\right)$, setting $-z_{1-\alpha/2} + \frac{\delta}{\sqrt{V(\hat{\bar{y}})}} = z_{1-\beta}$, and solving for the sample size give

$$n = \left[\sigma_y \frac{(z_{1-\alpha/2} + z_{1-\beta})}{\delta}\right]^2 = \left[\sigma_y \frac{(z_{1-\alpha/2} - z_\beta)}{\delta}\right]^2. \tag{4.9}$$

Note that to compute the sample size for a 2-sided test in (4.9), we just change α in (4.5) for the 1-sided test to $\alpha/2$. Comparing (4.9) with (4.5), we see that to obtain the same power to detect the alternatives $\mu_0 \pm \delta$ the required sample size will be larger than for detecting $\mu_0 + \delta$ alone because $z_{1-\alpha/2} > z_{1-\alpha}$. For example, $z_{0.975} = 1.96$ and $z_{0.95} = 1.645$. Some intuition for this is that a larger sample is needed to detect an alternative that can be on either side of the null value.

As in the 1-sided case, the R function `power.t.test` does a more refined version of the sample size calculation.

Example 4.3 (Sample size for a two-sided test). Continuing with Examples 4.1 and 4.2, suppose that power of 0.80 (i.e., 80 % power) is desired against either of the alternatives $H_A : \mu = \$50,000$ or $H_A : \mu = \$60,000$. As before, $H_0 : \mu = \$55,000$. Substituting in (4.8) gives

$$n = \left[74,000 \frac{(1.96 + 0.84)}{5,000}\right]^2 \doteq 1,720.$$

Adjusting this for a design effect of 1.6, the sample size is about $2,750$. If we want power of 0.80 against $H_A : \mu = \$52,500$ or $H_A : \mu = \$57,500$, then $5,000$ is replaced by $2,500$ in the above equation to give $n = 6,880$ or $n = 11,000$ adjusted for $deff = 1.6$. ∎

Section 4.4 illustrates how these computations can be done in R. They are also easily programmed in Excel. Figures 4.2 and 4.3 show screenshots of a spreadsheet that will compute the sample sizes in Examples 4.2 and 4.3. Figure 4.3 shows the formulas while Fig. 4.2 gives numerical results that match those in the examples. The spreadsheet is also available on the book's web site. Another excellent reference that combines R and Excel is Heiberger and Neuwirth (2009).

	A	B	C	D	E
		Example 4.2		Example 4.3	
1		**Example 4.2**		**Example 4.3**	
2		**(a)**	**(b)**	**(a)**	**(b)**
3	?-sided test?	1	1	2	2
4	a	0.05	0.05	0.05	0.05
5	$z_{(1-a/2)}$	1.645	1.645	1.960	1.960
6	β	0.200	0.200	0.200	0.200
7	$1-\beta$	0.800	0.800	0.800	0.800
8	$z_{(1-\beta)}$	0.842	0.842	0.842	0.842
9					
10	sigma	74,000	74,000	74,000	74,000
11	mean	55,000	55,000	55,000	55,000
12	*deff*	1.6	1.6	1.6	1.6
13	δ	5,000	2,500	5,000	2,500
14					
15	n.eff	1,354.2	5,416.9	1,719.2	6,876.9
16	n	2,166.8	8,667.1	2,750.7	11,003.0

Fig. 4.2: An excel spreadsheet for the computations in Examples 4.2 and 4.3.

4.3 Two-Sample Tests

Comparing the means of two different groups of units is a standard analytic goal. The term "two-sample" test stems from the aim of comparing parameters for two separate groups or populations with a sample being selected from each. This section describes the methods used for comparing means or proportions for two such groups.

4.3.1 Differences in Means

For a two-sample case, we may want to test that

$$H_0 : \mu_X \leq \mu_Y \text{ versus } H_A : \mu_X > \mu_Y$$

	A	B	C	D	E
		Example 4.2		**Example 4.3**	
		(a)	**(b)**	**(a)**	**(b)**
3	?-sided test?	1	1	2	2
4	α	0.05	0.05	0.05	0.05
5	$z_{(1-\alpha/2)}$	=NORMSINV(1-B4/B3)	=NORMSINV(1-C4/C3)	=NORMSINV(1-D4/D3)	=NORMSINV(1-E4/E3)
6	β	0.2	0.2	0.2	0.2
7	1-β	=1-B6	=1-C6	=1-D6	=1-E6
8	$z_{(1-\beta)}$	=NORMSINV(1-B6)	=NORMSINV(1-C6)	=NORMSINV(1-D6)	=NORMSINV(1-E6)
9					
10	sigma	74000	=B10	74000	74000
11	mean	55000	=B11	55000	55000
12	deff	1.6	=B12	1.6	1.6
13	δ	5000	2500	5000	2500
14					
15	n.eff	=(B10*(B5+B8)/B13)^2	=(C10*(C5+C8)/C13)^2	=(D10*(D5+D8)/D13)^2	=(E10*(E5+E8)/E13)^2
16	n	=IF(OR(B12="",B12=1), "",B15*B12)	=IF(OR(C12="",C12=1), ,C15*C12)	=IF(OR(D12="",D12=1), ,D15*D12)	=IF(OR(E12="",E12=1), ,E15*E12)

Fig. 4.3: An excel spreadsheet for the computations in Examples 4.2 and 4.3 with formulas shown.

at level α where X is the random variable associated with the first sample or group and Y is the random variable associated with the second. The sample test statistic is

$$t_d = \frac{\hat{\hat{d}}}{\sqrt{v\left(\hat{\hat{d}}\right)}}$$

with $\hat{\hat{d}} = \hat{\bar{x}} - \hat{\bar{y}}$, $v\left(\hat{\hat{d}}\right) = v\left(\hat{\bar{x}}\right) + v\left(\hat{\bar{y}}\right) - 2cov\left(\hat{\bar{x}}, \hat{\bar{y}}\right)$ where $v\left(\hat{\bar{x}}\right)$ and $v\left(\hat{\bar{y}}\right)$ are design-based estimates of the variances of the means and $cov\left(\hat{\bar{x}}, \hat{\bar{y}}\right)$ is a design-based estimate of their covariance. In a cross-sectional survey, we will usually be comparing the means for two nonoverlapping domains. If each domain is specific to different strata, then $cov\left(\hat{\bar{x}}, \hat{\bar{y}}\right) = 0$ by definition. But, if the design involves clustering, even nonoverlapping domains like male and female may have correlated estimates due to presence of domain members within the same PSUs.

The null hypothesis that the mean of Y is larger than or equal to the mean of X ($H_0 : \mu_X \le \mu_Y$) will be rejected in large samples if $t_d > z_{1-\alpha}$. If the true mean difference is some $|\delta| > 0$, then the mean of t_d is $\delta \left/ \sqrt{V\left(\hat{\hat{d}}\right)}\right.$ instead of 0. Letting $\mu_D = \mu_X - \mu_Y$, the probability that t_d is in the rejection region is then

$$\Pr\{t_d > z_{1-\alpha} \,|\, \mu_D = \delta\} = \Pr\left(t_d - \frac{\delta}{\sqrt{V\left(\hat{\hat{d}}\right)}} > z_{1-\alpha} - \frac{\delta}{\sqrt{V\left(\hat{\hat{d}}\right)}} \,\middle|\, \mu_D = \delta\right)$$

$$\doteq \Pr\left(Z > z_{1-\alpha} - \frac{\delta}{\sqrt{V\left(\hat{\hat{d}}\right)}}\right). \tag{4.10}$$

This is the power of the test against the alternative $\mu_D = \mu_X - \mu_Y = \delta$ and is similar to (4.4) for the one-sample case.

Suppose that the sample size in each domain is the same and that the variance of the difference can be written as $V\left(\hat{\hat{d}}\right) = \sigma_d^2/n$ where σ_d^2 is some population unit variance. For example, this will hold if the domain estimates are independent and their variances can be written as

$$V\left(\hat{\hat{d}}\right) = \frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{n} = \frac{1}{n}\left[\sigma_x^2 + \sigma_y^2\right]$$

as would be the case for *srswor*. If the domain estimates are correlated, then $\sigma_d^2 = \sigma_y^2 + \sigma_x^2 - 2\sigma_{xy}$ with σ_{xy} being the population covariance of X and Y.

If $\sigma_y^2 = \sigma_x^2 \equiv \sigma_0^2$, then the unit-level correlation between y and x is $\rho = \sigma_{xy}/\sigma_0^2$ and $\sigma_d^2 = 2\sigma_0^2(1 - \rho)$, which is a convenient form. To find the required sample size, we set z_β equal to $z_{1-\alpha} - \frac{\delta}{\sqrt{\sigma_d^2/n}}$ and solve for the sample size n to obtain

$$n = \left[\frac{\sigma_d (z_{1-\alpha} - z_\beta)}{\delta} \right]^2. \tag{4.11}$$

Note that this is the sample size in *each* domain. If $\sigma_x^2 = \sigma_y^2 \equiv \sigma_0^2$ and $\rho = 0$, then $\sigma_d^2 = 2\sigma_0^2$.

The calculation of power in a two-sided test leads to formulas analogous to (4.6) and (4.7). Figure 4.4 graphs the power in a two-sided test of $H_0 : \mu_D = 0$ versus $H_A : |\mu_D| = \delta$ for a test done at the 5 % level (i.e., $\alpha = 0.05$) assuming that $\sigma_d = 3$. Four different, group sample sizes are shown: 10, 25, 50, and 100. If $|\delta| = 2$, the power for $n = 10$ in each group is about 0.30. But, if $n = 50$, the power is over 0.90. For a given sample size, the power becomes larger as $|\delta|$ increases. The R function `power.t.test`, described later, was used for the power computations displayed in Fig. 4.4.
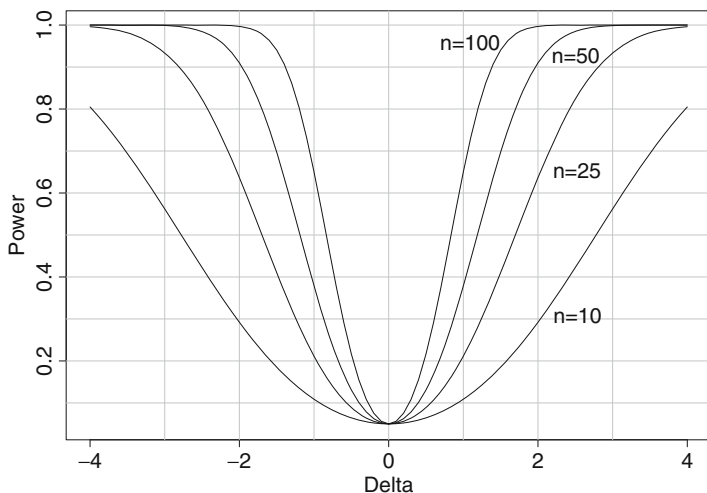
Fig. 4.4: Power for sample sizes of $n = 10, 25, 50, 100$ in a two-sided test of $H_0 : \mu_D = 0$ versus $H_A : |\mu_D| = \delta$ ($\alpha = 0.05$, $\sigma_d = 3$).

Partially Overlapping Samples

The case of partially overlapping samples can also be handled (e.g., see Woodward 1992). For example, persons may be surveyed at some baseline date and

then followed up at a later time. An estimate of the difference in population means may be desired, but the samples do not overlap completely because of dropouts, planned sample rotation, or nonresponse. Suppose that s_1 and s_2 are the sets of sample units with data collected only at times 1 and 2 and that s_{12} denotes the overlap. Thus, the full samples at times 1 and 2 are $s_1 \cup s_{12}$ and $s_2 \cup s_{12}$. Also, suppose that the samples at the two time periods are simple random samples. Assume that the samples at times 1 and 2 are not necessarily the same size, so that $n_1 = rn_2$ for some positive number r. The samples might be different sizes because of other survey goals or because the budget for data collection is different for the two times. A case that is covered by the analysis below is one where an initial sample of n_1 is selected, a portion of these respond at time 2, and additional units are selected to obtain a total sample of n_2 for time 2. Taking the case of simple random sampling, the difference in means can be written as

$$\hat{d} = \hat{\bar{x}} - \hat{\bar{y}} = \frac{1}{n_1} \sum_{s_1} x_i - \frac{1}{n_2} \sum_{s_2} y_i + \sum_{s_{12}} \left(\frac{x_i}{n_1} - \frac{y_i}{n_2} \right).$$

The variance can be expressed as

$$V\left(\hat{d}\right) = \frac{\sigma_x^2}{n_1} + \frac{\sigma_y^2}{n_2} - 2\sigma_{xy} \frac{n_{12}}{n_1 n_2}, \tag{4.12}$$

where n_{12} is the number of units in s_{12}. Writing $n_{12} = \gamma n_1$ and $r = n_1/n_2$, the variance becomes $V\left(\hat{d}\right) = \frac{1}{n_1} \left[\sigma_x^2 + r\sigma_y^2 - 2\gamma r\sigma_{xy}\right]$. For a 1-sided test of $H_0 : \mu_D = 0$ versus $H_A : \mu_D = \delta$, we set z_β equal to $z_{1-\alpha} - \delta \left/ \sqrt{V\left(\hat{d}\right)} \right.$ and solve for the sample size n_1 to give

$$n_1 = \frac{1}{\delta^2} \left[\sigma_x^2 + r\sigma_y^2 - 2\gamma r\rho\sigma_x\sigma_y\right] (z_{1-\alpha} - z_\beta)^2. \tag{4.13}$$

Using the simplification that $\sigma_y^2 = \sigma_x^2 \equiv \sigma_0^2$, the variance can be rewritten as $V\left(\hat{d}\right) = \frac{\sigma_0^2}{n_1} \left[1 + r\left(1 - 2\gamma\rho\right)\right]$. The sample size n_1 becomes

$$n_1 = \frac{\sigma_0^2}{\delta^2} \left[1 + r\left(1 - 2\gamma\rho\right)\right] (z_{1-\alpha} - z_\beta)^2. \tag{4.14}$$

If the samples are independent, then $\gamma = 0$ and the formula reduces to

$$n_1 = \frac{\sigma_0^2}{\delta^2} (1 + r) (z_{1-\alpha} - z_\beta)^2. \tag{4.15}$$

Note that if $n_1 = n_2$, then $r = 1$ and (4.15) equal (4.11) because σ_d^2 in (4.11) equals $2\sigma_0^2$. Given values of r, γ, and ρ, the sample size at time 1 can be found

via (4.14) and, in turn, n_2 solved for as $n_2 = n_1/r$. For the more general case, if estimates of the unit variances and covariance, or, equivalently, the unit correlation, are available, then (4.13) can be used. The R function `nDep2sam` in Sect. 4.4 will compute the sample sizes n_1 and n_2 based on (4.13).

4.3.2 Differences in Proportions

The test on the difference in two proportions is similar to that for the difference in the means for two quantitative variables. However, since the variance in a Bernoulli distribution is a function of the mean, the test statistic is specialized to account for this. Suppose we want to test the hypothesis $H_0 : P_1 = P_2$ where P_k is the population proportion in domain k. Assume that independent *srs*'s are selected from each domain, the estimated proportions are p_1 and p_2, and that the sample sizes in the two domains are n_1 and n_2. If the null hypothesis is true so that each population proportion is equal to the same value $\bar{P}$, then the variance of the difference is

$$V(p_1 - p_2) = \bar{P}\left(1 - \bar{P}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right) .$$

The test statistic is then

$$t_{\Delta p} = \frac{p_1 - p_2}{\sqrt{v\left(p_1 - p_2\right)}}, \qquad (4.16)$$

where $v\left(p_1 - p_2\right) = \bar{p}\left(1 - \bar{p}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$ with $\bar{p} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$ being the "pooled" estimate of $\bar{P}$. In large samples (4.16) is approximately normally distributed, which allows us to approximate the power at different alternatives and to compute sample sizes.

Because the variance of the estimated proportions depends on their means, the arithmetic needed to get a power formula is a little different from that used to arrive at (4.10). To simplify matters, we cover only the case of the same sample size n in each group. If the null hypothesis of equal proportions is true, then $v\left(p_1 - p_2\right) = 2\bar{p}\left(1 - \bar{p}\right)/n$. But, if $H_A : P_2 = P_1 + \delta$ is correct, the estimated variance of $p_1 - p_2$ does not depend on a pooled $\bar{p}$ but instead is $\left(p_1 q_1 + p_2 q_2\right)/n$. This is an estimate of the theoretical variance $\left(P_1 Q_1 + P_2 Q_2\right)/n$. The power of this test for a 1-sided alternative $H_A : P_2 = P_1 + \delta$ is then

$$\Pr\left(t_{\Delta p} > z_{1-\alpha} \,|\, P_2 - P_1 = \delta\right)$$

$$= \Pr\left(p_1 - p_2 > z_{1-\alpha}\sqrt{2\bar{p}\left(1-\bar{p}\right)/n}\,\Big|\, P_2 - P_1 = \delta\right)$$

$$\doteq \Pr\left(\frac{p_1 - p_2}{\sqrt{\left(P_1 Q_1 + P_2 Q_2\right)/n}} - \frac{\delta}{\sqrt{\left(P_1 Q_1 + P_2 Q_2\right)/n}} > \right.$$

$$\left.\frac{z_{1-\alpha}\sqrt{2\bar{P}\left(1-\bar{P}\right)/n} - \delta}{\sqrt{\left(P_1 Q_1 + P_2 Q_2\right)/n}}\,\Big|\, P_2 - P_1 = \delta\right)$$

$$= \Pr\left(Z > \frac{z_{1-\alpha}\sqrt{2\bar{P}\left(1-\bar{P}\right)/n} - \delta}{\sqrt{\left(P_1 Q_1 + P_2 Q_2\right)/n}}\right). \tag{4.17}$$

Power for a two-sided test is computed in a way similar to (4.6) and (4.7) beginning with $\Pr\left(|t_{\Delta p}| > z_{1-\alpha}\,|\, P_2 - P_1 = \delta\right)$ and following the steps in (4.17). The distribution of $t_{\Delta p}$ cannot be approximated as a t-distribution which requires normally distributed input data. Thus, only the normal approximation is used to assess power.

The sample size in each group needed to detect a difference of δ is found by setting the right-hand side of the inequality in the last line of (4.17) equal to z_β and solving for n to give

$$n = \left[\frac{z_{1-\alpha}\Delta_1 - z_\beta \Delta_2}{\delta}\right]^2, \tag{4.18}$$

where $\Delta_1 = \sqrt{2\bar{P}\left(1-\bar{P}\right)}$ and $\Delta_2 = \sqrt{P_1 Q_1 + P_2 Q_2}$. Advance estimates of P_1, P_2, and $\bar{P}$ are needed to evaluate (4.18). The R function `power.prop.test` in the `stats` package, described in Sect. 4.4, uses a search algorithm to solve for n which will give a similar answer to (4.18).

When samples overlap, computations similar to those for the difference in means in Sect. 4.3.1 can be made. Suppose that the variables X and Y are equal to 1 with probabilities P_x and P_y and that XY is equal to 1 with probability P_{xy}. The event, $XY = 1$, might correspond to a unit having some characteristic at both times 1 and 2. The conditional distribution of Y given X is $P_{y|x} = P_{xy}/P_y$; $P_{x|y}$ is defined similarly. The event that $Y = 1$ given that $X = 0$ could mean that a unit had a characteristic at time 2 given that it did not at time 1. With these definitions, $\sigma_x^2 = P_x\left(1-P_x\right)$, $\sigma_y^2 = P_y\left(1-P_y\right)$, $\sigma_{xy} = P_{xy} - P_x P_y$, and

$$\rho = \left(P_{xy} - P_x P_y\right)\Big/\left[P_x\left(1-P_x\right)P_y\left(1-P_y\right)\right]^{1/2}.$$

When the sample sizes in the two groups are n_1 and n_2, n_1 is found using (4.13). In this case, estimates (or educated guesses) are required for the

proportions at the two time periods, P_x and P_y, and the proportion, P_{xy}, that retains the characteristic from the first time to the second.

An R function, `nProp2sam`, that will compute the sample sizes is given in Sect. 4.4. There are two constraints on P_{xy} that are implemented in the function. First, since $P_{xy} = P_{y|x} P_x \leq P_x$ and $P_{xy} = P_{x|y} P_y \leq P_y$, it must be true that $P_{xy} \leq \min(P_x, P_y)$. Second, since the correlation must be in $[-1, 1]$, we must have $P_x P_y - \Delta \leq P_{xy} \leq P_x P_y + \Delta$ where $\Delta = [P_x (1 - P_x) P_y (1 - P_y)]^{1/2}$.

Arcsine Square Root Transformation

When a characteristic is extremely rare or highly prevalent, the normal approximation for (4.16) can be poor. One rule of thumb is that np and $n(1-p)$ should both be at least 5 to use the normal approximation. There are various fix-ups that can be used for small samples and rare (or highly prevalent) characteristics. Exact calculations using the binomial distribution are possible (Korn 1986), but even they have some peculiar anomalies (Brown et al. 2001). The Wilson method, which was one of the fix-ups used in Chap. 3 for computing sample sizes for proportions, does not appear to be amenable to two-sample power and sample size calculations.

Another method is to use a variance stabilizing transform to remove the dependence of the variance of an estimated proportion on the proportion itself. For p the transformation is the arcsine square root defined as

$$\phi = \arcsin \sqrt{p}$$

where arcsine is the inverse sine function. The variance of ϕ is approximately $1/4n$ radians. A radian is a unit of angle, e.g., a circle contains 2π radians and a right angle has $\pi/2$. Using this transform, a test of $H_0 : P_1 = P_2$ for independent samples is based on

$$t_\phi = \frac{\phi_1 - \phi_2}{\sqrt{V(\phi_1 - \phi_2)}} = \sqrt{2n}\,(\phi_1 - \phi_2). \tag{4.19}$$

This uses the approximation $V(\phi_1 - \phi_2) \doteq 1/4n + 1/4n = 1/2n$ for independent samples. If $H_A : P_2 = P_1 + \delta$ is correct, define $\delta_\phi = \arcsin \sqrt{P_1} - \arcsin \sqrt{P_1 + \delta}$. The power of a one-sided test is then

$$\Pr\left(t_\phi > z_{1-\alpha}\,|\,P_1 - P_2 = \delta\right) \tag{4.20}$$

$$= \Pr\left(t_\phi - \frac{\delta_\phi}{\sqrt{V(\phi_1 - \phi_2)}} > z_{1-\alpha} - \frac{\delta_\phi}{\sqrt{V(\phi_1 - \phi_2)}}\,\middle|\,P_1 - P_2 = \delta\right)$$

$$\doteq \Pr\left(Z > z_{1-\alpha} - \delta_\phi \sqrt{2n}\right).$$

(Note that $V(\phi_1 - \phi_2)$ is the same under H_0 and H_A since arcsine square root is the variance stabilizing transformation in both cases.) Setting $z_{1-\alpha} - \delta_\phi \sqrt{2n}$ equal to z_β leads to the sample size formula

$$n = \left(\frac{z_{1-\alpha} - z_\beta}{\sqrt{2}\delta_\phi} \right)^2 . \tag{4.21}$$

As with expression (4.11), this is the sample size required for *each* domain.

For a two-sided test of $H_0 : P_1 = P_2$ versus $H_A : P_2 = P_1 \pm \delta$, calculations like those in (4.8) and (4.9) give a sample size in each group of

$$n = \left(\frac{z_{1-\alpha/2} - z_\beta}{\sqrt{2}\delta_\phi} \right)^2 . \tag{4.22}$$

As when comparing means, a larger sample is needed for the two-sided test to have the same power to detect $H_A : P_2 = P_1 \pm \delta$ than the one-sided test needs to detect $H_A : P_2 = P_1 + \delta$.

Log-Odds Transformation

The log-odds transformation is another option that may be useful for a rare or highly prevalent characteristic. In this case, define

$$\phi = \log \left(\frac{p}{1-p} \right) .$$

The approximate variance of ϕ under $H_0 : P_1 = P_2$ is $\left(n\bar{P}\bar{Q} \right)^{-1}$ where $\bar{P}$ is the common value under H_0 and $\bar{Q} = 1 - \bar{P}$. The variances of the differences in the log-odds transforms for two independent samples are

$$V(\phi_1 - \phi_2) = \frac{2}{n} \frac{1}{\bar{P}\bar{Q}} \text{ under } H_0 \text{ and}$$

$$V(\phi_1 - \phi_2) = \frac{1}{n} \left(\frac{1}{P_1 Q_1} + \frac{1}{P_2 Q_2} \right) \text{ under } H_A,$$

assuming that the sample sizes are the same in both groups. The t-statistic has the same form as for the arcsine transformation, $t_\phi = (\phi_1 - \phi_2) \big/ \sqrt{V(\phi_1 - \phi_2)}$. Using the same steps that led to (4.17), the power against the alternative $H_A : P_2 = P_1 + \delta$ is

$$\Pr\left(t_\phi > z_{1-\alpha} \mid P_2 - P_1 = \delta\right) \doteq \Pr\left(Z > \frac{z_{1-\alpha}\sqrt{2\left[n\bar{P}\left(1 - \bar{P}\right)\right]^{-1}} - \delta_\phi}{\sqrt{n^{-1}\left[\left(P_1 Q_1\right)^{-1} + \left(P_2 Q_2\right)^{-1}\right]}}\right),$$

where $\delta_\phi = \log\left(P_1/Q_1\right) - \log\left(P_2/Q_2\right)$. Setting the term on the right-hand side of the inequality to z_β and solving for n give

$$n = \left(\frac{z_{1-\alpha}\sqrt{2V_0} - z_\beta\sqrt{V_A}}{\delta_\phi}\right)^2, \tag{4.23}$$

where $V_0 = \left[\bar{P}\left(1 - \bar{P}\right)\right]^{-1}$ and $V_A = \left(P_1 Q_1\right)^{-1} + \left(P_2 Q_2\right)^{-1}$. For a two-sided test of $H_0 : P_1 = P_2$ versus $H_A : P_2 = P_1 \pm \delta$, calculations like those in (4.8) and (4.9) give a sample size in each group of

$$n = \left(\frac{z_{1-\alpha/2}\sqrt{2V_0} - z_\beta\sqrt{V_A}}{\delta_\phi}\right)^2. \tag{4.24}$$

A numerical example using the arcsine and log-odds transformation is given in Sect. 4.4.

4.3.3 Special Case: Relative Risk

Epidemiologists and public health analysts often prefer the *relative risk*, $R = P_1/P_2$, for comparing two groups rather than the difference in proportions. A value of R much larger than 1.0 might mean that one group has a higher prevalence of some disease. The difference in proportions can be written in terms of the relative risk as

$$P_1 - P_2 = P_2\left(R - 1\right).$$

Consequently, if a sample size is desired to detect a relative risk of R^*, this corresponds to detecting a difference of $\delta = P_2\left(R^* - 1\right)$. With this value of δ, (4.18) can be used to compute the sample size for each group.

Notice that the method above is different from starting with a test statistic based on $\hat{R} = p_1/p_2$ to test the hypothesis $H_0 : R = 1$. In that case, an approximate variance would be needed in the denominator of the test statistic $t = \left(\hat{R} - 1\right)\big/\sqrt{v\left(\hat{R}\right)}$. Because of the direct linkage between the difference in proportions and the relative risk, a sample size can be computed from (4.18) that will be adequate regardless of which method of comparison you prefer.

4.3.4 Special Case: Effect Sizes

An *effect size* is usually defined as a measure of the standardized difference between two population values. When the difference is between means, one definition of the population effect size is $\delta_E = (\mu_x - \mu_y)/\sigma$ where the μ's are the means in two groups and σ is the common, unit standard deviation. This is a customary measure in meta-analysis and is also used in education research. An estimate of δ_E when simple random samples are selected from each group is

$$\hat{\delta}_E = \frac{\bar{x}_1 - \bar{x}_2}{s} . \qquad (4.25)$$

In (4.25), $\bar{x}_1$ and $\bar{x}_2$ are the sample means from each of the two groups and s is the pooled standard deviation

$$s = \sqrt{\frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{n_1 + n_2 - 2}}$$

where s_1^2 and s_2^2 are the group-specific sample variances. The form in (4.25) is known as Hedge's g (Hedges and Olkin 1985). The general idea of effect size is due to Cohen (1988). If the same size sample were used in each group and the groups are independent, then the methods from Sect. 4.3.1 can be used. In particular, if we want to detect an effect size of δ_E^*, this corresponds to a difference in means of $\delta = \delta_E^* \sigma$. Expression (4.11) applies for computing the sample size in each group with $\sigma_d^2 = 2\sigma^2$. The unit standard deviation σ could be estimated by the pooled estimate above if previous samples are available or by the square root of the sample variance if data from a single sample are in hand.

4.4 R Power Functions

The function `power.t.test`, included in the `stats` library, will calculate power or sample size for a given set of inputs. The form of the function call is

```
power.t.test(n = NULL, delta = NULL, sd = 1, sig.level = 0.05,
             power = NULL,
             type = c("two.sample", "one.sample", "paired"),
             alternative = c("two.sided", "one.sided"),
             strict = FALSE)
```

From the R help file:

> Exactly one of the parameters n, delta, power, sd, and sig.level must be passed as NULL, and that parameter is determined from the others. Notice that sd and sig.level have non-NULL defaults so NULL must be explicitly passed if you want to compute them.

The arguments are:

n	Number of observations (per group)
delta	True difference in means (i.e., desired detectable difference)
sd	Standard deviation σ_y for a one-sample test σ_x (or σ_y assuming the two are equal) for a two-sample test (more generally, $\sqrt{\sigma_d^2/2}$) σ_{x-y}, i.e., sd of differences within pairs for a paired test
sig.level	Significance level (Type I error probability)
power	Power of test (1 minus Type II error probability)
type	Type of t-test (two-sample, one-sample, paired) default is two-sample
alternative	One- or two-sided test
strict	Use strict interpretation in two-sided case. If strict = TRUE is used, the power will include the probability of rejection in the opposite direction of the true effect, in the two-sided case. Without this the power will be half the significance level if the true difference is zero

Calculations in `power.t.test` are based on a noncentral t-distribution rather than the normal approximation.

The function `power.prop.test` (`stats` library) will calculate power or sample size in a test of the difference of proportions for a given set of inputs. Calculations are based on the normal approximation; no t-distribution calculations are appropriate for this case. The form of the function call is

```
power.prop.test(n = NULL, p1 = NULL, p2 = NULL,
                sig.level = 0.05, power = NULL,
                alternative = c("two.sided", "one.sided"),
                strict = FALSE)
```

From the R help file:

> Exactly one of the parameters n, p1, p2, power, and sig.level must be passed as NULL and that parameter is determined from the others. Notice that sig.level has a non-NULL default so NULL must be explicitly passed if you want it computed.

The arguments are:

n	Number of observations (per group)
p1, p2	Probability in groups 1 and 2, respectively
sig.level	Significance level (Type I error probability)

power	Power of test (1 minus Type II error probability)
alternative	One- or two-sided test
strict	Use strict interpretation in two-sided case

Many other software packages will perform power calculations of different kinds. SAS, for example, has the procedures POWER and GLMPOWER. Stata has sampsi and various user-written functions for glm's and other specialized applications. There are also quite a few standalone packages that do nothing but power calculations (e.g., nQuery Advisor®, PASS®, Power and Precision®).

Example 4.4 (Continuation of Example 4.1). In that example, we were testing the hypothesis $H_0 : \mu = \$55,000$ and wanted the power of detecting that the mean was really $60,000 for a one-sided 0.05-level test. The CV of the estimated mean was specified to be 0.06 so that the standard error was 3,300. The R code to do this and its output are

```
power.t.test(
     n = 1000,
     power = NULL,
     delta = 5000,
     sd = 3300*sqrt(1000),    # results in sd/sqrt(n) = 3300
     type = "one.sample",
     alt = "one.sided",
     sig.level = 0.05
     )
```

The output from this function call is

```
     One-sample t test power calculation

              n = 1000
          delta = 5000
             sd = 104355.2
      sig.level = 0.05
          power = 0.4479952
    alternative = one.sided
```

This reproduces the power of 0.448 in Example 4.1. This function call uses a small trick to get power.t.test to calculate what we want. When the function computes sd/sqrt(n), the result is 3,300*sqrt(1,000)/sqrt(1,000) = 3,300, which is the standard error of the estimated mean. Using 1,000 is not critical—some other, large artificial sample size would have returned the same power. (Notice that $3,300\sqrt{1,000} = 104,355.2$ is not the unit standard deviation in the population.) ∎

Example 4.5 (Continuation of Example 4.2). In that example, we wanted 80 % power for a one-sided test to detect a difference of $5,000 when $\hat{\sigma} = 74,000$. The R code and output to compute the sample size (excluding a design effect adjustment) is

```
power.t.test(n = NULL,
  power = 0.8,
  delta = 5000,
  sd = 74000,
  type = "one.sample",
  alt = "one.sided",
  sig.level = 0.05
)
```

The output from this function call is

```
    One-sample t test power calculation

                n = 1355.581
            delta = 5000
               sd = 74000
        sig.level = 0.05
            power = 0.8
      alternative = one.sided
```

The resulting sample size is about the same as found earlier. There is a small difference due to the use of the noncentral t in `power.t.test`. ∎

Example 4.6 (Two-sample test on means). Suppose that we have two domains (males and females) and want to have equal size samples of men and women that are large enough to detect a difference in mean weights of 5 kg (i.e., $\mu_M = \mu_F + 5$) with power 0.80. We estimate that $\sigma_M^2 = \sigma_F^2 = 200$ and $\sigma_d^2 = 400$. Thus, `sd` in the input to `power.t.test` is $\sqrt{\sigma_d^2/2} = \sqrt{400/2} = \sqrt{200}$. If a 1-sided 0.05-level test is done, $z_{0.95} = 1.645$. For power of 0.80, we use $z_{0.20} = -z_{0.80} = -0.84$. The required sample size from (4.11) is then (treating 400 as if it were the true variance σ_d^2)

$$n = \frac{400 \, (1.645 + 0.84)^2}{5^2} \doteq 99.$$

On the other hand, if we wanted power of 0.90, then $z_{0.90} = 1.282$ and the sample would be 137. The same calculations can be made in R as follows:

```
power.t.test(power = 0.8,
  delta = 5,
  sd = sqrt(200),
  type = "two.sample",
  alt = "one.sided",
  sig.level = 0.05
  )
```

The output from this function call is

```
    Two-sample t test power calculation

                n = 99.60428
            delta = 5
               sd = 14.14214
        sig.level = 0.05
```

```
            power = 0.8
      alternative = one.sided

 NOTE: n is number in *each* group
```

For a power of 0.90 the function call and output are

```
power.t.test(power = 0.9,
      delta = 5,
      sd = sqrt(200),
      type = "two.sample",
      alt = "one.sided",
      sig.level = 0.05
      )
```

and

```
      Two-sample t test power calculation

                n = 137.7033
            delta = 5
               sd = 14.14214
        sig.level = 0.05
            power = 0.9
      alternative = one.sided

 NOTE: n is number in *each* group
```

∎

R does not have a built-in function to evaluate sample sizes in the two-sample case with partially overlapping samples. But, the function nDep2sam that was developed for the book and shown in Appendix C handles this case. The parameter names are S2x, S2y, g, r, rho, alt, del, sig.level, and pow and are designed to match those needed to evaluate (4.13). The parameters, sig.level, and pow, have default values of 0.05 and 0.80.

Example 4.7 (Two-sample test on means with overlapping samples). You would like to select a sample of women who are employees of a large company who also participate in a weekly yoga program. At the beginning and the end of the year the women will be weighed. Determine a sample that will allow a 5 kg difference in average weight to be detected with 80 % power. Assume that 25 % of the people in the initial sample will drop out of the program by the end of the year and that their weights cannot be measured. Also, suppose that additional women would be sampled at the end of the year to make up for the ones who dropped out but that the beginning of the year weights of these women are not available. These additional women may or may not have participated in the yoga classes all year. Thus, $n_1 = n_2$, $r = 1$, and $\gamma = 0.75$ in (4.13). As in Example 4.6, assume that $\sigma_F^2 = 200$ at both time periods. Let us also suppose that the correlation between weights at the beginning and end of the year is 0.9. The call to nDep2sam and its output are

```
nDep2sam(S2x=200, S2y=200,
         g=0.75, r=1, rho=0.9,
         alt="one.sided", del=5,
         sig.level=0.05, pow=0.80)

      Two-sample comparison of means
  Sample size calculation for overlapping samples

            n1 = 33
            n2 = 33
       S2x.S2y = 200, 200
         delta = 5
         gamma = 0.75
             r = 1
           rho = 0.9
           alt = one.sided
     sig.level = 0.05
         power = 0.8
```

That is, a sample of 33 should be selected at the beginning of the year. On
the other hand, if we wanted to detect a 5 kg difference in weight in either
direction (loss or gain), then we compute

```
nDep2sam(S2x=200, S2y=200, g=0.75, r=1, rho= 0.9,
         alt="two.sided", del=5, sig.level=0.05,
         pow=0.80)
```

resulting in the output

```
      Two-sample comparison of means
  Sample size calculation for overlapping samples

            n1 = 41
            n2 = 41
       S2x.S2y = 200, 200
         delta = 5
         gamma = 0.75
             r = 1
           rho = 0.9
           alt = two.sided
     sig.level = 0.05
         power = 0.8
```

Note that we implicitly estimated the difference in the means using all persons
available at each time period. An alternative would be to use only the women
who stayed in the program. This would be the correct approach if the goal
were to estimate the effect on weight of participating in the weekly yoga
classes for a year. In that case, nDep2sam could be used to compute a sample
size assuming complete overlap. The call for a 1-sided test would be

```
nDep2sam(S2x=200, S2y=200, g=1, r=1, rho= 0.9, alt="one.sided",
         del=5, sig.level=0.05, pow=0.80)
```

which yields $n_1 = 10$. Adjusting this for the 25 % dropout rate gives
about 14. Although this is much smaller than the 33 computed above, the

result is perfectly reasonable when we examine the variance of the difference in means in the two scenarios. As noted in the development leading to (4.13), the general formula for the variance of the difference in means is $V\left(\hat{\bar{d}}\right) = \frac{1}{n_1}\left[\sigma_x^2 + r\sigma_y^2 - 2\gamma r\sigma_{xy}\right]$. When only the overlapping cases are used, the variance is $V\left(\hat{\bar{d}}\right) = 2\sigma_x^2\left[1 - \rho\right]/n_1$ which evaluates to $40/n_1$ in Example 4.7. Using all cases available at each time period, the variance of the difference is $130/n_1$, which is 3.25 times as large as $40/n_1$. This is, in turn, about equal to the ratio of the sample sizes, $33/10$, we just computed.

Of course, there is also the important conceptual difference in what is being estimated when we use only matching cases compared to all cases. For the former, the argument can be made that the difference in means of matching cases estimates the effect of the exercise program on weight. In the latter, the difference in means is affected by the possibility that some women did not participate all year. ∎

Example 4.8 (Two-sample test on proportions with independent samples). One of the standard questions on the Defense Manpower Data Center's surveys of military personnel is:

> Taking all things into consideration, how satisfied are you, in general, with each of the following aspects of being in the (branch of service here, e.g., National Guard/Reserve)?

A list follows in the questionnaire, which includes compensation, opportunities for promotion, type of work, and other features of military life. One of the choices is "Your total compensation (i.e., base pay, allowances, and bonuses)." Suppose we would like to compare the proportions of Army and Marine personnel who say that they are "very dissatisfied" or "dissatisfied" with total compensation. If the percentages are 15 % of Army personnel and 18 % of Marines, we would like to be able to detect this with 80 % power. For a one-sided test, the R statements and output are:

```
power.prop.test(power = 0.8,
      p1 = 0.15,
      p2 = 0.18,
      alt = "one.sided",
      sig.level = 0.05
      )
```

and

```
      Two-sample comparison of proportions power calculation

            n = 1891.846
           p1 = 0.15
           p2 = 0.18
    sig.level = 0.05
        power = 0.8
```

```
     alternative = one.sided

 NOTE: n is number in *each* group
```

Thus, a sample of about $n = 1,900$ would be needed in each of the two services under study. If samples of 1,000 from each service have already been selected and the observed percentages are 15 and 18, then the power of detecting a 3 percentage point difference is only 0.56 as shown here:

```
power.prop.test(n = 1000,
     p1 = 0.15,
     p2 = 0.18,
     alt = "one.sided",
     sig.level = 0.05
     )

    Two-sample comparison of proportions power calculation

             n = 1000
            p1 = 0.15
            p2 = 0.18
     sig.level = 0.05
         power = 0.56456
   alternative = one.sided

 NOTE: n is number in *each* group
```

■

Example 4.9 (Effect of size of proportions). Note that the power is affected by the size of the proportions themselves because the pooled estimate of variance depends on the pooled p as shown in (4.16). If the percentages in Example 4.8 are 50 for Army and 53 for Marines, the power to detect an actual 3 percentage point difference is 0.38 rather than 0.56 above.

```
power.prop.test(n = 1000,
     p1 = 0.50,
     p2 = 0.53,
     alt = "one.sided",
     sig.level = 0.05
     )

    Two-sample comparison of proportions power calculation

             n = 1000
            p1 = 0.5
            p2 = 0.53
     sig.level = 0.05
         power = 0.3810421
   alternative = one.sided

 NOTE: n is number in *each* group
```

■

There is not a built-in R function to compute the sample size for a test in the difference in proportions when samples overlap. The function, nProp2sam, in Appendix C will evaluate (4.13) for proportions. The calling parameters are:

px	Probability in one group
py	Probability in other group
pxy	Probability that a unit in the overlap has the characteristic in both samples
g	γ in the relationship $n_{12} = \gamma n_1$
r	Ratio of group sample sizes, $r = n_1/n_2$
alt	Alternative hypothesis: "one.sided" or "two.sided"
sig.level	Significance level (Type I error probability)
pow	Power of test (1 minus Type II error probability)

The function returns a vector with n_1 and n_2 in the first two positions and other calling parameter information. As noted in Sect. 4.3.2, the function checks restrictions that must be satisfied on the probability P_{xy} of having the characteristic at both time periods.

Example 4.10 (Difference in proportions with overlapping samples). To take a concrete example, suppose that a baseline measurement is to be made of the proportion of registered voters who plan to vote for the incumbent in the next election which is six months away. A follow-up sample of voters is asked three months later for whom they plan to vote. Suppose that the advance estimates of the proportions of voters who will vote for the incumbent are $p_x = 0.5$ and $p_y = 0.55$. The proportion who say at both times that they will vote for the incumbent is estimated as $p_{xy} = 0.45$. You anticipate selecting the same size sample at each time period but that only half of the baseline sample will respond to the second survey. For a two-sided, 0.05-level test that will detect the difference of $\delta = 0.05$ with power of 0.80, the function call and output are

```
nProp2sam(px=0.5, py=0.55, pxy=0.45, g=0.5,
          r=1, alt="two.sided")
```

and

```
      Two-sample comparison of proportions
  Sample size calculation for overlapping samples

             n1 = 1013
             n2 = 1013
      px.py.pxy = 0.50, 0.55, 0.45
          gamma = 0.5
              r = 1
            alt = two.sided
      sig.level = 0.05
          power = 0.8
```

A total of 1,013 persons will be needed in each sample. ∎

Example 4.11 (Two-sample test on proportions with the arcsine and log-odds transformations). We will repeat Example 4.8 where the percentages are 15 % for Army personnel and 18 % for Marines, and we would like to be able to detect this with 80 % power. There is no built-in R function to do this, but the following code will evaluate (4.21) for a one-sided test.

```
p1 <- 0.15
p2 <- 0.18
alpha <- 0.05
power <- 0.80

phi1 <- asin(sqrt(p1))
phi2 <- asin(sqrt(p2))
d.phi <- phi1 - phi2
n <- ((qnorm(1-alpha) - qnorm(1-power)) / sqrt(2) / d.phi)^2
n
```

Program output:

```
[1] 1889.337
```

The following code uses the log-odds transformation to compute the sample size:

```
p1 <- 0.15
p2 <- 0.18
alpha <- 0.05
power <- 0.80

phi1 <- log(p1/(1-p1))
phi2 <- log(p2/(1-p2))
d.phi <- phi1 - phi2
p.bar <- mean(c(p1,p2))
V0 <- 1/p.bar/(1-p.bar)
VA <- 1/p1/(1-p1) + 1/p2/(1-p2)

n <- ( (qnorm(1-alpha)*sqrt(2*V0) - qnorm(1-power)*sqrt(VA)) /
       d.phi)^2
n
```

Program output:

```
[1] 1888.571
```

Both the arcsine and log-odds transformations give virtually the same answer. Both are very close to the value of about 1,892 as calculated in Example 4.8. ∎

4.5 Power and Sample Size Calculations in SAS

SAS has the procedure, `power`, which will do one- and two-sample calculations. We repeat some of the earlier examples to provide comparisons with the R functions.

Example 4.12 (Continuation of Example 4.5). The SAS code to do this one-sample calculation is

```
proc power;
  onesamplemeans
      mean = 60000
      ntotal = .
      stddev = 74000
      sides = 1
      nullmean = 55000
      power = 0.80;
run;
```

The parameters should be self-explanatory after referring to the earlier example. By specifying `ntotal = .`, we ask SAS to calculate a sample size needed for 0.80 power. Results are shown below; the total sample size of 1,356 is about the same as before.

```
        The POWER Procedure
    One-sample t Test for Mean

    Fixed Scenario Elements

 Distribution                  Normal
 Method                         Exact
 Number of Sides                    1
 Null Mean                      55000
 Mean                           60000
 Standard Deviation             74000
 Nominal Power                    0.8
 Alpha                           0.05

        Computed N Total

        Actual        N
        Power      Total

        0.800      1356
```

∎

Example 4.13 (Continuation of Example 4.8). Two-sample test on proportions: in this example, we want to find the sample size needed to obtain 80 % power to detect a 0.03 difference between two proportions. The SAS code to do this two-sample calculation is shown below. The option `test = pchi` results in the normal approximation being used, as described in Sect. 4.3.2. Unlike R `power.prop.test`, we do not specify each of the proportions, 0.15 and 0.18. SAS requires the two options `refproportion = 0.15` and `proportiondiff = 0.03` be used to do the same thing.

```
proc power;
  twosamplefreq
      test = pchi
      refproportion = 0.15
      proportiondiff = 0.03
      sides = 1
      power = 0.80
      npergroup = .
;
run;
```

The result for the sample size per group is $n = 1,892$ as in Example 4.8.

```
                    The POWER Procedure
        Pearson Chi-square Test for Two Proportions

                    Fixed Scenario Elements

  Distribution                         Asymptotic normal
  Method                             Normal approximation
  Number of Sides                                      1
  Reference (Group 1) Proportion                    0.15
  Proportion Difference                             0.03
  Nominal Power                                      0.8
  Null Proportion Difference                           0
  Alpha                                             0.05

                    Computed N Per Group
                       Actual    N Per
                       Power     Group

                        800      1892
```

■

Exercises

4.1. The average disposable employment income per worker in Mexico in 2002 was approximately \$6,100 US dollars (USD).[3] Suppose that a new survey is to be conducted in 2010 and you would like to determine the size of simple random sample that would permit you to detect that the average has risen to \$7,000. Assume that the unit relvariance of income in 2002 was 2.5 and that it will be about the same in 2010. Calculate a sample size for a 0.05-level test when the desired power is 0.80; treat the 2002 mean as a constant for this problem.

4.2. Consider Example 4.6 where one-sided tests were used to determine sample sizes with 80 % and 90 % power to detect differences in estimates for males and females.

(a) How does the sample size change if $\sigma_d^2 = 200$?
(b) How does a $\sigma_d^2 = 800$ affect your previous calculation?
(c) Compare your results.

4.3. Continuing with Exercise 4.2:

(a) What sample design is assumed under the calculations?
(b) How does your calculation change in 4.2(a) if the survey design results in an overall design effect of 1.0? A design effect of 3.2?
(c) How would you adjust your initial sample sizes in 4.2(b) to address differential response rates by gender, say a 75 % response rate for females and a 60 % response rate for males?

4.4. Your organization has been awarded a contract to conduct a study of obesity in children ages 6 to 14. Data on eating habits and levels of exercise are collected through a parent questionnaire; physical measurements are collected by trained nurse practitioners. Your task is to determine sample sizes under the following scenarios with 80 % power at a significance level of 0.05.

(a) The client is interested in determining if the average BMI for children in the first grade (ages 6–7) has increased by 1.5 % from a previously estimated average of 17.5. What is the sample size needed to detect this difference given that the population standard deviation is 0.70?
(b) How does the sample size change if the client is willing to accept being able to detect a 3.0 % increase?
(c) How does the sample size change if the client wants to detect a 0.5 % increase?
(d) Comment on the difference in your sample size calculations.

4.5. Rework the sample size calculations from the previous exercise assuming the client wants to detect either an increase or decrease in the average BMI.

---

[3] http://www.worldsalaries.org/employment-income.shtml

4.6. The average amount of taxable income reported by taxpayers to a country's revenue administration in 2008 was 44,000 in the local currency based on a tabulation of all tax returns. Due to an economic recession, it is speculated the average may have dropped by 10 % in 2010. Suppose that the unit relvariance of taxable income in the population is 3. What simple random sample size would be needed to detect a 10 % decline with a power of 0.90? How would your answer change if the unit relvariance were 6?

4.7. The relative risk of a person's having had malaria in the last five years is to be estimated for two villages in Liberia. You plan to select a simple random sample of the same size from each village. Because of their different proximities to bodies of water, village B is known to have a larger incidence rate than village A.

(a) You anticipate that village A will have an incidence of 20 % and village B will have an incidence of 30 %. You would like to be able to detect a relative risk of 1.5 with power of 0.90 using a 1-sided test. What size sample is needed in each village? Assume that the level of the test is 0.05.
(b) Suppose the desired power for part (a) is 0.8. What sample size is required?
(c) Last year samples of 50 were selected in each village and the 5-year incidence rates were 22 % in village A and 37 % in village B. What is the power for detecting a difference of 15 percentage points using a 1-sided 0.10-level test?
(d) Compute a 90 % 2-sided confidence interval on the difference in proportions for part (c).

4.8. A sample is to be selected from the population in a county that is age 18 or older. The proportion of persons that are unemployed will be measured. Three months later the proportion unemployed will again be recorded on a follow-up sample. It is anticipated that 75 % of the time 1 sample will cooperate at time 2. The same size sample will be maintained at time 2 by selecting additional persons.

(a) If the time 1 unemployment rate is anticipated to be 8 % and you want to be able to detect a decline of 1.5 percentage points with power 0.8 in a 1-sided, 0.05-level test, how large should the sample be at each time period? You will have to make some assumption about the proportion of persons unemployed at both times. Describe your reasoning for the value you assume.
(b) If you can only afford to sample 500 persons, what will be the power to detect a 1.5 percentage point change?

4.9. Students at public and private high schools are compared on a standardized achievement test. In previous years the average score has been about 600 (out of 800). Suppose you want to sample about twice as many public school students as private school students since there are some extra analyses you

plan for the public schools. The population relvariance of scores is known to be 0.6.

(a) What sample size of students for public and private is needed to detect an effect size of 0.10 with power 0.80? Assume that differences in either direction should be detected at a significance level of 0.05.
(b) What difference in means does this correspond to?

4.10. The Council of Governments (COG) is an organization in the Washington DC area that is funded by local governments from the District of Columbia and surrounding counties. The COG would like to fund a survey to compare crime rates in the central city to that of one of the suburban counties. It would like to select a sample of households from the two jurisdictions and conduct in-person interviews to determine whether central city residents are more likely to be victims of any type of crime than are the suburbanites. The overall metropolitan area rate of violent plus property crimes is 1,105 per 100,000 households. Analysts at COG think that the suburban crime rate is about 75 % of that of the overall rate. If the central city rate is twice the suburban rate, COG policymakers would like to be very sure that their sample will recognize that large difference. On the other hand, some COG analysts would like to know whether the central city rate is 1.5 times the suburban rate. To complicate matters, the amount of money available to do the survey is unclear because the local municipalities have not passed their budgets for the current fiscal year. Given that uncertainty, compute a range of sample sizes that you can discuss with COG. How will you describe the pros and cons of your alternatives to COG?

4.11. An organization surveys its employees in January and July to measure proficiency with the suite of data analytic software that the company supplies. Employees perform various tasks and receive an overall score between 0 and 100. Suppose that, based on past data, the average score is 72 and that the unit standard deviation of scores is 55 which is stable over time. The information technology department would like to know if the average score has changed 10 % or more from January to June. A simple random sample is selected of the employees in January. The same employees will be tested in July, if possible, but because of turnover, absenteeism, and scheduling conflicts, you expect that only 60 % of the initial sample will be retested in July. For cross-sectional analyses, the same size sample is desired at each time period. Assume that the correlation between individual scores at the two times is 0.76.

(a) Compute the sample size required in January (which will equal the size in July) that will be needed to detect a change of 10 % with power 0.80. Assume that all cases at each time period will be used to compute the difference and that the level of the test is 0.05.
(b) Repeat part (a) but assume that only the overlapping cases between January and July will be used.

(c) Calculate the variance of the estimated mean difference from parts (a)
and (b) and discuss how this relates to the sample sizes you computed in
parts (a) and (b).
(d) What assumption are you implicitly making to say that the difference in
means estimated in (a) and (b) are the same? Are there any reasons to
believe that this assumption is wrong? Explain your answer.

4.12. In the case of partially overlapping samples described in Sect. 4.3.1,
show that the variance of the difference in means, $\hat{d} = \hat{\bar{x}} - \hat{\bar{y}}$, is $V\left(\hat{d}\right) =$
$\frac{\sigma_x^2}{n_1} + \frac{\sigma_y^2}{n_2} - 2\sigma_{xy}\frac{n_{12}}{n_1 n_2}$ as shown in (4.12). When $n_{12} = \gamma n_1$ and $r = n_1/n_2$, show
that this reduces to $V\left(\hat{d}\right) = n_1^{-1}\left[\sigma_x^2 + r\sigma_y^2 - 2\gamma r\sigma_{xy}\right]$. When $\sigma_x^2 = \sigma_y^2 = \sigma_0^2$,
show that $V\left(\hat{d}\right) = \sigma_0^2 n_1^{-1}\left[1 + r\left(1 - 2\gamma\rho\right)\right]$.

Chapter 5
Mathematical Programming

Earlier chapters examined sample size determination and allocation to strata for a single variable. In reality, almost every survey of any size is multipurpose. Data on a number of different variables are collected on each sample unit. Estimates are made of population values for the full population and for various domains or subpopulations. In addition, a variety of types of estimates may be made, including means, totals, quantiles, and model parameters.

There is also a potentially long list of constraints that must be satisfied. Minimum sample sizes may be set for the domains based on, for example, a power calculation associated with a detectable difference. Targets for coefficients of variation (CV) may be set. A time schedule must be met, which dictates logistical decisions like mode of data collection and number of data collectors to hire. Above all, there is usually a limited amount of money available. Cost overruns are common, but the organization conducting the survey cannot count on getting a budget increase to cover them.

Multiple goals and constraints mean that the allocation problem is considerably more complicated than was presented earlier. In principle, these goals and constraints can be accommodated using the techniques of mathematical programming that are illustrated in this chapter. Mathematical programming is a general term that refers to choosing the best solution to some optimization problem from among the available alternatives. The term *programming* does not refer to computer programming although sophisticated computer algorithms have been developed for these problems. Instead, *program* refers to its use by the US military to refer to proposed training and logistics schedules (Freund 1994; Dantzig 1963). The term was coined by George Dantzig, who invented the area of linear programming.

As in the rest of this book, we concentrate on learning the methods of multicriteria optimization for single-stage design in this chapter and not the theory. Optimization for more complex designs is discussed in later chapters. The advantage of these methods is that they provide a formal way of solving what can be extremely complex allocation problems. The alternative is to rely on a crude diet of intuition and sense of smell. Although in the right hands

trial and error methods may eventually lead to efficient solutions, all sample designers are not equally good at this. Having a good, mathematical solution helps eliminate the guesswork. In addition, having sophisticated optimization software encourages us to carefully list all of the goals and constraints and, we can hope, produce better solutions.

5.1 Multicriteria Optimization

The general formulation of the problem is to minimize (or maximize) some (objective) function subject to constraints on cost, minimum sample size per stratum, minimum sample sizes in analytic domains, and CVs of stratum or other domain estimates. In general, an optimization problem has four parts:

1. *Objective function*—a function of one or several variables to be optimized
2. *Decision variables*—the quantities that are adjusted in order to find a solution, e.g., sample sizes
3. *Parameters*—fixed inputs that are treated as constants, e.g., stratum population counts, and variances
4. *Constraints*—restrictions on the decision variables or combinations of the decision variables, e.g., domain sizes and cost

A solution to these problems requires special algorithms and software. Some of the software options are Excel Solver, SAS `proc nlp`, SAS `proc optmodel`, and the R package `alabama` (Varadhan 2010), all of which are described in this chapter for single-stage designs.

When there are multiple variables and estimates, some judgment needs to be made about the relative significance of each to the goals of the survey. One option is to use a weighted combination of the relvariances for different estimators as the objective function. The weights could be selected based on the "importance" of each estimate to the survey goals. If, for example, the objective is a linear combination of the relvariance for the estimated proportion of employees who prefer flexible work hours and the estimated average number of sick days taken per employee, then the importance weights could each be 0.5, assuming that these two estimates are equally important. What the relative weights should be is a matter of opinion and assigning them will require conferring with the survey sponsor and, probably, some debate among the staff conducting the survey. In the end, some arbitrary assignments will probably be necessary. Sensitivity analyses can be done using different sets of importance weights.

Relvariances are convenient for forming the objective function since a relvariance is unitless, as noted in Sect. 3.1. The relvariance of the estimated mean number of employees, for example, has units $(\text{employees}^2)/(\text{employees}^2)$. If variances were used, a variable like number of employees would overshadow the effect of a 0–1 variable, like whether an establishment had laid off any workers in the last quarter.

Example 5.1 (Formal statement of an optimization problem). Suppose that a stratified single-stage sample is selected using *stsrswor*. Let $\hat{\bar{y}}_j = \sum_{h=1}^{H} W_h \bar{y}_{j,sh}$ be the stratified mean for variable j ($j = 1, 2, \ldots, J$). The stratum sample mean is $\bar{y}_{j,sh} = \sum_{i \in s_h} y_{jhi} / n_h$ with y_{jhi} being the value of variable j for sample unit hi. As in Sect. 3.1.2, the estimated domain mean for variable j is defined as

$$\hat{\bar{y}}_{dj} = \frac{\sum_h W_h p_{dh} \bar{y}_{dj,sh}}{\sum_h W_h p_{dh}},$$

where $\bar{y}_{dj,sh} = \sum_{i \in s_{dh}} y_{jhi} / n_{dh}$ is the sample mean of variable j in domain d within stratum h and $p_{dh} = n_{dh} / n_h$, the proportion of units in stratum h that are in domain d.

A formal statement of one mathematical programming problem might be the following. The terms CV_{0jh} and CV_{0dj} below are targets for the CVs of the estimated means for variable j for stratum h (a design domain as described in Chap. 3) and cross-strata domains (called a cross-class in Chap. 3).

- Find the set of sample sizes $\{n_h\}_{h=1}^H$ to minimize the weighted sum of relvariances (i.e., the *objective function*),

$$\Phi = \sum_{j=1}^{J} \omega_j \, relvar\left(\hat{\bar{y}}_j\right),$$

where $\{\omega_j\}_{j=1}^J$ are the importance weights assigned to estimates $j = 1, \ldots, J$ and $relvar\left(\hat{\bar{y}}_j\right) = V\left(\hat{\bar{y}}_j\right) / \bar{y}_{Uj}^2$.
- Subject to the constraints:

  (*i*) $n_h \leq N_h$ for all h
  (*ii*) $n_h \geq n_{\min}$, a minimum sample size in every stratum ($n_{\min} \geq 2$ in general)
  (*iii*) $\left[CV(\bar{y}_{j,sh})\right]^2 \leq (CV_{0jh})^2$ for certain strata and variables
  (*iv*) $\left[CV\left(\hat{\bar{y}}_{dj}\right)\right]^2 \leq (CV_{0dj})^2$ for certain domains and variables
  (*v*) $C = C_0 + \sum_{h=1}^H c_h n_h$

The decision variables to be adjusted in order to find a solution are $\{n_h\}_{h=1}^H$ in this case. ∎

Note that $\sum_{j=1}^J \omega_j$ need not equal 1 although normalizing them is sensible so that the relative sizes of the weights are easy to see. The vector $\{\omega_j\}_{j=1}^J$ may also contain some zero values to indicate a "relaxed" objective. This is especially useful when experimenting with inclusion or exclusion of some variables from the objective function.

The problem above is nonlinear in the decision variables because the n_h's are in the denominators of both the objective function, through $relvar\left(\hat{\bar{y}}_j\right)$,

and the constraints, through $\left[CV(\bar{y}_{j,sh})\right]^2$ and $\left[CV\left(\hat{\bar{y}}_{dj}\right)\right]^2$. In almost all non-linear problems, there are no closed-form, exact solutions like the ones we noted for stratified sampling in Sect. 3.1.2. Iterative, approximate solutions are needed but several software options are available, as described in the following sections.

Exactly how a problem is set up is important both (i) to get a solution that really addresses the goals of a survey and (ii) to formulate the problem in a way that is least burdensome for the solution algorithm. Some of the techniques for solving nonlinear optimization problems involve numerical approximations to partial derivatives of the objective function and to nonlinear constraints. How you phrase a problem can make finding a solution unnecessarily difficult for an algorithm. In Example 5.1, we could have defined the objective as the weighted sum of CVs instead of relvariances. Constraints (iii) and (iv) could also have been stated in terms of CVs. But, simpler is better. Stating the objective function and nonlinear constraints in terms of CVs makes both "more" nonlinear in the n_h's than does using relvariances because of the square root function required for CVs.

Setting up a problem that has no solution is certainly a possibility. Using constraints that are incompatible with each other is one mistake that can be made. For example, $n_h \leq N_h$ and $n_h \geq 100$ are incompatible for any strata with $N_h < 100$. More subtle errors are naturally possible. Tight constraints on relvariances may lead to a violation of a cost constraint, for example. Often the easiest way to discover these is to run the optimization and see what happens.

Good software will produce reports that inform whether or not a problem could be solved and whether any constraints were violated. The final value of the objective function should be reported along with a list of the constraining functions and their final values. A constraint that is satisfied exactly at the boundary or within some small tolerance of the boundary of the allowable value is labeled as *binding*; changing the constraint would have a direct effect on the objective function. Constraints that are easily met (and could be tightened in a subsequent optimization problem) are called *nonbinding*.

Many different algorithms have been developed to solve nonlinear optimization problems like the one in Example 5.1. The mathematics behind some of these is described in, for example, Winston and Venkataramanan (2003). Besides choosing an algorithm, software packages typically have a variety of tuning parameters that can be set to control the methods used for a solution. A user may be able to set the number of iterations before the algorithm terminates, the length of clock time the algorithm runs before stopping, the relative change between iterations in the objective function used to decide whether an optimum has been reached, and a tolerance used to determine whether a constraint is violated or not. We discuss four approaches for conducting an optimization for single-stage designs in the next sections.

5.2 Microsoft Excel Solver

Solver, a tool bundled with Microsoft Excel, is quite easy to use and can find solutions to problems as long as there are not too many decision variables or constraints. The standard Solver allows up to 200 decision variables (e.g., stratum sizes) and constraints on up to 100 cells in the spreadsheet. There are several upgraded versions that can be purchased separately from Frontline Systems, Inc.[1] The upgrades can handle much larger and more complex problems than addressed by the standard Solver and also work within Excel. A readable introductory text on the use of Solver and many other features of Excel is Powell and Baker (2003). Chapter 10 of their book, in particular, covers nonlinear optimization problems and the use of different versions of Solver.

This section describes how to set up a problem in Excel and find a solution using Solver. The example below is small but illustrates features that are common to sample allocation problems.

Example 5.2 (Optimizing a sample of business establishments). Table 5.1 gives stratum means, standard deviations, and proportions for an artificial population of business establishments. The US tax law in 2000 allowed a tax credit to be taken for certain expenses associated with doing scientific research. The column labeled "Claimed research credit" gives the proportion of establishments within business sector (classification area) that claimed the credit in a particular year. Suppose we want to find an allocation of an *stsrswor* to strata that will minimize the relvariance of estimated total revenue, $\hat{T}_{rev} = \sum_h N_h \bar{y}_{sh}$, subject to these constraints:

  (*i*) Budget on variable costs = \$300,000 US.
  (*ii*) $CV \leq 0.05$ on estimated total number of employees.
  (*iii*) At least 100 establishments are sampled in each sector, $n_h \geq 100$.
  (*iv*) The number sampled in each stratum is less than the population count, $n_h \leq N_h$.
  (*v*) $CV \leq 0.03$ on estimated total number of establishments claiming the research tax credit.
  (*vi*) $CV \leq 0.03$ on estimated total number of establishments with offshore affiliates

Offshore affiliates are companies or legal entities that are established to act as holding areas for investments. This may be a way to reduce tax liability and shield assets against future claims such as divorce proceedings, bankruptcy, creditors, and other litigation.

In this example, the population sizes in each stratum and overall are known so that optimizing for estimated totals and means will be the same (as discussed in Sect. 3.1). Recall that the relvariance of an estimated total in an *stsrswor* is

---

[1] http://www.solver.com/excel-solver.htm.

$$relvar\left(\hat{T}\right) = T^{-2} \sum_h N_h \left(\frac{N_h}{n_h} - 1\right) S_h^2$$

with T being the population total. This is a small-scale problem that is easy to unravel using Solver. The spreadsheet used in this example can be found in `Example 5.2.Solver.xls` on this book's web site. A screenshot of the spreadsheet showing row and column labels is in Fig. 5.1. The steps in using Solver are listed below.

1. Add columns to the spreadsheet that are used to calculate the statistics for the optimization. In this example, columns L, M, N, and O were added and contain the formula $N_h \left(\frac{N_h}{n_h} - 1\right) S_h^2$ for revenues, employees, the research credit, and offshore affiliates.
2. Add a column to hold the decision variables, $\{n_h\}_{h=1}^{H}$, (cells K3–K7).
3. Create a cell that contains a formula that computes the objective function. Here, the objective is $CV^2\left(\hat{T}\right) = T^{-2} \sum_{h=1}^{H} N_h \left(\frac{N_h}{n_h} - 1\right) S_h^2$ with the variable being total revenue (cell L11).
4. Add cells, if necessary, to hold formulas that compute the values that enter into the constraints. Here, the total budget is cell D12 and the computed cost for the particular sample allocation is D13. The CVs for employees, the research credit, and offshore affiliates are in M12, N12, and O12.
5. Open Solver by choosing Tools/Solver from the Data tab in Excel 2007 or 2010. If Solver is not listed, select Tools/Add-Ins and check Solver Add-in to activate the tool. In Excel 2010 select File/Options/Add-Ins/Manage Excel Add-ins.
6. Fill in the following boxes in the Solver Parameters screen: Set Target Cell, Equal to, By Changing Cells, and Subject to the Constraints. The contents of the cells for this example are (see Fig. 5.2):

   Set Objective: L11
   To: Min
   By Changing Variable Cells: K3–K7,

   Subject to the Constraints:
   $D$13 <= $D$12 (cost constraint)
   $K$3:K7 <= $C$3: $C$7 ($n_h \leq N_h$)
   $K$3: $K$7 >= 100 ($n_h \geq 100$)
   $M$11 <= 0.05^2 (relvariance of estimated total employees)
   $N$11 <= 0.03^2 (relvariance of estimated total number of establishments claiming the research tax credit)
   $O$11 <= 0.03^2 (relvariance on estimated number of establishments with offshore affiliates)

Note that Solver allows array notation so that, for example, K3 through K7 are constrained to be greater than 100 (i.e., K3:K7 >= 100) instead of constraining each cell separately. Figure 5.3 shows the Change Constraint screen in which the D13 <= D12 constraint is set. The other constraints are set in a similar way.

Tuning parameters that control how long the algorithm runs, when it stops, and methods used are set in the Solver Options screen (Fig. 5.4) which appears after clicking Options in the Solver Parameters screen. Max Time and Iterations are self-explanatory. Some of the other options relevant to sample allocation are:

Constraint Precision. This number in the All Methods tab determines how close the left-hand side value of a constraint should be to the right-hand bound in order to be satisfied. The default setting is 10^{-6}. Setting this value to an extremely small number can result in (a) Solver reporting that a constraint has been violated when for all practical matters it is simply binding without being violated or (b) Solver reporting that a solution cannot be found. Setting the Precision to too large a value can also result in "premature" convergence, i.e., a solution is found that satisfies all constraints but does not give the best value of the objective function. You can test this yourself by experimenting with different Precision values in the example in this section.

Convergence. This is in the GRG Nonlinear tab and represents the absolute value of the change in the objective function that is used to declare convergence. If the change between iterations is less than or equal to this number, then Solver stops.

Use Automatic Scaling. When this box is checked in the All Methods tab, Solver attempts to scale the values of the objective and constraint functions internally in order to minimize the effects of having values of the objective, constraints, or intermediate results that differ by several orders of magnitude.

Derivatives. This option in the GRG Nonlinear tab controls performance of the solution method. The default value of Forward can be used for most problems. At each iteration, values of derivatives of the objective and the constraints with respect to the decision variables are used. These derivatives are approximated by a technique known as differencing, the technique that is selected under the Derivatives choice. Central differencing requires more time per iteration than Forward differencing but may lead to a better search direction and fewer iterations.

The solution to this optimization problem is shown in Table 5.2. Three reports are available when a solution is found—the Answer Report, the Sensitivity Report, and the Limits Report. We will discuss the first two; the third appears to have little use in the problems we address.

The Answer Report summarizes the original and final values of the decision variables and constraints, with additional information about which constraints are binding. Figure 5.5 shows the Answer Report for this example. First, the original and final values of the objective function are listed. Initial

Table 5.1: Stratum population means, standard deviations, and proportions for an artificial population of business establishments

				Population means		Population standard deviation		Population proportion	
h	Business sector	Estab-lishments N_h	c_h	Revenue (millions)	Employees	Revenue (millions)	Employees	Claimed research credit	Had offshore affilates
1	Manufacturing	6,221	120	85	511	170.0	255.50	0.8	0.06
2	Retail	11,738	80	11	21	8.8	5.25	0.2	0.03
3	Wholesale	4,333	80	23	70	23.0	35.00	0.5	0.03
4	Service	22,809	90	17	32	25.5	32.00	0.3	0.21
5	Finance	5,467	150	126	157	315.0	471.00	0.9	0.77
	Pop Total	50,568		1,834,157	5,316,946			21,254	9,855

Fig. 5.1: Excel spreadsheet set-up for use with Solver

Table 5.2: Solution to the optimization problem in Example 5.2

Stratum	Sector	n_h
1	Manufacturing	413
2	Retail	317
3	Wholesale	119
4	Service	1,399
5	Finance	598
Total		2,846

values for the n_h's are needed to get the algorithm started; $n_h = 500$ was used for all strata in this case. An alternative would be to use our specified minimum value, $n_h = 100$. In this example, both starting values lead to almost the same solution. Next, the original and final values for the "adjustable cells," i.e., the decision variables, are listed.

The third section in Fig. 5.5 shows the constraints with their final cell values; a Formula column showing the spreadsheet formula entered by the user; a Status column showing whether the constraint was binding or nonbinding at the solution; and the slack value. The Name column is the combination of
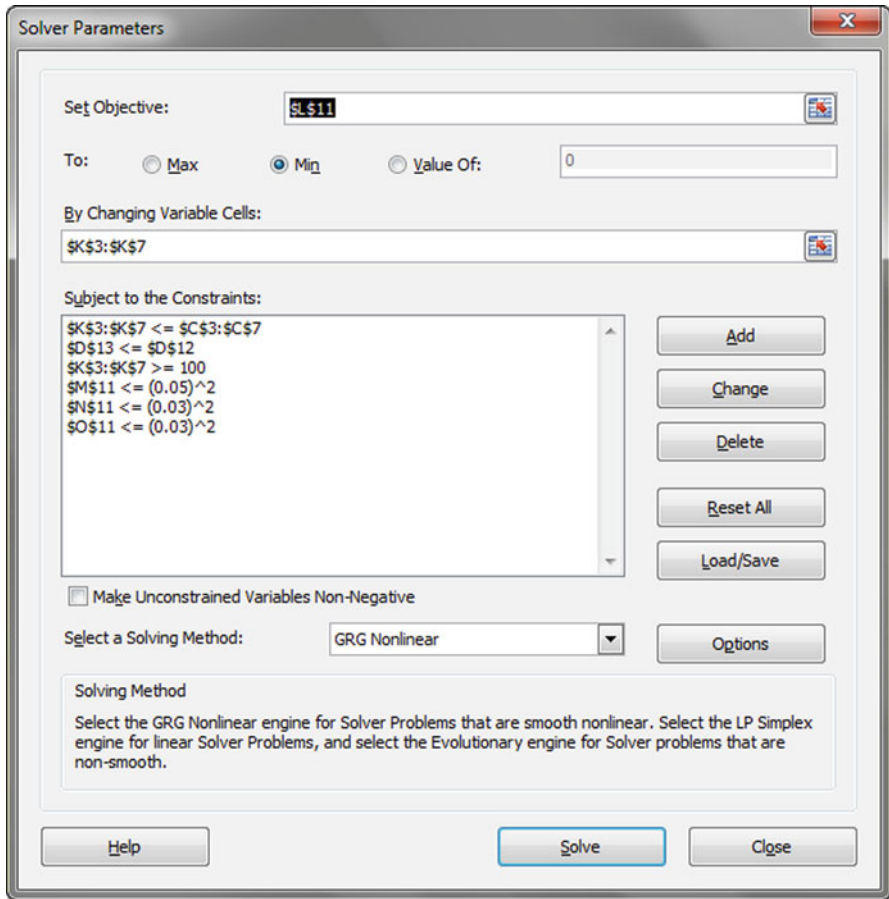
Fig. 5.2: Screenshot of the Excel Solver dialogue screen
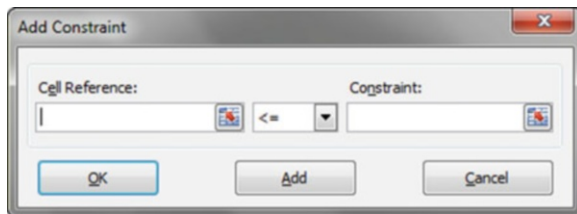
Fig. 5.3: Screenshot of the Change Constraint dialogue screen

the row and column label for the constraint, e.g., `relvariance of t.hat Offshore Nh*(Nh/nh - 1) * Sh^2`. The slack is the difference between the final value and the lower or upper bound imposed by that constraint. A binding constraint, which is satisfied with equality or with a negligible
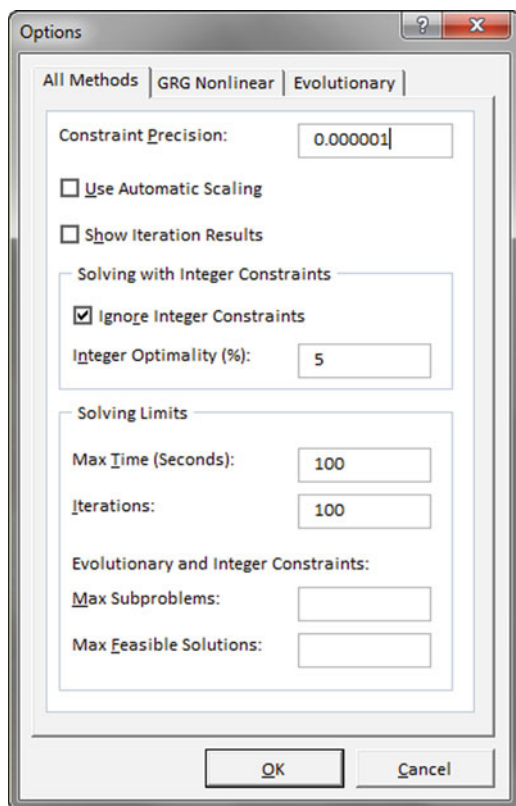
Fig. 5.4: Solver Options window where tuning parameters can be set and models saved

difference, will always have a slack of zero. Total sample cost and the relvariance of the proportion with offshore affiliates are both binding. Thus, the final allocation uses all of the available funds.

The Sensitivity Report in Fig. 5.6 provides information about how the solution would change for small changes in the constraints or the objective function. The two sections of the report are labeled Adjustable Cells and Constraints. The figures under the columns, Reduced Gradient and Lagrange Multiplier, are called *dual values*. For this example, the only interesting values are those under Constraints. The dual value for a constraint is nonzero only when the constraint is binding. Moving the value of the left-hand side of the constraint away from the bound will make the objective function's value worse; relaxing the bound will improve the objective. The dual value measures the increase in the value of the objective function per-unit increase in the constraint's bound.

In manufacturing applications where some number of products is built, interpretation of the dual value of a constraint can be fairly simple. For example, building one more of some electronic component might lead to a decrease in profits of \$100 if the Lagrange multiplier is negative. Interpretation in this example is less straightforward. The cost is constrained to be \$300,000. By relaxing this bound by 1 unit (i.e., increase the budget by \$1), the objective should change by $-1.644\mathrm{E}-08$ (i.e., the relvariance of estimated total revenue will reduce slightly). Since this is a minuscule change, a more meaningful approach would be to ask what would be the effect of increasing the budget by a substantial amount. For example, if the budget were increased by \$50,000, the relvariance would change by $50,000 \times (-1.644\mathrm{E}-08) = -0.00082$. That is, the relvariance would change to $0.002167 - 0.00082 = 0.001345$. This corresponds to a change in CV from $\sqrt{0.002167} = 0.0466$ to $\sqrt{0.001345} = 0.0367$.

The scale of the constraint is important when interpreting a Lagrange multiplier. For example, suppose the constraint on the relvariance of offshore affiliates was binding and its Lagrange multiplier was -4. A change of 1 unit to the relvariance constraint that leads to a change of -4 in the objective function would make the relvariance of total revenues negative, which is not possible. In such a case, the standard interpretation of the dual value can be made only for very small changes in the constraint. For example, suppose that the CV bound on the offshore estimate increases from 0.030 to 0.032. This implies that the change in the relvariance on that estimate is $0.001024 - 0.0009 = 0.000124$ (or, a 14 % increase in the offshore relvariance). This, in turn, means that the objective value should change by $-4 \times 0.000124 = -0.00049636$. Thus, the objective, which is the relvariance of total revenue, should change to $0.002167 - 0.000496 = 0.00167$; or, the CV of total revenue should change to $\sqrt{0.00167} = 0.0409$.

Rather than going through this sort of calculation, the simplest thing to do in an easy problem is to change the constraint and rerun the problem. The reader can verify by rerunning the optimization that changing the constraint on the budget to \$350,000 leads to a CV on estimated total revenues of 0.0387 rather than 0.0367 as predicted from the Lagrange multiplier analysis. ∎

When running variations on a problem by changing constraint values, importance weights in the objective, or something else, good practice is to save some or all of the variations so that they can be revisited if necessary. There are two ways of doing this. One is to save each variation as a new spreadsheet or a new tab within one spreadsheet. The other is to save more than one model in one tab of the spreadsheet. To save a model, click the Load/Save button in the Solver Parameters window in Fig. 5.2. Upon clicking the Load/Save Model button, a dialogue box appears where the range of cells can be specified where you want to save the model. The dialogue tells you to select an empty range of cells long enough to hold the information that Solver needs to store. In the example in this section, ten cells are needed. Putting a header cell over this range with a meaningful name is good documentation. To save another model, modify the Solver Parameter setup

as desired, then save the model in a different range of cells. To load one of
the models, open the Solver Parameters window, click Load/Save, and select
the range of cells that contains the model you want.

Section 5.6 gives some general remarks on how to track variations of opti-
mization problems that may be tried. As in all applications, good bookkeeping
is a critical part of good organization.

Solver Engine
 Engine: GRG Nonlinear
 Solution Time: 0.063 Seconds.
 Iterations: 19 Subproblems: 0
Solver Options
 Max Time 100 sec, Iterations 100, Precision 0.000001
 Convergence 0.0001, Population Size 100, Random Seed 0, Derivatives Forward, Require Bounds
 Max Subproblems Unlimited, Max Integer Sols Unlimited, Integer Tolerance 5%, Solve Without Integer Constraints

Objective Cell (Min)

Cell	Name	Original Value	Final Value
$L$11	relvariance of t.hat Revenue Nh*(Nh/nh - 1) * Sh^2	0.01298691	0.00216695

Variable Cells

Cell	Name	Original Value	Final Value	Integer
$K$3	Manufacturing nh	100	413	Contin
$K$4	Retail nh	100	317	Contin
$K$5	Wholesale nh	100	119	Contin
$K$6	Service nh	100	1,400	Contin
$K$7	Finance nh	100	598	Contin

Constraints

Cell	Name	Cell Value	Formula	Status	Slack
$D$13	Total sample cost ch	$ 300.000	$D$13<=$D$12	Binding	0
$M$11	relvariance of t.hat Employees Nh*(Nh/nh - 1) * Sh^2	0.000571	$M$11<=(0.05)^2	Not Binding	0.001928659
$N$11	relvariance of t.hat Research credit Nh*(Nh/nh - 1) * Sh^2	0.000437	$N$11<=(0.03)^2	Not Binding	0.000462853
$O$11	relvariance of t.hat Offshore Nh*(Nh/nh - 1) * Sh^2	0.000901	$O$11<=(0.03)^2	Binding	0
$K$3	Manufacturing nh	413	$K$3<=$C$3	Not Binding	5808.429995
$K$4	Retail nh	317	$K$4<=$C$4	Not Binding	11421.17068
$K$5	Wholesale nh	119	$K$5<=$C$5	Not Binding	4214.302446
$K$6	Service nh	1,400	$K$6<=$C$6	Not Binding	21409.37536
$K$7	Finance nh	598	$K$7<=$C$7	Not Binding	4869.111793
$K$3	Manufacturing nh	413	$K$3>=100	Not Binding	313
$K$4	Retail nh	317	$K$4>=100	Not Binding	217
$K$5	Wholesale nh	119	$K$5>=100	Not Binding	19
$K$6	Service nh	1,400	$K$6>=100	Not Binding	1,300
$K$7	Finance nh	598	$K$7>=100	Not Binding	498

Fig. 5.5: Solver's Answer Report for the business establishment example

Starting Values.   Finally, we note that the solution may be sensitive to
the starting values of the decision variables. In the business establishment
example, we started with $n_h = 100$ in each stratum, but other possibilities
would be proportional allocation, Neyman allocation for revenues, or one
of the other univariate allocations from Chap. 3. It is advisable to find solu-
tions using several different sets of starting values, which are substantially
different from each other. If the same, or a very similar solution is obtained
from each set, this provides some assurance that a global optimum has been
found. This is usually called a sensitivity analysis because you are evalu-
ating the sensitivity of the solution to, in this case, the starting values.

Variable Cells

Cell	Name	Final Value	Reduced Gradient
$K$3	Manufacturing nh	412.5700053	0
$K$4	Retail nh	316.8293242	0
$K$5	Wholesale nh	118.6975545	0
$K$6	Service nh	1399.624645	0
$K$7	Finance nh	597.888207	0

Constraints

Cell	Name	Final Value	Lagrange Multiplier
$D$13	Total sample cost ch	$300,000.00	-1.64405E-08
$M$11	relvariance of t.hat Employees Nh*(Nh/nh - 1) * Sh^2	0.000571341	0
$N$11	relvariance of t.hat Research credit Nh*(Nh/nh - 1) * Sh^2	0.000437147	0
$O$11	relvariance of t.hat Offshore Nh*(Nh/nh - 1) * Sh^2	0.000901	0

Fig. 5.6: Solver's Sensitivity Report for the business establishment example

You can also have Solver use multiple starting values automatically. In the Solver Parameters window, select Options. Then, in the Options window, choose the GRG Nonlinear tab and check the Use Multistart box. If this box is selected when you click Solve, the GRG Nonlinear method will run repeatedly, starting from different (automatically chosen) starting values for the decision variables. This process may find a better solution, but it will take more computing time than a single run of the GRG Nonlinear method.[2]

Limitations on Number of Decision Variables.   The standard Solver has a limit of 200 decision variables for both linear and nonlinear problems. By "linear" we mean that both the objective function and constraints are linear combinations of the decision variables. However, an upgraded version of Solver has limits of 2,000 decision variables for linear problems and 500 for nonlinear problems.

Limitations on Number of Constraints.   The standard Solver has a limit of 100 cells that can be constrained; the decision variables are not included in this list. Although this seems generous, exceeding this limit is not hard to do. If a population has 110 strata and the constraint is set that $n_h \leq N_h$ separately in each stratum, the limit is exceeded. A work-around is to set one cell equal to $\max_h(nh/Nh)$ and to constrain this cell to be less than or equal to 1. Thus, 110 constraint cells are converted to 1 constraint without changing the goals of the problem. Similarly, if a CV of 0.05 is desired for several different estimates, then a single cell can be defined that holds $\max\left(CV^2\right)$ over the set of estimates.

---

[2] More detailed help is available for this and all other options at www.solver.com/excel2010/solverhelp.htm.

Solver can also handle linear programming problems as illustrated in the next example of selecting a subsample from an existing sample. Example 5.3 is a particular instance of two-phase sampling which we cover in more detail in Chap. 17. The general idea in two-phase sampling is to select a second-phase sample from an initial (first-phase) sample based on information gained in the first phase.

Example 5.3 (Determining subsample sizes). Suppose that a sample of households is selected with the goal of obtaining specified numbers of children in the age groups 5–11 years, 12–14 years, and 15–17 years old. An initial sample of 27,400 households is selected and the numbers of children in each age group in each household is recorded based on a screening interview. The numbers of children in the initial sample and the target sample sizes for the subsample in each subgroup are:

	Age group	Number of children in initial sample	Target sample size of children in subsample
1	5–11 years	6,229	1,000
2	12–14 year	3,009	2,000
3	15–17 years	3,159	2,000
	Total	12,397	5,000

A household may contain no children in any of these groups or children in any combination of the three groups. We could simply list the children in each age group and sample each list separately to obtain 1,000, 2,000, and 2,000 in each group. However, this would not exercise any control over how many HHs were selected; nor would it control the number of children sampled per HH. We would like to sample only one child per household to limit reporting burden. Only households (HHs) that have children in one or more of the three age groups above will be eligible for the second-phase survey. Strata of households are indexed by the age groups of the children contained in the HHs: 1, 2, 3, 12, 23, 13, and 123. For example, stratum 13 is composed of HHs that have children in the age groups (1) 5–11 and (3) 15–17. To specify the problem clearly, we need some notation:

a_h = sampling rate of HHs in stratum h (to be determined using mathematical programming)
$C_{hi(k)}$ = number of children in HH i, stratum h in age group k (k=1, 2, 3)
$C_{hi(+)}$ = number of children in HH i, stratum h across all age groups
n_h = number of first-phase sample HHs that are in stratum h

If one child is selected at random without regard to age group in a HH, the selection probability within the HH is $1/C_{hi(+)}$. The expected number of children sampled from age group k in HH hi is also the proportion of children in that age group in the HH:

$$p_{hi(k)} = \frac{C_{hi(k)}}{C_{hi(+)}}.$$

The expected number of children selected from age group k across all HHs is

$$e_k = \sum_h a_h \sum_{i \in s_h} p_{hi(k)}$$

$$= \sum_h a_h n_h \bar{p}_{h(k)}, \qquad (5.1)$$

where s_h is the set of first-phase HHs in stratum h, and $\bar{p}_{h(k)} = \frac{1}{n_h} \sum_{i \in s_h} p_{hi(k)}$ is the average proportion of children per HH in stratum h that are in age group k. The total number of children subsampled is $e_+ = \sum_{k=1}^3 e_k$. The expected total number of HHs subsampled is

$$E_{HH} = \sum_h a_h n_h.$$

Since we only have seven HH strata and three age groups of children, the various parameters can be displayed in a short table.

Stratum	Sampling rate for HHs	No. of HHs	Average proportion of children $\bar{p}_{h(k)}$ per HH that are in age group k		
h	a_h	n_h	$k = 1$	2	3
1	a_1	n_1	1	0	0
2	a_2	n_2	0	1	0
3	a_3	n_3	0	0	1
12	a_{12}	n_{12}	$\bar{p}_{12(1)}$	$\bar{p}_{12(2)}$	0
13	a_{13}	n_{13}	$\bar{p}_{13(1)}$	0	$\bar{p}_{13(3)}$
23	a_{23}	n_{23}	0	$\bar{p}_{23(2)}$	$\bar{p}_{23(3)}$
123	a_{123}	n_{123}	$\bar{p}_{123(1)}$	$\bar{p}_{123(2)}$	$\bar{p}_{123(3)}$

Setting subsampling rates for HHs in each of the 7 strata can be formulated as a linear programming problem:

- Find the set of rates a_h that minimize the expected number of HHs, E_{HH}, selected.

- Subject to these constraints:

  (i) $e_1 = 1,000, e_2 = e_3 = 2,000$
  (ii) $\min_a < a_h \leq 1$ for all strata with $\min_a$ being the minimum sampling rate allowed for any stratum

Whether this problem can be solved or not depends, in part, in the value for $\min_a$. If it is set too high, it may not be possible to find a feasible solution, i.e., one that satisfies all constraints.

The Excel sheet, `Example 5.3 Subsampling age strata.xlsx`, which is on the web site for this book, has this problem set up in the Solver data analysis tool for the set of values for n_h and $\bar{p}_{h(k)}$ shown in Fig. 5.7. The solution is also given in Fig. 5.7. In this example, $\min_a = 0.1$. The values of $\bar{p}_{h(k)}$ are denoted by ph(k) in the spreadsheet. The decision variables a_h are in cells B11:B17 in the spreadsheet. In this case, the solution gives 5,000 total children subsampled and 5,003 HHs—a slight mismatch in the one child per HH requirement due to rounding. Strata 12, 13, and 23 are subsampled at the minimum allowed rate of 0.1. Strata 1, 2, 3, and 123 are subsampled at rates of about 0.161, 0.947, 0.793, and 1.00. ■

	A	B	C	D	E	F	G	H	I	J
1										
2	Age group	Target subsample sizes								
3	5-11	1000								
4	12-14	2000								
5	15-17	2000					Age groups			
6					Initial sample			Subsample		
7				5-11	12-14	15-17	5-11	12-14	15-17	
8	Stratum	Sampling rate for HH	Households		ph(k)			Expected no. of children		Expected no. of HHs
9	h	ah	nh	k=1	2	3	ah*nh*ph(1)	ah*nh*ph(2)	ah*nh*ph(3)	ah*nh
10	0	0	15000	0	0	0				
11	1	0.161	4700	1	0	0	758	0	0	758
12	2	0.947	1900	0	1	0	0	1800	0	1800
13	3	0.793	2300	0	0	1	0	0	1825	1825
14	12	0.10	1300	0.5	0.5	0	65	65	0	130
15	13	0.10	1300	0.6	0	0.4	78	0	52	130
16	23	0.10	600	0	0.6	0.4	0	36	24	60
17	123	1.00	300	0.33	0.33	0.33	99	99	99	300
18		Total children		6229	3009	3159		Total expected no. of HHs		5003
19										
20								Expected no. of children in sample		
21							k=1	2	3	All ages
22		Total HHs with children	12400				1000	2000	2000	5000

Fig. 5.7: Excel spreadsheet for finding subsampling rates via linear programming

Example 5.3 is a good illustration of the usefulness of mathematical programming in a problem different from the ones where minimizing variances is the goal. Be mindful that math programming can apply in many different situations and can provide better solutions to problems than crude trial-and-error approaches.

5.3 SAS PROC NLP

Multicriteria optimization in SAS can be conducted using the procedures `proc nlp` (<u>non</u>linear <u>p</u>rogramming) or the newer `proc optmodel`. We present details associated with the latter procedure in the next section. SAS `proc nlp` has fewer restrictions on factors such as the number of constraints than noted with the standard Solver. `Proc nlp` will solve problems of the form

$$\min_{x \in R^n} f(x), x = (x_1, \ldots, x_p)$$

subject to

$$
\begin{aligned}
c_i(x) &= 0 & i &= 1, \ldots, m_1, \\
c_i(x) &\geq 0 & i &= m_1, \ldots, m_1 + m_2, \\
\ell_j &\leq x_j \leq u_j & j &= 1, \ldots, p.
\end{aligned}
$$

The vector $\mathbf{x}$ contains the decision variables; the $c_i(\mathbf{x})$ are equality or inequality constraints. The decision variables have lower and upper bounds as specified by $\ell_j \leq x_j \leq u_j$. Note that a maximization problem, i.e., $\max_{x \in R^n} f(x)$ can be set up by using $-f(x)$ as the objective function; however, the user can specify whether an objective is to be minimized or maximized without worrying about the sign of $f(x)$. This general formulation fits for sample allocation problems with $\mathbf{x}$ being the sample sizes. Some of the advantages of `proc nlp` are:

- There are no specific limits on numbers of decision variables and constraints other than those imposed by computer memory and hard drive size.
- Detailed documentation is produced in a SAS log file for bookkeeping or for the project archive.
- Other features of SAS are available for data manipulation and analysis.

The set up for `proc nlp` differs from Solver though the formulation behind the optimization is the same. As an example, we revisit Example 5.2 with a simple SAS program. Detailed information on more advanced techniques in `proc nlp` (and other procedures) may be obtained from the SAS OnlineDoc web site.[3] Once at the SAS OnlineDoc web site, choose the set of online documents (HTML or pdf format) associated with your version of SAS. NLP is part of the operations research package SAS/OR. The pdf version of the documentation is best used for printing. The section on `proc nlp` gives descriptions of the various algorithms SAS offers along with some advice on what to consider when selecting an algorithm.

In any computer language in which program code is written to perform a task, it is good practice to document the program. This can be done through (i) comments within the program, (ii) a separate documentation "help" file, and/or (iii) in the case of more complicated general purpose programs, a

---

[3] http://support.sas.com/documentation/.

user's guide. For your own special purpose programs, choice (i) should be sufficient. The comments should include a header giving:

- Name of file that contains the program
- Purpose of the program
- Name of programmer
- Date written
- Date(s) revised and changes made in each revision

Choices (ii) and (iii) above are used by R, SAS, Stata, and other multipurpose packages. We discuss program documentation in more detail in Chap. 18.

Example 5.4 (Solve the business establishment allocation with SAS nlp). The SAS 9.2 `proc nlp` code, program log, and output file used in this example are located in the `Example 5.4 (NLP)` files (*.sas*, *.log*, and *.lst* files, respectively) on the book's web site. The code is also shown in Code 5.1.

Assign Initial Values. Initial values for the decision variables, $\{n_h\}_{h=1}^5$, are entered in a data set called `start500` that is then loaded by `proc nlp` via the `INEST` option. Each stratum sample size was initialized to 500 as in the Solver example. (The SAS code also creates a file called `start100` that can be used for comparison. Both starting points produce similar solutions, although initial values of 100 will lead to a misleading message that the algorithm did converge.) If initial values are not assigned the procedure will assign its own randomly selected values for n_h which are near zero. In this example, assigning all stratum sample sizes to be initially 500 does not lead to a better solution.

Load Optimization Parameters. The first step within `proc nlp` is to load the optimization parameter values by design stratum (business sector), as used in Solver, into a set of SAS variables. These include the population counts (`Nh[5]`, i.e., an array of length 5), cost values (`cost[5]`), population means and proportions (`p[4,5]`), and the population standard deviations (`sd[4,5]`) for the four analysis variables shown in Table 5.1. The order of the variables in the means and standard deviation arrays (i.e., matrices) is revenue, employees, research credit, and offshore affiliates so that, for example, the first rows (i.e., `p[1,]` and `sd[1,]`) correspond to the values for revenue. Note that the standard deviations for research credit and offshore affiliates are calculated using DO loops instead of "hard-coded" because estimates for the binary variables can be computed directly within the program.

Declare the Decision Variables. Our ultimate goal is to calculate the sample size to be selected within each business sector for the survey. The stratum sample sizes are loaded into an array of length five, i.e., `n[5]`, for use in the objective function and defined as the decision variables in the `DECVAR` statement. Note that the variables in the `start500` data set are named `n1`–`n5` to match the array in `DECVAR`.

Define the Constraints.   The first set of constraints is defined specifically for the decision variables. Based on the specifications of the problem, each stratum size must be bounded below by 100 (n1-n5 >= 100) and above by the corresponding frame count (e.g., n4 <= 22809). Additionally, the cost for the study must be linearly constrained (LINCON) to be less than or equal to the maximum budget of \$300,000 where cost is defined as $\sum_{i=1}^{4} cost[i] \times n[i]$.

   Additional nonlinear constraints (NLINCON) are imposed on the relvariance for the totals of three analysis variables—see constraints (ii), (v), and (vi) in Example 5.2. The relvariances (squares of the CVs) are calculated again using arrays in the later portion of the program and are constrained to be less than or equal to the values specified (e.g., relvar2 <= 0.0025 = 0.05^2). To facilitate the relvariance calculation, the five stratum means or proportions for each variable are converted to their corresponding estimate of the total (m1-m20) by multiplying the original values by the population size within each sector. As advised in Sect. 5.1, we constrain the relvariances not the CVs in order to make the form of the constraints simpler.

Specify the Objective Function.   The final step is to program the objective function Φ—the importance-weighted sum of the relvariances of estimated total of revenue, employees, total establishments claiming the research credit, and total establishments with offshore affiliates. This is accomplished in proc nlp by assigning the importance weight (impwts[j]) times the relvariance (relvar[j]) for each variable to the array elements, f1-f4. The statement MIN f1-f4 tells the procedure to minimize the sum of f1 through f4. Since impwt[1]=1 and the other importance weights are zero, the relvariance of only the estimated total revenues is minimized. The SAS code is written in general terms to illustrate how a problem would be set up for a multicomponent objective function.

The Optimization Procedure.   The final step prior to submitting the proc nlp code is to specify the optimization technique from among a list of 12 options (see the SAS OnlineDoc web site for more details). We chose the Nelder-Mead simplex technique (TECH=nmsimp) because of the problem has nonlinear constraints (see, e.g., constraint (iii) in Sect. 5.1). The other algorithm option that allows nonlinear constraints is the quasi-Newton method (TECH=quanew). After some experimentation, we found that Nelder-Mead was preferable for the examples in this chapter.

A Quick Note on the Program Log.   As with any program, viewing the program log is critical to determine if the code ran correctly. SAS notes include both compilation and execution messages. If there were syntax errors, illegal combinations of solving technique and options, or other violations, then such information would be displayed in the program log. The log also shows when the program was run and what the input and

output files were, if any. Retaining the log file as part of project records is an essential part of good documentation.

The Optimization Results. The output file (`Example 5.4(NLP).lst`) contains a lot of information, but we will focus only on certain sections. First, it is important to check the specifications for the optimization problem such as the summary statistics presented in Table 5.3).

The results for our optimization are located in the section entitled, `Optimization Results`. Table 5.4 summarizes the Solver and `nlp` results along with those from `proc optmodel`, which we cover in the next section. Summary results in the SAS `lst` file are listed on number of iterations, maximum constraint violations, and final value of the objective function, among other things. In this example, only 11 iterations were needed to find a solution. The sector-specific sample sizes (`n1-n5`) in Table 5.4 from `proc nlp` are almost the same as derived from the Solver optimization (see `Estimate` column in the `lst` file) and sum to an overall sample size of 2,848 after rounding up each value. This sample allocation satisfies the study budget constraint of $300,000 and the constraints on the CVs of estimated total number of employees and number of establishments claiming the research credit. There is a minor violation in the CV for the fourth variable, offshore affiliates (`relvar4_L 0.000900 -325E-21 Active NLIC`), but this is of no practical importance. We also note that the estimated relvariance for the total amount of revenue is given by the objective function (`Value of Objective Function = 0.0021705237`). Taking the square root gives the CV of total revenues of about 4.7 %, which is larger than that of the other estimates.

Table 5.3: Summary statistics from `PROC NLP` output

Summary statistics		Interpretation
Parameter estimates	5	Sample size per five sectors
Functions (observations)	4	Relvariances for four variables
Lower bounds	5	Sample sizes (5) greater than 100
Upper bounds	5	Sample sizes (5) less than pop sizes
Linear constraints	1	Cost model
Nonlinear constraints	3	Constraints on three CVs

Code 5.1: SAS 9.2 `proc nlp` code for the optimization problem in Example 5.2

```
/**************************************************************/
/* FILE:    Example 5.4 (NLP).sas                           */
/* PROJECT: Practical Tools for Designing and Weighting Survey */
/*          Samples                                          */
/* PURPOSE: Compare results from Solver for course example.  */
/* DATE:    10/17/2010                                       */
```

```
/* AUTHOR:  J.Dever, R.Valliant                                */
/***************************************************************/
options nocenter;

         * Initialize stratum sample sizes;
data start100 (type = est);
     input _type_ $ n1 n2 n3 n4 n5;
     datalines;
          parms 100 100 100 100 100
     ;
run;

data start500 (type = est);
     input _type_ $ n1 n2 n3 n4 n5;
     datalines;
          parms 500 500 500 500 500
     ;
run;

*****************************************************************;
** Optimization - Nelder-Mead Method.                        **;
*****************************************************************;
PROC NLP INEST=start500 TECH=nmsimp
     OUT=aa;
 *_____ LOAD PARAMETERS _____*;
                                    **   Population counts **;
 ARRAY Nh[5] 6221 11738 4333 22809 5467;
                                    ** Stratum cost values **;
 ARRAY cost[5] 120 80 80 90 150;
                                 ** Means and proportions **;
 ARRAY p[4,5]  85     11     23     17     126
              511     21     70     32     157
                0.8   0.2    0.5    0.3    0.9
                0.06  0.03   0.03   0.21   0.77;

                             ** Population Standard deviations **;
 ARRAY sd[4,5] 170   8.8  23 25.5 315
              255.5 5.25 35 32    471;
                                 ** Calculate for proportions **;
 DO J=3 TO 4;
   DO I=1 TO 5;
     sd[j,i] = sqrt(p[j,i] * (1 - p[j,i]) * Nh[i] / (Nh[i] - 1));
   END;
 END;


 *_____ DECISION VARIABLES _____*;
         ** Optimized Values = Stratum-specific Sample Sizes **;
 ARRAY n[5] n1-n5;
 DECVAR n1-n5;

 *_____ CONSTRAINTS _____*;
                  ** Bounds on Stratum-specific Sample Sizes **;
 BOUNDS n1-n5 >= 100,
```

```
       n1 <= 6221, n2 <= 11738, n3 <= 4333, n4 <= 22809, n5 <= 5467;

                    ** Linear Constraint = Overall Cost Constraint **;
  LINCON 120*n1 + 80*n2 + 80*n3 + 90*n4 + 150*n5
         <= 300000;
                    ** Calculate Stratum Components, Overall total **;
  ARRAY m[4,5] m1-m20;
  DO J=1 TO 4;
    DO I=1 TO 5;
      m[j,i] = p[j,i] * Nh[i];
    END;
  END;
                               ** Variable-specific relvariances **;
  ARRAY v[4,5]    v1 - v20;
  ARRAY var[4]    var1 - var4;
  ARRAY tot[4]    tot1 - tot4;
  ARRAY relvar[4] relvar1 - relvar4;

  DO J=1 TO 4;
    DO I=1 TO 5;
      v[j,i] = ((Nh[i]**2/n[i]) - Nh[i]) * (sd[j,i]**2);
    END;

    var[j] = v[j,1] + v[j,2] + v[j,3] + v[j,4] + v[j,5];
    tot[j] = m[j,1] + m[j,2] + m[j,3] + m[j,4] + m[j,5];
    relvar[j] = var[j] / tot[j]**2;
  END;

                    ** Non-Linear Constraints = Max Value for CV **;
  NLINCON relvar2 <= 0.0025, relvar3 <= 0.0009, relvar4 <= 0.0009;

  *_____ OBJECTIVE FUNCTION _____*;
  ARRAY impwts[4] 1 0 0 0;                      ** Importance weights **;
  ARRAY f[4] f1-f4;                    ** Function to be Minimized **;
  MIN f1-f4;

  DO J=1 TO 4;
    f[j] = impwts[j] * relvar[j];
  END;
  RUN;
/**************************************************************/
```

∎

5.4 SAS PROC OPTMODEL

SAS contains a number of options for multicriteria optimization. In addition to proc nlp, proc optmodel is very useful for allocating sample cases to design strata through a nonlinear optimization. The optmodel procedure has many of the same advantages noted for proc nlp. This newer SAS

Table 5.4: Summary of results for Solver, `proc nlp`, `proc optmodel`, and `constrOptim.nl` optimization solutions

h	Excel solver (init=500)	SAS NLP[a] (init=500)		SAS OPTMODEL[a] with SQP (init=100)		constrOptim.nl (init=1,100)	
Sector	n_h	n_h	$Diff^c$	n_h	$Diff$	n_h	$Diff$
1 Manufacturing	413	413	0	363	−50	430	17
2 Retail	317	318	1	372	55	233	−84
3 Wholesale	119	124	5	103	−16	114	−5
4 Service	1,399	1,397	−2	1,384	−15	1,535	136
5 Finance	598	596	−2	626	28	550	−48
	2,846	2,848	2	2,848	0	2,862	16
	CV	CV	$\%\ RelDiff$	CV	$\%\ RelDiff$	CV	$\%\ RelDiff$
1 Revenue (millions)[b]	4.65 %	4.66 %	−0.01 %	4.71 %	0.99 %	4.75 %	1.93 %
2 Employees	2.39 %	2.39 %	0.07 %	2.42 %	1.17 %	2.44 %	2.09 %
3 Research credit	2.09 %	2.08 %	−0.11 %	2.08 %	−0.38 %	2.19 %	4.78 %
4 Offshore affiliates	3.00 %	3.00 %	—%	3.00 %	—%	3.00 %	—%
Objective function	0.217 %	0.217 %	—%	0.221 %	2.07 %	0.226 %	4.15 %

[a] The procedures were implemented in SAS 9.2

[b] Minimized in the optimization

[c] Diff = difference from Solver solution

procedure uses "optmodel language" which is advertised as enabling a quick translation of an optimization "word problem" into executable program code. However, the nonlinear optimization techniques currently listed for this procedure are fewer than those specified for `proc nlp`.

Example 5.5 (Optimization with SAS optmodel). We recast the `proc nlp` code presented in Example 5.4 as SAS 9.2 `proc optmodel` code for comparison. The `proc optmodel` code, program log, and output file used in this example are located in the corresponding Example 5.5 (OptModel) files on the book's web site. The code is also shown in Code 5.2.

The program code follows the outline developed for the previous `proc nlp` example with a few exceptions. For example, the optimization parameters in this example are loaded from the Example_55 data file through a READ DATA statement. The `optmodel` PRINT statements throughout the code print the initial values to the output (`.lst`) file for verification purposes. Both linear and nonlinear constraints are specified with the CON statement. Additionally, we forgo the importance weights in this example and instead minimize only the relvariance for the revenue variable. In this case, initializing the stratum sample sizes to 100 rather than 500 produces a lower value

of the objective function. The initialization is done with the statement that
specifies the decision variables:

```
VAR NSamp{i in 1..5} init 100;
```

The "SOLUTION" section of the program contains statements that invoke
the optimization routine. The first `SOLVE` statement calculates an optimal
allocation with a default method that is appropriate for the specified opti-
mization problem. In this case, the default technique is the SQP, a general
nonlinear programming method. The subsequent `PRINT` statements display
the stratum sizes, the overall sample size, and the resulting relvariance for
the four analysis variables. The value of the objective function (relvariance
of revenue) is slightly higher (0.002214 vs. 0.002171) than the Nelder-Mead
method applied with `proc nlp`. Similar results were obtained using the
quasi-Newton method (`tech=quanew`) in the second `SOLVE` statement. The
overall sample size is similar between the `optmodel` and `nlp` procedures,
but there are differences for the stratum-specific sample sizes. This further
emphasizes that multiple solutions are possible to one optimization problem;
comparing the solutions under different optimization techniques (i.e., sensi-
tivity analysis) is always a useful practice.

The last section of the code, prior to the `QUIT` statement, outputs the
stratum ID (`Stratum`) and the optimization solution (`Resp_Alloc`) to a
text file called `OptModel.strata.out`. With this text file, a subsequent
SAS program can be constructed to inflate the number of respondents by
specified ineligibility and nonresponse rates to produce the final sample size
(see Chap. 6) and then to randomly select the cases from the sampling frame.
Without this text file, statisticians must, for example, cut-and-paste the opti-
mization results into the sampling program—a problem when the optimiza-
tion must be rerun multiple times with changes to the constraints and/or
when the number of strata is much larger than the example presented here.

Code 5.2: SAS proc optmodel code for the optimization problem in Example 5.2

```
/****************************************************************/
/* Program: Example 5.5 (OptModel).sas                         */
/* Date:    10/17/10                                           */
/* Author:  J.Dever                                            */
/* Purpose: Solve example optimization problem.                */
/****************************************************************/
options nocenter orientation=portrait

TITLE1 "Example 5.5";

****************************************************************;
Title2 "Load Information";
****************************************************************;
DATA Example_55;
  LENGTH Stratum 3 Nh UnitCost Revenue Emplyees Revnu_SD Emply_SD
         RCredit OffShore 8;
  LABEL Stratum  = "Stratum ID"
```

```
        Nh       = "Sampling Frame Counts per Stratum"
        UnitCost = "Unit-specific Data CollectionCost"
        Revenue  = "Pop. Mean Revenue (Millions)"
        Emplyees = "Pop. Mean Employees"
        Revnu_SD = "Pop. Standard Deviation Revenue (Millions)"
        Emply_SD = "Pop. Standard Deviation Employees"
        RCredit  = "Pop. Proportion Claimed Research Credits"
        OffShore = "Pop. Proportion Had Offshore Affiliates";
  INPUT Stratum Nh UnitCost Revenue Emplyees Revnu_SD Emply_SD
        RCredit OffShore;
  CARDS;
  1   6221  120   85  511  170.0  255.50  0.8  0.06
  2  11738   80   11   21    8.8    5.25  0.2  0.03
  3   4333   80   23   70   23.0   35.00  0.5  0.03
  4  22809   90   17   32   25.5   32.00  0.3  0.21
  5   5467  150  126  157  315.0  471.00  0.9  0.77
  ;
RUN;
                          *Standard deviations for proportions;
DATA Example_55;
  SET Example_55;
  ARRAY p_s  RCredit  OffShore;
  ARRAY sd_s RCrdt_SD OffSh_SD;

  DO OVER p_s;
    sd_s = SQRT(p_s * (1 - p_s) * Nh / (Nh - 1));
  END;
RUN;

PROC PRINT DATA=Example_55 UNIFORM NOOBS;  RUN;

****************************************************************;
Title2 "Sample Allocation - Initial Solution";
****************************************************************;
PROC OPTMODEL;

  *_____ LOAD PARAMETERS _____*;
                                        *Stratum frame counts;
    NUMBER Nh{1..5};
    READ DATA Example_55 INTO [_n_] Nh;
    PRINT Nh;
                                              *Per Unit Cost;
    NUMBER UnitCost{1..5};
    READ DATA Example_55 INTO [_n_] UnitCost;
    PRINT UnitCost;
                      *Population means & standard deviations;
    NUMBER Revenue{1..5},  Emplyees{1..5}, RCredit{1..5},
           OffShore{1..5}, Revnu_SD{1..5}, Emply_SD{1..5},
           RCrdt_SD{1..5}, OffSh_SD{1..5};
    READ DATA Example_55 INTO [_n_]
                         Revenue Emplyees RCredit OffShore
                         Revnu_SD Emply_SD RCrdt_SD OffSh_SD;
    PRINT Revenue Revnu_SD;
```

```
*_____ DECISION VARIABLES _____*;
          *Stratum sample sizes with initial value assignments;
   VAR NSamp{i in 1..5} init 100;
   PRINT NSamp;


*_____ CONSTRAINTS _____*;
                           *Stratum sizes >= 100, <= Frame Sizes;
   CON SampSize{i in 1..5}: 100 <= NSamp[i] <= Nh[i];


                                            *Survey Budget;
   CON Budget: (SUM{i in 1..5} UnitCost[i] * NSamp[i]) <= 300000;


                     *Relvariance for Mean Number of Employees;
   CON RelVar1:
       (SUM{i in 1..5} Nh[i]*(Nh[i]/NSamp[i]-1)*Emply_SD[i]^2)
       / ((SUM{i in 1..5} Nh[i] * Emplyees[i])^2)
       <= (0.05^2);
   PRINT ((SUM{i in 1..5} Nh[i]*(Nh[i]/NSamp[i]-1)*Emply_SD[i]^2)
        / ((SUM{i in 1..5} Nh[i] * Emplyees[i])^2));


           *Relvariance for Proportion of Claimed Research
               Credits;
   CON RelVar2:
       (SUM{i in 1..5} Nh[i]*(Nh[i]/NSamp[i]-1)*RCrdt_SD[i]^2)
       / ((SUM{i in 1..5} Nh[i] * RCredit[i])^2)
       <= (0.03^2);


        *Relvariance for Proportion Having Offshore Affiliates;
   CON RelVar3:
       (SUM{i in 1..5} Nh[i]*(Nh[i]/NSamp[i]-1)*OffSh_SD[i]^2)
       / ((SUM{i in 1..5} Nh[i] * OffShore[i])^2)
       <= (0.03^2);


*_____ OBJECTIVE FUNCTION _____*;

   MIN f = (SUM{i in 1..5} Nh[i] * (Nh[i]/NSamp[i] - 1) *
           Revnu_SD[i]^2) /
           ((SUM{i in 1..5} Nh[i] * Revenue[i])^2);


*_____ SOLUTION _____*;

   SOLVE;
     PRINT NSamp;
     PRINT (SUM{i in 1..5} NSamp[i]);
     PRINT (SQRT((SUM{i in 1..5} Nh[i] * (Nh[i]/NSamp[i] - 1) *
           Revnu_SD[i]^2) /
           ((SUM{i in 1..5} Nh[i] * Revenue[i])^2)));
     PRINT (SQRT((SUM{i in 1..5} Nh[i] * (Nh[i]/NSamp[i] - 1) *
           Emply_SD[i]^2) /
           ((SUM{i in 1..5} Nh[i] * Emplyees[i])^2)));
     PRINT (SQRT((SUM{i in 1..5} Nh[i] * (Nh[i]/NSamp[i] - 1) *
           RCrdt_SD[i]^2) /
           ((SUM{i in 1..5} Nh[i] * RCredit[i])^2)));
     PRINT (SQRT((SUM{i in 1..5} Nh[i] * (Nh[i]/NSamp[i] - 1) *
```

```
                    OffSh_SD[i]^2) /
                    ((SUM{i in 1..5} Nh[i] * OffShore[i])^2)));


    SOLVE with NLPC;
      PRINT NSamp;
      PRINT (SUM{i in 1..5} NSamp[i]);
      PRINT (SQRT((SUM{i in 1..5} Nh[i] * (Nh[i]/NSamp[i] - 1) *
             Revnu_SD[i]^2) /
             ((SUM{i in 1..5} Nh[i] * Revenue[i])^2)));
      PRINT (SQRT((SUM{i in 1..5} Nh[i] * (Nh[i]/NSamp[i] - 1) *
             Emply_SD[i]^2) /
             ((SUM{i in 1..5} Nh[i] * Emplyees[i])^2)));
      PRINT (SQRT((SUM{i in 1..5} Nh[i] * (Nh[i]/NSamp[i] - 1) *
             RCrdt_SD[i]^2) /
             ((SUM{i in 1..5} Nh[i] * RCredit[i])^2)));
      PRINT (SQRT((SUM{i in 1..5} Nh[i] * (Nh[i]/NSamp[i] - 1) *
             OffSh_SD[i]^2) /
             ((SUM{i in 1..5} Nh[i] * OffShore[i])^2)));

  *_____ OUTPUT SOLUTION _____*;

    NUMBER i;
    FILE "OptModel.strata.out";
      PUT @1 "Stratum"
          @10 "Resp_Alloc";
      DO i=1 TO 5;
        PUT @1 i
            @10 NSamp[i];
      END;
    CLOSEFILE "OptModel.strata.out";

    QUIT;
RUN;
```

■

5.5 R **alabama** Package

The R software has a number of different optimization routines. To date, most functions like `solve.QP`, `nlminb`, and `constrOptim` only allow constraints that are linear in the decision variables. The `alabama` package (Varadhan 2010) contains a modification of `constrOptim`, called `constrOptim.nl`, that will handle nonlinear constraints. It uses what is known as an augmented Lagrangian algorithm (Lange 2004; Madsen et al. 2004). This algorithm is different from the ones in Excel Solver and SAS. Code 5.3 shows R code that will repeat the optimization in Example 5.2.

The vector of decision variables, `nh`, the stratum population counts, `Nh`, the stratum unit costs, `ch`, the budget, and the stratum means of the four variables (revenues, employees, establishments claiming the research credit, and establishments with offshore affiliates) are assigned at the beginning of the program. As in the SAS `nlp` code, the stratum standard deviations are assigned for revenues and employees but computed for research credit and offshore affiliates. The functions, `relvar.rev`, `relvar.emp`, `relvar.rsch`, and `relvar.offsh`, compute the relvariances of estimated totals for each variable. Although each relvariance uses the same general formula, one of the restrictions of `constrOptim.nl` is that the objective function and functions that define nonlinear constraints can take only one parameter—`nh` in this case. Thus, separate functions were written for our example.

The function `constrOptim.nl` can take many input parameters, but only a few are needed for Example 5.2. The ones used here and their explanations from the help file are:

`par`	Vector of initial values of decision variables
`fn`	Objective function
`hin`	A vector function specifying inequality constraints such that `hin[j]>0` for all j
`heq`	A vector function specifying equality constraints such that `heq[j]=0` for all j
`control.outer`	
`eps`	Tolerance for convergence of outer iterations of the barrier and/or augmented Lagrangian algorithm
`mu0`	Parameter for barrier penalty
`method`	Algorithm in `optim()` to be used; default is `"BFGS"` variable metric method

Code 5.3: R `constrOptim.nl` code for the optimization problem in Example 5.2

```
#****************************************************************
# FILE:    constrOptim.example.R
# PURPOSE: Use constrOptim.nl to solve allocation business
#          establishment allocation problem
# DATE:    9/14/09
# AUTHOR:  R. Valliant
#****************************************************************

require(alabama)
require(numDeriv)  # alabama requires "numDeriv" package

    # Decision vars
nh <- vector("numeric", length = 5)

    # Stratum pop sizes
Nh <- c(6221, 11738, 4333, 22809, 5467)
```

```
    # Stratum costs
ch <- c(120, 80, 80, 90, 150)
    # Stratum means and SDs
    # Revenues
mh.rev <- c(85, 11, 23, 17, 126)
Sh.rev <- c(170.0, 8.8, 23.0, 25.5, 315.0)
    # Employees
mh.emp <- c(511, 21, 70, 32, 157)
Sh.emp <- c(255.50, 5.25, 35.00, 32.00, 471.00)
    # Proportion of estabs claiming research credit

ph.rsch <- c(0.8, 0.2, 0.5, 0.3, 0.9)
    # Proportion of estabs with offshore affiliates
ph.offsh <- c(0.06, 0.03, 0.03, 0.21, 0.77)
budget = 300000
n.min <- 100
    # Relvar function used in objective
relvar.rev <- function(nh){
    rv <- sum(Nh * (Nh/nh - 1)*Sh.rev^2)
    tot <- sum(Nh * mh.rev)
    rv/tot^2
}

  # Relvar functions used in nonlinear constraints
  # The nonlin constraints can take only 1 argument: in this case
  #    the vector of decision vars
relvar.emp <- function(nh){
    rv <- sum(Nh * (Nh/nh - 1)*Sh.emp^2)
    tot <- sum(Nh * mh.emp)
    rv/tot^2
}

relvar.rsch <- function(nh){
    rv <- sum( Nh * (Nh/nh - 1)*ph.rsch*(1-ph.rsch)*Nh/(Nh-1) )
    tot <- sum(Nh * ph.rsch)
    rv/tot^2
}
relvar.offsh <- function(nh){
    rv <- sum( Nh * (Nh/nh - 1)*ph.offsh*(1-ph.offsh)*Nh/(Nh-1) )
    tot <- sum(Nh * ph.offsh)
    rv/tot^2
}
constraints <- function(nh){
    h <- rep(NA, 13)
        # stratum sample sizes <= stratum pop sizes
    h[1:length(nh)] <- (Nh + 0.01) - nh
        # stratum sample sizes >= a minimum
    h[(length(nh)+1) : (2*length(nh)) ] <- (nh + 0.01) - n.min
    h[2*length(nh) + 1] <- 0.05^2 - relvar.emp(nh)
    h[2*length(nh) + 2] <- 0.03^2 - relvar.rsch(nh)
    h[2*length(nh) + 3] <- 0.03^2 - relvar.offsh(nh)
    h
}
heq <- function(nh){
```

```
    heq <- 1 - sum(nh*ch/budget)
    heq
}
ans <- constrOptim.nl(       # parameter and objective function
        par = rep(1100,5),   # using par = rep(100,5) gives error:
                             # "initial value violates inequality
                             #      constraints"

        fn = relvar.rev,
                     # parameter bounds
        hin = constraints,
        heq = heq,
        control.outer = list(eps = 1.e-10,
            mu0   = 1e-05,
            NMinit = TRUE,
            method = "BFGS"
            )
    )
ans
```

In this example, we wrote a function called `constraints` that returns a vector of length 13 containing the values of the inequality constraints. Since the inequality constraints must have the form `hin[j]>0`, the restrictions that each stratum sample size be less than the population size and greater than or equal to 100 and were written as

```
h[1:length(nh)]  <- (Nh + 0.01) - nh
h[(length(nh)+1) : (2*length(nh)) ] <- (nh + 0.01) - n.min
```

By adding 0.01 to N_h and n_h, we set up constraints where the inequality is strictly greater than 0 rather than greater than or equal to 0. The equality constraint `heq` sets the budget equal to \$300,000. A serious limitation of `constrOptim.nl` is that the initial value of `par` must be a feasible solution, i.e., one that does not violate any of the inequality constraints. If the value of `par` used to call the function is not feasible, the function will generate an error and terminate; it has no features for automatically correcting initial values that violate any of the inequality constraints. Some experimenting may be necessary to arrive at a trial allocation that is feasible. None of the previously discussed optimization software options had this requirement for the starting value of `nh`, which makes them simpler to use.

The function `constrOptim.nl` is also sensitive to the relative sizes of the values in the equality and inequality constraints. The relvariances in the example are small numbers, e.g., 0.03^2, while the budget of \$300,000 is large. If the equality constraint is set directly to be `sum(nh*ch)-budget`, the algorithm pays more attention to meeting the budget constraint than to minimizing the objective function, which is the relvariance of total revenue. By defining the equality constraint as `1-sum(nh*ch/budget)`, we had a quantity that was 0 when the budget was fully expended and whose range was in relative deviations from the budget and not in dollars. This scaling of the `heq` constraint helps achieve a smaller value of the objective function.

Results can be dumped to the screen or assigned to an object as in Code 5.3. The solution for the stratum sample sizes in this example is in `ans$par`; the value of the objective function is in `ans$value`. The output for the example above is

```
$par
[1]  429.7308  233.4132  113.5080 1534.6032   550.4323
$value
[1] 0.002260288
```

This solution is not quite as good as the one obtained earlier, although the difference is small. The objective value of 0.002260288 is about 4.3 % higher than the 0.00216695 obtained with Solver and `proc nlp`.

5.6 Accounting for Problem Variations

We conclude this chapter with a note on accounting. In the preceding sections, we stressed that multicriteria optimization in general is an iterative process. For example, constraints are set and then relaxed (or tightened) based on the initial allocation solution. Trying a series of options for costs and precision of estimates is an especially useful way to explore a problem. This is also often a good way of illustrating trade-offs to clients. We recommend that researchers establish and maintain an accounting system to document:

- Initial values set for the optimization problem
- Optimization results such as attained constraints and decision-variable values
- Reasons for changing the optimization components
- New values set for the optimization problem

Having a well-documented system will minimize the likelihood of repeating optimization problems implemented previously but discarded and will facilitate writing sampling documentation for the study at hand.

Exercises

5.1. A researcher would like to survey the mathematics teachers in the elementary and secondary schools in Montgomery, Howard, and Prince George's counties in the state of Maryland. The goals of the survey are to estimate the proportion of teachers who use computers in instruction and, among the teachers who do use computers, what proportion teach the use of spreadsheets. The estimates are desired for (i) each county separately, (ii) for domains defined by elementary and secondary combined across the three counties, and (iii) for elementary and secondary domains within each county. The researcher would also like to be able to recognize differences at the county level that are greater than 10 % points. The budget for the data collection part of the survey is \$100,000 and it is anticipated that surveying each teacher will cost about \$150.

How would you formulate the sample allocation problem as an optimization problem? List the population parameters that you would need in order to do the optimization problem. What would you do about parameter values if no previous, similar survey had been done?

5.2. Using the data in Example 5.2 calculate (a) the proportional allocation, (b) the Neyman allocation for estimating total revenue, and (c) the cost-constrained allocation for revenue, assuming a budget of \$300,000. Note that the proportional and Neyman allocations do not have a constraint on revenues; each should be found for the total sample size of $n = 2,848$ as in Example 5.2. For each of these allocations compute the CVs for estimated total revenue, total employees, total number of establishments claiming the research credit, and total number of establishments having offshore affiliates. Do allocations (a), (b), and (c) respect the constraints used in Example 5.2.

5.3. Resolve Example 5.2 with the following constraints:

   (i) Budget on variable costs = \$300,000.
   (ii) $CV \leq 0.05$ on estimated total number of employees.
   (iii) At least 100 establishments are sampled in each sector.
   (iv) The number sampled in each stratum is less than the population count, $n_h \leq N_h$.
   (v) $CV \leq 0.03$ on estimated total number of establishments claiming the research tax credit.
   (vi) $CV \leq 0.05$ on estimated total number of establishments with offshore affiliates.

In other words, change the constraint on the offshore affiliate CV to 0.05 and recalculate the allocation. Comment on the differences in the resulting allocation compared to that in Example 5.2.

5.4. Resolve Example 5.2 with the same CV constraints as in Exercise 5.3 (0.05 on employees, 0.03 on total establishments claiming the research credit,

0.05 on total establishments with offshore affiliates), but revise the objective to be minimizing the total cost. Retain the constraints that the sample in each stratum must be less than the population count and that at least 100 units be sampled in each stratum.

Discuss why there are differences in the solutions found in Exercises 5.3 and 5.4.

5.5. Determine the allocation to strata in Example 5.2 based on the following set-up. Minimize $\Phi = 0.75 \times relvar\left(\hat{T}_{rev}\right) + 0.25 \times relvar\left(\hat{T}_{emp}\right)$, where $\hat{T}_{rev}$ is the estimated total of revenues and $\hat{T}_{emp}$ is the estimated total of employees. The constraints in the problem are:

- Sample at least 200 establishments in each stratum.
- The number sampled from a stratum should be less than 20 % of the stratum population.
- The CVs on the estimated total numbers of establishments claiming the research credit and having offshore affiliates should be at most 0.02.
- The budget is $600,000.

Chapter 6
Outcome Rates and Effect on Sample Size

Outcome rates, such as the percent of sample units refusing to participate in a survey, generally have three uses. The first is to measure study performance and outcome rates are often also referred to as performance rates or process indicators. For example, a client might wish to know what proportion of the sample resulted in a completed interview. The second use is to inflate a calculated sample size for loss of sample units. For example, a survey statistician determines the number of sample units needed to detect a three percentage point difference in the estimates for males and females for a specified size and power of the test, as discussed in Chap. 4. Finally, study rates can also be incorporated into the design weights as adjustment factors to create final analysis weights.

There is much debate over study outcome rates, and it is important to note that those rates should not be seen as measures for data quality (Groves 2006). However, outcome rates guide field decisions, and the logic behind them helps in the planning stages of a survey. Thus, we will spend time in this chapter to explicate these first two uses. The incorporation of outcome rates into design weights will be discussed in Part III.

We begin our discussion by focusing on a common set of disposition codes that are needed to define the outcome rates. Much of the material summarized in this document follows the standard definitions given by the American Association for Public Opinion Research (AAPOR) in their document entitled "Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys" (AAPOR 2011). In some surveys, there may be differences in opinion across members of the project team about how rates should be computed. Using the AAPOR standards is a good way to avoid time-consuming debate about what should be done.

6.1 Disposition Codes

Numeric codes that describe the current or final data collection status of each sample unit are known as disposition codes. AAPOR provides a list of recommended disposition codes; however, these disposition codes are generally specific to each data collection agency. Thus, sometimes it might be necessary to negotiate with the data collection agency to expand their set of codes. For example, the sample disposition codes recorded for the May 2004 Status of Forces Survey of Reserve Component Members (SOFReserves), a survey conducted by Defense Manpower Data Center (2004) of Military Reservists, are provided in Table 6.1. If these disposition codes are also used to tailor fieldwork recruitment during the data collection, it would be advisable to differentiate between refusals and deployed personnel. Both of these codes are currently summed into category 8.

Table 6.1: Terminology: sample dispositions for the may 2004 SOF Reserves study.

Disposition code	Description
1	Ineligible—based on check of updated personnel records
2	Ineligible—Self-/Proxy report, deceased, ill, incarcerated, separated
3	Ineligible–Survey self-report
4	Complete eligible response
5	Incomplete eligible response
8	Refused—refusal, deployed, other refusal
9	Blank (returned questionnaire)
10	Postal nondelivery (PND)
11	Other nonrespondent

Depending on the survey and the mode of data collection, the number of disposition codes can be rather large. It is useful to specify ahead of time how disposition codes can be grouped to later compute study performance rates. AAPOR provides a map of the disposition codes to seven mutually exclusive categories used in the outcome rate calculations discussed in the next section. The general categories are described below (Table 6.2) along with the notation relevant to the rate calculations. We borrow some notation provided in the AAPOR document for consistency.

Sample units are assigned to category I (complete interview) if they provide responses to all appropriate questions in the questionnaire. If participants complete only a portion of the interview but the data are sufficient to address the analysis objectives, then the records are classified in the P group

(partially complete interview). Records with insufficient data (break-offs) or associated with those who refuse to participate (refusals) are collectively called category R if the participants are verified to be eligible for the study.

Table 6.2: Disposition code categories used in performance rate calculation.

Symbol in study rates category		Study eligibility
I	Complete interview	*Eligible*
P	Partially complete interview	*Eligible*
R	Refusal/break-off	*Eligible*
NC	Noncontact	*Eligible*
U	Unknown study eligibility	*Unknown*
NE	Not eligible	*Ineligible*
O	Other	*Eligible*

Category NC (noncontact) contains those sample members who were never contacted for the interview but were known to be eligible for the study, e.g., a "ring/no answer" in a telephone survey after an interview appointment was scheduled with another household member. Study participants classified as ineligible (category NE) are usually listed separately from those for whom the eligibility was never established (category U). In a telephone survey, for example, there may be many numbers that are classified as a ring/no answer whose study eligibility status is unknown. They might be residences or unassigned telephone numbers. How the unknowns (U) are handled can make a noticeable difference in the response rates, as discussed subsequently. All eligible cases that are not assigned to any of the previously mentioned categories are assigned to a "catch-all" other (O) category.

When used in a formula, the symbols in Table 6.2 represent the number of sample units that fall within each category; the sum of the categories ($I + P + R + NC + U + NE + O$) equals the overall sample size (n).

The primary task in calculating outcome rates is to map the disposition codes adopted for a particular survey into the AAPOR categories. One example of mapping is demonstrated in a paper by Abraham et al. (2006). An excerpt is provided below (Table 6.3) from their Table A-1 of American Time Use Survey disposition codes and corresponding AAPOR category designation.

A couple of issues are important to note regarding the mapping task. First, assignments may differ as a function of target populations. That is, for some studies, the finalized AAPOR category may differ from the ATUS assignment shown, because the target population excludes institutionalized persons. This would result in a change for code 19 (other: designated person institutionalized) from "other non-interview" to "not eligible." Second, researchers may express different preferences on how assignments should be executed.

Table 6.3: Concordance between AAPOR and internal disposition codes for the 2004 American Time Use survey.

Disposition code	Description	AAPOR category
1	Completed interview	*I*
2	Sufficient partial	*I*
14	Not eligible: designated person underage	*NE*
15	Not eligible: designated person not household member	*NE*
18	Other: designated person absent, ill, or hospitalized	*O*
19	Other: designated person institutionalized	*O*
21	Other: language barrier	*O*
23	Unknown eligibility: incorrect phone number	*U*
24	Not eligible: designated person in Armed Forces	*NE*
27	Unknown eligibility: privacy detector	*U*
29	Other: non-interview	*O*
100	Not eligible: miscellaneous	*NE*
104	Other: invalid input	*O*
108	Not eligible: case deleted as sample reduction	*NE*
109	Refusal: hostile break-off, interview progress achieved	*R*
112	Refusal: by parent/guardian/gatekeeper	*R*
113	Unknown eligibility: unproductive call counter	*U*
118	Noncontact: incomplete callbacks	*NC*
119	Noncontact: temporarily unavailable	*NC*
121	Other: unresolved language barrier	*O*
124	Noncontact: never contacted, confirmed number	*NC*
125	Unknown eligibility: never contacted, unconfirmed number	*U*
126	Other: instrument error	*O*
127	Unknown eligibility: never tried, no telephone number	*U*
130	Refusal: diary contains less than 5 activities	*R*
133	Refusal: other data quality issues	*R*

Note: Abbreviated Table A-1 from Abraham et al. (2006)

Abraham et al. (2006), for example, chose to not use partial interviews as a category in their assignments. For a survey statistician it is important to address these issues ahead of time.

6.2 Definitions of Outcome Rates

This section describes five general study rates which apply to most surveys—location, contact, eligibility, cooperation, and response. Variations of these rates that are specific to the mode of data collection are provided as examples.

All five rates we discuss here have upper and lower bounds depending on the treatment of ineligible cases and cases with unknown eligibility. Often an estimate of the proportion of eligibles among the unknowns is used to create more reasonable rates than the extreme bounds.

Location Rate

The *location rate* specifies the proportion of units for which contact information was obtained. The formula is expressed in words as follows along with a formulation of the rate:

$$LOC = \frac{\text{number of sample units located}}{\text{total number of sample units}}$$

$$= \frac{n - U}{n} = \frac{I + P + R + NC + NE + O}{I + P + R + NC + U + NE + O}$$

For example, a sample of female respondents to the 1993 National Health Interview Survey (NHIS) was selected for the National Survey of Family Growth (NSFG), Cycle 5 (Potter et al. 1998). Contact information collected during the NHIS was no longer valid for some sample members when the NSFG was fielded. Tracing procedures were used to locate many; however, the location rate was less than 100 %. Non-contacts (NC) may sound as if these are cases that were not located. But, in a household survey, a non-contact could be a case where a gatekeeper prevents an interviewer from entering a secure building. In that instance, the case has been located, but a direct contact was impossible.

An *address match rate* is an example of a location rate specific to telephone surveys. Typically, random telephone numbers selected under a detailed sampling plan are sent to a vendor for processing known as reverse matching to address lists. Some of the US companies that sell this service are Telematch[1], the GENESYS system within Marketing Systems Group (MSG)[2], and Survey Sampling International[3]. Sending an advance (or lead) letter through the mail to those with an address has been shown to improve telephone response rates (e.g., Traugott and Goldstein 1993). This is in comparison to those units without address information who are contacted without prior notification ("cold calls").

---

[1] http://www.relevategroup.com/.

[2] http://www.m-s-g.com/.

[3] http://www.surveysampling.com/.

Contact Rate

The *contact rate* is slightly different from the location rate and is the proportion of units with a successful initial contact. AAPOR provides three formulas for contact rates depending on the method used to deal with units without a known eligibility status. The first formula, *CON1*, is calculated after eliminating the known ineligibles and is considered a minimum value among the three contact rate formulas:

$$CON1 = \left.\frac{\text{number of sample units contacted}}{\text{number of sample units}}\right|_{\substack{excluding \\ inelgibles}}$$

$$= \frac{n - (NC + U + NE)}{n - NE} = \frac{I + P + R + O}{I + P + R + NC + U + O}$$

Note that the AAPOR rate *CON3*, which excludes those with unknown eligibility (U), is the maximum value because the denominator is smaller than in the formula for CON1:

$$CON3 = \frac{n - (NC + U + NE)}{n - (U + NE)} = \frac{I + P + R + O}{I + P + R + NC + O}$$

A successful contact is defined here to be a contact in which the location information was verified to be correct. In a telephone survey, sample units where no person in the household was ever reached may be considered as an "unsuccessful" contact as long as there is evidence that the telephone number actually belongs to a household. For example, a residential voicemail message is typically counted as sufficient evidence that a household has been reached. A third suggested contact rate (*CON2*) is not shown here but can be found in the AAPOR documentation (AAPOR (2011)). CON2 is similar to CON1 but includes in the base only the estimated eligible cases among the cases with unknown eligibility.

The definition of contact varies across survey research. For example, contact in a random digit dialing (RDD) survey may mean that the interviewer verified that the selected telephone number was linked with a residence regardless of the eligibility of the persons within the household. As with other project documents, project-specific definitions need to be detailed during the planning stages.

Eligibility Rate

The criteria to classify a sample unit as either eligible or ineligible (category NE) for the study are defined early in the planning process. The set of all eligibles defines the target population for which estimates will be produced.

The *eligibility rate*, referred to in the AAPOR document as "e", is calculated as

$$e = \frac{\text{number of study} - \text{eligible sample units}}{\text{number of sample units with an eligibility status}}$$

$$= \frac{n - (U + NE)}{n - U} = \frac{I + P + R + NC + O}{I + P + R + NC + NE + O}$$

Eligibility is determined through a set of preliminary questions—sometimes called a screener (questionnaire) for in-person and telephone surveys. Because the screener questions are themselves subject to nonresponse, the rate is calculated among those whose eligibility status is determined (i.e., eligible or ineligible for the study). The proportion of sample units who complete the screener is known as the *screening rate*.

Eligibility can be defined at more than one stage of sampling. For example, in a list-assisted RDD survey, the unit sampled from the list frame is a telephone number. However, not all telephone numbers are assigned to a household; some are unassigned and others are allocated to businesses or for business use. Therefore, the sampling frame contains two types of ineligibles for an RDD survey—nonworking numbers and nonresidential numbers. The *working-number rate* and the *residential (eligibility) rate* are defined as follows:

$$\text{working} - \text{number rate} = \frac{\text{number of working telephone numbers}}{\text{total number of telephone numbers}}$$

$$\text{residential rate} = \frac{\text{number of eligible telephone numbers}}{\text{total number of sample numbers}}$$

In addition to reverse matching, vendors such as MSG provide a service to prescreen telephone numbers to eliminate (i) all nonworking numbers identified by a computer through an electronic tritone and (ii) for residential surveys, all nonresidential numbers (e.g., businesses).

Cooperation Rate

The proportion of study-eligible sample units providing answers to a sufficient portion of the interview is called a *cooperation rate*. This rate has also been labeled as a *response rate among eligibles* prior to the standardized definitions. Four cooperation rates are provided in the AAPOR document depending on methods to deal with partially completed interviews and sample units with unresolved status codes (see Sect. 6.1 for further discussion). A general formula is expressed as

$$COOP2 = \frac{\text{number of completed/partial interviews}}{\text{number of contacted, eligible units}}$$

$$= \frac{n - (R + NC + U + NE + O)}{n - (NC + U + NE)} = \frac{I + P}{I + P + R + O}$$

Examining the formula in the AAPOR Standard Definitions documents shows that the upper and lower bound of the estimated cooperation rate could be derived by using *COOP1* and *COOP3*, respectively, based on how the cases in the "other" category (O) are classified. The lower and upper bounds on the cooperation are calculated as follows:

$$COOP1 = \frac{n - (P + R + NC + U + NE + O)}{n - (NC + U + NE)} = \frac{I}{I + P + R + O}$$

$$COOP3 = \frac{n - (P + R + NC + U + NE + O)}{n - (NC + U + NE + O)} = \frac{I}{I + P + R}$$

Sample members with a partially completed interview typically are classified as respondents if key information has been collected to address the primary analytic objectives for the study. Exactly which items in a survey instrument are considered key must be decided by the project staff and the sponsor of the study. In some cases, a few questions may be considered key; in others a long series may have to be answered before the case is considered a partial complete.

Response Rate

The *response rate* is likely the most familiar rate of those discussed. It is an extension of the cooperation rate to all sample cases that are (potentially) eligible. AAPOR standard definitions include six variants of the response formula. One that is often reported is

$$RR2 = \frac{\text{number of complete/partial interviews}}{\text{all sample members } - \text{ known ineligibles}}$$

$$= \frac{n - (R + NC + U + NE + O)}{n - NE} = \frac{I + P}{I + P + R + NC + U + O}$$

Three other formulas warrant special attention in our discussion—*RR1*, *RR6*, and *RR4*. The rates *RR1* and *RR6* bound the response rate below and above, respectively, due to the way in which the partial completes (P) are treated. This is seen by comparing the two formulas below:

$$RR1 = \frac{n - (P + R + NC + U + NE + O)}{n - (U + NE)} = \frac{I}{I + P + R + NC + O}$$

$$RR6 = \frac{n - (R + NC + U + NE + O)}{n - (U + NE)} = \frac{I + P}{I + P + R + NC + O}$$

One primary directive given to interviewers is to collect information in order to establish study eligibility. However, for many studies, eligibility is never verified for a proportion of the sample units. In 1982, the Council of American Survey Research Organizations (CASRO) recommended that an (estimated) eligibility rate (e) be applied to the number of sample units with unknown eligibility (U). An estimate of the number of ineligibles among the unknowns, $(1-e)U$, plus the number of known ineligibles (NE) is subtracted from the total number of sample cases leaving only the (estimated) total number of eligibles in the denominator. Hence, the CASRO response rate formula, such as $RR4$, was born:

$$RR4 = \frac{I + P}{I + P + R + NC + O + (e \times U)}$$

Note that the response rate as well as the other rates can be calculated for domains. For example, some clients are interested in the proportion of contacted, eligible participants who successfully complete the study interview. Some researchers call this a *completion rate*; it is calculated in the same manner as either $RR5$ in AAPOR (2011) or $RR6$ above.

6.3 Sample Units with Unknown AAPOR Classification

As shown in the definitions of the outcome rates in the previous section, the unknown status cases (U) can be handled in different ways. This section gives some examples that illustrate how the calculated rates can be affected by how the U's are treated. Because the decisions directly affect the numeric value of the study rate, they should be justified in the project documentation. Below, we provide two examples to illustrate this point.

Example 6.1. Table 6.4 contains the count of sample units by disposition code for a fictitious mail survey. How should the U sample units be treated, and how does this decision affect the location rate?

- *Scenario 1—All "No Response" Units Classified as Located*

Say, for example, that your client declares that the address list is updated on a regular basis so that the "no response" cases are actually refusals. The resulting location rate is calculated as

$$\frac{(3,000 - (120 + 75))}{3,000} = 93.5\,\%$$

- *Scenario 2—No "No Response" Units Classified as Located*

Table 6.4: Summary of sample-unit counts from a mail survey by disposition code.

Disposition	AAPOR category	Count
Completes	I	1,807
Refusals	R	642
Ineligible	NE	51
Unlocatable	NC	120
Postal nondelivery	NC	75
No response	U	305
Total		3,000

With this mail survey, you may suspect that the questionnaires were delivered to the wrong address for the "no response" units and that the household resident simply threw away the materials. The scenario-two location rate is much lower than the scenario-one rate:

$$\frac{(3,000 - (120 + 75 + 305))}{3,000} = 83.3\,\%$$

- *Scenario 3—A Portion of the "No Response" Units Classified as Located*

Similar to the eligibility rate "e" adjustment, you may wish to estimate the number of "no response" units that were located by using only those cases with a known location status (i.e., conditional on being known). The scenario-three location rate is closer in value to the scenario-one rate because of the high-conditional location rate:

$$\frac{3,000 - (120 + 75 + \{(1 - \ell) \times 305\})}{3,000} = 92.8\,\%,$$

where $\ell = (1,807 + 642 + 51) / (3,000 - 305) = 92.8\,\%$ is the location rate.

Example 6.2. Table 6.5 contains the count of sample units by disposition code for a fictitious RDD survey. The rates for RR2 and RR6 differ by less than $4\,\%$ points and are calculated as follows:

$$RR2 = \frac{I + P}{I + P + R + NC + O + U} = 37.3\,\%$$

$$RR6 = \frac{I + P}{I + P + R + NC + O} = 41.0\%$$

Note that there are no cases coded as noncontact or other; thus $NC = O = 0$ in *RR2* and *RR6*. Cases where the telephone was always busy are coded as *NE*. This is a matter of judgment and would not necessarily be done the same way in every survey.

Table 6.5: Summary of sample-unit counts from an RDD survey by disposition code.

Disposition	AAPOR category	Count
Complete interview	I	3,264
Partial complete interview	P	550
Voicemail	R	350
Language barrier	NE	75
Refusal	R	5,134
Ring—no answer	U	914
Always busy	NE	10
Not eligible	NE	3,181
Fax machine	NE	22
Total		13,500

The final response rate calculation is the CASRO response rate (AAPOR RR4) which includes the e adjustment factor discussed in the previous section. As with the above calculations, we assume that the nonrespondents (R cases) have been verified to be eligible. The e factor and the corresponding RR4 rate are calculated as follows:

$$e = \frac{I + P + R}{n - U} = \frac{13,500 - (75 + 914 + 10 + 3,181 + 22)}{13,500 - 914} = 0.7388$$

$$RR4 = \frac{I + P}{I + P + R + NC + (e \times U)} = 38.2\%.$$

Note that the $RR4$ value is located between the $RR2$ and $RR6$ values because (a) $RR2$ counts all of the $U{=}914$ unknowns as eligible while (b) $RR6$ counts none of them as eligible. Note that there are no cases coded as other; thus $O = 0$ in $RR4$.

6.4 Weighted Versus Unweighted Rates

One major question that arises is: should I calculate weighted or unweighted performance rates? The typical though potentially aggravating answer is: it depends. Note that weighted and unweighted rates are equivalent if the design contains equal weights—an equal probability sampling and estimation (EPSEM) design using the well-known acronym from Kish (1965). If members of the project team (including the client) wish to evaluate the particular sample in the current study, then you should calculate unweighted rates. For example, in developing the sample design, you estimate that 89 % of your sample units will be successfully located. An unweighted location rate either

well above or below the estimated rate might suggest the additional number of replicates to be released for data collection, as described in Sect. 6.5.

Conversely, a rate can also be viewed as an estimate of a population parameter. In this case, design weights (inverse inclusion probabilities) can be used to calculate the weighted rate. A weighted rate is viewed as an estimate of the rate that would be obtained if the entire target population was included in the study (i.e., a census). Another way to think about weighted rates is as follows. The unweighted rate is a function of the particular sample design which might include over- or undersampling of certain domains. The weighted rate is effectively adjusted back to the underlying distribution of the target population. One additional thing to note is that a confidence interval around the weighted rates can facilitate an analysis of how sensitive sample size calculations are to different assumed rates.

Given the contrast between the weighted and unweighted rates, however, it is our experience that the two values are typically close. Widely varying sampling and performance rates among subgroups of units can exacerbate the difference. This suggests that both the (design-)weighted and unweighted rates be calculated as a check on the weights (see Chap. 18 for a detailed discussion of weight checks).

6.5 Accounting for Sample Losses in Determining Initial Sample Size

If sample cases will be lost because they cannot be contacted, will not respond, or are lost for some other reason, a larger initial sample can be selected. This is especially important if the survey has a target number of responders. The adjustment to the initial sample size can use some of the outcome rates covered in Sect. 6.3, but those are generally more elaborate than are necessary (or useful). An example is given in Sect. 6.5.1. Another option, covered in Sect. 6.5.2, is to select subsamples (or replicates) that can be released for data collection one at a time until the target number of responders is reached.

6.5.1 Sample Size Inflation Rates at Work

Surveys often start with insufficient knowledge about the number of sample units that are eligible for the study, the number of units that can be contacted during the study period, or those that are willing to answer, to give just a few examples. Estimates of outcome rates (often based on other similar studies) can help in deciding how many cases should be sampled to achieve sufficient records to meet the targeted number of interviews. Exceeding the targeted number of interviews may unnecessarily reduce the study budget remaining

for, say, analysis and report writing. An insufficient number of interviews decreases the power of statistical tests or prevents certain analytic questions from being answered. Therefore, it is important to use rates that are as accurate as possible to inflate the number of target interviews. A good rule of thumb is to conduct a sensitivity analysis of the rates by using some upper and lower values (bounds) regardless of the source of your inflation factors.

Consider an in-person survey of households selected from a geographical area under a two-stage design. Residents of the sample household will be screened to determine if at least one of them is a study-eligible adult. One eligible adult is then selected from among the list of eligibles identified within the sample household. Your power calculations have determined that 200 completed (target) interviews will meet your analytic objectives. The following information on estimated study performance rates was gathered during the first week of the project. Note that project team communications have been added in italics:

- Approximately 3 % of the housing units (HUs) are vacant due to new construction and part-year residency. *As a conservative measure, the team decides to set a lower and an upper bound of 95 % and 97 % for the rate of eligible HUs.*
- Among those occupied HUs, between 92 and 95 % are expected to answer the door when the interviewers arrive.
- However, team members are unsure of the percent that will complete the 5 m screener. One project team member speculates that this rate could be between 70 and 87 %. *The team collectively decides to set the screening rate in the range of 70 to 82 %.*
- Census projections estimate that approximately 85 % of the households will contain at least one eligible person. *To ensure a sufficient number of eligible cases, you decide to compare eligibility rates in the range of 80 to 85 %.*
- Finally, the client stresses that <u>all</u> will want to participate in this survey and proposes a cooperation rate as high as 98 %. *Based on prior experience, the optimistic cooperation rate was lowered to the range of 70 to 75 % . You communicate to the client that you will release an initial random subsample from the full sample based on the 98 % assumption with the remaining cases released in replicates as needed. This procedure will ensure that the analytic objectives are met under the current project budget.*

Question. How many household addresses (also referred to as sample lines) should be selected to obtain in expectation the required number of in-person interviews?

Answer. Between 415 and 584 sample lines should be selected for the study. Table 6.6 inflates the 200 target interviews for the rates discussed above. The easiest way to do this is to work from the "bottom" up as shown in the table. Begin with the target number of interviews (i.e., the minimum number

required for the analytic objectives or the desired analytic sample size) and successively apply the inflation factors in the reverse of the temporal order in which they occur in the survey. The "*Occupied HU Rate*" has a different name, but it can be classified as an *eligibility rate* among housing units. Note that although no individual rate seems excessively low, in combination, they more than double the 200 target interviews. Only 318 to 417 sample lines are needed if the cooperation rate meets the high expectations of the client (i.e., 98 % is used in Table 6.6 for the cooperation rate). Therefore, the project team may consider randomly creating a replicate of 300 sample addresses for initial release followed by additional replicates of approximate size 100. We describe this technique more generally in the next section.

Table 6.6: Example of inflating target interviews for sample loss.

Target interviews	**200**
Cooperation rate	*0.7–0.75*
Number of eligibles	**267–286**
Eligibility rate	*0.8–0.85*
Number screened	**314–357**
Screening rate	*0.7–0.82*
Number contacted	**383–510**
Contact rate	*0.92–0.95*
Occupied HUs	**403–555**
Occupied HU rate	*0.95–0.97*
Sample addresses	**415–584**

6.5.2 Replicates

One technique that is also used in practice is to randomly select a large number of sample cases under a "worst-case scenario," randomly subdivide the full sample into data collection subsamples (sometimes called replicates), and release only the number of replicates necessary to meet the analytic objectives. As an uncomplicated example, suppose that a simple random sample of 500 members of a professional association is selected with the goal being to obtain 100 completed questionnaires (completes). The 500 might be divided randomly into 10 replicates of size 50. Initially, the first three replicates might be released. If necessary, additional replicates are released to obtain the desired 100 completes. Note that the replicates do not need to contain the same number of sample cases. This is done for convenience and

ease of accounting (i.e., to eliminate the need to keep track of differential sizes when deciding to release more sample).

Replicates are usually formed in a different way in a multistage sample. In an area sample, as described in Chap. 10, the replicates may be composed of geographic areas. The standard procedure in an area sample is to select primary sampling units (PSUs) which are counties or groups of counties, at the first stage, and smaller geographic areas as a second stage. The second stage units may be groups of city blocks and are often called segments. A large sample of segments can be selected initially and divided into replicates.

Because the replicates are randomly constructed, deciding to withhold any replicate does not negate the randomness of the sample; weights for the fielded cases are adjusted appropriately to reflect only those replicates that were released. The replicates identified for release are considered to be a simple random sample from the original sample for purposes of calculating the subsampling adjustment. However, once a replicate has been released for data collection, all cases in that replicate must be worked and be given a disposition code.[4] Otherwise, the full collection of released cases will not be a probability sample. The ultimate goal again is to ensure that you have a sufficient number of cases to meet the analytic objectives, keeping in mind any ramifications on the budget, time, and, if appropriate, the effects of unequal weighting (see Chap. 14 for further discussion of unequal weighting effects).

Creating the replicates in advance by subsetting a large sample is typically much easier than selecting an initial sample and then attempting to add to it later, depending on the sample design. Adding to a simple random sample by selecting another *srs* from the initial nonsample units is legitimate. But, if the initial sample is selected with probabilities proportional to size (*pps*), as might be the case in a school sample, selecting a supplemental sample in such a way that the overall sample is *pps* is not straightforward. (See the exercises.)

---

[4] Alternatively, the units could be worked in a random order, in which case, data collection could be stopped partway through a replicate. Working cases in a random order is typically impractical, however.

Exercises

6.1. Calculate the following study performance rates, unweighted and weighted, using data provided in the table below: location, contact (CON1 and CON3), eligibility, cooperation (COOP1 and COOP2), and response (RR2 and RR4).

Eligible	Complete?	Disp. code	Disp. description	Sample size	Sum of weights
Yes	Yes	1	Returned survey—complete	18,658	432,359
Yes	No	1	Returned survey—incomplete	754	18,046
No	No	2	Returned survey—deceased	52	1,281
No	No	3	Returned survey—incarcerated	18	300
Yes	Yes	8	Returned survey—complete	1,302	27,683
Yes	No	8	Returned survey—partial complete	102	2,507
Yes	No	14	Survey returned blank—active refusal	73	2,300
Yes	No	17	Survey returned blank—no reason	42	1,251
Yes	No	26	No return—no reason	2,500	25,000
Unknown	No	26	No return—no reason	143	3,072
Unknown	No	27	Postal nondelivery	1,313	35,576
Unknown	No	29	Original non-locatable	23	359
No	No	30	Inelig prior to contact—deceased	18	116
No	No	31	Inelig prior to contact—incarcerated	2	150
			Total	25,000	550,000

6.2. The following table on the next page contains an excerpt from a complete list of disposition codes developed for an RDD survey. Classify the following codes into the seven disposition code categories shown in Table 6.2. If you are unable to assign the disposition code to a single category with the description provided below, what additional information would need to be specified for you to choose among the categories?

6.3. Suppose that the number of units in a population is N and that an initial sample of n_1 is selected by *srswor*. A supplemental sample of n_2 is then selected from the $N - n_1$ remaining units, drawn with simple random sampling without replacement:

(a) Prove that the selection probability of each unit in the combined sample is $(n_1 + n_2)/N$.

Code	Description
1	Completed interview
2	Partially completed interview
3	Callback scheduled
4	Dataphone/fax number
5	Hospitalized
6	Language barrier
7	Refusal
8	Not available during data collection
9	Soft refusal—callback to be assigned
10	Ring/no answer
11	Deployed member of US military
12	No eligible respondent in HH
13	Hard/hostile refusal
14	Callback—eligible respondent not available, no interview
15	Other

(b) Show that if each initial sample unit has a response probability of r_1 and each supplemental sample unit has a response rate of r_2, then the inclusion probability of each unit, i.e., the probability accounting for sampling and response, is $(r_1 n_1 + r_2 n_2)/N$.

(c) How would you use the result in part (b) to select an initial sample large enough to produce a responding sample of some desired size, n^*?

6.4. The following is a population of 4 schools with their enrollments. A probability proportional to size sample of $n = 2$ schools is selected.

School	Students	Sample
1	110	x
2	58	
3	223	x
4	133	
Total	524	

(a) Compute the selection probabilities of all 4 schools in a sample of size 2.

(b) Suppose that school 3 refuses to cooperate. One replacement school is selected from schools 2 and 4 with probability proportional to their relative sizes. That is, select one school with *pps* with the relative size computed with respect to the remaining population size after schools 1 and 3 are removed. Show that the selection probabilities of schools 2 and 4 conditional on the initial sample of schools 1 and 3 are not equal to the selection probabilities computed in (a).

Calculate $\pi_2^* = \pi_{2,draw1} + (1 - \pi_{2,draw1}) \, \pi_{2,draw2}^*$ for school 2 with
$\pi_{2,draw1}$ = selection probability of school 2 in the first sample of size 2
$\pi_{2,draw2}^*$ = conditional selection probability of school 2, given that schools
1 and 3 were selected in the first sample.

(c) Make the same calculation as in (b) assuming that schools 1 and 4 were
    selected initially and that one replacement school is selected from schools
    2 and 3. Repeat the calculation of $\pi_2^* = \pi_{2,draw1} + (1 - \pi_{2,draw1}) \, \pi_{2,draw2}^*$
    for school 2 given that schools 1 and 4 were initially selected. Is your
    answer the same or different from that in part (b)?
(d) Explain why the value of π_2^* varies depending on which schools were
    selected in the initial sample. What is the implication of this for selecting
    substitutes in *pps* sampling?

6.5. You are to conduct a survey of retail business establishments in a large
metropolitan area to gauge their plans for hiring or laying-off employments
in the second half of the current calendar year. The frame will be purchased
from a commercial vendor, but any list the vendor provides is known to
have some problems. The list is updated once per calendar quarter to add
new businesses. The vendor does only a limited amount of work to purge its
database of establishments that have gone out of business. Contact informa-
tion (telephone numbers, mailing addresses, and physical location addresses)
are out-of-date for some establishments.

(a) List the types of sample losses that you may experience and that should
    be accounted for when determining an initial sample size.
(b) Discuss how you would attempt to assign percentages to these losses.

6.6. The proposal team has determined that data from 500 completed inter-
views will satisfy the analytic requirements for the list-assisted random digit
dialing (Brick et al. 1995) study detailed below. Your assignment is to com-
pute the size of the RDD sample to be selected in order to guarantee (in
expectation) 500 interviews. Specifically, you will accomplish the following
tasks:

(a) Identify the relevant study rates (e.g., response and eligibility rates) and
    values that need to be considered in order to arrive at the desired 500
    completed interviews.
(b) Estimate the number of telephone numbers to be selected for the study,
    and the corresponding number of interviewers required to complete the
    study on time.
(c) Determine the impact of your task two estimates on the study budget.
(d) Briefly summarize and justify your results.

The client has provided some assumptions in the study description section
that may be useful to your task. You should consider the ramifications of any
"overly optimistic" assumptions.

Table 6.7: Population estimates from the 2005 American Community Survey for the District of Columbia.

Subject (Year)	Total (%)	Margin of Error:	Male (%)	Margin of Error:	Female (%)	Margin of Error
Total population	515,118		242,560	+/− 593	272,558	+/− 593
Under 5	7.3	+/− 0.1	7.9	+/− 0.1	6.8	+/− 0.1
5 to 9	5.4	+/− 0.4	5.9	+/− 0.5	5.0	+/− 0.5
10 to 14	5.9	+/− 0.4	6.1	+/− 0.6	5.8	+/− 0.5
15 to 19	4.1	+/− 0.2	4.8	+/− 0.3	3.5	+/− 0.3
20 to 24	5.0	+/− 0.2	5.0	+/− 0.4	5.0	+/− 0.3
25 to 29	11.3	+/− 0.1	10.9	+/− 0.2	11.7	+/− 0.2
30 to 34	9.2	+/− 0.1	9.1	+/− 0.2	9.2	+/− 0.1
35 to 39	8.3	+/− 0.5	8.3	+/− 0.8	8.3	+/− 0.7
40 to 44	7.0	+/− 0.5	7.7	+/− 0.7	6.3	+/− 0.7
45 to 49	6.9	+/− 0.1	7.0	+/− 0.2	6.7	+/− 0.2
50 to 54	6.7	+/− 0.1	6.8	+/− 0.3	6.6	+/− 0.1
55 to 59	6.2	+/− 0.4	6.1	+/− 0.5	6.3	+/− 0.5
60 to 64	4.6	+/− 0.3	4.2	+/− 0.5	4.9	+/− 0.5
65 to 69	3.5	+/− 0.3	3.4	+/− 0.4	3.7	+/− 0.4
70 to 74	2.7	+/− 0.2	2.3	+/− 0.3	2.9	+/− 0.3
75 to 79	2.7	+/− 0.2	2.2	+/− 0.3	3.2	+/− 0.4
80 to 84	1.8	+/− 0.2	1.1	+/− 0.3	2.4	+/− 0.4
85 and over	1.4	+/− 0.2	1.1	+/− 0.3	1.7	+/− 0.3

Note: Data are limited to the household population and exclude the population living in institutions, college dormitories, and other group quarters

<u>Client Study Description</u>. The 2008 District of Columbia Social Interaction Study (DC-SIS) is sponsored by the D.C. Council of the Friendly Handshake (DCFH) to better understand the social dynamics of males in the District and how these dynamics change in the presence of alcohol. All noninstitutionalized males aged 20–34 who have lived in any of the eight D.C. wards[5] for at least six months are eligible for the DC-SIS. Population count estimates from the 2005 American Community Survey are given at the end of the problem in Table 6.7. The study is a two-phase sampling design with:

- Phase 1—a 5 minute CATI screening interview to identify eligible persons and to make an appointment for a face-to-face interview (no participant incentive)
- Phase 2—a 45 minute in-person interview conducted as soon as possible after the screener interview, with a \$50 incentive payment (\$25 more for refusal conversion)

The interviews will be conducted in either English or Spanish. An insignificant percent of the residents in D.C. speak a language other than English or Spanish.

Study cases will be selected from the 1+ 100-number blocks of landline telephone numbers supplied by a vendor of your choosing. A 100-block is a consecutive block of 100 telephone numbers. For example, 202-123-1200 through 202-123-1299 is a 100-block. A 1+ 100-block is a 100-block that contains at least 1 residential number. Assume that you will also choose a vendor to screen out nonworking telephone numbers (approximately 65 % of the sample) prior to phase-1 data collection and to reverse match the numbers to addresses. In other words, you will purchase an initial list of telephone numbers, and the vendor will determine which of these numbers are working residential numbers. For each working residential number, the vendor will supply a street address if one is available. The percentage of numbers for which an address can be supplied is typically about 65 %.

We anticipate at least a 50 % cooperation rate for the short screener interview among those with an available home address. An advance letter will be mailed to each household for which you have an address. Among numbers with no address, experience has shown that cooperation is poorer. For this exercise, assume that the cooperation rate is 25 % among the no-address numbers. Additionally assume that the rate of eligible persons is the same for address and no-address records. Among the eligible participants who complete the 5 m screener, we anticipate an 80 % overall response rate to the in-person interview. Approximately 10 % of the respondents will require refusal conversion.

The project should be completed within a six-month window—one month for sample design, sample selection, and pretesting; four months of data collection; and one month for post-survey processing and final reports.

---

[5]   http://planning.dc.gov/planning/frames.asp?doc=/planning/lib/planning/maps/docs/census_tract.pdf.

Additional In-house Assumptions. Telephone and field interviewers are paid $10.00 and $13.50 per hour, respectively, and work approximately 24 h per week. On average, approximately 4.5 phone calls will be required to complete a screening interview and 1.5 in-person visits to complete the 45 m interview. Nonproductive calls and contacts are expected to take 1.5 m and 30 m, respectively. We estimate that the telephone interviewers spend approximately 70 % of their weekly hours on tasks that are unrelated to interviewing such as address location, administrative duties, and documentation. The percent of time spent on scheduling interviews, administrative duties, uploading data, and other such tasks is higher at 85 %.

Chapter 7
The Personnel Survey Design Project: One Solution

An optimization problem was presented in Chap. 2 for a single-stage stratified sample design. In the following sections, we present a solution to the multipurpose design question borrowing from material presented in Chaps. 3–6. A series of solutions was generated for the sample allocation to test the sensitivity of the assumptions. Additionally, different software may produce different yet comparable results. Ultimately, a single solution must be chosen from this set for implementation as discussed below.

7.1 Overview of the Project

The Senior Council within the Verkeer NetUltraValid (VNUV) Corporation has tasked your design team with developing an optimal allocation for their annual employee climate survey—the VNUV Climate Survey, Cycle 5. The survey sample members will be randomly selected through a single-stage stratified design as employed in previous VNUV climate surveys. The analysis variables of interest for the survey include:

1. (Q5) Overall, I am satisfied with VNUV as an employer at the present time.
2. (Q12) There is a clear link between my job performance and my pay at VNUV.
3. (Q15) Overall, I think I am paid fairly compared with people in other organizations who hold jobs similar to mine.
4. The number of training classes attended by the employees in the past 12 months.

The design team met over a three-week period to develop the sample design. During this period, they:

(1) Finalized the assumptions used for the optimization
(2) Formulated the optimization problem

(3) Constructed and implemented computer programs to obtain multiple solution
(4) Developed a presentation to highlight the results to the Senior Council (not shown)
(5) Summarized the work in a final report (not shown)

7.2 Formulate the Optimization Problem

The first task for mathematical modeling as discussed in Chap. 5 is to translate the client's needs and constraints for a survey into a set of equations that can be solved. This is similar to the task of translating word problems into equations in our first algebra class, although often substantially more complicated. Following the components discussed in Sect. 5.1, we extract the necessary information from Chap. 2 to construct the multicriteria optimization problem.

7.2.1 Objective Function

The objective function is the equation that is minimized or maximized to develop a solution. Skimming back through Chap. 2, you will not locate an explicit definition for this function. Welcome to one of the many areas where creativity plays a role in the lives of survey statisticians. Through experience you may develop a preference for a particular type of objective function. Otherwise, the use of more than one objective function (and set of assumptions) can suggest the robustness of your final chosen solution.

Based on previous experience, the objective chosen by the design team was similar to the equation used for Example 5.2. Namely, the allocation should be constructed to minimize the sum of the relvariance of the estimated total ($\hat{T}_j$) for the four analysis variables (Sect. 2.1; repeated in Sect. 7.1 for convenience). In other words, the explicit formula for the first candidate objective function is

$$\Phi = \sum_{j=1}^{4} \omega_j relvar\left(\hat{T}_j\right), \tag{7.1}$$

where ω_j is the importance weights for variable j ($j = 1, \ldots, 4$), $relvar\left(\hat{T}_j\right)$ is the corresponding relvariance such that

$$relvar\left(\hat{T}_j\right) = T_j^{-2} \sum_h N_h \left(\frac{N_h}{n_h} - 1\right) S_{jh}^2$$

and S_{jh}^2 is the unit variance calculated within design stratum h ($h =$ 1, ..., 18). The design team had several discussions about the importance weights used in the objective function. After conferring with the Senior Council, the decision was reached that all of the analysis variables were of equal importance. Consequently, $\omega_j \equiv 1$ for all four variables so that expression (7.1) is rewritten as

$$\Phi = \sum_{j=1}^{4} relvar\left(\hat{T}_j\right).$$ (7.2)

Several objective functions could have been tested. However, because of the time commitments for the design time (a common constraint for researchers), the objective function discussed in Chap. 5 was borrowed for this project.

7.2.2 Decision Variables

The decision variables correspond to the solutions produced from the optimization problem, i.e., sample size and associated allocation to strata. For the VNUV Climate Survey, the allocation solution is required for the 18 design strata—business unit (3 levels) by salary grade (3 levels) by employment tenure (2 levels)—that were shown in Table 2.2. Note that the solution is derived to meet certain analytic objectives specified for the survey. Once the solution has been obtained, the values must be inflated to address sample loss associated with study ineligibility and nonresponse (Chap. 6).

7.2.3 Optimization Parameters

Three sets of parameters were defined for the optimization problem. First, HR provided counts of eligible employees by the sampling strata. These frame counts are shown in Table 2.2 of Chap. 2. Second, the design team incorporated the performance rates calculated from the Cycle 4 study (Table 2.4) to ensure that the analytic objectives could be met with the total number of respondents as well as their distribution across the sampling strata. The last set of parameters includes the population estimates, means/proportions, and standard errors, shown in Table 2.5. Prior to implementation, the design team constructed population standard deviations from the estimated standard errors using expression (3.40):

$$\hat{S}^2 = \frac{n_0 v\left(\hat{\bar{y}}\right)}{1 - f_0} \frac{1}{deff\left(\hat{\bar{y}}\right)}.$$ (7.3)

Note that $deff\left(\hat{\bar{y}}\right) = 1$ for the Cycle 5 calculations because the sample for the previous climate survey was selected by an *srs* design. We visit the design effect again for the Cycle 5 design in Sect. 7.3.

7.2.4 Specified Survey Constraints

Questions were posed to the VNUV Senior Council to finalize the optimization constraints on the sample size and on the precision for a set of estimates (Sect. 2.2). The first constraint was dictated by the survey budget—there are sufficient funds for the Cycle 5 climate survey to process responses from 600 sample members. In addition to constraining the sum of the respondent sizes generated from the allocation, the design team also required that the number in each stratum exceeds a specified minimum value in order to calculate a variance component. Because the actual number selected for the study was calculated as the respondent size inflated for sample loss (e.g., nonresponse) determined from the Cycle 4 survey, the inflated size was constrained to be less than the frame count within the stratum. In summary, the following set of equations was used to constrain the sample allocation:

$$\sum_{h=1}^{H} n_h \leq 600$$
$$2 \leq n_h$$
$$(n_h/r_h) \leq N_h$$

where n_h is the number of respondents within stratum h $(h = 1, \ldots, 18)$ derived from the optimum allocation, N_h is the total number of employees in stratum h calculated from the updated employee list provided by HR (see Table 2.2), and r_h is the sample-loss inflation rate from Cycle 4 calculated as the eligibility rate (= 1 – ineligibility rate) multiplied by the response rate (see Table 2.4).

A second set of constraints was placed on the coefficient of variation (CV) for four estimates (Q5, Q12, Q15, and the average number of training classes) within domains defined by business unit, salary grade within business unit, and categorized tenure within business unit (Table 2.1).

A third set of constraints was imposed by the design team prior to finalizing optimization. These constraints were derived from a power analysis discussed in the next section.

7.3 One Solution

7.3.1 Power Analyses

Having specified the known constraints for the optimization task, the design team next conducted a power analysis to establish a minimum sample size for the business unit domains to meet the desired detectable differences:

- A 5 percentage point difference (or larger) for the employee climate estimates
- A difference of two to three training classes for the average on-the-job education estimates

The design team, however, eventually determined from the power analysis that the desired difference levels were not attainable given the study budget, i.e., the funds used to edit and analyze data from 600 respondents.

The multivariable power analysis focused on four estimates. Beginning with the proportion of staff who (strongly) agrees with the three climate questions restated in Sect. 7.1, Table 2.6 in Chap. 2 showed that the fair compensation question (Q15) consistently had the lowest rate of agreement across the business units. The design team noted that the Q15 estimate had the strongest influence on the power calculations because it has the largest standard deviation. Thus Q5 and Q12 were set aside and not used in the minimum sample size analysis. The influence of Q15 in comparison of the average number of training classes was less clear, so two separate power calculations were made and subsequently combined.

The R function `power.prop.test` produced the results shown in Table 7.1. For example, the R code used to calculate the sample size for the Survey Research (SR) business unit with a detectable difference of 0.05 (or delta=5 percentage points) is

```
power.prop.test(p1=0.69, p2=0.74, sig.level = 0.05,
      power = 0.8, alternative = "two.sided").
```

Similar code was used to calculate the minimum analytic sample size for the CR and FO business units. Note that each value in Table 7.1 from the power analysis with delta=5 percentage points violates the constraint of 600 respondents. The team reran the analysis using several detectable differences; power results for 10, 13, and 15 % are included in the table for comparison. The values for 0.13 looked most promising because the total sample size was well below the maximum value and would hopefully allow the optimization algorithm some flexibility in allocating sample across the strata. Next, the team turned to a similar calculation for the average number of training classes.

The team accessed R again to calculate the minimum sample size per business unit for the average number of training classes with the `power.t.test`

Table 7.1: Minimum sample size by business unit and detectable difference produced by the R function `power.prop.test` for the fair salary question (Q15). Calculations were done for 80 % power and 0.05 level of significance for a two-sided test.

Business	Q15 Cycle 4	Detectable difference (%)			
unit	estimate	5	10	13	15
SR	0.69	1,278	300	171	124
CR	0.83	777	165	86	59
FO	0.60	1,470	356	206	152
Overall		3,526.1	821	463	335

Table 7.2: Minimum sample size by business unit and detectable difference produced by the R function `power.t.test` for the question on number of training classes. Calculations were done for 80 % power and 0.05 level of significance for a two-sided test.

Business	Cycle 4 estimates			Detectable difference			
unit	Mean	*se*	*std*	1.0	1.5	2.0	2.5
SR	18.10	0.98	12.02	1,037	462	261	168
CR	12.60	0.90	8.21	491	219	124	80
FO	8.94	0.60	7.74	432	193	109	71
Overall				1,959	874	494	318

Table 7.3: Minimum sample size by business unit for design optimization of the Verkeer NetUltraValid (VNUV) Climate Survey, Cycle 5. Design effect of 1.05 used to account for variation introduced through weighting.

Business	Minimum no.	*deff* adjusted no.
unit	of respondents	of respondents
SR	171	179
CR	86	90
FO	206	216
Overall	462	486

function. Table 7.2 contains the results from the second power analysis for a range of number of classes included as the desired detectable difference. The standard deviations (*std*) were calculated with expression (7.3) with $deff(\hat{\bar{y}}) = 1$. Detectable differences between 2 and 3 classes were classified by the Senior Council as meaningful. Differences less than these numbers were examined to evaluate the sample size requirements for higher levels of precision.

Having examined the results, the team decided to take the "best of both worlds." The maximum sample size required by business unit for a 13 percentage point difference in the climate estimates and a 2.5 difference in the average number of training classes combines the information, resulting in the values given in the "Minimum no. of respondents" column of Table 7.3.

Because the Cycle 5 post-data collection analysis will include the use of weights, unlike Cycle 4, a senior statistician was consulted on an appropriate

design effect. A conservative *deff* of 1.05 was used to inflate the analytic sample size to account for factors such as differential weights introduced from a nonresponse adjustment. These inflated values located in the last column of the table were used in the optimization routines discussed next.

7.3.2 Optimization Results

The sample allocation was optimized using both Excel Solver and SAS `proc optmodel` for comparison. The output files from the optimizations are located on the book's web site as discussed below.

Solver

The file containing the Solver output is named `Project 1.Solver.xlsx`. The workbook contains 14 worksheets, some corresponding to tabular information provided in Chap. 2:

- 1 Frame counts (Table 2.2)
- 2 Recode (Table 2.3)
- 3 Study rates (Table 2.4)
- 4_5 Estimates (Tables 2.5 and 2.6)
- 6 *CV*s (Table 2.1)

and some containing input or output from the optimizations:

- SAS data (input data for SAS proc optmodel)
- Power (summary tables from R power functions)
- Solver (Excel Solver optimization)
- Answer Report 1, Sensitivity Report 1, Limits Report 1 (output from Solver)
- Answer Report 2 (output from Solver with multistart option)
- Compare (comparison between Solver and proc optmodel solutions)
- Sensitivity (sensitivity of Solver solution to changes in the assumed response rates)

Details from the Solver optimization are summarized below:

1. The default settings for Solver were used including GRG nonlinear solving method, 0.0001 precision constraint, and 1,000 iterations. The optimization was computed both with and without the "multistart" option resulting in no difference.
2. As shown in Fig. 7.1, the objective function, Eq. (7.2), is tabulated in cell S36 within the Solver worksheet (or `'Solver'!$S$36` using Excel notation). The goal of the optimization is to minimize the sum of the four relvariances, one for each estimate. The change cells, or the respondent

sample size per strata, are located in column M, rows 10 through 27. The series of constraints that have been loaded into the third input box included the maximum sample size ($K$36 >= $M$36), the minimum sample size per business unit determined from the power calculations ($K$38 <= $M$38 through $K$40 <= $M$40), and an additional check to ensure that the allocation inflated for sample loss would not exceed the frame counts per strata ($N$10:$N$27 <= $E$10:$E$27).

3. A proportional allocation was used for the starting values—see 'Solver'!L9.

4. The original optimization was implemented using a maximum respondent sample size of 600. Because the constraints were easily met, the team evaluated a reduced respondent sample size in an attempt to save project time and funds. The final recommended sample size was 575 respondents.

SAS Proc Optmodel

The SAS programs, logs, and lst (output) files are identified by the label Project 1 OptModel n=* where the asterisk (*) indicates the maximum respondent sample size set for the optimization routine. The SAS programmer included the sample size constraints overall and by business unit along with the CV constraints as macro variables at the beginning of the program. Each section of the program either inputs tables specified in Chap. 2 or calculates components for the optimization. The SQP (default) procedure was used in the optimization as shown in the log files.

The design team initially produced an optimization for a 600 respondents to mirror the work completed with Solver (see the files Project 1 OptModel n=600.*). Two additional SAS programs were created corresponding to n=575 and n=550 respondents. As shown in the file Project 1 OptModel n=500.log, a feasible solution with a maximum of 550 respondents was not found.

Comparison of Solver and SAS Proc Optmodel

A comparison of the allocation results from Excel Solver and SAS proc optmodel is shown in Table 7.4. The *solution* to both algorithms satisfied the revised respondent sample size of 575. However, after inflating the allocation for sample loss (*adjusted solution*) and randomly rounding[1] the adjusted values (*random round*), the Solver solution required the selection of 36 (=1,025–1,061) fewer sample cases. Additionally, the Solver objective function was

---

[1] Random numbers from the uniform distribution are generated for each value requiring rounding. If the random number is less than or equal to 0.5, then the integer portion of the value is used as the rounded value. Otherwise, the integer portion plus one is used as the rounded value.
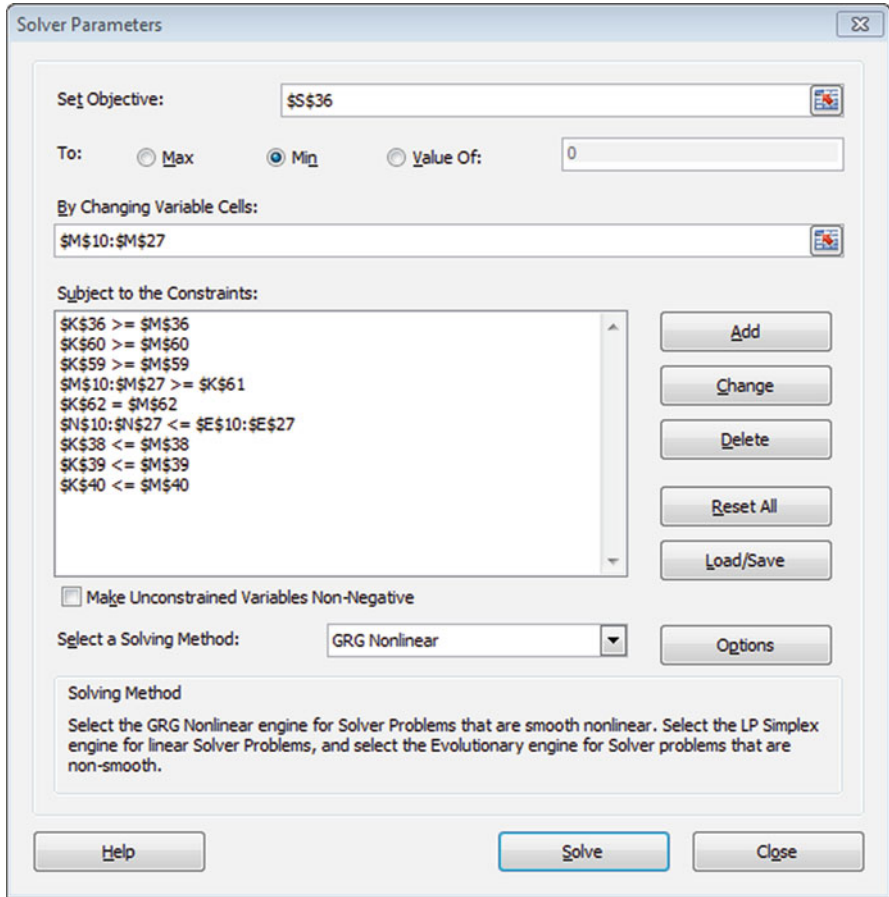
Fig. 7.1: Excel Solver optimization parameter input box.

slightly lower, indicating a somewhat better solution. Therefore, the design team chose the Solver solution as the sample allocation included in the report.

7.4 Additional Sensitivity Analysis

The design team completed one last analysis prior to finalizing the report to the VNUV Senior Council to address the concerns about the estimated response rates (see the response to question #8 in Sect. 2.2). Without detailed information on the likely differential rates by the stratifying characteristics, the team evaluated the impact of an overall reduction in the response rates to identify subgroup estimates that would be most affected. In summary,

Table 7.4: Comparison of optimization results from Excel Solver and SAS `proc optmodel` for the VNUV Climate Survey, Cycle 5

Strata	Business unit	Salary grade	Tenure (Years)	Solver			OptModel (SQP)		
				Solution[a]	Adjusted solution[b]	Random round[c]	Solution	Adjusted solution[a]	Random round
1	SR	A1-A3	<5	12.7	14.3	15	10.9	12.3	12
2			5+	20.1	23.8	23	17.8	21.0	21
3		R1-R5	<5	34.1	73.2	74	33.5	71.9	71
4			5+	65.1	83.8	83	81.4	104.8	104
5		M1-M3	<5	27.1	29.7	30	23.1	25.3	26
6			5+	20.1	28.1	28	12.5	17.5	17
7	CR	A1-A3	<5	12.8	23.5	23	54.6	100.3	100
8			5+	20.3	40.7	41	42.8	85.7	86
9		R1-R5	<5	26.9	53.8	53	43.0	86.0	85
10			5+	24.9	46.2	46	34.7	64.4	64
11		M1-M3	<5	11.9	11.9	12	10.5	10.5	10
12			5+	17.9	22.6	22	30.3	38.4	38
13	FO	A1-A3	<5	59.2	215.6	215	60.9	222.1	222
14			5+	34.3	87.2	87	28.0	71.1	72
15		R1-R5	<5	103.8	162.8	162	53.3	83.6	84
16			5+	45.7	65.3	65	21.8	31.2	31
17		M1-M3	<5	19.3	19.3	20	8.1	8.1	8
18			5+	18.9	26.4	26	7.5	10.5	10
		Total		575.0	1,028.0	1,025	575.0	1,064.9	1,061
Objective function (RelVar)				0.0023			0.0024		
Objective function (Pct CV)				4.82			4.87		

[a] Optimized solution from the package
[b] Solution adjusted for sample loss, i.e., optimized solution divided by the eligibility rate times the response rate
[c] Adjusted solution randomly rounded to whole numbers

that group is the SR division. The following three points are the take-away messages:

(1) If the Cycle 5 response rates are less than 5 percentage points lower than the Cycle 4 values used in the optimization, then there will be a negligible difference in the results.
(2) If the difference in the actual and estimated Cycle 5 response rates is approximately 5 percentage points, then the precision of the estimates within the business units will likely fall below the desired $CV = 0.10$.
(3) If the actual response rates are more than 5 percentage points lower than the estimated values, then the precision of the business unit estimates will approach a CV of 70 %. This is especially true of the SR division estimates because the binding constraint on the sample size as shown in worksheet="Answer Report 2."

7.5 Conclusion

The design team then proceeded to develop a design report around the recommended allocation produced by Excel Solver (Table 7.4). This report included a discussion of the optimization constraints including the need to increase the values for the meaningful detectable differences given the constraint of no more than 600 respondents (i.e., budget). The design team also justified the lowering of the respondent sample size from 600 to 575 by (1) demonstrating the convergence of the optimization system under the reduced sample size and (2) suggesting that the cost savings could be used on methods to increase participation such as a small incentive.

Part II
Multistage Designs

Chapter 8
Project 2: Designing an Area Sample

In this project you will design a sample of census tracts, block groups, and persons from Anne Arundel County in the state of Maryland in the USA. Considering the analytic subgroups, the desired precision of estimates, and the available budget, it has been determined that these sample sizes are to be selected:

Age group (years)	Sample sizes
18–24	200
25–44	200
45–54	200
55–64	200
65+	200
Total	1,000
Sample tracts	25
Sample block groups per tract	1

The sample design will use tracts as PSUs, block groups as SSUs, and persons as elements. The goals of the sample design are to select a sample of the sizes above while (1) achieving a self-weighting sample in each of the age groups above and (2) obtaining an equal workload in each sample PSU. You should pay particular attention to geographic areas that have small population counts and decide how they should be handled in the frame. The tools you need to complete this project are covered in Chaps. 9 and 10.

Use Sampford's method to select the PSUs and SSUs. This is one of several options for selecting probability proportional to size samples. Sampford works for samples of any size and permits joint selection probabilities to

be computed, a requirement for some of the variance estimators described in Chap. 15. This method of selection is available in R pps and `sampling` packages and in SAS `proc surveyselect`. In order to reproduce the solution given later in Chap. 11, include the statement

```
set.seed(-741881304)
```

at the beginning of your program if you use R. If you use SAS surveyselect, use the procedure option

```
seed = 1953.
```

The deliverables for the project will be:

- A sampling report
- SAS or text files giving the units used for the area frame and relevant census counts and measures of size
- SAS or text file for the selected sample along with relevant census counts, measures of size, selection probabilities, and weights.

Contents of the Sampling Report

Below is a list of topic areas that should be included in your report. The order of the sections in your report does not have to be the same as that given below. You should construct your report in a way that presents topics in an order that seems logical to your team.

The report should be written to a client whose staff includes managers and technical personnel. Managers will be more interested in understanding the broad outline of the steps used in weighting. Technical personnel will be interested in understanding the details of sample selection and weight computation, including appropriate formulas. You should consider how to structure your report to serve these audiences.

- Topic areas for the sampling report
- Title page (project title, date of submission, and name of project contact person)
- Introduction (overview of the document)
- Sample design

    Goals of the sample design
    Area sampling frame
    Units, data available, source of the data
    Assigning measures of size to units

- Sample selection

    Method of selection
    Selected units and characteristics of each
    Selection probabilities of units at each stage of selection
    Description of how persons should be selected from area listings

- Maps

    Anne Arundel County
    Selected tracts and block groups

- Appendix

    PROC CONTENTS or codebook of frame and sample files
    Listing of the sample PSUs and sample SSUs with their selection probabilities and census data. On each sample SSU, list the sampling rate you will use to select persons in each domain.

Data Files and Other Information

- *AnneArundel.MD.xls*—Census 2000 tract and block group data for Anne Arundel County
- *Census.glossry2.pdf* —Defines geographic terms used by Census Bureau
- Census tract and block maps for Maryland from the Census Bureau;

    American Factfinder at Census Bureau: factfinder.census.gov
    www.census.gov/geo/www/maps/CP_MapProducts.htm

- Maps of the county are also in

    Anne Arundel.blkgrps(streets).pdf
    Anne Arundel.tracts(streets).pdf
    Anne Arundel.tracts(no streets).pdf

Chapter 9
Designing Multistage Samples

Previous chapters have covered the design of samples selected in a single stage. However, sampling is often done using more than one stage. There are a number of reasons why cluster or multistage sampling may be used. Using multistage samples can often be a practical and cost-efficient solution in situations where a list of elementary (or analytic) units is not available for direct sampling. In those cases, a list of elementary units can be compiled within just the sample clusters rather than for the whole frame. This is especially useful in household samples if a list of every household in a country, state, county, etc., is not available. In other cases, permission to do a survey may have to be obtained at the cluster level. For example, if the goal is to administer a standardized test to a sample of students, administrators in the school district or in the school may have to grant permission to do the survey.

The mode of data collection will also affect the decision on whether to use cluster sampling. If data are to be collected by personal interview, then clustering the sample cases can be a way to save travel costs. This is true regardless of whether a complete list of population members is available. If interviewing will be done by telephone, then clustering sample cases may be unnecessary and statistically inefficient.

Some comments on terminology are in order. *Cluster sampling* means that a group of units is selected at the first stage of sampling. The clusters can be geographic areas, establishments, schools, or some other type of aggregate unit. We will also use the terms *primary sampling unit* (PSU) or *first-stage unit* to be synonymous with *cluster*. Within a sample cluster, elementary units are sampled. Some texts reserve the term "cluster sample" for a single-stage sample in which all elementary units within a cluster are included in the sample. In this book, a cluster sample will include both the cases of complete enumeration of a cluster and of subsampling in a cluster. If subsampling is used within a cluster, we will also call this *multistage sampling*. There can be two or more stages of selection, depending on the application. The term *ultimate cluster* denotes the aggregate of the elementary units across the stages of selection within a sample PSU.

In designing samples of PSUs and subsamples within PSUs, there are two situations to consider. The first is designing a PSU sample from scratch. In that case the issues are how to form the PSUs; how they should be stratified, the number of sample PSUs; how the sample is allocated to strata, the method of sampling the PSUs; and finally how the sampling is to be done within the selected PSUs.

The second case is using an existing PSU sample. The focus then is on how to efficiently design a sample of secondary sampling units (SSUs) and, for a three stage design, elements within SSUs. Decisions must be made on the sample size and method of sampling SSUs and the number of elements to sample within each PSU and SSU. The sample allocation must be determined conditional on the sample of PSUs. Theory for much of the material here can be found in Hansen et al. (1953a, vol. I, Chaps. 6–9), Hansen et al. (1953b, vol. II, Chap. 6), and in Särndal et al. (1992, Chap. 4). Hereafter we will refer to Hansen, Hurwitz, and Madow books as HHM. Despite being almost 60 years old at this point, HHM still has a wealth of valuable information about many of the practical problems encountered in sample design.

Section 9.1 describes some of the units that can be used as PSUs. Section 9.2 presents some of the basic variance formulas for two- and three-stage sampling. These are used in the third section to determine optimal allocations in which cost is a consideration. The fourth section of the chapter discusses estimation of the variance components that are required for sample allocation. Sections 9.5 and 9.6, respectively, briefly cover stratification of PSUs and criteria for identifying PSUs that are selected with certainty, i.e., with probability one.

9.1 Types of PSUs

The types of units that constitute a PSU depend on the survey. In an area probability sample, the units are usually geographic areas like counties, sub-county areas, or other local administrative units. A survey designer may have some freedom in how areas are combined to form PSUs. We discuss these options in depth in Chap. 10. In other cases, the PSUs are naturally occurring units that are forced on the designer. Changing them would be either infeasible or inefficient. When surveying schools, the hierarchy of school districts, schools, classrooms, and students is common in the USA. Trying to use another type of aggregation as a cluster would require defining units that are unnatural to school administrators and would probably be in conflict with the way school records are kept. Other types of natural hierarchies are:

- Business establishments—Employees or accounts might be the elements to be sampled.
- Hospitals—Departments like emergency room, intensive care, and long-term care might be an SSU. Patient records may be considered as nested

within the department where the patient was last treated or might be
sampled directly within a hospital.

- Military personnel—In the American model, some of the levels of hierarchy
  in descending order of size are corps, division, brigade, regiment, company,
  platoon, and squad. Any of these might be used as SSUs. On the other
  hand, these may not be convenient for sampling since all personnel in a
  given level (brigade, say) may not be stationed in the same place. In that
  case, military bases, which are specific geographic locations, may be more
  useful as PSUs.

9.2 Basic Variance Results

To allocate a sample among different stages of sampling, the contributions of
the different stages to the variance of an estimator must be considered. These
components of variance generally depend on the analysis variable and also on
the form of the estimator. In Sects. 9.2.1–9.2.3, we cover some basic results for
linear and nonlinear estimators in two-stage sampling. Section 9.2.4 presents
similar results for three-stage samples.

9.2.1 Two-Stage Sampling

Consider a two-stage sample design in which the first-stage units are selected
using πps sampling, i.e., with varying probabilities and without replacement.
Elements are selected at the second stage via *srswor*. Quite a bit of notation
is needed, even in this fairly simple case:

U = universe of PSUs
M = number of PSUs in universe
U_i = universe of elements in PSU i
N_i = number of elements in the population for PSU i
$N = \sum_{i \in U} N_i$ is the total number of elements in the population
π_i = selection probability of PSU i
π_{ij} = joint selection probability of PSUs i and j
m = number of sample PSUs
n_i = number of sample elements in PSU i
s = set of sample PSUs
s_i = set of sample elements in PSU i
y_{ik} = analysis variable for element k in PSU i
$\bar{y}_U$ = mean per element in the population
$\bar{y}_{Ui}$ = mean per element in the population in PSU i

The π-estimator of the population total, $t_U = \sum_{i \in U} \sum_{k \in U_i} y_k$, of an analysis variable Y is

$$\hat{t}_\pi = \sum_{i \in s} \frac{\hat{t}_i}{\pi_i}$$

where $\hat{t}_i = (N_i/n_i) \sum_{k \in s_i} y_k$, which is the estimate of the total for PSU i with a simple random sample. The design variance of the estimated total can be written as the sum of two components:

$$V\left(\hat{t}_\pi\right) = \sum_{i \in U} \sum_{j \in U} (\pi_{ij} - \pi_i \pi_j) \frac{t_i}{\pi_i} \frac{t_j}{\pi_j} + \sum_{i \in U} \frac{N_i^2}{\pi_i n_i} \left(1 - \frac{n_i}{N_i}\right) S_{U2i}^2 \qquad (9.1)$$

where

$$S_{U2i}^2 = \sum_{k \in U_i} (y_k - \bar{y}_{Ui})^2 \bigg/ (N_i - 1)$$

is the unit variance of Y among the elements in PSU i.

Formula (9.1) is difficult or impossible to use for sample size computations because the number of PSUs in the sample is not exposed. One fallback is to assume with-replacement selection of PSUs, as we did in Chap. 3. Another is to analyze *srs* sampling of PSUs and SSUs as in Example 9.1 below. Determining sample sizes this way does not mean that you are necessarily locked into selecting PSUs and elements within PSUs via *srswr*. Basing sample sizes on a design that is less complicated than the one that will actually be used is a common approach, although, as we will illustrate, it can be deceptive for some analysis variables.

Example 9.1 (Special case: srswor at first and second stages). Suppose the first stage is an *srswor* of m of M PSUs and the second stage is a sample of n_i elements selected by *srswor* from the population of N_i. The π-estimator is

$$\hat{t}_\pi = \frac{M}{m} \sum_{i \in s} \frac{N_i}{n_i} \sum_{k \in s_i} y_{ik}.$$

Its variance is equal to

$$V\left(\hat{t}_\pi\right) = \frac{M^2}{m} \frac{M - m}{M} S_{U1}^2 + \frac{M}{m} \sum_{i \in U} \frac{N_i^2}{n_i} \frac{N_i - n_i}{N_i} S_{U2i}^2$$

where $S_{U1}^2 = \frac{\sum_{i \in U} (t_i - \bar{t}_U)^2}{M-1}$ with t_i being the population total of Y in PSU i and $\bar{t}_U = \sum_{i \in U} t_i / M$ is the mean total per PSU. The relvariance of $\hat{t}_\pi$, $V\left(\hat{t}_\pi\right)/t_U^2$, is

$$V\left(\hat{t}_\pi\right)/t_U^2 = \frac{1}{m} \frac{M - m}{M} B^2 + \frac{1}{t_U^2} \frac{M}{m} \sum_{i \in U} \frac{N_i^2}{n_i} \frac{N_i - n_i}{N_i} S_{U2i}^2 \qquad (9.2)$$

where $B^2 = S_{U1}^2/\bar{t}_U^2 = M^2 S_{U1}^2/t_U^2$ is the unit relvariance among PSU totals.
∎

If $\bar{n}$ elements are selected in each PSU and the sampling fractions of PSUs and elements within PSUs are all small, then the relvariance can be written as

$$\frac{V(\hat{t}_\pi)}{t_U^2} = \frac{B^2}{m} + \frac{W^2}{m\bar{n}} \tag{9.3}$$

with $W^2 = M \sum_{i \in U} N_i^2 S_{U2i}^2/t_U^2$. Expression (9.3) is the form used in the R function, BW2stageSRS, presented later in this section. Textbooks often list a specialized form of Eq. (9.2) that requires that all PSUs have the same size, $\bar{N}$, and that $\bar{n}$ elements are selected in each. In that case, the second-stage sampling fraction is $\bar{n}/\bar{N}$. This implies that the sample is self-weighting: $\pi_i \pi_{k|i} = m\bar{n}/M\bar{N}$. The relvariance in Eq. (9.2) simplifies to

$$V\left(\hat{t}_\pi\right)/t_U^2 = \frac{1}{m}\frac{M-m}{M}B^2 + \frac{1}{m\bar{n}}\frac{\bar{N}-\bar{n}}{\bar{N}}W^2 \tag{9.4}$$

where $W^2 = \frac{1}{M\bar{y}_U^2} \sum_{i \in U} S_{U2i}^2$.

Assuming that $\bar{n}$ elements are selected in each sample PSU, and m/M and $\bar{n}/N_i$ are both small, the more general form of the relvariance, $V\left(\hat{t}_\pi\right)/t_U^2$ in Eq. (9.2), can also be written in terms of a measure of homogeneity δ as follows:

$$\frac{V\left(\hat{t}_\pi\right)}{t_U^2} \doteq \frac{\tilde{V}}{m\bar{n}}k\left[1 + \delta\left(\bar{n} - 1\right)\right] \tag{9.5}$$

where $\tilde{V} = S_U^2/\bar{y}_U^2$, $k = (B^2 + W^2)/\tilde{V}$, and

$$\delta = \frac{B^2}{B^2 + W^2}. \tag{9.6}$$

With some algebra (see Exercise 9.10), it can be shown that when $N_i = \bar{N}$ and both M and $\bar{N}$ are large,

$$\frac{S_U^2}{\bar{y}_U^2} = \frac{1}{\bar{y}_U^2}\frac{\sum_{i \in U}\sum_{k \in U_i}\left(y_{ik} - \bar{y}_U\right)^2}{(N-1)} \doteq B^2 + W^2 \tag{9.7}$$

i.e., the population relvariance can be written as the sum of between and within relvariances. If $k = 1$, Eq. (9.5) equals the expression found in many textbooks. However, when the population count of elements per cluster varies, k may be far from 1, as will be illustrated in Example 9.2. In those cases, Eq. (9.5) with an estimate of the actual k should be used for determining sample sizes and computing advance estimates of coefficients of variation.

With single-stage *srs* sampling of clusters, δ is an intraclass correlation [see (Cochran, 1977, Chap. 8)] but not with two-stage sampling. Nonethe-

less, practitioners do habitually refer to δ as an intraclass correlation. An *ad hoc fpc*, $1 - m\bar{n}/M\bar{N}$, is sometimes inserted in Eq. (9.5) although this does not follow directly from rewriting Eq. (9.2). See Exercise 9.11 for the details needed to obtain Eq. (9.5). Hansen et al. (1953a, Sect. 6.6) and Hansen et al. (1953b, Sect. 6.5) use a more elaborate form of δ, but Eq. (9.6) is more than adequate in practice.

Expression (9.5) is useful for sample size calculation since the number of sample PSUs and sample units per PSU are explicitly in the formula. We will apply the formula in some examples in Sect. 9.4. Equation (9.5) also connects the variance of the estimated total to the variance that would be obtained from a simple random sample since $\tilde{V}/m\bar{n}$ is the relvariance of the estimated total in an *srswor* of size $m\bar{n}$ when the sampling fraction is small. The product $k[1 + \delta(\bar{n} - 1)]$ is a type of design effect. When $k = 1$, the term $1 + \delta(\bar{n} - 1)$ is the approximate design effect found in many textbooks.

Expression (9.5) with $k = 1$ seems often to be treated as if it applies regardless of how the samples of PSUs and elements within PSUs are selected and without regard to the kind of estimator that is used. If, for example, a *pps* sample of PSUs is selected and a poststratified estimator of a total is used, Eq. (9.5) reflects neither of those features. A practitioner needs to realize that it is a specialized formula that does not apply well when methods of sampling other than *srs* are used at different stages. Section 9.2.3 covers a more general two-stage design in which PSUs are selected with varying probabilities and gives relvariance formulas that apply to that case.

Table 9.1 lists some values of $1 + \delta(\bar{n} - 1)$ for a range of δ's and within cluster sample sizes. Even when the measure of homogeneity is small, the effect on the variance of an estimated total can be substantial if many elements are sampled per cluster. For instance, if $\delta = 0.05$, the variance can be 20 % larger than the *srs* variance when $\bar{n} = 5$ but will be almost six times as large when $\bar{n} = 100$. The intuition for this is simply that increasing the sample within each cluster is adding correlated (i.e., more of the same) information which is less effective than adding uncorrelated (new) information from different clusters.

The size of δ is affected by the size of a cluster. Although this is not always true, the elements in a cluster may be more alike when the cluster size is small. This is especially true when clusters are based on geographic areas. Hansen et al. (1953a, Chap. 6, Table 6) give some examples of variables that have different sizes of δ. For clusters of 3 nearby households, the value of δ for number of persons in the household was 0.430 in their example. For clusters of 9, 27, and 62 households, the values of δ were 0.439, 0.243, and 0.112, respectively. These are high compared to many variables. For the indicator variable, unemployed male, the δ's for clusters of 3, 9, 27, and 62 households were 0.060, 0.070, 0.045, and 0.034. For agricultural variables, like whether a farm reports raising a specific crop (e.g., barley, potatoes, or wheat), δ's of 0.4 or larger may be common as long as cluster sizes are 4 or 5 nearby farms. These data are old (1940 US Census), but the fact that δ decreases as geographic cluster size increases is a standard phenomenon.

Table 9.1: Approximate design effects for different sizes of homogeneity measure δ and number of sample elements per cluster.

	$1 + \delta\,(\bar{n} - 1)$		
$\bar{n}$	$\delta = 0.01$	$\delta = 0.05$	$\delta = 0.20$
5	1.04	1.20	1.80
10	1.09	1.45	2.80
25	1.24	2.20	5.80
50	1.49	3.45	10.80
100	1.99	5.95	20.80

Considering two extreme examples may help in understanding when δ is large or small. First, think of an indicator for whether a person has a college degree or not. Suppose that PSUs are city blocks.

High intraclass correlation. Assume that most or all persons on any given block (PSU) either have a college degree or not. In that case, the within-PSU variance component is near 0. The between-PSU variance component is approximately equal to the total variance, which implies that $\delta \doteq B^2 / \left(B^2 + W^2 \right)$ is close to 1. A large sample of blocks will be needed to get a precise estimate of the proportion of people with a college degree. Sampling more than 1 person per block would be inefficient because the people on a block all tend to have the same education level.

Zero intraclass correlation. Suppose blocks are the same size and the proportion with college degrees is the same, $\bar{p}$, in every block in the population. The total of persons with degrees in each PSU is $t_i = \bar{N}\bar{p}$, which is some constant. The between variance is 0, implying that $\delta = 0$. Only 1 block needs to be sampled to estimate the proportion with a college degree because every block is the same. (Notice that if the N_i's varied, δ would not be zero, even if $\bar{p}$ was the same in every block.)

The Maryland Area Population

The next example uses the `MDarea.pop` data set which contains three continuous and two binary variables. This data set is based on the US Census counts from the year 2000 for Anne Arundel County in the state of Maryland. The geographic divisions used in this data set are called tracts and block groups; these will be explained in more detail in Chap. 10. Tracts are constructed to have a desired population size of 4,000 people. Block groups (BGs) are smaller with a target size of 1,500 people. Counts of persons in the data set are the same for most tracts and block groups as in the 2000 Census; five BGs were augmented to have at least 50 persons each. Obtain-

ing microdata for persons within small areas like block groups is generally difficult because of confidentiality restrictions. Thus, we have used models to generate values for persons. The analysis variables in `MDarea.pop` are denoted by `y1`, `y2`, `y3`, `ins.cov` and `hosp.stay` and are generated from models. The first three variables are continuous and positively skewed. The binary variables, `ins.cov`, and `hosp.stay`, are based on the rates of insurance coverage and hospital stay in a 12-month period, as reported in the US National Health Interview Survey (NHIS). We created these variables by fitting models for several variables in the US National Health and Nutrition Examination Survey (NHANES) and NHIS data sets to get regression means that depended on whether a person was Hispanic and on the person's gender and age. Person-level values were created using random effects models that had error terms for tracts, block groups, and persons. These variables are intended to illustrate a range of potential measures of homogeneity while being somewhat realistic.

Because the tracts and block groups in the Maryland population are extremely variable in size, we created two other variables called PSU and SSU and appended them to the data set. Each PSU has approximately the same number of persons; likewise the SSUs were created to have about the same number of persons. The PSUs and SSUs were formed after sorting the file by tract and block group within tract, thus, retaining geographic proximity of persons grouped together. Each PSU has about 5,000 persons while an SSU has about 1,000. Recall that the assumption to obtain expression (9.5) for the variance of an estimator in two-stage sampling is that all PSUs do have the same number of elements, $\bar{N}$. Similar assumptions will be made to simplify the variance in three-stage sampling. Although the assumption of equal PSU size, and later equal SSU size, may seem innocuous, it is far from that as we will illustrate in the next example.

Example 9.2 (Between and within variance components in srs/srs design). The R function `BW2stageSRS` will calculate the unit relvariance of a population, $B^2 + W^2$ for comparison, the ratio $k = (B^2 + W^2)/(S_U^2/y_U^2)$, and the full version of δ in Eq. (9.6). The function assumes that the entire frame is an input. The R code for this example is in *Example 9.2.R*; the code for `BW2stageSRS` is in a separate file. We first compute the results using the PSU and SSU variables as clusters. For the variable `y1` in the Maryland population, the code is

```
 BW2stageSRS(MDarea.pop$y1, psuID=MDarea.pop$PSU,
        pop=MDarea.pop)
 BW2stageSRS(MDarea.pop$y1, psuID=MDarea.pop$SSU,
        pop=MDarea.pop)
```

	B^2	W^2	$S_U^2/\bar{y}_U^2$	$B^2 + W^2$	k	δ
PSUs as clusters						
y1	0.0079	1.4553	1.4627	1.4631	1.0003	0.0054
y2	0.0069	1.0097	1.0163	1.0166	1.0003	0.0068
y3	0.0090	0.1048	0.1136	0.1137	1.0012	0.0787
ins.cov	0.0012	0.2599	0.2611	0.2611	1.0003	0.0046
hosp.stay	0.0175	12.8831	12.8979	12.9006	1.0002	0.0014
SSUs as clusters						
y1	0.0365	1.4277	1.4627	1.4642	1.0010	0.0249
y2	0.0169	1.0004	1.0163	1.0173	1.0010	0.0166
y3	0.0184	0.0954	0.1136	0.1137	1.0012	0.1615
ins.cov	0.0032	0.2581	0.2611	0.2613	1.0010	0.0124
hosp.stay	0.0558	12.8549	12.8979	12.9107	1.0010	0.0043

Values of δ range from 0.0014 to 0.0787 when PSUs are clusters. Deltas are somewhat larger when SSUs are clusters, reflecting the common phenomenon that smaller geographic areas are somewhat more homogeneous than large ones in household populations. The fourth through the sixth columns show that the approximation that $S_U^2/\bar{y}_U^2 \doteq B^2 + W^2$ in Eq. (9.7) works well in this case.

Next, to illustrate the dramatic effect that varying sizes of clusters can have, we compute the same statistics as above using tracts and block groups within tracts as clusters. A variable called `trtBG` is computed since the values of the variable, `BLKGROUP`, are nested within each tract:

```
trtBG <- 10*MDarea.pop$TRACT + MDarea.pop$BLKGROUP
```

	B^2	W^2	$S_U^2/\bar{y}_U^2$	$B^2 + W^2$	k	δ
Tracts as clusters						
y1	0.2605	1.8390	1.4627	2.0995	1.4353	0.1241
y2	0.2687	1.2662	1.0163	1.5349	1.5103	0.1750
y3	0.2707	0.1253	0.1136	0.3960	3.4856	0.6836
ins.cov	0.2624	0.3260	0.2611	0.5884	2.2538	0.4460
hosp.stay	0.3078	16.3171	12.8979	16.6249	1.2890	0.0185
Tract/block groups as clusters						
y1	0.3489	1.9499	1.4627	2.2987	1.5715	0.1518
y2	0.3485	1.3338	1.0163	1.6823	1.6553	0.2072
y3	0.3492	0.1220	0.1136	0.4712	4.1478	0.7411
ins.cov	0.3408	0.3426	0.2611	0.6834	2.6180	0.4987
hosp.stay	0.4246	17.2695	12.8979	17.6941	1.3719	0.0240

Values of δ range from 0.0185 to 0.6836 when tracts are clusters. When block groups are clusters, δ's range from 0.0240 to 0.7411. The measures of homogeneity increase substantially when tracts or block groups are clusters. For example, when PSUs were clusters, $\delta = 0.0054$ for y1 but is 0.1241 when tracts are clusters. This is entirely due to the increase in B^2 when units with highly variable sizes are used. For example, $B^2 = 0.0079$ for y1 when PSU is a cluster but is 0.2605 when tract is a cluster. The fourth through the sixth columns above show that the approximation $S_U^2 / \bar{y}_U^2 \doteq B^2 + W^2$ does not work well when either tracts or block groups are clusters. For y3 and ins.cov, $B^2 + W^2$ is much larger than S_U^2 / y_U^2, implying that setting $k = 1$ in Eq. (9.5) may not be very accurate for some variables if clusters vary in size. ∎

Create Clusters with Equal Sizes If Possible

The variation of the tract sizes in the Maryland population is considerably more than practitioners would prefer when defining PSUs. The range of the number of persons per tract is 86–13,579. Having such large variation in PSU sizes, leads to large differences in the cluster sizes (N_i's) and totals (t_i's). This causes the between variance component B^2 to be large, which in turn leads to the high measures of homogeneity seen above and inefficiencies if a clustered sample is selected. This is also the reason that the approximation $S_U^2 / \bar{y}_U^2 \doteq B^2 + W^2$ is poor in Example 9.2. Standard practice would be to combine the small tracts or block groups so that all PSUs have some prescribed minimum number of persons. Although variation in cluster sizes has a dramatic effect on the factors, like δ, needed to design a sample, this seems to be rarely emphasized in sampling texts. If the designer has some flexibility in forming the clusters, as would usually be the case in a household survey, clusters with nearly equal numbers of elements should definitely be created. In some surveys, the clusters are naturally occurring units, like schools, classrooms, or establishments. In those cases, you may have to live with the pre-defined units, but considering the variation in cluster size will be important when determining sample sizes.

9.2.2 Nonlinear Estimators in Two-Stage Sampling

The between and within variance components can be written down for more complicated designs and estimators. With some simplifying assumptions, the formulas for a two-stage design are analogous to those in the preceding section.

If a nonlinear estimator, like the ratio of two estimated totals or means is used, a general approach to getting variance components is to construct a linear approximation to the nonlinear estimator and then write down the

variance of the approximation. We will cover this technique in more depth in Chap. 15 when variance estimation is discussed. One of the options described there is the *linear substitute* method, which we sketch here. The reader should consult Sect. 15.2 for more details.

Consider an estimator like the ratio of two estimated means, $\hat{\theta} = \hat{\bar{y}}_1/\hat{\bar{y}}_2$, where $\hat{\bar{y}}_j = \hat{t}_{j\pi}/\hat{N}_\pi$ ($j=1,2$) with $\hat{t}_{j\pi} = \sum_{i\in s}\sum_{k\in s_i} d_k y_{jk}$, d_k is the inverse of the selection probability of element k, and $\hat{N}_\pi = \sum_{i\in s}\sum_{k\in s_i} d_k$. Because of the cancellation of $\hat{N}_\pi$, $\hat{\theta}$ is a function of two estimated totals, $\hat{t}_{1\pi}$ and $\hat{t}_{2\pi}$. With some manipulation, the linear approximation to $\hat{\theta}$ can be written as

$$\hat{\theta} - \theta \doteq \sum_{i\in s}\sum_{k\in s_i} d_k z_k + \text{constants}$$

where θ is the population ratio to be estimated, $z_k = \sum_{j=1}^2 \frac{\partial f(\hat{t})}{\partial \hat{t}_{j\pi}} y_{jk}$ ($k \in s_i$), $\hat{t} = (\hat{t}_{1\pi}, \hat{t}_{2\pi})$, and $\partial f(\hat{t})/\partial \hat{t}_j$ is the partial derivative of $\hat{t}$ with respect to the jth estimated total. The term z_k is referred to as a *linear substitute*. The "constants" above do not enter into the variance calculation. The variance of $\hat{\theta}$ can be approximated by computing the variance of $\sum_{i\in s}\sum_{k\in s_i} d_k z_k$.

In the case of simple random sampling at both stages, as in Sect. 9.2.1, $d_k = \frac{M}{m}\frac{N_i}{n_i}$ and $z_k = y_{1k} - \theta y_{2k}$. The ratio can be approximated as

$$\hat{\theta} \doteq \frac{M}{m}\sum_{i\in s}\hat{t}_{zi}$$

where $\hat{t}_{zi} = N_i/n_i \sum_{k\in s_i} z_k$. Thus, the approximate ratio $\hat{\theta}$ can be written in the same way as the estimated total in Example 9.1. Consequently, the relvariance of $\hat{\theta}$ can be expressed in exactly the same way as in Eq. (9.5), assuming that $n_i = \bar{n}$:

$$\frac{V(\hat{\theta})}{\theta^2} \doteq \frac{\tilde{V}}{m\bar{n}}k[1 + \delta(\bar{n}-1)]$$

where $\tilde{V}$ is the unit relvariance of the z_k's, $k = (B^2 + W^2)/\tilde{V}$ and $\delta = B^2/(B^2 + W^2)$. The between and within relvariance components are written in terms of the z_k rather than y_k. Specifically,

$B^2 = S_{U1}^2/\bar{t}_U^2$ with $S_{U1}^2 = \frac{\sum_{i\in U}(t_{zi} - \bar{t}_{Uz})^2}{N-1}$,
$t_{zi} = \sum_{k\in U_i} z_k$, and $\bar{t}_{Uz} = \sum_{i\in U} t_{zi}/M$;
$W^2 = \frac{M}{\theta^2}\sum_{i\in U} N_i^2 S_{U2i}^2$ with $S_{U2i}^2 = \frac{\sum_{k\in U_i}(z_k - \bar{z}_{Ui})^2}{N_i-1}$ and
$\bar{z}_{Ui} = \sum_{k\in U_i} z_k/N_i$;
$\tilde{V} \doteq B^2 + W^2$.

Other nonlinear estimators can be handled by this same method. For example, an estimated mean like $\hat{\bar{y}} = \hat{t}_\pi/\hat{N}_\pi$ or odds ratio in a 2×2 table can

both be linearized and written, approximately, as an estimated total of linear substitutes.

Example 9.3 (Ratio of two totals). Suppose that the proportion of Hispanics with insurance coverage is to be estimated. Define y_{2k} to be 1 if a person is Hispanic and 0 if not; $\alpha_{1k} = 1$ if a person has insurance coverage. Then, $y_{1k} = \alpha_{1k}y_{2k}$ is 1 if person k has insurance and is Hispanic and is zero otherwise. The linear substitute is $z_k = y_{1k} - \theta y_{2k}$ where θ is the proportion of Hispanics with insurance coverage. In this case, z_k can take only three values: $-\theta$, 0, and $1 - \theta$. If a simple random sample of clusters and persons within clusters is selected, BW2stageSRS can be used to compute B^2, W^2, and δ using the linear substitutes as inputs. Assuming that the full population is available, the R code is the following. We do the calculation for clusters defined as either tracts or BGs:

```
    # recode Hispanic to be 1=Hispanic, 0 if not
 y2 <- abs(MDarea.pop$Hispanic - 2)
 y1 <- y2 * MDarea.pop$ins.cov
    #  proportion of Hispanics with insurance
 p <- sum(y1) / sum(y2)
    # linear sub
 z <- y1 - p*y2
 trtBG <- 10*MDarea.pop$TRACT + MDarea.pop$BLKGROUP
 BW2stageSRS(z, psuID=MDarea.pop$TRACT)
 BW2stageSRS(z, psuID=trtBG)
```

The results are $\delta = 0.00088$ for tracts and $\delta = 0.00276$ for BGs. Thus, the effect of clustering on this estimated proportion is inconsequential—a two-stage sample will estimate the proportion almost as precisely as an *srs* would. In contrast, if the estimate is the total number of Hispanics with insurance, then we call BW2stageSRS this way:

```
BW2stageSRS(y1, psuID=MDarea.pop$TRACT)
BW2stageSRS(y1, psuID=trtBG)
```

which return $\delta = 0.02251$ for tracts and $\delta = 0.04026$ for BGs. These are still far less than the δ's in Example 9.2 which also uses tracts and BGs as clusters. Thus, the effect of clustering can be quite different depending on the variable. ∎

In Sect. 9.3, we give formulas for the optimal allocation of a sample to clusters and elements within clusters. The allocations depend, in part, on the value of δ. Examples 9.2 and 9.3 show that sample design decisions on the number of sample clusters and persons per cluster could be quite different depending on which type of estimate we consider. This will be especially true for calibration estimators, which are covered in Chap. 14. Calibration estimators use auxiliary variables to reduce variances. Similar to what we just saw in Example 9.3, the effect of clustering on calibration estimators can be much less than for π-estimators.

9.2.3 More General Two-Stage Designs

Variances of estimators in designs more complicated than simple random sampling at each stage can also be written as a sum of components. However, these have limited usefulness in determining sample sizes. Expression (9.1) is an example of the component variance formula for a design in which PSUs are selected with varying probabilities and without replacement. The first term in Eq. (9.1) has the problem that the number of PSUs is not explicit in the formula.

A more useful formulation is the case where PSUs are selected with varying probabilities but with replacement, and the sample within each PSU is selected by *srswor*. As noted in Chap. 3, with-replacement designs may not often be used in practice but have simple variance formulae. The *pwr*-estimator of a total is

$$\hat{t}_{pwr} = \frac{1}{m} \sum_{i \in s} \frac{\hat{t}_i}{p_i}$$

where $\hat{t}_i = \frac{N_i}{n_i} \sum_{k \in s_i} y_{ik}$ is the estimated total for PSU i from a simple random sample and p_i is the one-draw selection probability of PSU i. The variance of $\hat{t}_{pwr}$ from Cochran (1977, pp. 308–310) is

$$V\left(\hat{t}_{pwr}\right) = \frac{1}{m} \sum_{i \in U} p_i \left(\frac{t_i}{p_i} - t_U\right)^2 + \sum_{i \in U} \frac{N_i^2}{m p_i n_i} \left(1 - \frac{n_i}{N_i}\right) S_{U2i}^2. \qquad (9.8)$$

Making the same assumption as in Sect. 9.2.1 that $\bar{n}$ elements are selected in each PSU, the variance reduces to

$$V\left(\hat{t}_{pwr}\right) = \frac{S_{U1(pwr)}^2}{m} + \frac{1}{m\bar{n}} \sum_{i \in U} \left(1 - \frac{\bar{n}}{N_i}\right) \frac{N_i^2 S_{U2i}^2}{p_i}$$

where, in this case, $S_{U1(pwr)}^2 = \sum_{i \in U} p_i \left(\frac{t_i}{p_i} - t_U\right)^2$. Dividing this by t_U^2 and assuming that the within-PSU sampling fraction, $\bar{n}/N_i$, is negligible, we obtain the relvariance of $\hat{t}_{pwr}$ as, approximately,

$$\frac{V\left(\hat{t}_{pwr}\right)}{t_U^2} \doteq \frac{B^2}{m} + \frac{W^2}{m\bar{n}} = \frac{\tilde{V}}{m\bar{n}} k \left[1 + \delta\left(\bar{n} - 1\right)\right] \qquad (9.9)$$

with $\tilde{V} = S_U^2/\bar{y}_U^2$, $k = (B^2 + W^2)/\tilde{V}$,

$$B^2 = \frac{S_{U1(pwr)}^2}{t_U^2}, \qquad (9.10)$$

$$W^2 = \frac{1}{t_U^2} \sum_{i \in U} N_i^2 \frac{S_{U2i}^2}{p_i}, \tag{9.11}$$

$$\delta = B^2 / \left(B^2 + W^2 \right). \tag{9.12}$$

For later use in variance component estimation, we can also write Eq. (9.8) as

$$V \left(\hat{t}_{pwr} \right) = V_{PSU} + V_{SSU}. \tag{9.13}$$

Expression (9.9) has the same form as Eq. (9.5) but with different definitions of B^2 and W^2. Expression (9.9) also has the interpretation of an *srs* variance or an unclustered variance, $\tilde{V}/m\bar{n}$, times a design effect, $k[1 + \delta(\bar{n} - 1)]$, in the same way that Eq. (9.5) did.

Example 9.4 (ppswr at first stage, srs at second). This example repeats the calculations in Example 9.2 for the variables in the Maryland area population. Assume that clusters will be selected proportional to the count of persons in each cluster. The function `BW2stagePPS` computes the population values of B^2, W^2, and δ shown in Eqs. (9.10)–(9.12), which are appropriate for *ppswr* sampling of clusters. The code for `y1` using PSUs or SSUs as clusters is shown below. The variables, `pp.PSU` and `pp.SSU`, hold the one-draw probabilities p_i that appear in Eq. (9.8):

```
pp.PSU <- table(MDarea.pop$PSU) / nrow(MDarea.pop)
pp.SSU <- table(MDarea.pop$SSU) / nrow(MDarea.pop)
BW2stagePPS(MDarea.pop$y1, pp=pp.PSU,
            psuID=MDarea.pop$PSU)
BW2stagePPS(MDarea.pop$y1, pp=pp.SSU,
            psuID=MDarea.pop$SSU)
```

The code using tracts or BGs as clusters is similar and is in the file `Example 9.4.R`. The results are:
With this design, the between term is much smaller than the within, compared to the results in Example 9.2. This is true whether PSU and SSU are used as clusters or tracts and BGs are used. When clusters are selected by *srs*, S_{U1}^2 is the variance of the cluster totals around the average cluster total. In contrast, with *pps* sampling of clusters, $S_{U1(pwr)}^2$ is the variance of the estimated population totals, t_i/p_i, around the population total, t_U. When clusters are selected with probability proportional to N_i, then $t_i/p_i = N_i \bar{y}_{Ui}/ \left(N_i/N \right) = N \bar{y}_{Ui}$. If these one-cluster estimates of the population total are fairly accurate, as they are here, the B^2 term can be quite small. This leads to much smaller values of δ in *pps* sampling of clusters, implying that the effect of clustering is less important in this population for a design that selects clusters with probabilities proportional to their population counts. ∎

	B^2	W^2	$B^2 + W^2$	k	δ
PSUs as clusters					
y1	0.0078	1.4553	1.4630	1.0002	0.0053
y2	0.0068	1.0097	1.0165	1.0002	0.0067
y3	0.0088	0.1048	0.1136	1.0002	0.0778
ins.cov	0.0012	0.2599	0.2611	1.0002	0.0046
hosp.stay	0.0173	12.8831	12.9004	1.0002	0.0013
SSUs as clusters					
y1	0.0364	1.4277	1.4642	1.0010	0.0249
y2	0.0169	1.0004	1.0173	1.0010	0.0166
y3	0.0183	0.0954	0.1137	1.0008	0.1611
ins.cov	0.0032	0.2581	0.2613	1.0010	0.0124
hosp.stay	0.0557	12.8549	12.9106	1.0010	0.0043

	B^2	W^2	$B^2 + W^2$	k	δ
Tracts as clusters					
y1	0.0092	1.4539	1.4631	1.0002	0.0063
y2	0.0107	1.0058	1.0165	1.0002	0.0106
y3	0.0136	0.1001	0.1136	1.0002	0.1194
ins.cov	0.0018	0.2593	0.2611	1.0002	0.0069
hosp.stay	0.0223	12.8786	12.9009	1.0002	0.0017
Tract/block groups as clusters					
y1	0.0160	1.4478	1.4638	1.0007	0.0109
y2	0.0176	0.9994	1.0171	1.0007	0.0173
y3	0.0211	0.0926	0.1137	1.0006	0.1857
ins.cov	0.0039	0.2574	0.2612	1.0007	0.0148
hosp.stay	0.0509	12.8567	12.9076	1.0008	0.0039

Practitioners habitually gravitate toward *pps* sampling of clusters rather than *srs*. This example makes it clear why this choice is often a good one. Of course, accurate values of the cluster sizes are needed for *pps* to be effective, and these are not always available.

9.2.4 Three-Stage Sampling

A common design in household surveys is to select PSUs, SSUs within PSUs, and households within SSUs. In the USA, SSUs are typically subcounty geographic areas like census tracts or block groups. These are described in detail in Chap. 10. In such a three-stage design, there, naturally, are three variance components. We first present the variance formula for an estimated total when simple random sampling is used at all three stages.

Simple Random Sampling at All Three Stages

There is, regrettably, even more notation in three-stage sampling to specify the situation. Suppose that N_i is the population number of SSUs in PSU i and that n_i is the number selected by *srswor*; $N = \sum_{i \in U} N_i$ is the total number of SSUs in the population; Q_{ij} is the population number of elements in SSU j within PSU i; and q_{ij} is the number of elements selected by *srswor* from PSU/SSU ij. The total number of elements in PSU i is Q_i and in the population is Q. The population of SSUs in PSU i is U_i; the population of elements in PSU/SSU ij is U_{ij}.

If an *srswor* is selected at each stage, the selection probabilities of PSUs, SSUs, and elements are m/M, n_i/N_i, and q_{ij}/Q_{ij}. The π-estimator of the total is

$$\hat{t}_\pi = \frac{M}{m} \sum_{i \in s} \frac{N_i}{n_i} \sum_{j \in s_i} \frac{Q_{ij}}{q_{ij}} \sum_{k \in s_{ij}} y_k,$$

where s_i is the set of sample SSUs in PSU i and s_{ij} is the set of sample elements in PSU/SSU ij. The relvariance of the π-estimator is (Hansen et al. 1953b, Sect. 7.4)

$$\frac{V\left(\hat{t}_\pi\right)}{t_U^2} = \frac{1}{t_U^2} \left\{ \frac{M^2}{m} \frac{M-m}{M} S_{U1}^2 + \frac{M}{m} \sum_{i \in U} \frac{N_i^2}{n_i} \frac{N_i - n_i}{N_i} S_{U2i}^2 \right. \tag{9.14}$$

$$\left. + \frac{M}{m} \sum_{i \in U} \frac{N_i}{n_i} \sum_{j \in U_i} \frac{Q_{ij}^2}{q_{ij}} \frac{Q_{ij} - q_{ij}}{Q_{ij}} S_{U3ij}^2 \right\},$$

where

$S_{U1}^2 = \frac{\sum_{i \in U}(t_i - \bar{t}_U)^2}{M-1}$ as in Example 9.1

$S_{U2i}^2 = \frac{1}{N_i - 1} \sum_{j \in U_i} \left(t_{ij} - \bar{t}_{Ui}\right)^2$ is the unit variance of SSU totals in PSU i with $t_{ij} = \sum_{k \in U_{ij}} y_k$ being the population total for PSU/SSU ij,

$\bar{t}_{Ui} = \sum_{j \in U_i} t_{ij} / N_i$ is the average total per SSU in PSU i

$S_{U3ij}^2 = \frac{1}{Q_{ij} - 1} \sum_{k \in U_{ij}} \left(y_k - \bar{y}_{Uij}\right)^2$ is the unit variance among elements in PSU/SSU ij with $\bar{y}_{Uij} = \sum_{k \in U_{ij}} y_k / Q_{ij}$

To write Eq. (9.14) in a form more useful for sample size calculation, assume that the same number of SSUs, $\bar{n}$, is selected from each PSU and the same number of elements, $\bar{\bar{q}}$, is selected from each SSU. Further suppose that the number of SSUs in each PSU is the same, $N_i = \bar{N}$, and that the number of elements in each SSU is the same, $Q_{ij} = \bar{\bar{Q}}$. In that special case, Eq. (9.14) can be rewritten as

$$\frac{V\left(\hat{t}_\pi\right)}{t_U^2} = \frac{1}{\bar{\bar{y}}_U^2} \left\{ \frac{1}{m} \frac{M-m}{M} S_1^2 + \frac{1}{m\bar{n}} \frac{\bar{N}-\bar{n}}{\bar{N}} S_2^2 + \frac{1}{m\bar{n}\bar{\bar{q}}} \frac{\bar{\bar{Q}}-\bar{\bar{q}}}{\bar{\bar{Q}}} S_3^2 \right\}, \quad (9.15)$$

where

$$\bar{\bar{y}}_U = \sum_{i\in U} \sum_{j\in U_i} \sum_{k\in U_{ij}} y_k \Big/ M\bar{N}\bar{\bar{Q}}$$

$$S_1^2 = (M-1)^{-1} \sum_U \left(\bar{\bar{y}}_{Ui} - \bar{\bar{y}}_U\right)^2$$

$\bar{\bar{y}}_{Ui} = t_i \Big/ \bar{N}\bar{\bar{Q}}$ is the mean per element in PSU i

$$S_2^2 = \sum_{i\in U} \sum_{j\in U_i} (\bar{y}_{Uij} - \bar{\bar{y}}_{Ui})^2 \Big/ M\left(\bar{N}-1\right)$$

$\bar{y}_{Uij} = \sum_{k\in U_{ij}} y_k \Big/ \bar{\bar{Q}}$ is the mean per element in SSU ij

$$S_3^2 = \sum_{k\in U_{ij}} (y_k - \bar{y}_{Uij})^2 \Big/ M\bar{N}\left(\bar{\bar{Q}}-1\right)$$

Expression (9.15) is also found in Cochran (1977, Eq. (10.26)). Although Eq. (9.15) is relatively simple, the assumptions that each PSU has the same population number of SSUs and that each SSU has the same population number of elements are limitations. Assuming that the sampling fractions of PSUs, SSUs within PSUs, and elements within SSUs are all small, a more general relvariance formula that does allow for varying population sizes and still requires that $\bar{n}$ SSUs and $\bar{\bar{q}}$ elements be selected is

$$\frac{V\left(\hat{t}_\pi\right)}{t_U^2} \doteq \frac{B^2}{m} + \frac{W_2^2}{m\bar{n}} + \frac{W_3^2}{m\bar{n}\bar{\bar{q}}}, \quad (9.16)$$

where $B^2 = M^2 S_{U1}^2/t_U^2$, $W_2^2 = M\sum_{i\in U} N_i^2 S_{U2i}^2/t_U^2$, and $W_3^2 = M\sum_{i\in U} N_i \sum_{j\in U_i} Q_{ij}^2 S_{U3ij}^2/t_U^2$.

Varying Probabilities at First Stage, Simple Random Sampling at Later Stages

In the case of with-replacement sampling of PSUs with varying probabilities and *srswor* at the second and third stages, the relvariance can be written (with a few assumptions) in a form useful for sample size calculations. Treating the case where SSUs are selected via *srs* (either with or without replacement) is not too unrealistic since SSUs (like block groups) are often created to have about the same population sizes.

The relvariance of the *pwr*-estimator of a total is derived in Hansen et al. (1953b, Chap. 9, p. 211) and Särndal et al. (1992, p. 149):

$$
\frac{V\left(\hat{t}_{pwr}\right)}{t_U^2} = \frac{1}{t_U^2}\left\{ \frac{S_{U1(pwr)}^2}{m} + \frac{1}{m}\sum_{i\in U}\frac{N_i^2}{p_i n_i}\frac{N_i - n_i}{N_i}S_{U2i}^2 \right.
$$

$$
\left. + \frac{1}{m}\sum_{i\in U}\frac{1}{p_i}\frac{N_i}{n_i}\sum_{j\in U_i}\frac{Q_{ij}^2}{q_{ij}}\frac{Q_{ij} - q_{ij}}{Q_{ij}}S_{U3ij}^2 \right\} \quad (9.17)
$$

$$
\equiv \frac{1}{t_U^2}\left\{V_{PSU} + V_{SSU} + V_{TSU}\right\},
$$

where V_{PSU}, V_{SSU}, and V_{TSU} are defined by the last equality. ("TSU" stands for third-stage unit.) In Eq. (9.17) $S_{U1(pwr)}^2$ is the same as defined below expression (9.8) and

$S_{U2i}^2 = \frac{1}{N_i - 1}\sum_{j\in U_i}\left(t_{ij} - \bar{t}_{Ui}\right)^2$ is the unit variance of SSU totals in PSU i with $t_{ij} = \sum_{k\in U_{ij}} y_k$ being the population total for PSU/SSU ij

$\bar{t}_{Ui} = \sum_{j\in U_i} t_{ij}\Big/ N_i$ is the average total per SSU in PSU i

$S_{U3ij}^2 = \frac{1}{Q_{ij} - 1}\sum_{k\in U_{ij}}\left(y_k - \bar{y}_{Uij}\right)^2$ is the unit variance among elements in PSU/SSU ij with $\bar{y}_{Uij} = \sum_{k\in U_{ij}} y_k\Big/ Q_{ij}$

Expression (9.17) also applies if the inputs are linear substitutes, as defined earlier in Sect. 9.2.2.

HHM present a more complex version of Eq. (9.17) in which PSUs are stratified and SSUs are substratified, but we have not added that complication here. Another complication that is omitted here is the selection of domain elements at different rates. For example, a goal may be to equalize the sample sizes from different race/ethnicity groups.

Expression (9.17) is complex enough that it is not useful for sample size planning. To obtain a simpler formula, suppose that $\bar{n}$ SSUs are sampled in each sample PSU, the sampling fractions of SSUs in each PSU, $\bar{n}/N_i$, are small, and $\bar{\bar{q}}$ elements are selected in each sample SSU. By specializing Eq. (9.17), the relvariance of the *pwr*-estimator is then

$$
\frac{V\left(\hat{t}_{pwr}\right)}{t_U^2} = \frac{B^2}{m} + \frac{W_2^2}{m\bar{n}} + \frac{W_3^2}{m\bar{n}\bar{\bar{q}}}, \quad (9.18)
$$

where $B^2 = S_{U1(pwr)}^2\Big/ t_U^2$ is given by Eq. (9.10),

$$
W_2^2 = \frac{1}{t_U^2}\sum_{i\in U} N_i^2 S_{U2i}^2/p_i; \quad (9.19)
$$

$$
W_3^2 = \frac{1}{t_U^2}\sum_{i\in U}\frac{N_i}{p_i}\sum_{j\in U_i} Q_{ij}^2 S_{U3ij}^2. \quad (9.20)
$$

Expression (9.18) has the same form as Eq. (9.16) for an *srs/srs/srs* design but with different definitions for B^2, W_2^2, and W_3^2. In some applications, an *ad hoc* second-stage *fpc*, $(\bar{N} - \bar{n})/\bar{N}$ where $\bar{N}$ is the average number of SSUs in each PSU, and an *ad hoc* third-stage *fpc*, $(\bar{\bar{Q}} - \bar{\bar{q}})/\bar{\bar{Q}}$ where $\bar{\bar{Q}}$ is the average number of elements in each TSU, may be used in Eq. (9.18) to get a better approximation.

The relvariance in Eq. (9.18) can also be written in terms of two measures of homogeneity:

$$\frac{V\left(\hat{t}_{pwr}\right)}{t_U^2} = \frac{\tilde{V}}{m\bar{n}\bar{\bar{q}}} \left\{ k_1 \delta_1 \bar{n}\bar{\bar{q}} + k_2 \left[1 + \delta_2 \left(\bar{\bar{q}} - 1 \right) \right] \right\} \tag{9.21}$$

where

$k_1 = (B^2 + W^2)/\tilde{V}$ with $\tilde{V} = \frac{1}{Q-1} \sum_{i \in U} \sum_{j \in U_i} \sum_{k \in U_{ij}} (y_k - \bar{y}_U)^2 \big/ \bar{y}_U^2$ is the unit relvariance of y in the population

$k_2 = (W_2^2 + W_3^2)/\tilde{V}$

$\delta_1 = B^2/(B^2 + W^2)$

$W^2 = \frac{1}{t_U^2} \sum_{i \in U} Q_i^2 S_{U3i}^2 \big/ p_i$ with $S_{U3i}^2 = \frac{1}{Q_i - 1} \sum_{j \in U_i} \sum_{k \in U_{ij}} (y_k - \bar{y}_{Ui})^2$ and $\bar{y}_{Ui} = \sum_{j \in U_i} \sum_{k \in U_{ij}} y_k \big/ Q_i$, i.e., S_{U3i}^2 is the element-level variance among all elements in PSU i

$\delta_2 = W_2^2/(W_2^2 + W_3^2)$

Hansen et al. (1953b, Chap. 9) give more elaborate versions of δ_1 and δ_2, but the simpler ones above are adequate for sample size planning.

Note that the term W^2 in δ_1 does not enter the variance in Eq. (9.18) but is defined by analogy to the term in two-stage sampling in Eq. (9.10). If elements were selected directly from the sample PSUs (instead of first sampling SSUs), then W^2 above would be the appropriate within-PSU component.

The term δ_1 is a measure of the homogeneity among the PSU totals. If the estimate of the population total from each PSU total, t_i/p_i, was exactly equal to the population total, t_U, then $B^2 = 0$ and $\delta_1 = 0$. That is, if the variation within PSUs is much larger than the variation among PSU totals, then δ_1 will be small; this is the typical situation in household surveys *if PSUs all have about the same number of elements.* As we saw in Example 9.2, the condition of equal-sized PSUs can be critically important to insure that B^2 is small.

If the SSUs all have about the same totals, t_{ij}, then W_2^2 will be small and $\delta_2 \doteq 0$. Although attempts may be made to create SSUs that have about the same number of elements Q_{ij}, the totals t_{ij}, of other variables tend to vary, leading to values of δ_2 that are larger than those of δ_1, as discussed below.

HHM note that in some applications k_1 and k_2 will be near 1 so that this simpler version of the relvariance can be used for planning:

$$\frac{V\left(\hat{t}_{pwr}\right)}{t_U^2} \doteq \frac{\tilde{V}}{m\bar{n}\bar{\bar{q}}} \left\{ \delta_1 \bar{n}\bar{\bar{q}} + \left[1 + \delta_2 \left(\bar{\bar{q}} - 1 \right) \right] \right\}. \tag{9.22}$$

The term in braces is the increase in relvariance over simple random sampling due to using three-stage sampling. If both δ_1 and δ_2 are 0, then three-stage sampling will be as efficient as *srs*. Usually, δ_1 and δ_2 will be positive so that there will be some increase in relvariance compared to *srs*.

In the USA, household survey PSUs are typically counties or groups of counties. These vary in population size but can contain hundreds of thousands or even millions of persons. SSUs may be census tracts which are smaller areas with 1,500–8,000 persons. An alternative for an SSU is a block group, which generally has 600–3,000 people. Household survey PSUs are often large, fairly heterogeneous areas implying that δ_1 tends to be very small for many variables, say, 0.01 or less. SSUs are smaller areas where persons tend to be more alike, leading to δ_2 being a larger number like 0.10. In school surveys, δ_2 may also be larger than δ_1 if PSUs are districts, SSUs are schools, elements are students, and the analysis variables are different kinds of achievement tests. As we have noted several times before, whether the PSUs and SSUs have the same numbers of elements or not can have a major impact on the measures of homogeneity. As always, the sizes of parameters like δ_1 and δ_2 depend on the population and the analysis variables. Having relevant data is important in order to make realistic advance estimates for sample design.

The R function, `BW3stagePPS`, will calculate B^2, W^2, W_2^2, W_3^2, δ_1, and δ_2 defined above for *ppswr/srs/srs* and *srs/srs/srs* sampling. The function is appropriate if an entire frame is available and takes the following parameters:

X	data vector
pp	vector of one-draw probabilities for PSUs; length is number of PSUs in the population
psuID	vector of IDs for PSUs; length is number of units in the population
ssuID	vector of IDs for SSUs; length is number of units in the population. The parameter `ssuID` should have the form `psuID`‖(`ssuID` within PSU)

If the parameter `pp` is set equal to $1/M$ for all PSUs, then the B^2 is computed as $M^{-1} \sum_U \left(t_i - \bar{t}_U \right)^2 \big/ \bar{t}_U^2$, which is approximately equal to the *srswr* value of B^2.

Example 9.5 (Three stages srs/srs/srs). In the Maryland population suppose that the PSU and SSU variables define the first- and second-stage units and that persons are elements in a three-stage design. PSUs, SSUs, and persons are selected by simple random sampling. The call to `BW3stagePPS` for the variable `y1` and the results for `y1`, `y2`, `y3`, `ins.cov`, and `hosp.stay` are listed below:

```
M <- length(unique(MDarea.pop$PSU))
pp.PSU <- rep(1/M,M)

BW3stagePPS(X=MDarea.pop$y1, pp=pp.PSU,
      psuID=MDarea.pop$PSU, ssuID=MDarea.pop$SSU)
```

PSUs and SSUs are first- and second-stage units								
	B^2	W^2	W_2^2	W_3^2	k_1	k_2	δ_1	δ_2
y1	0.0078	1.4553	0.0358	1.4277	1.0002	1.0006	0.0053	0.0245
y2	0.0068	1.0097	0.0125	1.0004	1.0002	0.9967	0.0067	0.0124
y3	0.0088	0.1048	0.0119	0.0954	1.0002	0.9439	0.0778	0.1105
ins.cov	0.0012	0.2599	0.0025	0.2581	1.0002	0.9983	0.0046	0.0098
hosp.stay	0.0173	12.8831	0.0480	12.8549	1.0002	1.0004	0.0013	0.0037

The values of δ_1 are almost the same as in Example 9.2 where PSUs were also selected using *srs*. The values of δ_2 range from 0.0037 to 0.1105, which are fairly small. The values of k_1 and k_2 are near 1, meaning that $B^2 + W^2$ and $W_2^2 + W_3^2$ are close to the unit relvariance in the population. Next, suppose that tracts and BGs within tracts are the first- and second-stage units and that all three stages are again selected by *srs*. The code for the y1 variable is

```
M <- length(unique(MDarea.pop$TRACT))
trtBG <- 10*MDarea.pop$TRACT + MDarea.pop$BLKGROUP
pp.trt <- rep(1/M,M)
BW3stagePPS(X=MDarea.pop$y1, pp=pp.trt,
      psuID=MDarea.pop$TRACT, ssuID=trtBG)
```

As in Example 9.2, the variable trtBG holds a unique identifier for block groups. The results are listed below. Notice that the values of δ_1 and δ_2 are much larger when tracts and BGs are used for sampling units than when the PSU and SSU variables were used. As in Example 9.2, this is due to the highly variable sizes of tracts and BGs.

Tracts and BGs are first- and second-stage units								
	B^2	W^2	W_2^2	W_3^2	k_1	k_2	δ_1	δ_2
y1	0.2577	1.8390	0.2699	2.1084	1.4334	1.6259	0.1229	0.1135
y2	0.2658	1.2662	0.2613	1.4442	1.5075	1.6781	0.1735	0.1532
y3	0.2678	0.1253	0.2609	0.1323	3.4605	3.4615	0.6813	0.6635
ins.cov	0.2597	0.3260	0.2584	0.3730	2.2432	2.4185	0.4434	0.4092
hosp.stay	0.3046	16.3171	0.3155	18.6391	1.2887	1.4696	0.0183	0.0166

The values of δ_1 are about the same as in Example 9.2 where tracts were also selected in the first stage using *srs*. The values of k_1 and k_2 are much larger than 1, implying that $B^2 + W^2$ and $W_2^2 + W_3^2$ are different from the unit relvariance in the population. This is due to the varying sizes of tracts and BGs. ■

Example 9.6 (Three stages ppswr/srs/srs). We repeat the calculation in Example 9.5 but assuming *ppswr* sampling of PSUs. The calculation for y1 using PSU and SSU as the first- and second-stage sampling units is done via this call:

```
pp.PSU <- table(MDarea.pop$PSU) / nrow(MDarea.pop)
BW3stagePPS(X=MDarea.pop$y1, pp=pp.PSU,
      psuID=MDarea.pop$PSU, ssuID=MDarea.pop$SSU)
```

The values of δ_1 and δ_2 are at most 0.0236 with the exception of y3 which has $\delta_1 = 0.0776$ and $\delta_2 = 0.1097$.

	B^2	W^2	W_2^2	W_3^2	k_1	k_2	δ_1	δ_2
PSUs and SSUs are first- and second-stage units								
y1	0.0078	1.4553	0.0358	1.4277	1.0002	1.0006	0.0051	0.0236
y2	0.0068	1.0097	0.0125	1.0004	1.0002	0.9967	0.0065	0.0114
y3	0.0088	0.1048	0.0119	0.0954	1.0002	0.9439	0.0776	0.1097
ins.cov	0.0012	0.2599	0.0025	0.2581	1.0002	0.9983	0.0044	0.0088
hosp.stay	0.0173	12.8831	0.0480	12.8549	1.0002	1.0004	0.0011	0.0027

The situation changes substantially when tracts and BGs are used for stages one and two. The call for y1 is

```
trtBG <- 10*MDarea.pop$TRACT + MDarea.pop$BLKGROUP
pp.trt <- table(MDarea.pop$TRACT) / nrow(MDarea.pop)
BW3stagePPS(X=MDarea.pop$y1, pp=pp.trt,
      psuID=MDarea.pop$TRACT, ssuID=trtBG)
```

	B^2	W^2	W_2^2	W_3^2	k_1	k_2	δ_1	δ_2
Tracts and BGs are first- and second-stage units								
y1	0.0092	1.4539	0.2499	1.6873	1.0002	1.3243	0.0060	0.1284
y2	0.0107	1.0058	0.2379	1.1619	1.0002	1.3774	0.0103	0.1694
y3	0.0136	0.1001	0.2376	0.1073	1.0002	3.0356	0.1192	0.6889
ins.cov	0.0018	0.2593	0.2321	0.3011	1.0002	2.0424	0.0067	0.4351
hosp.stay	0.0223	12.8786	0.2728	14.8946	1.0002	1.1760	0.0015	0.0173

The results are shown above. The values of k_1 are essentially 1, but k_2 is larger than 1 for all variables. The values of δ_2 are much larger than when PSU was used as the first-stage unit, ranging from 0.0173 to 0.6889. Once again, this illustrates the huge effect that varying unit size can have on the measures of homogeneity. ∎

In the next section, we discuss how to determine optimum allocations of the numbers of sample PSUs, SSUs, and elements in both two- and three-stage samples.

9.3 Cost Functions and Optimal Allocations for Multistage Sampling

In determining the allocation of a multistage sample, there are two common situations. One is designing a PSU sample from scratch in which both the number of sample PSUs and the number of elements per PSU are to be determined. The second case is one in which an existing PSU sample will be used and the task is to determine how many elements to sample per PSU. In both cases, the cost of having a PSU in the sample and the cost of collecting and processing data from each element should be considered.

9.3.1 Two-Stage Sampling When Numbers of Sample PSUs and Elements per PSU Are Adjustable

A simple cost function for two-stage sampling assumes that there is a cost per sample PSU and a cost per sample element. As in the case of stratified sampling in Chap. 3, this cost structure is probably an oversimplification, but a simple model can be of some practical use as long as the relative sizes of the unit costs are reasonable. Take the case of an equal number $\bar{n}$ of elementary units sampled from each PSU. We model the total cost as

$$C = C_0 + C_1 m + C_2 m\bar{n},$$

where

$C_0 =$ costs that do not depend on the number of sample PSUs and elements
$C_1 =$ cost per sample PSU
$C_2 =$ cost per element

Groves (1989) is a good source for the many facets of surveys that contribute to costs. Per-PSU costs in a household survey can include recruiting and training interviewers, paying field supervisors, and field listing costs. Per-element costs could include personnel time for conducting an interview, printing costs if paper questionnaires are used, and clerical staff time to review special problems with completed or partially completed questionnaires. The C_0 component can include personnel time for central office staff, e.g., a project manager, computer scientists to program the instrument if computer-assisted personal interviewing is used, programmers to edit the data, and statisticians to design the sample, devise nonresponse follow-up procedures, and develop weighting schemes. As we have pointed out elsewhere, tracking these costs is difficult and often does not mesh well with survey accounting practices. As a result, you may have to be satisfied with fairly rough unit cost estimates.

As shown in Eqs. (9.5) and (9.9), the form of the relvariance of the estimated total is the same when the design is *srs/srs* or *ppswr/srs* and $\bar{n}$ elements are selected in each sample PSU:

$$\frac{V\left(\hat{t}_\pi\right)}{t_U^2} \doteq \frac{\tilde{V}}{m\bar{n}}k\left[1+\delta\left(\bar{n}-1\right)\right]. \tag{9.23}$$

Thus, the formulas below apply to both *srs/srs* and *ppswr/srs* as long as B^2 and W^2 are appropriately defined. The optimal number of units to select per PSU, i.e., the number that minimizes the approximate relvariance, is

$$\bar{n}_{opt} = \sqrt{\frac{C_1}{C_2}\frac{1-\delta}{\delta}}. \tag{9.24}$$

Note that only the ratio of the unit costs needs to be known in order to compute $\bar{n}_{opt}$. The more a PSU costs, the more elements should be selected within each PSU. On the other hand, the larger δ is, the fewer elements should be selected per PSU.

To find the optimal m for a fixed total cost, we substitute $\bar{n}_{opt}$ into the cost function to obtain

$$m_{opt} = \frac{C - C_0}{C_1 + C_2\bar{n}_{opt}}. \tag{9.25}$$

Alternatively, to find the optimal m for a fixed relvariance, CV_0^2, we substitute $\bar{n}_{opt}$ into the relvariance formula to obtain

$$m_{opt} = \frac{\tilde{V}k}{\bar{n}_{opt}CV_0^2}\left[1+\delta\left(\bar{n}_{opt}-1\right)\right]. \tag{9.26}$$

In either the case of finding sample sizes for a fixed total cost or for a target CV, the total sample size is simply $m_{opt}\bar{n}_{opt}$ where the number of sample PSUs and elements per PSU are computed using Eqs. (9.24) and either (9.25) or (9.26). If $k = 1$, Eq. (9.26) reduces to the formula found in most texts.

Figure 9.1 graphs the coefficient of variation based on Eq. (9.23), assuming that $k = 1$ of an estimated total versus a range of values of $\bar{n}$ for $\delta = 0.01$, 0.05, 0.10, and 0.20. A dot is placed on each curve at the point where the CV is a minimum. In some situations, the conventional wisdom is that "the optimum is flat" in the sense that a range of sample sizes will give a CV that is near the minimum value. That is often the case in stratified sampling where the allocation to the strata can depart from the optimal allocation and still be reasonably efficient. In contrast, Fig. 9.1 illustrates that the "flatness" of the optimum clearly depends on the size of δ. If $\delta = 0.01$, sampling anywhere from about 7 to 30 elements per PSU is fairly efficient. For larger δ's, the optimum is more sharply defined. For example, when $\delta = 0.20$, $\bar{n}$ of 2, 3, or 4 gives a CV near the optimum, but allocating more than that to each PSU quickly becomes inefficient.
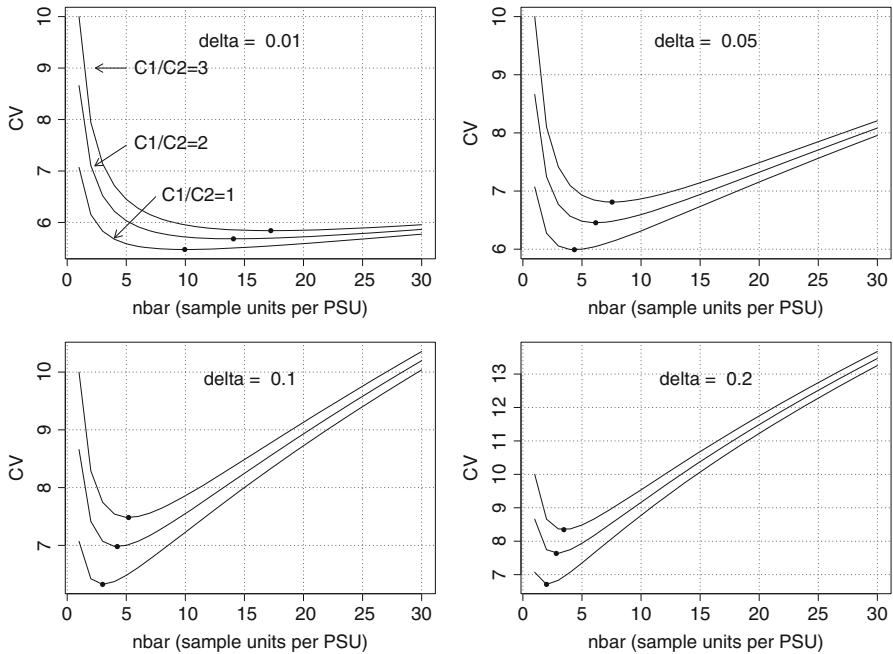
Fig. 9.1: Coefficients of variation for an estimated mean for different numbers of sample elements per PSU. The three curves in each panel correspond to cost ratios of $C_1/C_2 = 3, 2, 1$ (*top to bottom*). The unit relvariance $\hat{V}$ is assumed to be 1. The total budget is $C - C_0 = \$100,000$ with $C_1 = 750, 500,$ and 250 and $C_2 = 250$. Each *dot* is at the optimum.

Example 9.7 (An R function for optimal sample sizes). The R function `clusOpt2` will compute m_{opt} and $\bar{n}_{opt}$ for a two-stage sample which uses simple random sampling at each stage or *ppswr* at the first stage and *srs* at the second. The arguments to the function are

C1	unit cost per PSU
C2	unit cost per element
delta	homogeneity measure δ
unit.rv	unit relvariance
k	ratio of $B^2 + W^2$ to unit relvariance
CV0	target CV
tot.cost	total budget for variable costs, $C - C_0$
	Only one of CV0 and tot.cost can be
	entered in one call to the function.
cal.sw	1, find optimal m_{opt} for fixed total budget
	2, find optimal m_{opt} for target CV0

The code below will compute the optimal number of PSUs and elements per PSU assuming that $C_1 = 750$, $C_2 = 100$, $\delta = 0.05$, the unit relvariance and k are 1, and the total budget for variable costs is \$100,000:

```
clusOpt2(C1=750, C2=100, delta=0.05, unit.rv=1, k=1,
         tot.cost=100000, cal.sw=1)

            C1 = 750
            C2 = 100
         delta = 0.05
   unit relvar = 1
             k = 1
        budget = 1e+05
         m.opt = 51.4
         n.opt = 11.9
            CV = 0.0502
```

In clusOpt2, and in all the functions in this chapter, we have not rounded sample sizes to integers. You can either do the rounding yourself or use a method of sample selection where the expected size can be specified as a non-integer. The function will also accept vector input for one of the parameters at a time. For example, we can see the effect of a range of δ's with

```
clusOpt2(C1=750, C2=100,
         delta=c(0.01, 0.05, 0.10, 0.20),
         unit.rv=1, k=1,
         tot.cost=100000, cal.sw=1)

            C1 = 750
            C2 = 100
         delta = 0.01, 0.05, 0.10, 0.20
   unit relvar = 1
             k = 1
        budget = 1e+05
         m.opt = 28.8, 51.4, 63.6, 77.1
         n.opt = 27.2, 11.9, 8.2, 5.5
            CV = 0.0401, 0.0502, 0.0574, 0.0670
```

Sending the function vectors for more than one parameter (e.g., C2=c(100, 120) and delta=c(0.01, 0.05) will generate an error. ∎

9.3.2 Three-Stage Sampling When Sample Sizes Are Adjustable

A cost function for three-stage sampling, analogous to the one for two-stage sampling in Sect. 9.3.1, is

$$C = C_0 + C_1 m + C_2 m\bar{n} + C_3 m\bar{n}\bar{\bar{q}}. \tag{9.27}$$

The term C_0 is again costs that do not depend on the sample sizes at different stages; C_1 is the cost per PSU; C_2 is the cost per SSU; and C_3 is the cost per element. The function in Eq. (9.27) is, by no means, unique. The cost function for three-stage sampling can potentially be more complicated than for two-stage sampling because more types of costs may have to be considered. For example, in a household survey, travel between SSUs within a PSU may be a consideration, especially if only one or two interviewers cover an entire PSU. Hansen et al. (1953a, Chap. 9, Sect. 18) consider this cost function for three-stage samples:

$$C = C_0\sqrt{m} + C_1 m + C_2 m\bar{n} + C_3 m\sqrt{\bar{n}} + C_4 m\bar{n}\bar{\bar{q}}, \tag{9.28}$$

where $C_0\sqrt{m}$ represents the cost of traveling between PSUs, C_1 is the cost per PSU, C_2 is the cost per SSU, $C_3 m\sqrt{\bar{n}}$ is the total cost of traveling among SSUs, and C_4 is the cost per element. This formulation was found to be useful in Census Bureau work several decades ago, but may not be applicable to surveys with a more modern cost structure. Here we present the results for optima with the simpler cost function (9.27) as illustration.

Minimizing the *ppswr/srs/srs* relvariance in Eq. (9.21) subject to a fixed total cost gives the following optima (Hansen et al., 1953b, p. 225):

$$\bar{\bar{q}}_{opt} = \sqrt{\frac{1 - \delta_2}{\delta_2} \frac{C_2}{C_3}}, \tag{9.29}$$

$$\bar{n}_{opt} = \frac{1}{\bar{\bar{q}}} \sqrt{\frac{1 - \delta_2}{\delta_1} \frac{C_1}{C_3} \frac{k_2}{k_1}}, \tag{9.30}$$

$$m_{opt} = \frac{C - C_0}{C_1 + C_2\bar{n} + C_3\bar{n}\bar{\bar{q}}}. \tag{9.31}$$

If a target relvariance is set at CV_0^2, then the equations for finding the optima for $\bar{\bar{q}}_{opt}$ and $\bar{n}_{opt}$ are the same as above, but the optimum number of elements to sample from each SSU is

$$m_{opt} = \frac{\tilde{V}}{CV_0^2 \bar{n}_{opt}\bar{\bar{q}}_{opt}} \{k_1\delta_1\bar{n}_{opt}\bar{\bar{q}}_{opt} + k_2\left[1 + \delta_2\left(\bar{\bar{q}}_{opt} - 1\right)\right]\}. \tag{9.32}$$

In either the case of finding sample sizes for a fixed total cost or for a target CV, the total sample size is $m_{opt}\bar{n}_{opt}\bar{\bar{q}}_{opt}$, where the number of sample PSUs,

SSUs, and elements per SSU are computed using Eq. (9.31) or (9.32). The R function `clusOpt3` provides a solution for both the problems of minimizing the approximate variance for a fixed total cost and minimizing total cost for a target CV.

The function `clusOpt3` also can be used for *srs* sampling at all three stages. The values of δ_1 and δ_2 that are defined for Eq. (9.21) must be computed with formulas appropriate to simple random sampling. In particular, p_i, the one-draw PSU probability, would be set equal to $1/M$.

If the more complicated cost function in Eq. (9.28) is appropriate, explicit solutions for m_{opt}, $\bar{n}_{opt}$, and $\bar{\bar{q}}_{opt}$ cannot be obtained. HHM give an iterative procedure for arriving at approximate values of m_{opt}, $\bar{n}_{opt}$, and $\bar{\bar{q}}_{opt}$.

Example 9.8 (Optimal sample sizes in a three-stage sample). The R function `clusOpt3` accepts the following parameters:

C1	unit cost per PSU
C2	unit cost per element
C3	cost per element
delta1	homogeneity measure within PSUs, δ_1
delta2	homogeneity measure within SSUs, δ_2
unit.rv	unit relvariance
k1	ratio of $B^2 + W^2$ to the unit relvariance
k2	ratio of $W_2^2 + W_3^2$ to the unit relvariance
CV0	target CV
tot.cost	total budget for variable costs, $C - C_0$
	Only one of CV0 and tot.cost can be entered
	in one call to the function.
cal.sw	1, find optima for a fixed total budget
	2, find optima for a target CV0

The function computes the optima based on Eqs. (9.29), (9.30) and either (9.31) or (9.32). Suppose that $C_1 = 500$, $C_2 = 100$, $C_3 = 120$, $\delta_1 = 0.01$, $\delta_1 = 0.10$, the unit relvariance is 1, as are k_1 and k_2, and the total budget for variable costs is \$100,000. The call to `clusOpt3` is

```
clusOpt3(unit.cost=c(500, 100, 120), delta1=0.01,
        delta2=0.10, unit.rv=1, k1=1, k2=1,
        tot.cost=100000, cal.sw=1)

                C1 = 500
                C2 = 100
                C3 = 120
            delta1 = 0.01
            delta2 = 0.1
      unit relvar = 1
                k1 = 1
                k2 = 1
            budget = 1e+05
```

```
cost check = 1e+05
     m.opt = 28.3
     n.opt = 7.1
     q.opt = 2.7
        CV = 0.0499
```

The function echoes back the input parameter values, returns the optima, and computes the CV that will be achieved with the optimal allocation. In the output, budget is the value of tot.cost while cost check is the value of variable costs found by substituting the optima into Eq. (9.27).

   The function will also accept a vector input for one non-cost parameter at a time. For example, we can see the effect of a range of δ_1's with

```
clusOpt3(unit.cost=c(500, 100, 120),
      delta1=c(0.01,0.05,0.10), delta2=0.10,
      unit.rv=2, k1=1,k2=1,tot.cost=100000,cal.sw=1)

         C1 = 500
         C2 = 100
         C3 = 120
     delta1 = 0.01, 0.05, 0.10
     delta2 = 0.1
 unit relvar = 2
         k1 = 1
         k2 = 1
     budget = 1e+05
 cost check = 1e+05, 1e+05, 1e+05
      m.opt = 28.3, 53.9, 68.6
      n.opt = 7.1, 3.2, 2.2
      q.opt = 2.7
         CV = 0.0706, 0.0830, 0.0922
```

∎

9.3.3 Two- and Three-Stage Sampling with a Fixed Set of PSUs

In some applications, a fixed set of PSUs is used for multiple surveys and the main flexibility in the design is deciding how many elements to select from those PSUs. This is often the case for household samples where an organization may have a "master" sample of PSUs that it uses for different household surveys. Reusing a given sample of PSUs saves the cost of recreating a frame of PSUs, designing the sample, and making the selections. Having a master sample of PSUs may also allow the same set of trained and trusted field personnel to be employed for data collection.

If the total cost, $C = C_0 + C_1 m + C_2 m \bar{n}$, is fixed along with the set of sample PSUs, the number of elements per PSU is determined by the cost constraint only:

$$\bar{n} = \frac{C - C_0 - C_1 m}{C_2 m}. \tag{9.33}$$

If this sample size is not large enough to achieve the desired CV targets, then two choices are (1) to be satisfied with lower precision than desired or (2) to increase the number of sample PSUs. The latter may be difficult to do in a way that it has a design-based justification, depending on how the initial sample of PSUs was selected. The general idea would be to add PSUs but decrease the number of sample elements per PSU in a way that stays within the allotted budget. This may or may not be possible. A final option, which may not be feasible, is to increase the budget and the total sample size.

If a target CV is set and we minimize the cost, then the number of elements to sample per PSU is found by solving for $\bar{n}$ in the approximate relvariance formula, $V\left(\hat{t}_\pi\right)/t_U^2 = \frac{1}{m\bar{n}}\tilde{V}k\left[1 + \delta\left(\bar{n} - 1\right)\right]$, which gives

$$\bar{n} = \frac{1 - \delta}{CV_0^2 m / \tilde{V}k - \delta}. \tag{9.34}$$

The R function, `clusOpt2fixedPSU`, will compute $\bar{n}$ using either Eq. (9.33) or (9.34). The function takes as input the fixed number of PSUs m in addition to the same parameters as `clusOpt2` shown in Example 9.7.

Example 9.9 (Elements per PSU for a fixed set of PSUs and fixed total cost). The following code determines the number of sample elements per PSU for unit costs of $C_1 = 500$ and $C_2 = 100$ when the number of PSUs is fixed at $m = 100$. Budgets of \$100,000, \$500,000, and 1 million dollars are used:

```
clusOpt2fixedPSU(C1=500, C2=100, m=100, delta=0.05,
        unit.rv=2, k=1, CV0=NULL,
        tot.cost=c(100000, 500000, 10^6), cal.sw=1)

             C1  = 500
             C2  = 100
              m  = 100
          delta  = 0.05
    unit relvar  = 2
              k  = 1
         budget  = 1e+05, 5e+05, 1e+06
              n  = 5, 45, 95
             CV  = 0.0693, 0.0377, 0.0346
```

∎

If the sample is three stage, there is some flexibility in how many SSUs and elements per SSU to sample. When the PSU sample is fixed, the term B^2/m in Eq. (9.18) is fixed. The values of $\bar{n}$ and $\bar{\bar{q}}$ can be adjusted to achieve either a budget constraint or a CV target. In either case, the optimal value of $\bar{\bar{q}}$ is

$$\bar{\bar{q}}_{opt} = \sqrt{\frac{1 - \delta_2}{\delta_2} \frac{C_2}{C_3}}.$$

If the PSU sample is fixed and the budget is given by Eq. (9.27), then the implied number of SSUs per PSU is

$$\bar{n} = \frac{C'}{C_2 + C_3 \bar{\bar{q}}} \text{ with } C' = m^{-1} (C - C_0) - C_1 = C_2 \bar{n} + C_3 \bar{n} \bar{\bar{q}}. \qquad (9.35)$$

If a target coefficient of variation, CV_0, is set, then the number of SSUs is

$$\bar{n} = \frac{k_2}{\bar{\bar{q}}} [1 + \delta_2 (\bar{q} - 1)] \left(\frac{m}{\tilde{V}} CV_0^2 - k_1 \delta_1 \right)^{-1}. \qquad (9.36)$$

Notice that Eq. (9.36) involves a subtraction in the denominator. Thus, there is no guarantee that the computed $\bar{n}$ is positive. If Eq. (9.36) produces a negative number, this is an obvious signal that the target CV cannot be achieved with the fixed PSU sample. The R function clusOpt3fixedPSU will compute the optimum numbers of sample SSUs and elements in a three-stage sample when the PSU sample is fixed. The function takes as input the fixed number of PSUs m as well as the parameters defined for clusOpt3 in Example 9.8.

Example 9.10 (Number of SSUs and elements per SSU for a fixed set of PSUs and fixed total cost). Suppose that an existing area sample contains 100 PSUs and that the cost per PSU is \$500. The survey has a total budget for variable costs of $C - C_0 = \$500,000$. The unit costs of having SSUs and persons in the sample are $C_2 = 100$ and $C_3 = 120$. This implies that we have \$500,000 – \$500*100 = \$450,000 to cover the cost of sampling within PSUs. The optimal number of SSUs and persons are found with the function clusOpt3fixedPSU, assuming that the unit relvariance is 1 and that the measures of homogeneity are $\delta_1 = 0.01$ and $\delta_2 = 0.05$:

```
clusOpt3fixedPSU(unit.cost=c(500, 100, 120), m=100,
        delta1=0.01, delta2=0.05, unit.rv=1,
        k1=1, k2=1, tot.cost=500000, cal.sw=1)

                C1  = 500
                C2  = 100
                C3  = 120
```

```
              m = 100
         delta1 = 0.01
         delta2 = 0.05
    unit relvar = 1
             k1 = 1
             k2 = 1
variable budget = 450000
     total cost = 5e+05
              n = 7.8
              q = 4
             CV = 0.0217
```

Thus, the numbers of SSUs per PSU and persons per SSU are 7.8 and 4. Now, suppose that a target CV of 0.05 is set. Other parameters are the same, but the unit relvariance is 4. In this case, we call the function with cal.sw=2. The number of SSUs per PSU is 5.5 and the number of sample persons per SSU is 4:

```
clusOpt3fixedPSU(unit.cost=c(500, 100, 120), m=100,
      delta1=0.01, delta2=0.05, unit.rv=4, k1=1, k2=1,
      CV0=0.05, cal.sw=2)

             C1 = 500
             C2 = 100
             C3 = 120
              m = 100
         delta1 = 0.01
         delta2 = 0.05
    unit relvar = 4
             k1 = 1
             k2 = 1
variable budget = 317617.8
     total cost = 367618
              n = 5.5
              q = 4
             CV = 0.05
       CV check = 0.05
```

In this case the CV target can be achieved for a total cost of about \$368 thousand. (In the output, CV check is a calculation of the CV from the relvariance formula using the optimal sample sizes. This is done to verify that the computed optima yield the target CV.) ∎

Finally, before moving to estimation of the ingredients for the sample size formulas, we note that anticipated sample losses should be accounted for just as they were in Chap. 6. For example, if the response rate for elements is expected to be 60 %, then the number of sample elements computed

from functions like `clusOpt3` and `fixedPSUclusOpt3fixedPSU` should be inflated by $1/0.60$. Depending on the application, the number of sample PSUs or SSUs may also have to be inflated if those units can be lost to ineligibility, nonresponse, or some other reason.

9.4 Estimating Measures of Homogeneity and Variance Components

The parameters in the preceding variance formulas are never known exactly. Either a survey designer must guess their values based on experience or must estimate them from prior, similar surveys. In this section, we cover several ways of estimating the variance components needed for sample allocation.

9.4.1 Two-Stage Sampling

Expressions (9.5) and (9.9) suggest a quick way of estimating the measure of homogeneity δ in a two-stage sample. Suppose that $v\left(\hat{t}_\pi\right)$ is an estimate of the variance of the π-estimator appropriate for the sample design that was used. (In the case of *ppswr* sampling at the first stage, we use $v\left(\hat{t}_{pwr}\right)$.) There are several alternative ways of doing this, which we cover in Chap. 15. Dividing $v\left(\hat{t}_\pi\right)$ by an estimate of the variance of a π-estimator from a simple random sample of the same size gives the design effect, $deff(\hat{t}_\pi) = v(\hat{t}_\pi)/v_{srs}$. Setting $deff(\hat{t}_\pi)$ equal to $1 + \delta\left(\bar{n} - 1\right)$ and solving for δ gives

$$
\begin{aligned}
\hat{\delta} &= \frac{1}{\bar{n} - 1}\left[\frac{1}{k}\frac{v\left(\hat{t}_\pi\right)}{v_{SRS}} - 1\right] \\
&= \frac{k^{-1}deff\left(\hat{t}_\pi\right) - 1}{\bar{n} - 1}.
\end{aligned}
\tag{9.37}
$$

In most cluster samples, the design effect will be greater than 1. In that circumstance, $\hat{\delta} > 0$ is expected, for elements within a cluster are somewhat alike. In Eq. (9.37) it is important to use a variance estimate appropriate for the design that was used to select the sample. Since expressions (9.5) and (9.9) have the same general form, Eq. (9.37) offers a rough estimate of the measure of homogeneity whether PSUs are selected with equal probability or varying probabilities.

The estimates of δ are sensitive to the values of $\bar{n}$ as illustrated in Table 9.2, which uses $k = 1$. For example, if the design effect is 1.6, $k = 1$, and $\bar{n} = 20$, then $\hat{\delta} = 0.032$. But, if $\bar{n} = 5$, $\hat{\delta}$ is much higher at 0.15.

Direct estimates of the components can also be made in some special cases. Särndal et al. (1992, Result 4.3.1, p. 137) give general formulas for estimates

Table 9.2: Indirect estimates of δ based on design effects and average number of sample elements per cluster.

$\bar{n} = 20$		$\bar{n} = 5$	
deff	$\hat{\delta}$	*deff*	$\hat{\delta}$
1.1	0.005	1.1	0.025
1.2	0.011	1.2	0.050
1.3	0.016	1.3	0.075
1.4	0.021	1.4	0.100
1.5	0.026	1.5	0.125
1.6	0.032	1.6	0.150
1.7	0.037	1.7	0.175
1.8	0.042	1.8	0.200
1.9	0.047	1.9	0.225
2.0	0.053	2.0	0.250

of variance components in two-stage designs. We can specialize these to the case of *ppswr* sampling of PSUs and simple random sampling of elements within PSUs in which case $V\left(\hat{t}_{pwr}\right) = V_{PSU} + V_{SSU}$ as shown in Eq. (9.13). The estimators of V_{PSU} and V_{SSU} are

$$v_{SSU} = \sum_{i \in s} \frac{\hat{V}_i}{\left(\pi_i^*\right)^2}$$

$$v_{PSU} = \frac{1}{m\left(m-1\right)} \sum_{i \in s} \left(\frac{\hat{t}_{i\pi}}{p_i} - \hat{t}_{pwr}\right)^2 - \sum_{i \in s} \frac{1 - \pi_i^*}{\left(\pi_i^*\right)^2} \hat{V}_i$$

with $\hat{V}_i = \dfrac{N_i^2}{n_i} \left(1 - f_i\right) \hat{S}_{2i}^2$

where $\pi_i^* = m p_i$, $\hat{S}_{2i}^2 = (n_i - 1) \sum_{k \in s_i} (y_k - \bar{y}_{si})^2$, and $\bar{y}_{si} = \sum_{k \in s_i} y_k / n_i$. (Notice that π_i^* is not the selection probability of PSU i in with-replacement sampling, but will be close to it if p_i is small.) The first term in v_{PSU} is an estimator of the variance of $\hat{t}_{pwr}$ and is called the *ultimate cluster* variance estimator. We cover this estimator in more detail in Chap. 15. If w_k is the full sample weight for element k, the first component of v_{PSU} can also be written as

$$\frac{1}{m\left(m-1\right)} \sum_{i \in s} \left(\frac{\hat{t}_{i\pi}}{p_i} - \hat{t}_{pwr}\right)^2 = \frac{m}{m-1} \sum_{i \in s} \left(\sum_{k \in s_i} w_k y_k - m^{-1} \sum_{i \in s} \sum_{k \in s_i} w_k y_k\right)^2.$$

Software packages often use the ultimate cluster estimator since it requires only the full sample weights. In the case where the same number of elements, $n_i = \bar{n}$, is sampled in each PSU, we can factor out $\bar{n}$ in v_{SSU}. The corresponding estimators of B^2 and W^2 in Eqs. (9.10) and (9.11) are then

$$\hat{B}^2 = \frac{1}{\hat{t}_{pwr}^2} \left\{ \frac{1}{(m-1)} \sum_{i \in s} \left(\frac{\hat{t}_{i\pi}}{p_i} - \hat{t}_{pwr} \right)^2 - \sum_{i \in s} \frac{1 - \pi_i^*}{mp_i^2} \hat{V}_i \right\}, \qquad (9.38)$$

$$W^2 = \frac{1}{\hat{t}_{pwr}^2} \sum_{i \in s} \frac{N_i^2 \hat{S}_{2i}}{mp_i^2}. \qquad (9.39)$$

For the estimator of W^2 we assume that the second-stage sampling fraction, $\bar{n}/N_i$, is small in every PSU. The implied estimator of δ is then

$$\hat{\delta} = \hat{B}^2 \Big/ \left(\hat{B}^2 + \hat{W}^2 \right).$$

A practical difficulty with the estimator for B^2 is that it involves a subtraction. There is no guarantee that $\hat{B}^2$ will be positive. This is similar to the well-known problem with analysis of variance (ANOVA) estimators of variance components. If $\hat{B}^2$ is negative, this is probably evidence that component is small. An option that may be less prone to this defect is to compute the anticipated variance of the estimated total, as described later in Sect. 9.4.3.

Example 9.11 (Variance component estimates in two-stage samples). The function, BW2stagePPSe, will estimate variance components using Eqs. (9.38) and (9.39) for a design in which PSUs are selected with *ppswr* and SSUs with *srswor*. The code below selects a two-stage sample from the Maryland population and then does the calculation for the variable y1. The sampling package is used to systematically select a cluster sample of 20 tracts with probabilities proportional to the count of persons in each tract (Ni below). Notice that this selection of PSUs is without replacement, but we use the standard practice of applying a with-replacement variance estimator. The function, cluster, returns all units in the sample clusters with the cluster selection probability stored in the field Prob. The function, rename, from the reshape package (Wickham 2011) renames Prob to be pi1. Then, the sample tracts are treated as strata, and an *srswor* of $\bar{n} = 50$ persons is selected from each tract. The conditional selection probability of persons within tracts is renamed from Prob to pi2:

```
require(sampling)
require(reshape)         # has function that allows
                         # renaming variables
Ni <- table(MDarea.pop$TRACT)
m <- 20
probi <- m*Ni / sum(Ni)

    # select sample of clusters
set.seed(-780087528)
```

```
sam <- cluster(data=MDarea.pop, clustername="TRACT",
               size=m, method="systematic",
               pik=probi, description=TRUE)

    # extract data for the sample clusters
samclus <- getdata(MDarea.pop, sam)
samclus <- rename(samclus, c(Prob = "pi1"))

    # treat sample clusters as strata and select
    # srswor from each
s <- strata(data = as.data.frame(samclus),
            stratanames = "TRACT",
            size = rep(50,m), method="srswor")
    # extracts the observed data
samdat <- getdata(samclus,s)
samdat <- rename(samdat, c(Prob = "pi2"))

    # extract pop counts for PSUs in sample
pick <- names(Ni) %in% sort(unique(samdat$TRACT))
Ni.sam <- Ni[pick]
d1 <- Ni.sam / sum(Ni)
wt <- 1/samdat$pi1/samdat$pi2

BW2stagePPSe(Ni = Ni.sam, ni = rep(50,20), X=samdat$y1,
             psuID = samdat$TRACT, w = wt,
             m = 20, pp = pp)
```

The function `BW2stagePPSe` accepts seven parameters:

Ni	number of elements in pop in PSU i
ni	number of elements in sample in PSU i
X	data vector
psuID	vector PSU IDs (length is same as length of X)
w	full sample weight
m	number of sample PSUs
pp	vector of one-draw PSU selection probabilities (length is same as that of X)

The results for the variables in the Maryland data set are shown below. Tracts are treated as clusters.

Tracts as clusters			
	B^2	W^2	δ
y1	0.0418	1.3934	0.0291
y2	0.0208	1.0416	0.0196
y3	0.0101	0.1028	0.0894
ins.cov	0.0007	0.3051	0.0023
hosp.stay	0.1056	13.9161	0.0075

These estimates compare to the population calculations in Example 9.4 where tracts were used as clusters. The estimates of the between and within variance components above differ noticeably from the population values. This leads to sample estimates of δ that are different in this sample from the population δ's. Variance component estimates are themselves inherently unstable, and it is no surprise that the estimation error is relatively large here. ∎

If variance component estimates are used for planning, they should be scrutinized to decide whether their sizes seem reasonable. Sensitivity analyses of computed sample sizes should be performed to see what the sizes would be for a range of δ's and other design parameters.

9.4.2 Three-Stage Sampling

Direct estimates of the components of Eq. (9.17) can also be made from a sample. The estimates presented below are based on the ones in Hansen et al. (1953b, Chap. 9, Sect. 10) for the case of *ppswr* sampling of m PSUs and simple random sampling of n_i SSUs in PSU i and q_{ij} elements in SSU *ij*. First, define

$\bar{y}_{sij} = \sum_{k \in s_{ij}} y_k \big/ q_{ij}$, the sample mean of elements in SSU *ij*

$\hat{t}_{ij} = Q_{ij} \bar{y}_{sij}$, the estimated total for SSU *ij*

$\hat{t}_{i\pi} = \frac{N_i}{n_i} \sum_{j \in s_i} \hat{t}_{ij}$, the estimated total for PSU i

$\hat{S}^2_{2ai} = (n_i - 1)^{-1} \sum_{j \in s_i} \left(\hat{t}_{ij} - \bar{\hat{t}}_i \right)^2$, the sample variance among estimated SSU totals, where $\bar{\hat{t}}_i = \sum_{j \in s_i} \hat{t}_{ij} \big/ n_i$

$\hat{S}^2_{3ij} = (q_{ij} - 1)^{-1} \sum_{k \in s_{ij}} (y_k - \bar{y}_{sij})^2$, the sample variance among elements in SSU *ij*

$\hat{V}_{3ij} = \frac{Q^2_{ij}}{q_{ij}} \frac{Q_{ij} - q_{ij}}{Q_{ij}} \hat{S}^2_{3ij}$, the estimated variance of the estimated total $\hat{t}_{ij}$ for SSU *ij*

$\hat{S}^2_{2bi} = \frac{1}{n_i} \sum_{j \in s_i} \hat{V}_{3ij}$

$\hat{S}^2_{2i} = \hat{S}^2_{2ai} - \hat{S}^2_{2bi}$

$$\hat{S}_{1a}^2 = \frac{1}{m-1} \sum_{i \in s} \left(\frac{\hat{t}_{i\pi}}{p_i} - \hat{t}_\pi \right)^2$$

$$\hat{S}_{1b}^2 = \frac{1}{m} \sum_{i \in s} \frac{N_i^2}{p_i n_i} \left[(1 - f_{2i}) \hat{S}_{2ai}^2 + f_{2i} S_{2bi}^2 \right] \text{ where } f_{2i} = n_i/N_i$$

$$\hat{S}_1^2 = \hat{S}_{1a}^2 - \hat{S}_{1b}^2$$

As shown in Hansen et al. (1953b), $\hat{S}_1^2$ estimates the finite population parameter $S_{U1(pwr)}^2$ in Eq. (9.8) or (9.17), $\hat{S}_{2i}^2$ estimates S_{2Ui}^2, and $\hat{S}_{3ij}^2$ estimates S_{U3ij}^2. The estimators of V_{TSU}, V_{SSU}, and V_{PSU} are

$$v_{TSU} = \sum_{i \in s} \frac{1}{(mp_i)^2} \frac{N_i^2}{n_i^2} \sum_{j \in s_i} \hat{V}_{3ij}$$

$$v_{SSU} = \sum_{i \in s} \frac{1}{(mp_i)^2} \frac{N_i^2}{n_i} (1 - f_{2i}) \hat{S}_{2i}^2$$

$$v_{PSU} = \hat{S}_1^2 \Big/ m$$

The relvariance of the *ppswr*-estimator is then estimated by

$$\frac{v\left(\hat{t}_{pwr}\right)}{\hat{t}_{pwr}^2} = \frac{1}{\hat{t}_{pwr}^2} \left(v_{PSU} + v_{SSU} + v_{TSU} \right).$$

When the same number of sample SSUs, $\bar{n}$, is selected in each PSU, the same number of sample elements, $\bar{\bar{q}}$, is selected in each SSU, and the sampling fractions of PSUs, SSUs, and elements are all small, the estimated relvariance can be written as

$$\frac{v\left(\hat{t}_{pwr}\right)}{\hat{t}_{pwr}^2} = \frac{\hat{B}^2}{m} + \frac{\hat{W}_2^2}{m\bar{n}} + \frac{\hat{W}_3^2}{m\bar{n}\bar{\bar{q}}},$$

where

$$\hat{B}^2 = \frac{\hat{S}_1^2}{\hat{t}_{pwr}^2},$$

$$\hat{W}_2^2 = \frac{1}{\hat{t}_{pwr}^2} \sum_{i \in s} \frac{N_i^2}{mp_i^2} \hat{S}_{2i},$$

and

$$\hat{W}_3^2 = \frac{1}{\hat{t}_{pwr}^2} \left\{ \sum_{i \in s} \frac{1}{mp_i^2} \frac{N_i^2}{n_i} \sum_{j \in s_i} Q_{ij}^2 \hat{S}_{3ij}^2 \right\}.$$

Each of these estimates the components in Eqs. (9.10), (9.19), and (9.20). Similar to the case for two-stage sampling, $\hat{B}^2$ and $\hat{W}_2^2$ can be negative since both involve a subtraction. Computing the anticipated variance of the estimated total and using model-based estimators of variance components may

remedy this problem, as described in Sect. 9.4.3. Plug-in estimators of the measures of homogeneity are

$\delta_1 = \hat{B}^2/(\hat{B}^2 + \hat{W}^2)$ and $\hat{W}^2 = \frac{1}{\hat{t}_{pwr}^2} \sum_{i \in s} \frac{Q_{ij}^2 \tilde{S}_i^2}{mp_i^2}$,

where $\tilde{S}_i^2 = \left(\sum_{j \in s_i} \sum_{k \in s_{ij}} w_k \right)^{-1} \sum_{j \in s_i} \sum_{k \in s_{ij}} w_k \left(y_k - \hat{\bar{y}}_i \right)^2$,

$\hat{\bar{y}}_i = \sum_{j \in s_i} \sum_{k \in s_{ij}} w_k y_k \Big/ \sum_{j \in s_i} \sum_{k \in s_{ij}} w_k$, and

$\hat{\delta}_2 = \hat{W}_2^2/(\hat{W}_2^2 + \hat{W}_3^2)$.

Example 9.12 (Variance components in a three-stage sample). The function, `BW3stagePPSe`, will estimate variance components for a *ppswr/srs/srs* design. The Maryland population is used in this example with PSUs being tracts; SSUs block groups; and elements persons. The full sets of code, which are lengthy, are in the files `Example 9.12a.R` and `Example 9.12b.R` and are not shown here. The `sampling` package is used to systematically select a cluster sample of 30 tracts with probabilities proportional to the count of persons in each tract. As in Example 9.9, PSUs are selected without replacement, but a with-replacement variance estimator is used. The selected PSUs are treated as strata and a sample of 2 SSUs is selected from each PSU. The selected SSUs are, in turn, treated as strata and an *srswor* of 50 persons selected from each sample SSU. The function `BW3stagePPSe` accepts six parameters:

dat	Data frame for sample elements with PSU and SSU identifiers, weights, and analysis variable(s). The data frame should be sorted in hierarchical order: by PSU and SSU within PSU
	Required names for columns:
	`psuID` = PSU identifier
	`ssuID` = SSU identifier. These must be unique, i.e., numbering should not restart within each PSU
	`w1i` = vector of weights for PSUs
	`w2ij` = vector of weights for SSUs (PSU weight*SSU weight within PSU)
	`w` = full sample weight
v	Name or number of column in dat with variable to be analyzed
Ni	*m*-vector of number of SSUs in the population in the sample PSUs
Qi	*m*-vector of number of elements in the population in the sample PSUs
Qij	Vector of numbers of elements in the population in the sample SSUs
m	Number of sample PSUs

The three-stage sample must be selected outside the function. Given the input values above, `BW3stagePPSe` returns the values of v_{PSU}, v_{SSU}, v_{TSU}, $\hat{B}^2$, $\hat{W}^2$, $\hat{W}_2^2$, $\hat{W}_3^2$, $\hat{\delta}_1$, and $\hat{\delta}_2$. The function call for the variable `y1` is

```
BW3stagePPSe(dat=samdat, v="y1", Ni=Ni.sam, Qi=Qi.sam,
         Qij=Qij.sam, m=30)
```

Consult the file with the code for this example to see how the input values are constructed. Using the field PSU as the first-stage unit, SSU as the second-stage, and persons as the TSUs, part of the output of `BW3stagePPSe` is:

	PSU as first-stage unit, SSU as second-stage unit					
	$\hat{B}^2$	$\hat{W}^2$	$\hat{W}_2^2$	$\hat{W}_3^2$	$\hat{\delta}_1$	$\hat{\delta}_2$
y1	0.0375	1.3660	0.0016	0.0544	0.0265	0.0282
y2	0.0153	0.9481	0.0005	0.0381	0.0157	0.0111
y3	0.0084	0.0961	0.0004	0.0037	0.0804	0.1081
ins.cov	0.0041	0.2709	0.0000	0.0109	0.0147	0.0018
hosp.stay	0.1633	14.5448	0.0026	0.5867	0.0109	0.0034

These estimates compare to the population figures in Example 9.6. The estimated measures of homogeneity are similar to the population values. However, the estimates of V_{SSU} (not shown here) are negative.

The estimates using tract as the first-stage unit and block group (BG) as the second-stage unit are shown below:

	Tract as first-stage unit, BG as second-stage unit					
	$\hat{B}^2$	$\hat{W}^2$	$\hat{W}_2^2$	$\hat{W}_3^2$	$\hat{\delta}_1$	$\hat{\delta}_2$
y1	0.0150	1.4613	0.1835	1.7326	0.0099	0.0952
y2	0.0085	0.9680	0.1914	1.1528	0.0085	0.1418
y3	0.0096	0.0935	0.1891	0.1051	0.0930	0.6426
ins.cov	0.0119	0.2742	0.1659	0.3014	0.0414	0.3547
hosp.stay	0.0855	14.7706	0.1321	16.6331	0.0055	0.0071

These also compare to the population values in Example 9.6. For example, the population values for the measures of homogeneity for `y1` were $\delta_1 = 0.0060$ and $\delta_2 = 0.1284$ and the estimates are 0.0099 and 0.0952, respectively. Although the relative sizes of the population values and the sample estimates are similar, their absolute sizes are noticeably different. This illustrates a point that we have made before—the estimates of variance components are variable and may be distant from the underlying population values in a particular sample. If the estimates of δ_1 and δ_2 are used to determine sample sizes, do a sensitivity analysis. Compute sample sizes for a range of values around $\hat{\delta}_1$ and $\hat{\delta}_2$. ∎

Discussion of variance component estimation in a three-stage sample can also be found in Särndal et al. (1992, p. 149). They derive the design variance of the π-estimator in three-stage sampling for a general, probability

sample design. Where convenient, we will refer to the Särndal, Swensson, and Wretman book as SSW. The SSW formulas are quite general but require knowledge of joint selection probabilities at each stage. Exercise 9.12 asks you to specialize their results for the theoretical design variance to the case of *srswor* at each stage. There is some potential for confusion when comparing the HHM and SSW results. HHM assume that the PSUs are selected with probabilities proportional to size and *with replacement*. They then use a *ppswr* variance estimator for the PSU variance component. On the other hand, SSW present variance component estimators for the π-estimator of a total (not a *pwr*-estimator). Consequently, the HHM estimators discussed here are not the same as those in SSW. We feel that the HHM formulation is closer to standard practice in the way the PSU sample is handled and will often be more computationally feasible.

9.4.3 Using Anticipated Variances

The formulas in the previous sections for estimation of variance components are specialized and somewhat complex. Being able to use the many software routines that are available for variance component estimation would be a real advantage. Design-based variance component estimators found in, e.g., Särndal et al. (1992) can be negative, depending on the configuration of the data. Using anticipated variances permits the variance of the *pwr*-estimator to be written in terms of model variance components. The model components can be estimated using algorithms that can avoid the numerical problems that the basic design-based, analysis of variance formulas have. Searle et al. (1992) review the methods that are available, including minimum variance quadratic unbiased estimation (MIVQUE0), maximum likelihood, and restricted maximum likelihood (REML). The use of anticipated variances will also clarify the key role that PSU and SSU sizes have in determining measures of homogeneity. However, integrating model variance components needs to be done with care as we show in this section.

To incorporate a variance component model, we use an anticipated variance (Isaki and Fuller 1982) defined as

$$AV\left(\hat{t}\right) = E_M \left\{ E_\pi \left[\left(\hat{t} - t_U \right)^2 \right] \right\} - \left[E_M \left\{ E_\pi \left(\hat{t} - t_U \right) \right\} \right]^2.$$

If the estimator is design unbiased or approximately so, i.e., $E_\pi(\hat{t}) \doteq t_U$, then the AV is $AV\left(\hat{t}\right) = E_M \left[var_\pi \left(\hat{t} - t_U \right) \right]$. Thus, the model expectation of a formula like Eq. (9.8) can be computed, giving model variance components that can be estimated using standard software.

In a clustered population, the simplest model to consider is one with common mean, μ, and random effects for clusters, α_i, and elements, ε_{ij}:

$$y_k = \mu + \alpha_i + \varepsilon_{ik}, k \in U_i, \tag{9.40}$$

with $\alpha_i \sim \left(0, \sigma_\alpha^2\right)$, $\varepsilon_{ik} \sim \left(0, \sigma_\varepsilon^2\right)$, and the errors being independent. The model expectation of the design variance can be computed under this model, but for sample size calculation, only the approximate expectations of B^2 and W^2 for two-stage sampling are needed. In this section, we only consider the variance components in *srs/srs* sampling for a two-stage design. Similar calculations can be done for a *ppswr/srs* design. After some algebra, the model expectations of S_{U1}^2 and S_{U2i}^2 from Eq. (9.2) are (see Exercise 9.16)

$$E_M\left(S_{U1}^2\right) = \left(\sigma_\alpha^2 + \mu^2\right) S_N^2 + \bar{N}^2 \sigma_\alpha^2 + \sigma_\varepsilon^2,$$

$$E_M\left(S_{U2i}^2\right) = \sigma_\varepsilon^2,$$

where $\bar{N} = \sum_{i \in U} N_i / M$ is the average number of elements per cluster, $S_N^2 = \sum_{i \in U} \left(N_i - \bar{N}\right)^2 \big/ (M-1)$, and M is assumed to be large. The anticipated measure of homogeneity is then

$$E_M\left(\delta\right) \doteq \frac{(\sigma_\alpha^2 + \mu^2)\nu_N^2 + \sigma_\alpha^2 + \sigma_\varepsilon^2/\bar{N}^2}{(\sigma_\alpha^2 + \mu^2)\nu_N^2 + \sigma_\alpha^2 + \sigma_\varepsilon^2 \left[1 + (\nu_N^2 + \bar{N}^{-2})\right]}, \tag{9.41}$$

where $\nu_N^2 = S_N^2 / \bar{N}^2$ is the relvariance of the PSU sizes. If $N_i = \bar{N}$, i.e., all the clusters are the same size, then $\nu_N^2 = 0$. In that case, if $\bar{N}$ is large,

$$E_M\left(\delta\right) \doteq \frac{\sigma_\alpha^2 + \sigma_\varepsilon^2/\bar{N}}{\sigma_\alpha^2 + \sigma_\varepsilon^2 \left(1 + 1/\bar{N}^2\right)} \doteq \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\varepsilon^2}. \tag{9.42}$$

Expression (9.42) is the correlation under model (9.40) of any two elements in the same cluster. If the model holds for the population and a sample is selected from it, non-survey software can be used to estimate the variance components, as shown in the next example.

Example 9.13 (Anticipated variance components from a model in two-stage sampling). The R package lme4 (Bates et al. 2012) will estimate the variance components for model (9.40) and for models that are more elaborate. Using the full Maryland population as in Example 9.4, the code to fit the model using the PSU and SSU variables as clusters is

```
require(lme4)
m.y1a <- lmer(y1 ~ (1 | PSU), data = MDarea.pop)
m.y1b <- lmer(y1 ~ (1 | SSU), data = MDarea.pop)
tt <- summary(m.y1a)
```

Part of the summary for `m.y1a` is

```
Random effects:
 Groups    Name            Variance Std.Dev.
 PSU       (Intercept)    36.801   6.0664
 Residual                 7072.180 84.0963

Number of obs: 403997, groups: PSU, 80
```

The variance component estimates are in the object `tt@REmat`. The function `lmer` returns what are known as S4 class objects. The `@` syntax is used to access components. You can see the names of all components of `tt` using `slotNames(tt)`. The estimate of the model correlation in Eq. (9.42) can be computed as

```
vmat <- data.frame(tt@REmat)
vc <- as.numeric(as.character(vmat[,3]))
delta <- vc[1] / sum(vc)
```

The results for all variables using PSUs, SSUs, tracts, and block groups as clusters are shown below. The estimates for δ when PSUs and SSUs are clusters are almost the same as in Example 9.2 where *srs* is used at each stage. But, when tracts and BGs are clusters, the δ's here are much different from those in Example 9.2. As seen in Eq. (9.42), the design-based formula for $B^2 / (B^2 + W^2)$ will estimate the same thing as the model-based calculation if the clusters have the same size but not otherwise. Thus, the big differences we see between Example 9.2 and this example for tracts and BGs are due to the highly varying sizes of those units in the Maryland population. Using the formula for the anticipated δ in Eq. (9.41) in the lower bank of the table below yields values much closer to those in Example 9.2.

	Values of model correlation			
Variable	PSUs as clusters	SSUs as clusters	Tracts as clusters	Tract/block groups as clusters
y1	0.0052	0.0240	0.0082	0.0117
y2	0.0066	0.0157	0.0129	0.0172
y3	0.0786	0.1608	0.1476	0.1906
ins.cov	0.0044	0.0114	0.0076	0.0144
hosp.stay	0.0012	0.0033	0.0016	0.0032
	Values of δ from expression (9.41)			
y1	0.0082	0.0117	0.1306	0.1561
y2	0.0139	0.0172	0.1790	0.2115
y3	0.1476	0.1906	0.6908	0.7464
ins.cov	0.0076	0.0144	0.4454	0.4970
hosp.stay	0.0016	0.0032	0.0173	0.0222

∎

In a population where three-stage sampling is appropriate, the simplest model to consider is one with common mean (μ) and random effects for PSUs (α_i), SSUs (β_{ij}), and elements (ε_{ijk}):

$$y_k = \mu + \alpha_i + \beta_{ij} + \varepsilon_{ijk}, k \in U_{ij}, \tag{9.43}$$

with $\alpha_i \sim \left(0, \sigma_\alpha^2\right)$, $\beta_{ij} \sim \left(0, \sigma_\beta^2\right)$, $\varepsilon_{ijk} \sim \left(0, \sigma_\varepsilon^2\right)$, and the errors being independent. Below, we consider only the *srs/srs/srs* design. Similar calculations can be done for the *ppswr/srs/srs* design. In expression (9.16) and the following, the model expectations of B^2, W^2, W_2^2, and W_3^2 can be found as follows (see Exercise 9.17):

$$E_M\left(t_U^2 B^2\right) = M\left[M\sigma_\alpha^2 \bar{Q}^2(\nu_Q^2+1) + \sigma_\beta^2 \sum_{i \in U} N_i \bar{Q}_i^2 \left(\nu_{Qi}^2+1\right) + \sigma_\varepsilon^2 M \bar{N}\bar{\bar{Q}}\right] + M^2\mu^2\bar{Q}^2\left[\nu_Q^2+1\right], \tag{9.44}$$

$$E_M\left(t_U^2 W^2\right) = M\left(\sigma_\beta^2 + \sigma_\varepsilon^2\right)\bar{Q}^2\left(\nu_Q^2+1\right), \tag{9.45}$$

$$E_M\left(t_U^2 W_2^2\right) = M\left(\sigma_\alpha^2 + \mu^2\right)\sum_{i \in U} N_i^2 \bar{Q}_i^2 \nu_{Qi}^2 + M\sigma_\beta^2 \sum_{i \in U} N_i^2 \bar{Q}_i^2\left(\nu_{Qi}^2+1\right) + M\sigma_\varepsilon^2 \sum_{i \in U} N_i Q_i, \tag{9.46}$$

$$E_M\left(t_U^2 W_3^2\right) = M\sigma_\varepsilon^2 \sum_{i \in U} N_i^2 \bar{Q}_i^2\left(\nu_{Qi}^2+1\right), \tag{9.47}$$

where $\nu_Q^2 = S_Q^2/\bar{Q}^2$ is the relvariance of PSU sizes Q_i, $S_Q^2 = \sum_{i \in U}\left(Q_i - \bar{Q}\right)^2 / (M-1)$, $\bar{Q} = Q/M$; $\nu_{Qi}^2 = S_{Qi}^2/\bar{Q}_i^2$ is the relvariance of the SSU sizes Q_{ij}, $S_{Qi}^2 = \sum_{j \in U_i}\left(Q_{ij} - \bar{Q}_i\right)^2 / (N_i - 1)$, and $\bar{Q}_i = Q_i/N_i$.

Expressions (9.44) and (9.45) can be used to evaluate $E_M(\delta_1) \doteq E_M\left(B^2\right) / \left[E_M\left(B^2\right) + E_M\left(W^2\right)\right]$. Note that these expectations depend on the variances of both Q_i and Q_{ij}. Suppose that all SSUs have the same number of elements, $Q_{ij} = \bar{\bar{Q}}$, and that all PSUs contain the same number of SSUs, $N_i = \bar{N}$. These restrictions imply that $S_{Qi}^2 = S_Q^2 = 0$ and $Q_i = \bar{N}\bar{\bar{Q}}$. In that case, the approximate model expectation of δ_1 is

$$E_M\left(\delta_1\right) \doteq \frac{\sigma_\alpha^2 + \frac{\sigma_\beta^2}{M\bar{N}} + \frac{\sigma_\varepsilon^2}{M\bar{N}\bar{\bar{Q}}}}{\sigma_\alpha^2 + \sigma_\beta^2\left(1 + \frac{1}{M\bar{N}}\right) + \sigma_\varepsilon^2\left(1 + \frac{1}{M\bar{N}\bar{\bar{Q}}}\right)} \doteq \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\beta^2 + \sigma_\varepsilon^2} \tag{9.48}$$

assuming that $M\bar{N}$ and $M\bar{N}\bar{\bar{Q}}$ are large. This is the model correlation of two elements that are in the same SSU, but the reduction in Eq. (9.48) occurs only when the PSUs and SSUs all have the same sizes.

Expressions (9.46) and (9.47) can be used to evaluate $E_M(\delta_2) \doteq E_M(W_2^2)/[E_M(W_2^2) + E_M(W_3^2)]$. In the special case of equal-sized PSUs and SSUs ($Q_{ij} = \bar{\bar{Q}}$ and $N_i = \bar{N}$), the approximate expectation of δ_2 is

$$E_M(\delta_2) \doteq \frac{\sigma_\beta^2 + \frac{\sigma_\varepsilon^2}{M\bar{N}\bar{\bar{Q}}}}{\sigma_\beta^2 + \sigma_\varepsilon^2\left(1 + \frac{1}{M\bar{N}\bar{\bar{Q}}}\right)} \doteq \frac{\sigma_\beta^2}{\sigma_\beta^2 + \sigma_\varepsilon^2}. \tag{9.49}$$

Note that Eq. (9.49) is not the model correlation of two elements in the same SSU, which would be $(\sigma_\alpha^2 + \sigma_\beta^2)/\left(\sigma_\alpha^2 + \sigma_\beta^2 + \sigma_\varepsilon^2\right)$. However, the output from standard variance component estimation software can be used to evaluate Eqs. (9.48) and (9.49). These will be appropriate estimates of δ_1 and δ_2, but only when all PSUs and all SSUs have the same sizes. Otherwise, the variance components from standard routines can be ingredients to the evaluation of Eqs. (9.44)–(9.47).

Example 9.14 (Anticipated variance components from a model in three-stage sampling). Using the full Maryland population, we computed the anticipated measures of homogeneity using PSU/SSU and tracts/BGs as the primary and secondary units. This example gives the results of using expressions (9.48) and (9.49), which are appropriate if each primary unit has the same population number of secondary units and each secondary unit has the same number of elements. We compare these to the results of using expressions (9.44)–(9.47), which account for differing sizes. The R code is in the file `Example 9.14.R`. The results for the same variables as in Example 9.13 are listed below.

Variable	PSUs, SSUs		Tracts, BGs	
	δ_1	δ_2	δ_1	δ_2
Computed from Eqs. (9.48) and (9.49)				
y1	0.0005	0.0235	0.0040	0.0078
y2	0.0044	0.0114	0.0089	0.0092
y3	0.0579	0.1097	0.1116	0.1002
ins.cov	0.0027	0.0088	0.0017	0.0128
hosp.stay	0.0006	0.0027	0.0004	0.0028
Computed from Eqs. (9.44)–(9.47)				
y1	0.0053	0.0245	0.1282	0.1130
y2	0.0067	0.0124	0.1769	0.1545
y3	0.0762	0.1105	0.6822	0.6650
ins.cov	0.0046	0.0098	0.4415	0.4044
hosp.stay	0.0013	0.0037	0.0171	0.0164

When the PSU and SSU variables are used as the first- and second-stage units, the values of δ_1 and δ_2 are almost the same as in Example 9.5 where an $srs/srs/srs$ was assumed. This is true when either Eqs. (9.48) and (9.49)

or Eqs. (9.44)–(9.47) are used to evaluate δ_1 and δ_2. When tracts and BGs are used for first- and second-stage units, the correspondence to the Example 9.5 results is not close at all when Eqs. (9.48) and (9.49) are used. This is due to the fact that the assumptions do not hold well that the number of SSUs, $\bar{N}$, in each tract and the number of elements, $\bar{\bar{Q}}$, in each BG are constants. On the other hand, when Eqs. (9.44)–(9.47), which account for varying sizes of units, are used, the measures of homogeneity are very similar to the values in Example 9.5. ∎

Examples in the literature of using model-based variance component estimates in survey design seem limited, even though practitioners often use the technique. A few examples are Chromy and Myers (2001), Hunter et al. (2005), Judkins and Van de Kerckhove (2003), Valliant et al. (2003), and Waksberg et al. (1993). How to arrive at component formulas using anticipated variances seems to rarely be explained in the literature.

The lme4 package in R is the successor to the earlier nlme (Pinheiro and Bates 2000). We have encountered some examples where a variance component is somewhat close to zero and lmer will not find the correct answer directly. In any case where lmer returns a zero variance component, it is advisable to call the algorithm with a number of random starting values and select the solution with the largest AIC (Akaike information criterion) or log-likelihood. Another option is to use the lme function in the nlme package, which does not seem to be so susceptible to this problem.

Informative Sampling and Variance Component Estimation

Biases of variance component estimators are affected by whether sampling is *informative* or *non-informative*. The idea of informativeness applies to estimation of model parameters. For example, suppose that the random effects model in Eq. (9.40) holds for the population. A sample is non-informative when the same model holds for both the sample and the population. In that case, the sample design can be ignored and unweighted variance component estimators can be used. The R package lme4, the SAS procedure proc mixed, and the xtmixed routine in Stata will provide the unweighted estimates. The weighted estimators that we covered in this section will also provide approximately model-unbiased estimators of the model parameters, σ_α^2 and σ_ε^2 in two-stage sampling, and σ_α^2, σ_β^2, and σ_ε^2 in three stage assuming that the units used at different stages are all equally sized.

Pure probability samples are non-informative. By "pure" we mean a sample in which some probability mechanism that is completely under control of the sample designer is used to select the sample. If that control is lost, estimation is harder. A sample can be informative if selective nonresponse or measurement error occurs that is out of control of the sample designer. For example, if the probability of response depends on the y variable in the

model and this cannot be corrected through some type of nonresponse adjustment, the sample will be informative. (We cover some of the methods used to attempt to correct for nonresponse in Chaps. 13 and 14.)

However, even in a pure probability sample, some features of a sample design may need to be considered when fitting a variance component model (or any other type of model). For example, in a stratified sample, different models may be appropriate for the different strata. This could be described either as "accounting for the design" or "using the appropriate model."

Pfeffermann et al. (1998) and Korn and Graubard (2003) address the problem of estimating variance components from survey samples. The weighted variance component estimates from earlier in this section can be biased when the sample is informative. Korn and Graubard illustrate the biases with some artificial examples and propose some alternative estimators. The alternatives may not be feasible in many survey data sets because they require various conditional weights that may not be available. However, they provide an example from a real survey in which some practical work-arounds appear to be have some advantages over the types of estimators we covered earlier. We will not deal with these alternatives here, although they may be worth considering for some applications.

9.5 Stratification of PSUs

In most designs, PSUs are stratified. The reasons for stratification are the same as those covered in Chap. 3, Sect. 3.1.2, which we recapitulate briefly here. Stratification is, in general, a good way to restrict the distribution of the sample. By selecting a sample of PSUs from each stratum, some maldistributed samples are eliminated. Separate estimates may be needed for some or all strata. For example, in a household survey, regions of the country may be strata or regions crossed with population density (urban, suburban, rural). In a school survey, the PSUs might be schools and the elements, students within the schools. Strata may be based on grade levels of a school, which are usually related to age of the children.

There may also be administrative reasons for stratifying PSUs. In a school survey in one region within a state, it may be necessary to contact the superintendent of each district in order to get permission to survey schools and students. Assuming that the number of districts is limited, the schools might be stratified by district to control the number that must be contacted for permission to collect data.

Other considerations are the number of strata and the allocation of PSUs to strata. If estimates are needed for certain strata, that may determine the number that is created. If strata are mainly created to restrict the distribution of the sample PSUs, then the same techniques can be invoked as in Chap. 3. If PSUs are to be selected with probabilities proportional to a measure of

size (MOS), strata can be created to have approximately equal totals of the MOS or of some power of the MOS as in Example 3.13.

In area samples, the number of sample PSUs is determined and enough strata are usually created so that either 1 or 2 PSUs are selected in each stratum. Selecting one PSU per stratum allows a large amount of control over the achieved distribution of the sample, but does create some variance estimation problems. We will address these in Chap. 15.

Another important consideration in some survey designs is having flexibility to expand or contract the PSU sample. If the survey is longitudinal, the budget may not be the same for every round of the survey. If the budget is cut, the easiest way to reduce costs may be to drop entire PSUs from the sample. This may also be reasonably efficient statistically if the between PSU component of variance is small. In a 2-PSU per stratum design, one PSU can be randomly deleted from the sample in some strata to achieve the reduction. In a one-PSU per stratum design, the strata should be paired in advance for variance estimation, as discussed in Chap. 15. One PSU could be randomly dropped from one or more pairs to reduce the sample.

Having a preset path for expansion of the PSU sample is also useful when the sample must be accumulated over time to make estimates. In the NHANES, extensive physical examinations are given to survey respondents. Mobile examination centers (MECs) carrying diagnostic equipment are ferried from one PSU to another. Moving the MECs is time consuming and expensive and only a subset of the full national PSU sample can be done each year. Two or more years of sample must be accumulated to make reliable national estimates.

9.6 Identifying Certainties

In probability proportional to size sampling, the sizes of some PSUs may be so large that they would be selected with probability 1. These PSUs would be designated as certainties. Sometimes the rule is relaxed so that any PSU that would be selected with probability greater than, say, 0.80 is made a certainty. In area samples, PSUs are often selected with probabilities proportional to their population sizes. Extremely large metropolitan areas will usually be certainties. However, there is some flexibility in how PSUs are defined. Different types of geographic areas (e.g., metropolitan statistical areas, counties, tracts, or block groups) can be used as PSUs. We cover this issue further in Chap. 10.

Exercises

9.1. Using the Maryland population, plot the PSU totals (y-axis) versus PSU population counts. Do this for PSUs defined as (a) tracts and (b) block groups. Do the plots for the five variables in the Maryland data set: `y1`, `y2`, `y3`, `ins.cov`, and `hosp.stay`. Discuss whether simple random sampling of PSUs or *pps* sampling will be more efficient. Explain your reasoning.

9.2. Suppose that a sample of tracts, block groups, and persons is to be selected from the Maryland population to estimate the proportion of persons with some characteristic. Assume that the proportion of the population having the characteristic is 0.32 and that the values of δ_1 and δ_2 in Eq. (9.22) are the same as those for the insurance coverage variable. (b) Compute the coefficient of variation that you would anticipate from a sample of 20 PSUs, 2 SSUs per PSU, and 10 persons per sample SSU.

9.3. Evaluate $\bar{n}_{opt}$ and $\bar{m}_{opt}$ from formulas (9.24) and (9.25) and the following combinations of parameters: $\delta = (0.01, 0.10, 0.20, 0.40)$; $C_1 = 100$, $C_2 = 200$, 400, 600. Assume that the total budget for variable costs is \$275,000. Discuss the results.

9.4. Suppose that a two-stage sample is selected and the π-estimator of the total is used for a series of analysis variables. The average number of sample elements per cluster is 23. What are approximate estimates of the measure of homogeneity for design effects equal to 1.1, 1.2, 1.3, ..., 2.7, 2.8, 2.9, and 3.0? How do your answers change if $\bar{n} = 13$?

9.5. Explore the effects of different sizes of δ_1 and δ_2 on the allocation of a three-stage sample with a total budget of \$500,000 and cost components $(C_1, C_2, C_3) = (1000, 200, 120)$. Assume that the π-estimator is used, that the number of sample PSUs is m, the same number of SSUs, that $\bar{n}$ is allocated to each PSU, and that $\bar{\bar{q}}$ elements are selected from each SSU. Calculate the optimum values of m, $\bar{n}$, and $\bar{\bar{q}}$ for all combinations of $\delta_1 = (0.001, 0.01, 0.05)$ and $\delta_2 = (0.05, 0.10, 0.25)$. Compute the anticipated CVs for each combination assuming that the unit relvariance of the analysis variable is 2.

9.6. Repeat the calculations in Example 9.11 for two-stage sampling using block groups as PSUs in the Maryland population. Use set.seed(-780087528) in R. Select 20 BGs with probabilities proportional to number of persons per tract and 50 persons per BG using *srswor*. Compare your results to those in Example 9.9 where tracts were used as PSUs.

9.7. Use the full Maryland population and the function `BW3stagePPS` to answer the following:

(a) Compute $B^2, W^2, W_2^2, W_3^2, \delta_1$, and δ_2 for the variables Hispanic, Gender, and Age. Recode Hispanic and Gender so that they are (0,1) variables.

Treat Age as continuous for this exercise (even though it is coded into 23-ordered categories). Do the calculations assuming that three-stage sampling will be used with tracts as PSUs, block groups within tracts as SSUs, and persons as elements. The sample at all three stages will be selected using *srswr*.
(b) Repeat the calculations for a design in which PSUs are selected via *ppswr* rather than *srswr*.
(c) Discuss the differences in results. In particular, comment on why the values of δ_1 are different in the two designs.

9.8. Use the Maryland population and the function `BW3stagePPSe` to compute variance components from a sample of 30 PSUs (tracts), 2 SSUs (block groups) per tract, and 50 persons per sample SSU. Assume that tracts are selected with probabilities proportional to the number of persons in the tract and that SSUs and persons are selected via *srs*. Use `set.seed(1696803792)` in R.

(a) Do the computation for the variables `y2`, `y3`, `ins.cov`, and `hosp.stay`.
(b) How do your answers compare to the full population results in Example 9.12?
(c) Use the estimated values of δ_1 and δ_2 to compute the optimum values of m, $\bar{n}$, and $\bar{\bar{q}}$ in a three-stage sample where $C_1 = 500$, $C_2 = 100$, $C_3 = 120$, and the total budget for variable costs is \$100,000. How can you estimate the unit relvariance for each variable?
(d) Discuss your results in (c). Is the same allocation optimal for each of the five variables? Which allocation would you use in practice?

9.9. Use the Labor force population to compute between and within variance components and the measure of homogeneity, δ in a two-stage sample for the variables, `HoursPerWk` and `WklyWage`. The variable cluster defines the first-stage units.

(a) Do the calculation using the function `BW2stageSRS` and `BW2stagePPS`. How do the answers compare? What are the assumptions for the sample designs in these functions?
(b) Repeat the calculations using `lmer` in the `lme4` R package. Which results do you expect the `lmer` results to be closest to—`BW2stageSRS` or `BW2stagePPS`? (Several random starts should be used for `lmer` and the best (minimum AIC) solution selected; the function will not give correct estimates without searching.)

9.10. Consider a population that is divided into M clusters, each of which has $\bar{N}$ elements as in Example 9.1. Show that when both M and $\bar{N}$ are large, the unit relvariance of a variable y can be written as $S_U^2/\bar{y}_U^2 \doteq B^2 + W^2$. All terms are defined in Example 9.1. Use the form $W^2 = \frac{1}{M\bar{y}_U^2}\sum_{i \in U} S_{U2i}^2$ to derive the result.

9.11. Show that $V\left(\hat{t}_\pi\right)/t_U^2 = \frac{1}{m}\frac{M-m}{M}B^2 + \frac{1}{m\bar{n}}\frac{\bar{N}-\bar{n}}{N}W^2$ in Eq. (9.3) can be rewritten as $\tilde{V}k\left[1 + \delta\left(\bar{n}-1\right)\right]\Big/m\bar{n}$, i.e., equal to Eq. (9.5) with $k = (B^2 + W^2)/\tilde{V}$. You will need to make the substitution, $(M-m)/M \approx (M-1)/M$, to get the result.

9.12. Särndal et al. (1992, p. 149) derive the design variance of the π-estimator in three-stage sampling for a general sample design. Suppose that U is the population of PSUs; U_{IIi} is the population of SSUs within PSU i; U_{ij} is the population of elements within PSU/SSU ij; π_{Ii} is the selection probability of PSU i in the first stage; $\pi_{Iii'}$ is the joint selection probability of PSUs i and i'; $\pi_{IIj|i}$ is the conditional selection probability of SSU j given that PSU i is selected; $\pi_{IIjj'|i}$ is the joint conditional probability that SSUs j and j' are selected within PSU i; $\pi_{k|ij}$ is the conditional selection probability of element k within PSU/SSU ij; and $\pi_{kk'|ij}$ is the joint selection probability of elements k and k' within PSU/SSU ij. The variance of the π-estimator is then $V\left(\hat{t}_\pi\right) = V_{PSU} + V_{SSU} + V_{TSU}$ where

$V_{PSU} = \sum_{i\in U}\sum_{i'\in U}\Delta_{Iii'}\frac{t_i}{\pi_{Ii}}\frac{t_{i'}}{\pi_{Ii'}}$ with t_i being the population total of the analysis variable for PSU i and $\Delta_{Iii'} = \pi_{Iii'} - \pi_{Ii}\pi_{Ii'}$

$V_{SSU} = \sum_{i\in U}V_{IIi}/\pi_{Ii}$ with $V_{IIi} = \sum_{j\in U_i}\sum_{j'\in U_i}\Delta_{IIjj'|i}\frac{t_{ij}}{\pi_{IIj|i}}\frac{t_{ij'}}{\pi_{IIj'|i}}$, $\Delta_{IIjj'|i} = \pi_{IIjj'|i} - \pi_{IIj|i}\pi_{IIj'|i}$, and t_{ij} being the population total for PSU/SUU ij

$V_{TSU} = \sum_{i\in U}\frac{1}{\pi_{Ii}}\sum_{U_{IIi}}\frac{V_{ij}}{\pi_{IIj|i}}$ with $V_{ij} = \sum_{k\in U_{ij}}\sum_{k'\in U_{ij}}\Delta_{IIkk'|ij}\frac{y_k}{\pi_{k|ij}}\frac{y_{k'}}{\pi_{k'|ij}}$, $\Delta_{IIkk'|ij} = \pi_{kk'|ij} - \pi_{k|ij}\pi_{k'|ij}$

(a) Specialize this formula to the case of simple random sampling at each stage. In particular, suppose that m PSUs are selected from M using *srswor*. In PSU i suppose that n_i SSUs are selected from N_i in PSU i and that q_{ij} elements are selected from Q_{ij} in PSU/SSU ij. That is, show that
$V_{PSU} = \frac{M-m}{M}\frac{M^2}{m}S_{U1}^2$ with $S_{U1}^2 = \sum_{i\in U}\left(t_i - \bar{t}_U\right)^2\Big/(M-1)$ where $\bar{t}_U = \sum_{i\in U}t_i/M$

$V_{SSU} = \frac{M}{m}\sum_{i\in U}\frac{N_i^2}{n_i}\frac{N_i-n_i}{N_i}S_{U2i}^2$ with $S_{U2i}^2 = \frac{1}{N_i-1}\sum_{j\in U_i}\left(t_{ij} - \bar{t}_{Ui}\right)^2$ is the unit variance of SSU totals in PSU i with $t_{ij} = \sum_{k\in U_{ij}}y_k$ being the population total for PSU/SSU ij, $\bar{t}_{Ui} = \sum_{j\in U_i}t_{ij}\Big/N_i$ is the average total per SSU in PSU i

$V_{TSU} = \frac{M}{m}\sum_{i\in U}\frac{N_i}{n_i}\sum_{j\in U_i}\frac{Q_{ij}^2}{q_{ij}}\frac{Q_{ij}-q_{ij}}{Q_{ij}}S_{U3ij}^2$ with $S_{U3ij}^2 = \frac{1}{Q_{ij}-1}\sum_{k\in U_{ij}}\left(y_k - \bar{y}_{Uij}\right)^2$

(b) Show that, if $n_i = \bar{n}$ and $q_{ij} = \bar{\bar{q}}$, i.e., the same number of sample SSUs is selected from each sample PSU, the same number of sample elements is

selected from each SSU, and the number of SSUs is $\bar{N}$ in every PSU and the number of elements in each SSU is $\bar{\bar{Q}}$, then the relvariance of $\hat{t}_\pi$ can be written as

$$\frac{V(\hat{t}_\pi)}{t_U^2} = \frac{M-m}{M}\frac{B^2}{m} + \frac{\bar{N}-\bar{n}}{\bar{N}}\frac{W_2^2}{m\bar{n}} + \frac{\bar{\bar{Q}}-\bar{\bar{q}}}{\bar{\bar{Q}}}\frac{W_3^2}{m\bar{n}\bar{\bar{q}}} \text{ with}$$

$$B^2 = M^2 S_{U1}^2 / t_U^2, \ W_2^2 = M\bar{N}^2 \sum_{i\in U} S_{U2i}^2 / t_U^2, \text{ and}$$

$$W_3^2 = M\bar{N}\bar{\bar{Q}}^2 \sum_{i\in U}\sum_{j\in U_i} S_{U3ij}^2 \Big/ t_U^2.$$

9.13. Suppose that a simple random sample of m PSUs and $\bar{n}$ elements is selected per sample PSU. Assume that the cost of the survey can be modeled as $C = C_0 + C_1 m + C_2 m\bar{n}$ and that the relvariance of the π-estimator is $\frac{V(\hat{t}_\pi)}{t_U^2} \doteq \frac{B^2}{m} + \frac{W^2}{m\bar{n}} = \frac{1}{m\bar{n}}\tilde{V}\left[1 + \delta\left(\bar{n}-1\right)\right].$

(a) Show that the number of sample elements per PSU that minimizes the relvariance is

$$\bar{n}_{opt} = \sqrt{\frac{C_1}{C_2}\frac{W^2}{B^2}} \doteq \sqrt{\frac{C_1}{C_2}\frac{1-\delta}{\delta}}.$$

(b) Use the total cost constraint to show that $m_{opt} = \frac{C-C_0}{C_1+C_2\bar{n}_{opt}}$. (Hint: Use a Lagrange function defined as $\phi = V\left(\hat{t}_\pi\right)\big/t_U^2 + \lambda\left(C - C_0 - C_1 n - C_2 m\bar{n}\right)$

9.14. Consider the situation in a two-stage sample where the PSU sample is fixed.

(a) Show that, if the total cost, $C = C_0 + C_1 m + C_2 m\bar{n}$, is fixed, then the number of elements to be sampled per PSU is $\bar{n} = \frac{C-C_0-C_1 m}{C_2 m}$.
(b) If a target CV is set, then the number of elements to sample per PSU is $\bar{n} = \frac{1-\delta}{CV_0^2 m/\tilde{V}-\delta}$.

9.15. (a) In a three-stage sample where the set of PSUs is fixed show that if either the budget is fixed or a target CV is set, the optimal number of elements to sample is $\bar{\bar{q}} = \sqrt{\frac{1-\delta_2}{\delta_2}\frac{C_2}{C_3}}$.
(b) If the budget is fixed, show that the optimal number of SSUs per PSU is $\bar{n} = \frac{C'}{C_2+C_3\bar{\bar{q}}}$ with $C' = m^{-1}\left(C-C_0\right) - C_1 = C_2\bar{n} + C_3\bar{n}\bar{\bar{q}}$.
(c) If a target coefficient of CV_0 is set, then the number of SSUs is $\bar{n} = \frac{1}{\bar{\bar{q}}}\left[1 + \delta_2\left(\bar{\bar{q}}-1\right)\right]\left(\frac{m}{\tilde{V}}CV_0^2 - \delta_1\right)^{-1}.$

9.16. In a clustered population, consider this model with common mean and random effects for clusters and elements:

$$y_k = \mu + \alpha_i + \varepsilon_{ik}, k \in U_i,$$

with $\alpha_i \sim \left(0, \sigma_\alpha^2\right)$, $\varepsilon_{ik} \sim \left(0, \sigma_\varepsilon^2\right)$, and the errors being independent. Define $S_{U1}^2 = \sum_{i\in U}\left(t_i - \bar{t}_U\right)^2/(M-1)$ as for the case of *srswr* sampling of clusters

and $S^2_{U2i} = \frac{\sum_{k \in U_i}(y_k - \bar{y}_{Ui})^2}{N_i - 1}$ as for simple random sampling of elements within sample clusters. Other terms are defined in Sects. 9.2.1 and 9.2.2.

(a) Show that under the model above

$$E_M\left(S^2_{U1}\right) \doteq \left(\sigma^2_\alpha + \mu^2\right)S^2_N + \bar{N}^2\sigma^2_\alpha + \sigma^2_\varepsilon,$$

$$E_M\left(S^2_{U2}\right) = \sigma^2_\varepsilon,$$

where $\bar{N} = \sum_{i \in U} N_i/M$ is the average number of elements per cluster, $S^2_N = \sum_{i \in U}\left(N_i - \bar{N}\right)^2 / (M-1)$, and M is assumed to be large.

(b) If $N_i = \bar{N}$, then $E_M\left(S^2_{U1}\right) \doteq \bar{N}^2\sigma^2_\alpha + \sigma^2_\varepsilon$, $E_M\left(B^2\right) \doteq \left(\bar{N}^2\sigma^2_\alpha + \sigma^2_\varepsilon\right) / (N\mu)^2$, $E_M\left(W^2\right) \doteq \sigma^2_\varepsilon/\mu^2$, and that

$$E_M\left(\delta\right) \doteq \frac{\sigma^2_\alpha + \sigma^2_\varepsilon/\bar{N}}{\sigma^2_\alpha + \sigma^2_\varepsilon\left(1 + 1/\bar{N}^2\right)} \doteq \frac{\sigma^2_\alpha}{\sigma^2_\alpha + \sigma^2_\varepsilon}.$$

9.17. In a population where three-stage sampling is appropriate, consider this model with common mean and random effects for PSUs, SSUs, and elements:

$$y_k = \mu + \alpha_i + \beta_{ij} + \varepsilon_{ijk}, k \in U_{ij},$$

with $\alpha_i \sim \left(0, \sigma^2_\alpha\right)$, $\beta_{ij} \sim \left(0, \sigma^2_\beta\right)$, $\varepsilon_{ijk} \sim \left(0, \sigma^2_\varepsilon\right)$, and the errors being independent. Using the formulas for B^2, W^2_2, W^2_3 below Eq. (9.16) and the formula for W^2 defined below expression (9.21) with $p_i = 1/M$, verify that their model expectations are given by Eqs. (9.44)–(9.47). Use these to show that if the number of SSUs in every PSU is $\bar{N}$ and that the number of elements in each SSU is $\bar{\bar{Q}}$, then the approximate expectations of δ_1 and δ_2 are

$$E_M\left(\delta_1\right) \doteq \frac{\sigma^2_\alpha}{\sigma^2_\alpha + \sigma^2_\beta + \sigma^2_\varepsilon},$$

$$E_M\left(\delta_2\right) \doteq \frac{\sigma^2_\beta}{\sigma^2_\beta + \sigma^2_\varepsilon}.$$

l Show that $E_M\left(\delta_2\right)$ is not the model correlation of two elements in the same SSU, which would be $\left(\sigma^2_\alpha + \sigma^2_\beta\right) / \left(\sigma^2_\alpha + \sigma^2_\beta + \sigma^2_\varepsilon\right)$.

Chapter 10
Area Sampling

Area sampling is a catchall term for a set of procedures in which geographic areas are selected as intermediate units on the way to sampling lower-level units that are the targets of a survey. Area sampling is just an example of multistage sampling, but because special data sources and methods are used, we devote a separate chapter to it. Calculations for determining sample allocations to the different stages are the same as those covered in Chap. 9.

There are several reasons that multistage sampling is used. One is that clustering can reduce costs if field listing is needed or in-person interviews are conducted. Having the sample units clustered in fairly small geographic areas allows data collectors to be hired in a limited number of areas and reduces travel costs. Another reason is that a complete list of the target units in the survey may not be available. By sampling small areas, a list can be compiled in the field and used for sampling. In some surveys, like school samples, permission to collect data may have to be obtained from a high-level administrative unit, like a school district. In that case, sampling districts is a way of limiting the number of organizational units that have to be negotiated with. A major application of area sampling is in household surveys where data are collected by personal interview. In the USA a complete list of persons and households is not maintained by either the government or private organizations. Even if one were available, an unclustered sample would be extremely inefficient for personal interviewing because the area of the country is so large. Area sampling is certainly not limited to household sampling. Other target populations where area sampling may be efficient are business establishments, schools, bodies of water, and the like—any population requiring that data be collected where the units are physically located.

The description of area sampling presented in this chapter is primarily US-centric. We concentrate on the types of geographic areas that have been developed by the US Census Bureau, primarily for household surveys (Sect. 10.1). However, the general techniques are applicable to other countries where various levels of geographic areas have been defined for administrative and statistical purposes. Therefore, we include a few non-US examples for comparison.

Population counts, demographic distributions, and detailed estimates are summarized within the various geographic areas for use in constructing the (multistage) area sample design in lieu of a population registry. These data are obtained through various sources including the US Census (a census, mandated through the US Constitution to be conducted every 10 years, of the population residing in the 50 states, the District of Columbia, and Puerto Rico) and a large household survey known as the American community survey or just ACS (Sect. 10.2). Because the counts and other information used in the various stages of sampling (Sect. 10.3) are time sensitive, we include a discussion of procedures to address shifts in the population distribution after the initial sample of units has been drawn (Sect. 10.6). In addition to the ACS, design details of a few example surveys are discussed including the sampling frame and stages of the design (Sect. 10.4).

Many multistage surveys, including one of our example studies, are designed to meet sample size and other criteria for several domains simultaneously within the lowest stage of sampling. Unlike an *stsrs* where strata can be designed to reflect the domains, multistage surveys sometimes rely on *pps* sampling with *composite size measures* to accomplish the design goals while keeping cost in check (Sect. 10.5).

Finally, area sampling has many benefits and some drawbacks. For example, it is important to have timely and accurate population information prior to selecting the multistage sample. However, migration and/or length of time since the last census introduce differences between the frame data or estimates and what can be found "in the field." Techniques implemented to address these population shifts are discussed in Sect. 10.6. Another less than desirable trait for area samples is the amount of time and funds required to develop and select units at lower sampling stages. A relatively new type of sampling methodology, known as *address-based sampling* (ABS), is reviewed as a remedy for surveys with limited resources (Sect. 10.7).

10.1 Census Geographic Units

The US Census Bureau uses several layers of geographic areas for its survey operations. These are also in common use by private survey organizations. The areas consist of existing administrative divisions and other units constructed for statistical use. Figure 10.1 shows the hierarchy of the areas.

From the largest to the smallest in terms of population size and geographic area, the hierarchy of areas is state, metropolitan area, county, census tract, block group, and block. In some parts of the US, terms other than county, like parish or minor civil division, are used to denote local government jurisdictions that are equivalent to counties, but we do not need to be concerned with that here.
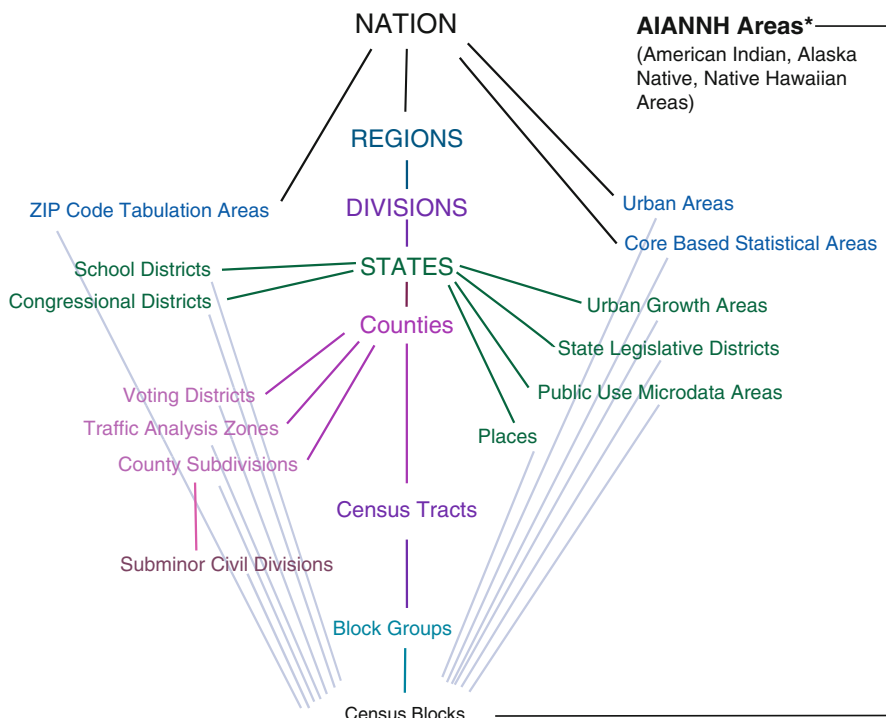
Fig. 10.1: Geographic hierarchy of units defined by the US Census Bureau. See U.S. Census Bureau (2011).

Metropolitan areas are defined by office of management and budget (OMB), a US federal agency. Part of OMB's job is to provide consistent definitions for collecting, tabulating, and publishing federal statistics for a set of geographic areas. Four of the larger areas defined by OMB are:

Metropolitan statistical area (MSA)—contains at least one urbanized area of at least 50,000 people, plus any adjacent territory that has a high degree of social and economic integration with the core as measured by commuting ties. There were 374 MSAs in 2009 (366 in the US and 8 in Puerto Rico), just prior to the 2010 US Census (OMB Bulletin No. 10-02).[1] Commuting is part of the definition because for work some people may travel a considerable distance into a central city, thus, tying an area together. Approximately 84 % of the US population resides within an MSA.

Metropolitan division—a county or group of counties within a MSA that has a population core of at least 2.5 million.

---

[1] http://www.census.gov/population/www/metroareas/metrodef.html.

Micropolitan statistical areas—an area containing one or more urban clusters of at least 10,000 but less than 50,000 population, plus adjacent territory.

Combined statistical area—adjacent metropolitan and micropolitan statistical areas; combinations are based on commuting ties.

There are 3,141 counties in the US, a map of which for the entire US can be found at

```
ftp2.census.gov/geo/maps/general_ref/us_base/
   stco2003/stco2003.pdf.
```

Choropleth maps of the US with counties marked by percent of population in poverty and median household income in 2008 are at

www.census.gov/did/www/saipe/data/statecounty/maps/2008.html.

Census tracts, blocks, and block groups are the units most often used in sampling within primary sampling units (PSUs) for household surveys. Tracts are small, statistical subdivisions of a county or equivalent entity. Tracts generally have between 1,500 and 8,000 people, with a desired size of 4,000 people. Counties and equivalent entities with fewer than 1,500 people have a single census tract. Tracts do not cross state boundaries. The first decennial census for which the entire United States was covered by census tracts was in 2000.

Census blocks are areas bounded on all sides by visible features, such as streets, roads, streams, and railroad tracks, and by invisible boundaries, such as city, town, township, and county limits, property lines, and short, imaginary extensions of streets and roads. Blocks are usually small in area but in sparsely settled areas may contain many square miles of territory. All territory in the 50 United States, the District of Columbia, Puerto Rico, and the Island Areas governed by the US has been assigned block numbers.

A block group (BG) is a cluster of census blocks. BGs generally contain between 600 and 3,000 people, with a target size of 1,500 people. BGs on American Indian reservations, off-reservation trust lands, and special places must contain a minimum of 300 people. Special places include correctional institutions, military installations, college campuses, worker's dormitories, hospitals, nursing homes, and group homes. Such special places are also called group quarters. There are typically three BGs per tract. The counts of the various areas for 2010 census were:

Counties	3,141
Census tracts	74,002
Block groups	217,740
Blocks	11,078,297

Tallies by state of the number of tracts, block groups, and blocks used in the 2000 Census can be found at www.census.gov/geo/www/2010census/. The number of counties and other administrative divisions by state are at www.census.gov/geo/www/tallies/ctytally.html. Boundary files for areas are available at www.census.gov/geo/www/cob/index.html in what is known as the TIGER (topologically integrated geographic encoding and referencing) database. These are a collection of cartographic database files that are available to the public and are used in a variety of commercial geographic information system (GIS) or mapping software products. The boundary files define geographic areas using polygons with sides based on longitude and latitude coordinates.

10.2 Census Data and American Community Survey Data

In the US, extensive demographic information has traditionally been collected on a large sample of persons as part of each decennial census. In the 2000 Census, approximately one-sixth of the population living in the US received a "long form." Since then, the long-form sample has been replaced by the ACS, which collects this same information on a continuously updated sample (www.census.gov/acs/www/). The 2010 Census collected the following:

- Address-level items:
  Number of persons living at the address on April 1, 2010
  Tenure: whether the residence was owned or rented
- Items collected for each person:
  Age
  Gender
  Ethnicity (whether the person is Hispanic, Latino, or Spanish origin)
  Race (14 choices are listed, plus a person can fill in an unlisted choice)

Counts of persons for every block in the US are available from the 2010 census. In addition, block-level counts will be available for all of the characteristics listed above.

In the ACS, detailed questions are asked about each person's socioeconomic status and housing unit characteristics, including:

Age	Income
Birthplace	Language spoken at home
Citizenship	Marital status
Highest level of education	Number of rooms in the housing unit
Employment status	Presence of indoor plumbing
Ethnicity	Race
Gender	Time spent traveling to work
Housing value	Year when the housing unit was built

The Census Bureau tabulates sample estimates from the ACS at a variety of geographic levels. The ACS publishes one-year, three-year, and five-year moving averages since the sample in any single month is small. As of 2010, the planned, total ACS sample size was about three million addresses in a year. Estimates at all geographic levels in Fig. 10.1 down to BGs are published for overlapping five-year estimates, e.g., 2006–2010 and 2007–2011. One-year and overlapping three-year estimates are published only for higher levels of geography. The importance of this for sample design is that statistics for the small geographic areas that are often used as sampling units will not refer to a particular point in time but instead will be average values over extended periods of time. This may actually be advantageous for designing a sample since the way in which the population is distributed is always in flux. Population counts from the decennial census become progressively more out-of-date as a decade wears on. Waksberg et al. (1997) analyze the effects of using such out-of-date census information when doing geographic oversampling aimed at improving estimates for small demographic domains. The further removed from the census date a survey is, the less accurate the census counts for small areas are. Consequently, the moving averages from the ACS will give a more current picture of the population.

10.3 Units at Different Stages of Sampling

Multistage samples can use PSUs, secondary sampling units (SSUs), and, in some cases, units at later stages. In area samples, the first two stages are geographic areas with SSUs nested within PSUs. Units at the third or later stages are typically households or persons.

10.3.1 Primary Sampling Units

PSUs in area samples are geographic areas that collectively cover the entire area in-scope of a survey. PSUs are used as the first stage of sampling in an area probability sample. The PSUs are usually stratified by geography to ensure representation of regions or other types of subareas. The number of sample PSUs may be based on rough optimality calculations to account for between PSU variances contributions to simple estimators (like the π-estimator), as described in Chap. 9. Conversely, rules of thumb may be used; 100 PSUs is a common sample size, but some surveys like the current population survey (CPS; the US labor force survey) use hundreds of sample PSUs. The number is mainly affected by whether subnational estimates such as regional or local areas are needed. The sample is usually allocated to strata accounting for the desire to make regional estimates. PSU samples are often used for long stretches of time, e.g., ten years between decennial censuses, and for many different surveys.

Rules for Defining PSUs

There are some general rules that are useful when defining PSUs for the area sampling frame. These are used in many household surveys, like CPS; other types of surveys might use different rules:

1. PSUs are contained within state boundaries. This facilitates tabulations by state.
2. Each PSU is a county or group of counties, except in the New England states where other equivalent areas used.
3. MSAs are sometimes defined as separate PSUs. Exceptions may be made to this rule because some MSAs are too big to be efficient for fieldwork and/or could result in being selected multiple times with some sampling methods.
4. The area of a PSU should not exceed some maximum area (e.g., 3,000 square miles or about 7,770 square kilometers in CPS). This helps limit the distance interviewers must travel.
5. The PSU population must be greater than some minimum (e.g., 7,500 in CPS) as long as rule 4 is not violated. The idea is to allow the sample to be large enough to provide a reasonable workload for interviewers as well as the calculation of efficient estimates within the PSU. For example, a PSU with a small number of sample members might require all to be selected for the survey, introducing unequal weights into the design as well as potentially unstable estimates if 100 % response is not achieved. Another consideration may be to support a longitudinal survey where persons are in the survey for a set number of waves and are then rotated

out and replaced by other sample persons. A PSU that is too small might
not permit the desired number of rotations to be made.

6. Avoid extreme length. Like rule 4, this is designed to limit travel in surveys
   done by personal interview. For example, PSUs designed to be roughly
   square in distance (e.g., 50 square kilometers) are more cost-effective than
   excessively rectangular clusters of the same size (e.g., 5 km by 500 km).

Rule 3 is applied when a metro area is used for tabulation and publication.
For the US consumer price index (CPI), separate indexes are published for
some metro areas.[2] For example, the following MSAs have their own indexes:
Detroit-Ann Arbor-Flint, Michigan; New York-Northern New Jersey-Long
Island, covering parts of the states of New Jersey, New York, and Connecti-
cut; San Francisco-Oakland-San Jose, California; Washington DC-Baltimore,
covering the District of Columbia and parts of Maryland, Virginia, and West
Virginia.

Rule 3 is one that has become less beneficial over time in the US because
of the large geographic areas covered by some MSAs. The Washington-
Baltimore MSA is a case where the area is extremely large, being about
150 miles (240 km) from the northwest to southeast corner. Figure 10.2 is a
map of this MSA. Covering the entire MSA by automobile would involve a
lot of driving by a field-worker doing personal interviews over areas that can
be extremely congested with traffic. If metro areas are of no interest to the
goals of the survey, they do not have to be used as PSUs. For example, in
the commercial building energy consumption survey (CBECS), conducted by
the US Department of Energy, counties are PSUs without regard to MSA
definitions.[3] Climate zones are more important in defining PSUs and strata
in that survey.

10.3.2 Secondary Sampling Units

SSUs are units first selected within each sample PSU. These are also geo-
graphic areas but are much smaller than the typical PSU. SSUs may be
tracts, block groups, or ZIP code (postal delivery) areas. ZIP code areas are
not often used for household sampling in the US because statistics and maps
are readily available for tracts and BGs. However, ZIP codes can be useful
for establishment sampling if they are the smallest areas for which business
data are available. In that case, the ZIP code data can be used for assigning
measures of size to the areas.

Some large-scale area surveys include as many as five stages of sam-
pling. For example, the National Household Survey on Drug Use and Health
(NSDUH), discussed in more detail in Sect. 10.4.2, has a total of four sampling

---

[2] http://www.bls.gov/cpi/.

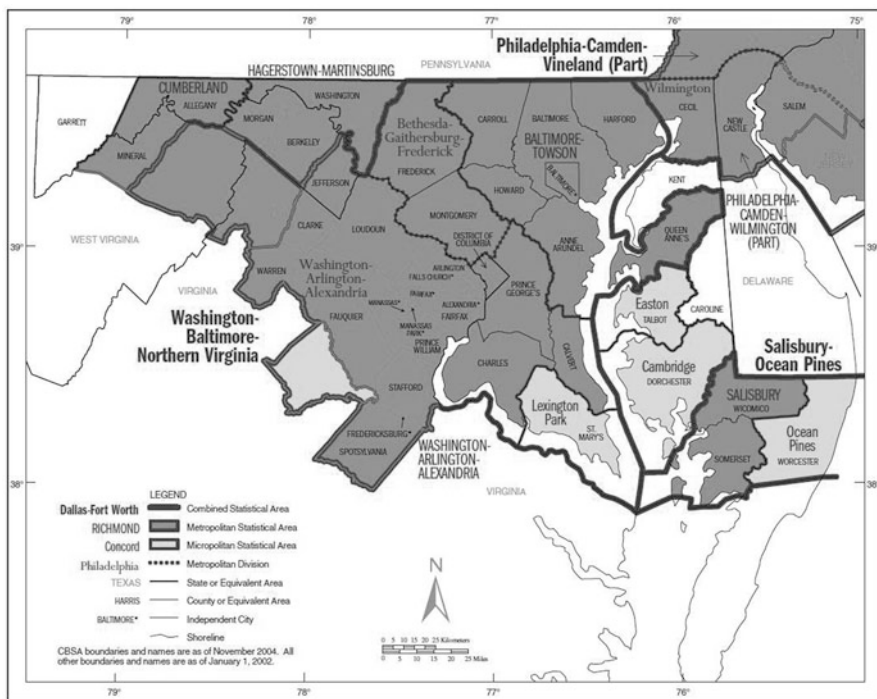[3] http://www.eia.gov/emeu/cbecs/.

Fig. 10.2: A map of the Washington–Baltimore metropolitan statistical area and smaller subdivisions. Source: Census Bureau: Metropolitan and Micropolitan Statistical Areas, http://www.census.gov/population/www/metroareas/metroarea.html.

stages and the NSDUH mental health surveillance study (NSDUH-MHSS) has five. Typically, the third and subsequent stages of sampling just prior to the last stage do not have a special label and are identified only by the stage number, i.e., fourth stage of sampling.

SSUs and stages prior to the ultimate sampling stage (Sect. 10.3.3) are usually used when lists of the units eligible for the survey are not available in advance and field listings have to be made. Most US organizations, for example, do not have a complete list of households from which to sample, and these frames historically have been constructed by field staff canvassing a randomly selected area. This procedure is typically referred to as *counting and listing* and the field staff as *field (household) listers*; see, e.g., Eckman and O'Muircheartaigh (2011) and Wright and Marsden (2010, chap. 2). With the NSDUH, for example, area segments (SSUs) are constructed to have at least either 150 urban dwelling units or 100 rural dwelling units. Trained listers are sent into the field with maps and recording material (global positioning system, GPS, recorders, or hard-copy rosters) to obtain identifying information for all dwelling units contained within the established boundaries of the randomly chosen area segments. Once compiled, this list is used as the household

sampling frame and again to isolate the chosen household for the survey. The information collected includes physical location (e.g., street address; cross street and a description of the house in lieu of a physical address; location on the property of another address; GPS coordinates) and other paradata (e.g., identification of toys or equipment associated with young children; likelihood that a house is currently occupied; graffiti and trash in the neighborhood as a predictor of survey cooperation). Listers may be provided with a partial list of addresses from a prior survey or from administrative records; all are given a starting point on the segment map showing the area boundaries as well as the direction the lister should travel in order to record the dwelling units.

10.3.3 Ultimate Sampling Units

The last stage of sampling (regardless of its number) is of particular importance and is called an *ultimate sampling unit* (USU). Examples of USUs are a small group of housing units (e.g., 4); individual households or persons; business establishments, and buildings. Additional examples specific to a particular survey are examined in the next section. Note that the last stage of sampling may not be equivalent to the smallest unit of analysis. In the case of a household survey such as the National Crime Victimization Survey (NCVS), the household is labeled as the USU because all household members 12 years of age and older are asked to participate in the study (i.e., certainty selections within household), but a person is the unit of analysis.[4] The USU for the NSDUH is a group of persons because zero, one, or two persons are selected from the sample household.[5]

10.4 Examples of Area Probability Samples

To better understand how area samples are implemented in practice, we summarize the designs of some major household surveys in the US and Germany. These designs have similarities, but each has some unique goals and features.

---

[4] The NCVS is conducted by the Bureau of Justice Statistics at the US Department of Justice. http://bjs.ojp.usdoj.gov/index.cfm?ty=dcdetail&iid=245.

[5] No persons may be sampled from households with common characteristics so that study funds can be used to oversample households with more rare traits (e.g., households with children or of a certain minority race/ethnicity).

10.4.1 Current Population Survey

The CPS is the US labor force survey and is the primary source of data to estimate various unemployment rates and characteristics of the labor force.[6] The details of the methodology for the survey are in U.S. Census Bureau (2006). The survey is paid for by the Bureau of Labor Statistics with the sample being selected and the data collected by the Census Bureau. The target population is the civilian, noninstitutionalized population aged 16 years or older residing within the 50 states and the District of Columbia. The survey is designed to produce national and state estimates, and also substate estimates in California (Los Angeles and the rest of state) and in New York state (New York City and the rest of the state).

The Census Bureau maintains a master address file (MAF) that attempts to cover all HUs in the USA. In practice, it is impossible to have a list that is completely up-to-date, but the MAF is closer to current than any other list, including the aforementioned DSF. The MAF is not available to any private survey organization nor to other governmental agencies unless the Census Bureau itself does the sampling and data collection. Having the MAF gives the CPS and other surveys done by Census some options that are not available to other survey organizations.

Using the MAF, efforts are made to coordinate CPS sampling with nine other surveys done by Census; see U.S. Census Bureau (2006, pp. 3–7). The aim is to avoid having households selected for multiple surveys, which would increase burden and probably decrease response rates. The entire sample of SSUs for a decade is selected at once. This makes sample selection more specialized than for many area samples.

The precision targets for the survey are to achieve a 1.9% CV on the national, monthly unemployment rate, assuming that the rate is 6%. Also, a difference of 0.2% points in the national unemployment rate in two consecutive months should be significant at 0.10 level. The goal for each state and substate area and the District of Columbia is to have an 8% CV on the estimates of average annual unemployment rate, assuming that the rate is 6%.

Stages of Sampling: PSUs and Groups of Households

The total national sample size is about 72,000 housing units (HUs), although this number can fluctuate depending on the budget allocated for the survey. Independent samples are selected for each state and substate area for which separate estimates are made. At the first stage, 824 PSUs are selected, which are MSAs or combinations of counties in non-MSA areas. The design has 446 certainty PSUs and 378 non-certainties. One PSU from each non-certainty

---

[6] http://www.census.gov/cps/.

stratum is selected with probability proportional to the most recent census population count.

At the second stage the SSU is a group of four adjacent HUs. (Groups of HUs or groups of geographic blocks are sometimes referred to as a *segment*.) Notice that a group of four HUs is much smaller than the block groups or blocks mentioned earlier. Directly sampling such a small SSU relies on census having the MAF. Since no further subsampling is done, the SSU can also be considered the USU in most PSUs. There are a few exceptions to this. In cases where addresses are "not recognizable on the ground," then area sampling is used to select USUs. Then, a third stage is sometimes used if an SSU is large in area or number of households. This third stage is mainly used in rural areas. Interviews are conducted with all members of the household who are at least 15 years of age.

The CPS also uses a building permit sample within sample PSUs to cover dwellings constructed after the address list was compiled. The method used to compile the permit frame is similar to option 1 in Sect. 10.6.

Formation of PSUs

PSUs are formed using the rules in Sect. 10.3 with some adaptations specific to CPS. MSAs are used for PSUs, except that PSUs do not cross state boundaries. When an MSA crosses state boundaries, which is common in the eastern part of the US, the MSA is split into two or more PSUs. The minimum population size of a PSU is 7,500 except where this would require creating a PSU having an area of 3,000 square miles. After the 2000 census, a total of 2,025 PSUs were created from the 3,141 counties.

Stratification and Selection of PSUs

PSUs are stratified within state. The key variables used for stratification are number of males that are unemployed, number of females unemployed, number of families with female head of household, and ratio of occupied HUs with three or more people, of all ages, to total occupied housing units. Strata are created to contain PSUs that are similar to each other on these variables, using a clustering algorithm. Any PSU that is part of the 151 largest metropolitan areas is a certainty. This, in conjunction with the *CV* requirement for state estimates, leads to nine entire states being certainty PSUs. These are geographically small but densely populated: Connecticut, Delaware, Hawaii, Massachusetts, New Hampshire, New Jersey, Rhode Island, Vermont, and the District of Columbia.

Strata of non-self-representing PSUs are formed to have about the same total population. One PSU is selected from each stratum with probability proportional to the total population. Each PSU is constructed to supply a

sample of 35–55 HUs, which is large enough to be a workload for one data collector.

Selection of USUs

The frame for selecting USUs in each PSU has four pieces: (1) the housing unit (HU) address list frame which is the MAF, (2) an area frame, which is used where a census address cannot be used to locate an HU, (3) a group quarters (GQ) frame, and (4) a building permit frame. For survey operations HUs and group quarters are defined as:

- **Housing units (HUs)**—a group of rooms or a single room occupied as a separate living quarter (or intended to be). About 98 % of the US population enumerated in a census resides in an HU; the rest are in group quarters or are homeless.
- **Group quarters (GQs)**—residents share common facilities or receive formally authorized care. Examples of group quarters are college dormitories, nursing homes, retirement homes, and communes. Since CPS covers only the non-institutionalized population, institutional group quarters like prisons and military facilities are not in-scope.

Military and institutional GQs are left on frame in case they convert to an in-scope unit before interviewing.

As a brief aside, note that some surveys the term *dwelling unit* (DU) is also used to mean housing unit (HU). Many times a frame of USUs will have on it units that turn out to be more than one DU or HU. For example, the house at 104 Cherry Street may actually contain a family on the first floor and a renter in a basement apartment. Most surveys would classify these are two HUs, even though the sampling frame showed it as one. For this reason some organizations will refer to the address listings on the frame as *lines* rather than HUs or DUs because their status is not fully determined until the time of interview. After the status of the line is determined, screening and multiphase sampling is often used in order to target certain demographic groups, types of establishments, or buildings as described in Chap. 17.

The sample of HUs is designed to have an overlap across time periods. Data are collected monthly. An HU is in the sample for four months, is out of the sample for the next eight months, and then is back in the sample for another four months. HUs are rotated in such a way that 3/4 of the HUs overlap between consecutive months; 1/2 of the HUs overlap between samples 12 months apart. The monthly overlap helps when estimating monthly change while the overlap between samples 12 months apart reduces the variance for an estimate of annual change. Many countries use some form of overlap sampling in their labor force surveys. The 4-8-4 pattern in the CPS is just one of many possibilities. Canada, for example, retains HUs for six months and

then rotates them out. This leads to there being no overlap between samples 12 months apart.

Housing units and USUs are selected to be self-weighting within a state. That is, each HU has the same selection probability. Thus, there is no differential sampling within domains (e.g., gender) defined within each state. Although estimates for domains are published, the sample design does not directly control domain sample sizes.

10.4.2 National Survey on Drug Use and Health

Another large household survey that has some different features from the CPS is the National Survey on Drug Use and Health (NSDUH) in the USA sponsored by the Substance Abuse and Mental Health Services Administration and collected by RTI International.[7] A detailed description of the 2006 sample design is in Morton et al. (2006). The sample is selected in four stages and publishes estimates for a variety of demographic domains. The target population is the civilian, noninstitutionalized population aged 12 years or older residing within the 50 states and the District of Columbia. The survey also covers residents of noninstitutional group quarters (e.g., shelters for the homeless, rooming houses, dormitories, and group homes) and civilians residing on military bases.

The total targeted sample size of 67,500 persons is equally allocated across three age groups: persons aged 12–17, 18–25, and 26 or older. The large sample size allows the survey to obtain enough cases in other major demographic groups to make separate national estimates without oversampling them. Separate estimates are also made for each state.

In the first stage of the design, each state is handled separately and, thus, can be considered a stratum. The PSUs in each state were census tracts. The PSUs were themselves stratified within each state. State sampling (SS) regions were formed. Based on a composite size measure, states were geographically partitioned into regions that had about the same total population. The use of a composite measure of size (MOS) is an interesting technique that we will cover in more depth in Sect. 10.5. The effect of using the composite MOS was that regions could be formed so that each area yielded, in expectation, about the same number of interviews during each data collection period. The smaller states were partitioned into 12 SS regions, while the eight largest states were divided into 48 SS regions. A total of 900 SS regions were formed across all states.

In some cases, small census tracts were combined to obtain a minimum number of dwelling units (DUs). In urban areas the minimum was 150 DUs; in rural areas, it was 100. Within each SS region, 48 PSUs were selected

---

[7] http://oas.samhsa.gov/nsduh.htm.

with probability proportional to the composite MOS, giving a total of 43,200 PSUs. Thus, about two-thirds of the tracts in the USA are in the sample. The 48 SSUs in each state were randomly assigned to six rotation groups of four to be used as primary sample SSUs while the other 24 were a reserve sample to be used if needed to compensate for nonresponse. The 24 primary SSUs were assigned to years and calendar quarters using a simple rotation plan shown in Fig. 10.3. Subsamples 1 and 2 were assigned to year 1 for data collection. Subsample 1 was rotated out in year 2; subsample 2 was retained and subsample 3 was rotated in for year 2; and so on. Other surveys like the CPS use more elaborate rotation plans, but Fig. 10.3 conveys the main idea behind rotation. Looking down a column for a given year, the combination of subsamples across all states for that year must represent the nation. For example, in year 4, estimates for each state can be made from subsamples 4 and 5, and the combination of subsamples 4 and 5 across all states can be used to estimate national statistics.

The SSUs (or segments) in the design are block groups aggregated to meet the same minimum sizes of DUs as for the PSUs, i.e., 150 DUs in urban areas and 100 in rural. One SSU was selected with *pps* in each sample PSU. The MOS for each SSU is a composite MOS based on 2000 census data adjusted to more recent data from a commercial vendor.

The third stage consists of selecting an equal probability sample of DUs within each sample SSU. In most SSUs, field personnel listed all DUs and a sample was selected from the list. As noted in Sect. 10.1, a tract contains a target of about 4,000 people but can have as many as 8,000. For large tracts where it was uneconomical to list all DUs, a rough count was made of DUs on the streets of a tract. Central office personnel then split the SSU into two or more parts and one was selected for full listing. The fourth stage of selection was of persons within a sample HU. The interviewer constructed a roster of all eligible persons in the HU, and persons were selected with different rates depending on their age (12–17, 18–25, and 26 or older). Sampling rates were preset by age group and state. In a given household, 0, 1, or 2 persons were selected using predefined sampling rates for five age groups within state established during the design phase of the project. The roster information was entered directly into an electronic screening instrument, which automatically implemented the fourth stage of selection based on the state and age group rates.

10.4.3 Panel Arbeitsmarkt und Soziale Sicherung

The Panel Arbeitsmarkt und soziale Sicherung (PASS) is a labor force survey in Germany conducted by the Institute for Employment Research, a federal

Subsample	Year 1	2	3	4	5
1	●				
2	●——————●				
3		●——————●			
4			●——————●		
5				●——————●	
6					●

Fig. 10.3: Rotation plan for SSUs in the National Survey on Drug Use and Health.

agency.[8] Its goals are to assess the effects of various unemployment and social assistance programs. The data also allow the examination of questions like (1) which factors help persons move from being unemployed to having a job, (2) which pathways lead people into unemployment, and (3) how does the personal situation (e.g., health, financial, integration in society) change for people who receive such benefits.

PASS is a dual frame, longitudinal survey that provides a good example of the use of administrative records that are available for sampling in some European countries. PASS combines an area probability sample with a sample of benefit recipients from an administrative database. A total of 300 postal code areas (PSUs) are selected with probability proportional to population. Within each sample PSU, two parallel samples are selected—one of recipients of assistance from an administrative record list and a second using an address list that covers all persons. In each PSU a commercial database of addresses provides the frame for the second sample. The commercial database is comparable to the US Delivery Sequence File mentioned later in Sect. 10.7. Various sources are used to construct indicators at the level of buildings for recipiency status, residential mobility, predominant age groups, and type of building. At least five households are aggregated to define a building. These indicators are then appended to the frame of addresses.

From the list of recipients in each PSU, a sample is selected and information collected for the entire household to which a recipient belongs. From the address frame a sample is selected, the number of households at the address determined, and one of the households is sampled. Information for the entire household is then collected.

Being a dual-frame and longitudinal sample, PASS has some special weighting issues that we will not pursue here. One issue is that the address list frame includes the places where recipients live. Thus, a decision needs to be made on whether a recipient should be allowed to enter the sample through both sources or only through the recipient sample. In addition, PASS tracks persons over time to determine how long they stay on assistance programs and what may cause people to move into one of the programs. Tracking leads to

---

[8] http://fdz.iab.de/de/FDZ_Individual_Data/PASS.aspx.

some difficult operational problems since 15 % or more of persons may move their residences in a typical year.

10.5 Composite MOS for Areas

As we noted in Sect. 10.4.2, the NSDUH uses a composite MOS for both PSUs and SSUs. The purposes of the composite MOS are to get:

1. Self-weighting samples from each of several domains
2. Equal workload in each PSU, i.e., same total sample size in each PSU (across all domains)

When all elements are to be sampled at the same rate, the calculation to determine within-PSU sampling rates is simple. Suppose that the desired overall rate is f and the selection probability of PSU i is π_i. To obtain a self-weighting sample, the within-PSU rate must be set so that $\pi_i \pi_{k|i} = f$ where k denotes any element in the PSU. This implies that $\pi_{k|i} = f/\pi_i$. Thus, the within-PSU sampling rate is adjusted, depending on the selection probability of the PSU, but the adjustment does not depend on domain membership. The composite MOS technique refines this to allow self-weighting samples to be selected within different domains while obtaining the same sample size within each PSU across all domains.

10.5.1 Designing the Sample from Scratch

The presentation of the method here is somewhat simplified compared to what is implemented in NSDUH and is based on Folsom et al. (1987). We need the following notation:

N_d = Number of elementary units (i.e., the smallest unit of analysis) in a unique domain d in population, e.g., the number of persons in an age group

$N = \sum_{d=1}^{D} N_d$, the total number of elementary units in the population

n_d = Desired sample size in domain d (values based on precision, power, and budget considerations)

$n = \sum_{d=1}^{D} n_d$, total sample size across all domains

$f_d = n_d/N_d$, desired sampling rate for units in domain d

$N_i(d)$ = Number of elementary units in domain d in PSU i in the population $i \in U$, the universe of PSUs

$\bar{n}$ = Desired sample size in each PSU across all domains. This is the equal workload requirement

m = Number of sample PSUs

Note that the domains must partition the target population into mutually exclusive groups. The age groups used in the 2005 NSDUH composite size measures were 12–17, 18–25, 26–34, 35–49, and 50 years or older where the people within the two youngest categories were oversampled. The composite MOS for PSU i is defined to be

$$S_i = \sum_{d=1}^{D} f_d N_i (d).$$ (10.1)

This is the expected sample size from domain d in the PSU if the desired overall sampling rate for that domain were used. Suppose the PSUs are sampled with probabilities proportional to S_i so that $\Pr(i \in s) = \pi_i = mS_i/S_+$ where $S_+ = \sum_{i \in U} S_i$, the sum of the MOSs across all PSUs. The total MOS can be written as

$$S_+ = \sum_{i \in U} \sum_{d=1}^{D} f_d N_i (d) = \sum_{d=1}^{D} f_d \sum_{i \in U} N_i (d).$$

That is, the sum of the MOSs equals the total desired sample size. Next, set the desired sample size in PSU i and domain d to be

$$n_i^* (d) = \bar{n} f_d \frac{N_i (d)}{S_i}.$$ (10.2)

The within-PSU sampling rate for units in domain d is required to be $\pi_{k|i} (d) = \frac{n_i^* (d)}{N_i (d)} = \frac{\bar{n}}{S_i} f_d$. Thus, the within-PSU rate is a modification of the overall rate. This is possible as long as $f_d \leq S_i/\bar{n}$ for all PSUs. Checking whether any PSU violates this requirement is an important step in quality control (see Chap. 18 for more details on quality checks).

Next, we can check the workload. Using the fact that $S_i = \sum_d f_d N_i (d)$, the expected total number of sample units in PSU i is

$$\sum_d n_i^* (d) = \frac{\bar{n}}{S_i} \sum_d f_d N_i (d).$$

That is, the within-PSU, domain sample sizes sum to the desired workload in each PSU. The sample for a domain is also self-weighting in each domain. Assuming that PSUs are picked $pp(S_i)$, the overall selection probability for a unit k in domain d is

$$\pi_i \pi_{k|i} (d) = m \frac{S_i}{S_+} \frac{n_i^* (d)}{N_i (d)} = m \frac{S_i}{S_+} \frac{\bar{n}}{S_i} f_d = f_d,$$

where we use the fact that $m\bar{n} = n = S_+$.

Example 10.1 (Composite measures of size). Suppose the sampling frame has 10 PSUs and that there are two domains. Table 10.1 lists the population

counts of units by PSU and domain. We want to sample four PSUs and 45 units per sample PSU. The desired domain sampling rates and sample sizes are given in Table 10.2.

Table 10.1: Population counts and composite MOSs for Example 10.1.

PSU	$N_i(d)$ Domain		Total no. of units in PSU	Composite MOS	PSU probability	Within-PSU probability, $\pi_{k|i}(d)$ Domain	
	$d=1$	$d=2$		S_i	π_i	$d=1$	$d=2$
1	50	50	100	17.5	0.389	0.643	0.257
2	50	30	80	15.5	0.344	0.726	0.290
3	50	90	120	21.5	0.478	0.523	0.209
4	50	40	90	16.5	0.367	0.682	0.273
5	50	25	75	15.0	0.333	0.750	0.300
6	70	40	110	21.5	0.478	0.523	0.209
7	50	80	130	20.5	0.456	0.549	0.220
8	50	65	115	19.0	0.422	0.592	0.237
9	50	30	80	15.5	0.344	0.726	0.290
10	50	50	100	17.5	0.389	0.643	0.257
Totals	520	500	1,000	180			

Table 10.2: Desired sampling rates and sample sizes in Example 10.1.

Domain	Sampling rate, f_d	Desired domain sample size, n_d
1	0.25	130
2	0.10	50
Total		180

In a small example like this, a spreadsheet is convenient for doing the calculations. For example, the expected total sample size in domain 1 is 0.25*520 = 130. From (10.1), the composite MOS for PSU 1 is 50*0.25 + 50*0.10 = 17.5; for PSU 6 it is 70*0.25 + 40*0.10 = 21.5. The selection probability for PSU 1 is 4*17.5/180 = 0.389. The within-PSU sampling rate for units in PSU 1 and domain 1 is 45*0.25/17.5 = 0.643 using $\pi_{k|i}(d) = \bar{n}f_d/S_i$. The expected sample size in PSU 1 is 50*0.643 + 50*0.257 = 32.1 + 12.9 = 45, which is the desired workload. Although the PSU-level workload is an integer, notice that the expected sample sizes in each domain are not. Thus, it is important to sample the domain units at the specified *rates*—not by sampling a fixed number of units based on rounded off sample sizes. We can also check that the sample for each domain will be self-weighting. Taking PSU 8 as an

illustration, the selection probability for units in domain 1 is 0.422*0.592 = 0.25; for domain 2, we have 0.422*0.237 = 0.10.

More Than Two Stages of Sampling

The calculations above are for a two-stage sample—PSUs followed by units within domains. If we interpose other stages of sampling between PSUs and elementary units, the selection probabilities still work out. The key requirement is that both PSUs and subareas must be selected with probabilities proportional to the composite MOS. Suppose that the design uses PSUs, SSUs, and elements as the stages of selection. Define:

$Q_{ij}(d)$ = Number of elements in PSU i, SSU j that are in domain d
$Q_{i+}(d)$ = Number of elements in PSU i that are in domain d
$S_{ij} = \sum_d f_d Q_{ij}(d)$ to be the composite MOS for SSU j in PSU i
U_i = Universe of SSUs within PSU i
m = Number of sample PSUs
n = Number of sample SSUs in each PSU
$\bar{\bar{q}}$ = Average number of elements selected and interviewed per SSU (i.e., inflated for sample loss associated with ineligibility and nonresponse)

Assuming that both PSUs and SSUs are sampled *pps* using the composite MOS, the selection probability of SSU ij is $\pi_i \pi_{j|i} = m\frac{S_{i+}}{S_{++}} n\frac{S_{ij}}{S_{i+}} = mn\frac{S_{ij}}{S_{++}}$ where $S_{i+} = \sum_{j \in U_i} S_{ij} = \sum_d f_d Q_{i+}(d)$ and $S_{++} = \sum_{i \in U} S_{i+} = \sum_d f_d Q_{++}(d)$. Note also that $S_{++} = mn\bar{\bar{q}}$, the total sample size. Next, set the expected number to be sampled from domain d in PSU/SSU ij to be $q_{ij}^*(d) = \bar{\bar{q}} f_d Q_{ij}(d)/S_{ij}$ and the sampling rate within PSU/SSU ij to be $\pi_{k|ij}(d) = \bar{\bar{q}} f_d/S_{ij}$. The overall selection probability is then

$$\pi_i \pi_{j|i} \pi_{k|ij}(d) = mn\frac{S_{ij}}{S_{++}} \frac{\bar{\bar{q}}}{S_{ij}} f_d = \frac{mn\bar{\bar{q}}}{S_{++}} f_d = f_d$$

so that the desired overall rate is obtained. The workloads per SSU and PSU are

$$\sum_d q_{ij}^*(d) = \bar{\bar{q}} \frac{1}{S_{ij}} \sum_d f_d Q_{ij}(d) = \bar{\bar{q}}$$

and

$$\sum_d \sum_{j \in s_i} q_{ij}^*(d) = \bar{\bar{q}} \sum_{j \in s_i} \frac{1}{S_{ij}} \sum_d f_d Q_{ij}(d) = n\bar{\bar{q}}.$$

That is, the workload is the same in each SSU and PSU. Note that the population count of elements in SSU ij does not have to be nonzero in every

domain for this to be true. The SSU MOS S_{ij} may be based on only the subset of domains that have elements in the SSU.

As in the case of two-stage sampling, quality control checks are important. For example, determine whether:

1. $q_{ij}^*(d) \leq Q_{ij}(d)$ for every SSU and domain. Since $q_{ij}^*(d) = \bar{\bar{q}} f_d Q_{ij}(d)/S_{ij}$, this is equivalent to $f_d \leq S_{ij}/\bar{\bar{q}}$. Note that if $Q_{ij}(d) = 0$, then $n_{ij}^*(d) = 0$ so that an attempt will not be made to sample something from nothing. But, the algebra may well result in an attempt to sample more from a domain in an SSU than the population can support.
2. $\bar{\bar{q}} \leq Q_{ij}$ in every SSU.
3. $n\bar{\bar{q}} \leq Q_{i+}$ in every PSU.
4. π_i, $\pi_{j|i}$, and $\pi_{k|ij}$ are all less than or equal to 1.

If any of these conditions are violated, then small PSUs or SSUs can be combined with others.

10.5.2 Using the Composite MOS with an Existing PSU Sample

Some survey organizations select a general purpose PSU sample once a decade using new census information and use that sample for ten years (or more). This "every x-year selection approach" was commonplace in the past because designing and selecting a national PSU sample was difficult and time-consuming. Now, some organizations are set up to do this more efficiently. When a PSU sample is used for a number of years, the PSUs will probably not be selected with the composite MOS appropriate for a new survey. SSUs can be selected with the composite MOS that reflects the desired domain subsampling rates. In this situation, only one of these goals can still be met:

1. Select a self-weighting sample for units in each domain, but have different sampling rates for the domains.
2. Obtain a constant workload in each PSU.

Suppose the sample design uses PSUs, SSUs, and elements with SSUs as the stages of sampling. The notation here is the same as above. We will also use

$U_{ij}(d)$ = the universe of elements in domain d in SSU ij.

For the sample to be self-weighting we need $\pi_i \pi_{j|i} \pi_{k|ij}(d) = f_d$, which implies that $\pi_{k|ij}(d) = f_d / (\pi_i \pi_{j|i})$. This conditional sampling rate can be used regardless of how the SSUs are selected as long as $f_d \leq \pi_i \pi_{j|i}$ for every PSU and segment.

The expected sample size in PSU i, i.e., the workload, is

$$\sum_d \sum_{j \in U_i} \pi_{j|i} \sum_{k \in U_{ij}(d)} \pi_{k|ij}(d) = \sum_d \sum_{j \in U_i} \pi_{j|i} \sum_{k \in U_{ij}(d)} \frac{f_d}{\pi_i \pi_{j|i}}.$$

For an equal workload in every PSU, we need this to equal $m\bar{n}$. The expected sample size S_{i+}/π_i depends on the PSU i and, in general, cannot be made a constant. In Sect. 10.5.1, the math worked out to have the same workload in each PSU when the PSUs were selected with $pp\,(S_i)$ because $\pi_i = mS_i/S_+$ and $S_i/\pi_i = S_+/m$ which is a constant. However, the PSU selection probability does not have this special form in all samples.

The variation in workload will depend on how much S_{i+}/π_i varies. In practice, it's desirable to have integer multiples of some minimum workload per PSU. The PSU sample sizes might be set to have enough workload for 1, 2, or 3 interviewers. Typically the workload is set to be large enough for at least two interviewers. Having only one interviewer is chancy because if that interviewer quits or cannot work for some other reason (sick, family reasons, etc.), there is no backup. A new replacement person would have to be hired and trained. Alternatively, an interviewer from another area might travel to the PSU to collect data.

Two Ways to Implement the Design

Suppose that the MOS for SSU j in PSU i is computed as $S'_{ij} = \sum_{d=1}^{D} f_d N_{ij}\,(d)\,/\pi_i$. This is the expected sample size if elements in domain d, SSU i were sampled directly at rate f_d/π_i. The sum of these MOSs is

$$S'_{i+} = \sum_{j\in U_i} S'_{ij}.$$

This is the expected workload in PSU i when elements are sampled at rate f_d/π_i. If the SSUs within PSU i are selected with probabilities proportional to S'_{ij}, then $\pi_{j|i} = mS'_{ij}/S'_{i+}$. Option 1 is to set the sampling rate in SSU ij to be $\pi_{k|ij}\,(d) = f_d/\left(\pi_i\pi_{j|i}\right)$. This is self-weighting since $\pi_i\pi_{j|i}\pi_{k|ij}\,(d) = \pi_i\pi_{j|i}f_d/\left(\pi_i\pi_{j|i}\right) = f_d$. The workload is not constant because

$$\sum_{d}\sum_{j\in U_i}\pi_{j|i}\sum_{k\in U_{ij}(d)}\pi_{k|ij}\,(d) = S_{i+}/\pi_i$$

as shown above.

A second option is to set the sample size in SSU ij to be

$$n^*_{ij}\,(d) = \frac{\bar{n}f_d}{\pi_i S'_{ij}}N_{ij}\,(d) = \frac{\bar{n}f_d}{S_{ij}}N_{ij}\,(d)\,.$$

Assuming an equal probability sample is selected from the elements in SSU ij in domain d, the selection probability is then

$$\pi_{k|ij}\,(d) = \frac{n^*_{ij}\,(d)}{N_{ij}\,(d)} = \frac{\bar{n}f_d}{\pi_i S'_{ij}} = \frac{\bar{n}f_d}{S_{ij}}\,.$$

For this to be feasible, we must have $f_d \leq S_{ij}/\bar{n}$. The sample does not achieve the desired overall sampling rate in each domain because

$$\pi_i \pi_{j|i} \pi_{k|ij} \, (d) = \pi_i \frac{mS'_{ij}}{S'_{i+}} \frac{\bar{n}f_d}{\pi_i S'_{ij}} = \frac{m\bar{n}}{S'_{i+}} f_d \neq f_d. \qquad (10.3)$$

Note that the sampling rate for units in domain d is the same in every SSU in PSU i since (10.3) does not depend on j. The problem is that $S'_{i+} = \sum_{d=1}^{D} f_d N_i \, (d) \big/ \pi_i$ is not a constant in each PSU. However, the workload is constant because

$$\sum_d \sum_{j \in U_i} \pi_{j|i} \sum_{k \in U_{ij}(d)} \pi_{k|ij} \, (d) = \sum_d \sum_{j \in U_i} \frac{mS'_{ij}}{S'_{i+}} \sum_{k \in U_{ij}(d)} \frac{\bar{n}f_d}{\pi_i S'_{ij}}.$$

For a given PSU sample, making calculations using the two options will allow you to weigh the alternatives.

Example 10.2 (Subsampling with an existing PSU sample: obtaining a self-weighting sample). Table 10.3 shows the counts and composite measures of size in two PSUs from a larger PSU sample. Two SSUs are to be selected from each PSU. The first PSU contains four SSUs; the second has five SSUs. The desired rates for domains 1 and 2 are 0.015 and 0.035. The SSU composite MOS is calculated as $S'_{ij} = \sum_{d=1}^{D} f_d N_{ij} \, (d) / \pi_i$. For example, the MOS for SSU 3 in PSU 1 is $(5*0.015 + 90*0.035)/0.1117 = 28.9$. Sampling SSUs with $pp \, (S'_{ij})$ gives $\pi_{j|i} = 0.0.643$ for PSU 1, SSU 3. The within-SSU sampling rate for domain d is $\pi_{k|ij} \, (d) = f_d / \left(\pi_i \pi_{j|i} \right)$. For PSU 1, SSU 3, this is $0.015/0.1117/0.643 = 0.209$ for domain 1. This combination produces a self-weighting sample in each domain. In PSU 1, SSU 3, the selection probability for elements in domain 1 is $\pi_i \pi_{j|i} \pi_{k|ij} \, (1) = 0.1117*0.643*0.209 = 0.015$. For domain 2 in that SSU we have $0.1117*0.643*0.487 = 0.035$. However, the expected sample size is not the same in each PSU. In PSU 1, the expected workload is $2*44.9=89.8$ while it is $2*39.4=78.8$ in PSU 2, which the reader can verify. These are different from the expected workload per PSU based on the overall domain rates: $(0.15*4000 + 0.035*2000)/2 = 65$.

Example 10.3 (Subsampling with an existing PSU sample: obtaining equal workloads). Table 10.4 repeats the counts and composite measures of size from Example 10.2. The SSU composite MOSs are the same as in the preceding example. The desired rates for domains 1 and 2 are again 0.015 and 0.035. The desired sample size in each SSU is $\bar{n} = 65/2 = 32.5$. Sampling two SSUs with $pp \, (S'_{ij})$ gives $\pi_{j|i} = 0.0.618$ for PSU 1, SSU 3, as in Example 10.2. The within-SSU sampling rate for domain d is $\pi_{k|ij} \, (d) = \bar{n}f_d / \left(\pi_i S'_{ij} \right)$. For PSU 1, SSU 3, this is $32.5*0.015/0.1117/28.9 = 0.151$ for domain 1. For domain 2 in PSU 1,

Table 10.3: Population counts and composite MOSs for Example 10.2.

PSU	PSU probabil-ity π_i	SSU	Domain population size $Q_{ij}(d)$		Total count Q_{ij}	Comp. MOS S'_{ij}	SSU probabil-ity $\pi_{j\mid i}$	Within-SSU probability $\pi_{k\mid ij}$		Expected size within PSU, SSU			sample
			Domain					Domain		Domain			
			$d=1$	$d=2$				$d=1$	$d=2$	$d=1$	$d=2$	Total	
1	0.1117	1	20	80	100	27.8	0.618	0.217	0.507	4.3	40.5	44.9	
1	0.1117	2	25	45	70	17.5	0.389	0.345	0.806	8.6	36.3	44.9	
1	0.1117	3	5	90	95	28.9	0.643	0.209	0.487	1.0	43.8	44.9	
1	0.1117	4	35	35	70	15.7	0.349	0.385	0.898	13.5	31.4	44.9	
PSU total					335	89.8							
2	0.1567	1	40	80	120	21.7	0.551	0.174	0.406	7.0	32.5	39.4	
2	0.1567	2	40	100	140	26.2	0.664	0.144	0.336	5.8	33.6	39.4	
2	0.1567	3	20	35	55	9.7	0.247	0.388	0.905	7.8	31.7	39.4	
2	0.1567	4	45	30	75	11.0	0.279	0.343	0.800	15.4	24.0	39.4	
2	0.1567	5	60	20	80	10.2	0.259	0.370	0.862	22.2	17.2	39.4	
PSU total					470	78.8							
Population totals (includes all PSUs)			4,000	2,000	6,000								

Table 10.4: Population counts and composite MOSs for Example 10.3.

PSU	Within probability SSU / SSU	PSU probability π_i	SSU	Domain population size $Q_{ij}(d)$ d=1	d=2	Total count Q_{ij}	Comp. MOS S'_{ij}	SSU probability $\pi_{j\mid i}$	Within-SSU probability $\pi_{k\mid ij}$ Domain d=1	d=2	Expected size within PSU, SSU Domain d=1	d=2	Total	sample	Overall sel. prob. Domain d=1	d=2
1		0.1117	1	20	80	100	27.8	0.618	0.157	0.367	3.1	29.4	32.5		0.011	0.025
1		0.1117	2	25	45	70	17.5	0.389	0.250	0.583	6.3	26.3	32.5		0.011	0.025
1		0.1117	3	5	90	95	28.9	0.643	0.151	0.353	0.8	31.7	32.5		0.011	0.025
1		0.1117	4	35	35	70	15.7	0.349	0.279	0.650	9.8	22.8	32.5		0.011	0.025
PSU total						335	89.8									
2		0.1567	1	40	80	120	21.7	0.551	0.143	0.335	5.7	26.8	32.5		0.012	0.029
2		0.1567	2	40	100	140	26.2	0.664	0.119	0.277	4.8	27.7	32.5		0.012	0.029
2		0.1567	3	20	35	55	9.7	0.247	0.320	0.746	6.4	26.1	32.5		0.012	0.029
2		0.1567	4	45	30	75	11.0	0.279	0.283	0.659	12.7	19.8	32.5		0.012	0.029
2		0.1567	5	60	20	80	10.2	0.259	0.305	0.711	18.3	14.2	32.5		0.012	0.029
PSU total						470	78.8									
Population totals (includes all PSUs)		4,000 2,000 6,000														

SSU3, the rate is 32.5*0.035/0.1117/28.9 = 0.353. This combination produces the same workload in each SSU. For instance, in PSU 1, SSU 3, the expected sample size is 5*0.151 + 90*0.353 = 32.5. On the other hand, the sample is not self-weighting. The selection probabilities for domains 1 and 2 are 0.011 and 0.025 in PSU 1 but are 0.012 and 0.029 in PSU 2. In this example, the sample is not far from self-weighting, but this is not always true.

10.6 Effects of Population Change: The New Construction Issue

SSUs (BGs, tracts, etc.) for area samples and multistage designs in general are almost always selected with probabilities proportional to some MOS. Examples of MOSs are:

- Total population
- Total households
- A weighted combination of domain population counts (e.g., the composite MOS discussed in the previous sections)

Generally speaking, the larger the relative MOS of an area, the larger its selection probability will be in most designs. If these are based on decennial census data, the farther the date of sample selection is from the census, the more out-of-date these counts get. Table 10.5 shows the growth rates for population and housing units between censuses from 1960 to 2000. The country increased in population by about one hundred million over this period. Notice that the growth in housing units (HUs), which are usually sampled at some stage of a household survey, does not equal the growth in population. The 1960s and 1980s saw a relatively large increase in HUs compared to population. As Table 10.6 illustrates, there can be a lot of regional variation in the growth rates. Nevada and Arizona are popular retirement destinations. During the 1990–2000 period, this led to a boom in construction with many new housing units being built. Between 2000 and 2010, the growth continued in the same states. Other areas, like the District of Columbia, had low growth or lost population from 1990 to 2000 but resumed growing between 2000 and 2010.

Small areas may be especially affected by population changes between censuses. Some MOSs will be too big (due to demolitions, vacancies, and natural disasters); some will be too small due to new construction. Either of these can lead to some severe inefficiencies in a sample design. For example, in 2005, a hurricane destroyed large sections of New Orleans on the coast of the Gulf of Mexico in the southern USA. Entire residential neighborhoods were destroyed and not rebuilt for years. As of 2011, some neighborhoods were still

Table 10.5: Change in US population between decennial censuses.

Year	Population Total	Percent change from last census	Housing units Total	Percent change from last census
1960	179,323,175	18.5	58,326,357	26.4
1970	203,302,031	13.4	68,704,315	17.8
1980	226,542,199	11.4	88,410,627	28.7
1990	248,709,873	9.8	102,263,678	15.7
2000	281,421,906	13.2	115,904,641	13.3
2010	308,745,538	9.7	131,704,730	13.6

Sources: U.S. Census Bureau (1991, 2001a,b) and Bell et al. (1999)

Table 10.6: Percentage change in population and housing units, 1990–2000, for selected states.

State	Percent change 1990–2000 Population	housing units	Percent change 2000–2010 Population	housing units
Nevada	66.3	59.5	35.1	41.9
Arizona	40.0	31.9	24.6	29.9
Utah	29.6	28.4	23.8	27.5
New York	5.5	6.3	2.1	5.6
Connecticut	3.6	4.9	4.9	7.4
District of Columbia	−5.7	−1.3	5.2	8.0

vacant. Using Census 2000 data on population counts would lead to tracts and BGs being sampled that have virtually no people living in them. This would reduce sample size and lead to expensive and unproductive fieldwork if personnel are sent out to attempt interviews in vacated areas.

In other cases, construction of new housing will result in the census counts being far too small. Consequently, an SSU may be selected with a smaller probability based on out-of-date population counts than it deserves based on its actual size. To illustrate the problem, suppose that the design calls for HUs within a sample SSU to be selected at rate 1/4 and the census count of HUs (the MOS) is 100. Field staff arrive at the SSU to count and list all HUs contained within the area and discover that a new apartment complex has been built so that the actual total of housing units is 500. With the rate of 1/4, the expected sample size is 25 (= 100 × 0.25). If we use the planned rate, the actual sample size will be 125 (= 500 × 0.25). The larger sample size is probably statistically inefficient because the intraclass correlation, discussed in Chap. 9, will be high for at least some variables. It is also likely that neither the budget nor the time schedule can tolerate an extra 100 interviews.

There are several approaches to handling this problem. One is to use the initially planned sampling rate, which has the disadvantages just noted. Another

is to reduce the sampling rate to follow the initial plan of selecting 25 sample units. In that case, the weights for each unit within the SSU will be 20 instead of 4. This may create an undesirable disparity with weights in other SSUs, which in turn may increase variances (see Sect. 14.4.1). Other modifications of the sample design can be used that may be preferable. The Census Bureau does publish annual updated population estimates for all counties and for incorporated places. [9] These may be of some use in partially updating population counts for some subcounty areas. But, since places do not necessarily coincide with census tracts or block groups, the population estimates cannot be used to update the measures of size of the units normally used to construct SSUs.

This issue is referred to as the *new construction problem.* The problem is one of efficiency rather than bias. If nothing special is done, existing neighborhoods with new construction will have some chance of selection in an area sample and will probably not be missed entirely. Since all areas have some chance of being selected, some new construction is likely to be selected by chance. But, in the one sample you do select, major developments could be missed. This causes face validity problems. As a result, methods have been devised to avoid glaring omissions. The ones we cover here are based on Bell et al. (1999) and Montaquila et al. (1999).

Option 1: Sample Building Permits

Local governments in the USA usually require that building permits be obtained when new construction projects are undertaken. Permits are required to insure that the planned construction does not violate zoning ordinances. The general idea in this option is to get lists of permits issued by local jurisdictions and sample from those lists. One source of information is the Census Bureau's building permit survey (BPS), which is a monthly survey of building permit offices. Aggregate statistics are published for counties and places (e.g., cities, and incorporated places) on the number of permits issued, HUs authorized, and valuations. From these statistics, a judgment can be made about whether individual sample PSUs will require special, new construction samples. Once a jurisdiction has been identified as needing a new construction sample, local permit offices must be visited to obtain the addresses for new construction projects. One could compile a complete list of permits to use as a sample frame. If the number of permits is large, an option is to form new construction "segments" defined by local permit issuing office and time period. For example, suppose the desire is to select a sample of permits for the period July 2003 through June 2006 in Montgomery County,

---

[9] www.census.gov/popest/estimates.html.

Maryland (MD). Segments, created to have permits for about the same number of housing units, might be:

- All residential permits issued by Gaithersburg, MD, permit office the periods (1) July 2003–June 2004, (2) July 2004–June 2005, and (3) July 2005–June 2006 (The Gaithersburg office is the main issuer of permits in the county)
- All permits issued by (4) the other Montgomery County, MD, permit offices between July 2003 and June 2006

A sample of two of these four segments might be selected with probability proportional to number of permits. The selected local permit offices would be visited and lists of the permits themselves obtained. The HUs corresponding to the permits would be listed and a subsample selected. The new construction HUs might be selected only from the permit frame, not from the area sample. Alternatively, overlap in the permit and area frame units may be allowed; however, the associated selection probability for a unit in both lists is somewhat more complicated to calculate. The implications of these two scenarios for weighting are described below.

The advantages of the building permit option include (1) some new construction is guaranteed to be sampled and (2) unplanned variation in the size of area segments is controlled. Disadvantages of this option are:

1. About 5 % of new HUs are built in areas that do not require permits.
2. Building permits are not required for placement of mobile homes. These are usually in mobile home parks, which are common in some areas, but nonexistent in others.
3. The fact that a permit is issued does not mean that an HU was ever built; however, only about 1 % are not built. Thus, the permit list may contain a small number of ineligibles.
4. There may also be a delay between the time that a permit is issued and an HU is actually built. In the example above, some HUs may be built in the period July 2003–June 2006 may have had their permits issued before July 2003. On the other hand, some construction projects associated with permits issued in July 2003–June 2006 will not be built until later.
5. Some permit offices are uncooperative or have poor records. This can make dealing with them expensive and unproductive because of the extra personnel time needed to access and process the records.
6. Sample permit cases are not necessarily clustered. This may increase travel costs for interviewing.

An operational problem is handling the HUs on both the area and the permit sample properly for weighting. If the rule is used that an HU on the permit frame can only be selected from that frame, then it must be determined whether an HU in the area sample could have been picked through the permit sample. If it could, then the unit would be removed from the area sample frame. To decide whether an area sample HU was on the permit frame is

easier said than done. Matching an address for an area sample HU is usually error-prone because of variations of addresses that can be used.

An alternative is to have the field personnel guess whether an HU was built in the period covered by the permit frame. This is easy for older housing but not for newer ones. The respondents in the area sample can be asked when their HU was built. Any area sample HUs would be excluded that are reported to have been built in the period covered by the permit frame. Field personnel are not equally accurate at estimating the age of a housing unit, and not everyone knows the construction date of their HU. Thus, this method, too, is error-prone.

Another possibility is to allow new construction to be selected in either the area or permit samples. The selection probability could be computed as

$$\text{Pr}\,(\text{HU selected in area sample}) + \text{Pr}\,(\text{HU selected in permit sample})$$
$$- \,\text{Pr}\,(\text{HU selected in both}).$$

This does not unambiguously solve the problem either since we must still identify those units that had a chance of selection from both frames. With this option, less than perfect control over both the outcome of the fieldwork and the weighting is inevitable (although one can say this about almost every survey).

Option 2: Two-Phase Sample of Segments

As should be clear from the considerations above, using a permit sample is not straightforward. A somewhat simpler procedure is to use a variation on two-phase sampling. We will cover the general topic of multiphase sampling in Chap. 17, but the application to new construction is easy to understand. The general idea is to select an extra large sample of area segments or SSUs and update the MOS for each. Then, select a subsample of the first-phase segment sample using the updated MOSs. More specific steps are:

1. Use the Census BPS data to update population counts in individual places in a PSU.
2. Convert the HU counts from the BPS to counts of persons using a conversion factor for persons per HU, e.g., 2.6 persons per HU might be used.
3. Apply the place-specific population adjustment to every SSU contained in the place. This gives a new set of measures of size for the SSUs in the PSU.
4. Select a large sample of SSUs. Montaquila et al. (1999) suggest that the first-phase sample be 5–10 times larger than the number of SSUs ultimately desired. This is a large multiple, and the cost of handling SSUs will determine how large the first-phase sample of SSUs can be.
5. "Counters," i.e., experienced field listers (Sect. 10.3.2), then travel to the first-phase SSUs and count the HUs. Satellite pictures may help in this

task. These do not have to be perfect but should identify areas where the HU count has changed substantially since the date to which the frame MOSs refer.

6. Update the measures of size for each SSU based on the field counts.
7. Select a second-phase sample of segments from all the first-phase units using the new MOSs.

This method has several advantages. It does lead to the SSU sample being selected with MOSs that are a better reflection of the current population sizes in areas where there has been considerable change. This approach will identify areas where there has been growth and also ones where demolitions were common. This may be especially relevant where natural disasters have hit. All SSUs become regular area SSUs. There is no need to deal with permit offices, which can be a costly nuisance in some areas. Travel time is reduced for interviewers since all sample HUs are clustered by area SSU. The screening question about whether an HU was built in the last decade (or some other specified time period) is eliminated.

There are, of course, disadvantages. The main one is that counting of first-phase segments is an extra cost, exhibited both in calendar time and project funds. This added field cost may even be more than for permit sampling. Training materials have to be developed for counters, the training conducted, and the field counting must be done. A data processing system is needed to incorporate the updated information on SSUs.

Finally, we note that a combination of permit and two-phase sampling of SSUs could be used—permit in high-growth areas and two phase (or neither) in the low growth.

Option 3: Half-Open Interval Technique

The final option we discuss is one that has been in practice since the 1960s. The "half-open interval" (HOI) procedure is attributed to Kish (1965) as a method to ensure HU frame coverage in a geographic area already selected for a survey. With the HOI procedure, field staff members are provided with a frame listing for a particular area, sorted in some order. Their task is to identify any HUs not on the frame (e.g., new construction) that exist between the sampled HU and the next HU on the list. Any newly discovered units are automatically included in the sample. The HOI procedure, as noted by many researchers, is effective only when field interviewers are trained extensively and the technique properly implemented (Eckman 2010; Eckman and O'Muircheartaigh 2011).

10.7 Special Address Lists

Area household sampling may be impractical for surveys with a limited budget or a small data collection window. As noted previously, the development of an HU frame through counting-and-listing procedures (either with or without HOI) may take months. Iannacchione et al. (2003) turned to a residential mailing list frame as a cost-effective way to randomly select and survey approximately 15,000 households in 2000 for the Dallas Heart Study (Victor et al., 2004). A random subset of the respondent pool were asked to provide blood and urine samples at the end of the interview, hence the need for in-person data collection. This seminal project opened a new area of research known widely now as *address-based sampling* or just ABS. The research to date including coverage of this type of frame is synthesized in a paper by Iannacchione (2011) and summarized below.

All ABS sampling frames in the USA generally derive from a single source—the US postal service (USPS) address management system (AMS). Information contained in this system include: street name and number, mailbox number (if appropriate), city, state, nine-digit ZIP code, the delivery-sequence number (order that a USPS carrier delivers mail), and address vacancy indicators. Only commercial vendors that apply and qualify for a license may access the USPS-AMS data through an electronic file called the computerized delivery sequence (CDS) file. Other mailing-address files, referred to as data products, are available to the vendors. However, the CDS in combination with the CDS No-Stat file containing, among other things, the addresses of HUs under construction has been shown to have nearly complete coverage of the US household population (Iannacchione, 2011).

In addition to the USPS-AMS characteristics, commercial vendors of survey samples sell "enhanced" versions of the CDS. The enhancements can include landline telephone numbers, a name associated with the address, Spanish surname indicator, estimated age of the head of household, as well as some geocoded (i.e., latitude and longitude) and census tract information. The geocoding information is required to map ABS geography to census geography (Sect. 10.1). The quality of this information depends on factors such as the address-telephone number match rate, age of the household contact information, geocoding errors attributed to the approximate mapping of the mailing address to a physical location, and the number of HUs linked to a particular mailing address.

At least four advantages of the CDS are noted in the literature. First, data collection costs may be reduced either by using the list frame information directly in combination with an HOI technique in lieu of counting-and-listing procedures or by providing listing personnel with starting information. Researchers have noted problems with this first advantage. For example, a list of articles cited in Iannacchione (2011) point to varying levels of undercoverage in ABS frames for rural areas of the USA. Additionally Eckman (2010) found that field staff tends to ratify that the starter list is correct rather than

checking for omissions—an issue for ABS as well as when the HOI procedure is implemented regardless of the sampling frame.

A second advantage is that unclustered samples can be selected within a PSU, which will reduce the SE of many estimates. The cost implications can be minimal only if the size of the PSUs, measured in terms of travel distance for field staff, is not large. Third, if accurate, the household-level items, like race/ethnicity and age, can improve the efficiency of a sample by enabling the targeting of different groups. Fourth, again if the contact information is correct, then several modes may be used to contact sample members with the ultimate goal of increasing response rates.

Even though the moniker ABS may be specific to the USA, the use of residential mailing lists as survey sampling frames is not unique. For example, samples for the British Household Panel Survey (http://www.iser.essex.ac.uk/bhps) have been drawn from the *postcode address file* (PAF),[10] a file containing approximately twenty-eight million UK residential and business addresses. As part of the 2008 European Social Survey, researchers selected random samples using "postal delivery points" supplied by the Croatian Postal Service.[11]

---

[10] www.postcodeaddressfile.co.uk.

[11] ess.nsd.uib.no/ess/round4/surveydoc.html.

Exercises

10.1. A survey of the hamlet of Loon Lake will be conducted to determine the health status of the local population. The town has four census tracts and two will be sampled. The Census 2000 population counts and the number of permits issued since the year 2000 are shown below. Suppose that two tracts are to be sampled with probabilities proportional to the census counts. A self-weighting sample of 300 persons will selected.

Tract	2000 Census population	Permits issued since 2000
1	6,000	0
2	5,200	100
3	2,120	875
4	3,700	6
Total	17,020	981

(a) Determine the selection probabilities of the tracts using the census population counts, the within-tract sampling rates of persons, and the expected number of persons selected per tract.
(b) Next, use the number of permits issued to get an updated estimate of the number of persons in each tract. Assume that there are 2.6 persons associated with each housing unit and that a permit is associated with one HU.
(c) Using the sampling rates computed in (a) and the updated population counts from (b), how many persons do you expect to sample in each tract if it is one of the two selected? Discuss the effects on workload of using the out-of-date population counts.
(d) Using the updated population estimates, compute selection probabilities for tracts, within-tract sampling rates of persons, and expected number of sample persons in each tract. Discuss how these compare to those in (a) and (c).

10.2. The following table shows a population of four PSUs with the counts of persons in each of two domains in each PSU. Suppose that the desired overall sampling rates for the domains are $f_1 = 0.05$ and $f_2 = 0.10$. You want to select a sample of two PSUs with probabilities proportional to the composite MOS described in Sect. 10.7.

PSU	$N_i(1)$	$N_i(2)$	N
1	50	50	100
2	20	100	120
3	90	60	150
4	160	70	230
Totals	320	280	600

Compute the following:

(a) Total expected sample sizes for the two domains.
(b) Composite MOS for each PSU and the total across PSUs. Verify that the
    grand total equals the total expected sample size.
(c) Selection probability for each PSU.
(d) Expected domain sample size and domain sampling rate within each PSU.
    Are the expected sample sizes integers? If not, what method can be used
    for sampling within a PSU that will achieve the desired rate?
(e) Verify that the expected sample sizes for any two of the PSUs sum to the
    total expected sample size you computed in (a).

10.3. A two-stage survey of persons is to be done in which $5\,\%$ of persons age
35 and under and $15\,\%$ of persons over 35 will be sampled. Four PSUs will
be selected using the composite MOS defined in Sect. 10.5. A self-weighting
sample is to be selected within each domain, and the workload should be the
same in each selected PSU. Compute the following:

PSU	Domain $N_i\,(d)$		Total count N_i
	$\leq 35,\ d{=}1$	$>35,\ d{=}2$	
1	80	20	100
2	60	20	80
3	50	90	140
4	80	10	90
5	50	25	75
6	90	20	110
7	50	80	130
8	50	65	115
9	55	25	80
10	50	50	100
Totals	615	405	1,020

(a) Total expected sample sizes for the two domains and the total sample size
    across domains.
(b) Composite MOS for each PSU and the total across PSUs. Verify that the
    grand total equals the total expected sample size.
(c) Selection probability for each PSU.
(d) Domain sampling rate and expected domain sample size within each PSU.
    Are the expected sample sizes integers? If not, what method can be used
    for sampling within a PSU that will achieve the desired rate?
(e) Verify that the expected sample sizes for any four of the PSUs sum to the
    total expected sample size you computed in (a).

10.4. The two PSUs below are an existing PSU sample selected some years ago. A new survey is to be done in these PSUs. The selection probabilities for PSU 1 and PSU 2 were 0.5 and 0.3. These are fixed and cannot be altered. The goal is to select a sample from domains 1 and 2 at rates 0.030 and 0.125. Within each domain, the sample is to be self-weighting. Two sample SSUs will be selected in each PSU.

	PSU probability		Domain tion size $N_{ij}(d)$	popula-	Total population
PSU	π_i	SSU	$d=1$	$d=2$	N_{ij}
1	0.1263889	1	40	80	120
1	0.1263889	2	25	45	70
1	0.1263889	3	35	90	125
1	0.1263889	4	105	35	140
2	0.2805556	1	80	180	260
2	0.2805556	2	40	200	240
2	0.2805556	3	20	85	105
2	0.2805556	4	85	150	235
2	0.2805556	5	110	60	170
Pop totals (includes all PSUs in frame)			5,000	2,200	7,200

(a) Compute the expected sample sizes in each domain in each SSU and the total sample size in each SSU across the domains. Assume that rates of 0.03 and 0.01 are used for domains 1 and 2. Note that the population totals for the domains are 5,000 and 2,200 as shown in the table above.
(b) Compute the composite MOS for each SSU using the method in Sect. 10.5.
(c) Compute the SSU selection probabilities assuming that the SSU sample will be selected with probabilities proportional to the composite MOS.
(d) Calculate the within-SSU probabilities required for the sample in each domain to be self-weighting.
(e) Compute the expected workload in each SSU if it were to be sampled. Are these equal? If not, explain why.
(f) Verify that the SSU and within-SSU probabilities computed in (c) and (d) do yield a self-weighting sampling in each domain.
(g) Determine a sampling scheme for SSUs and units within SSUs that will give an equal workload in each SSU. Carry out the calculations for SSU and within-SSU selection probabilities, and verify that the total expected sample size across the two domains is the same in every SSU.
(h) Does the scheme you designed in (g) lead to a self-weighting sample? Why or why not? Support your answer with calculations.

Chapter 11
The Area Sample Design: One Solution

The project in Chap. 8 requested that you design a sample of twenty-five census tracts ($m = 25$) and one block group per sample census tract ($n=1$). The desired total sample size is 1,000 persons which was split equally among five age groups. Thus, the requirement for an equal workload per block group (BG) leads to $\bar{q} = 1,000/(mn) = 40$ persons in each BG. Table 11.1 shows the population counts from the 2000 census for the five age domains. Each domain was to receive a sample size of 200. The implied sampling rates range from about 0.12 % for ages 25–44 to 0.51 % for ages 18–24.

Table 11.1: Population, sample size, and overall sampling rate for five age domains in Anne Arundel County, Maryland.

Age domain d	Population	Percent of population (%)	Sample size	Domain sampling rate f_d (%)
18–24	39,448	10.76	200	0.5070
25–44	160,940	43.92	200	0.1243
45–54	71,657	19.55	200	0.2791
55–64	45,637	12.45	200	0.4382
65+	48,765	13.31	200	0.4101
Total	366,447	100.00	1,000	0.2729

Since a self-weighting sample within each age group is desired along with the same workload in each PSU, the composite measure of size (MOS) method, described in Sect. 10.5, can be used. In particular, the composite MOS for BG j in tract i is

$$S_{ij} = \sum_d f_d Q_{ij}(d),$$

where $Q_{ij}(d)$ is the number of persons in age group d in tract i and BG j. The MOS for tract i is then $S_{i+} = \sum_{j \in U_i} S_{ij} = \sum_d f_d Q_i(d)$ where U_i is the set of all BGs in tract i. The total MOS across all tracts and BGs is S_{++}. The project assignment asks you to select tracts and BGs using Sampford's procedure, which is one method of probability proportional to size selection in which joint selection probabilities can be computed. If we select a $pp(S_{i+})$ sample of tracts followed by a $pp(S_{ij})$ sample of 1 BG in each tract, then the selection probability of that BG is

$$\pi_{ij} = \pi_{j|i} \pi_i \pi_{j|i} = 25 \frac{S_{i+}}{S_{++}} \frac{S_{ij}}{S_{i+}} = 25 \frac{S_{ij}}{S_{++}}.$$

This is the same selection probability that would be obtained by selecting a sample of 25 BGs directly from the frame of BGs. However, notice that this sample design of selecting tracts first, followed by a single BG per tract, is not the same as selecting BGs directly. If we selected BGs directly using Sampford, all pairs of BGs would have non-zero joint selection probabilities. Since we select tracts and then 1 BG per tract, the joint selection probability of any two BGs in a given tract is zero.

The spreadsheet, `AnneArundel.MD.solution.xls`, shows the value of the composite MOS for each tract and BG, along with population counts by age group, and a variety of other calculations. Note that some ages are out of scope for this survey (0–5 years, 6–11, 12–17). These are excluded from the composite MOS.

A number of quality control checks must be made to determine if some small BGs should be combined with others. Among the checks are whether each BG will provide an adequate workload and whether some BGs will have relatively small selection probabilities and, therefore, large weights relative to other BGs. Combining of BGs terminates when each workload is adequate and no weight will be extremely different from the ones. Creating PSUs that are geographically large is undesirable because limiting interviewer travel may also be a goal.

The first check is whether sampling at the desired rates is possible in all BGs. As outlined in Sect. 10.7, the expected number of persons sampled in each domain in each SSU (BG) should be less than the population count in the SSU. Also, the sum of these expected counts in a BG across the domains must be less than the population in the BG. There are six BGs that violate the requirement that $q_{ij}^*(d) \leq Q_{ij}(d)$ where $q_{ij}^*(d)$ is the expected number of sample persons in BG ij from domain d. The six are shown in Table 11.2. Each violates the sample size constraint in at least one age group. For example, tract 701400, block group 3 has a population of 16 in the 25–44 group, but the sampling algorithm requires an expected sample size of 16.4; the population is 7 in the 65+ group, but the desired sample size is 23.6. A borderline case is tract 741100 where the population of 18–24 is 10 and the sample is to be 10.1.

Two other BGs are shown in Table 11.2 that have no population in any of the in-scope age groups. These could be left in the frame in case some eligible people have moved in since the 2000 census. Or, if we are confident that the entire BG is out of scope, it could be classified as ineligible and dropped from the frame. In fact, inspection of the map in `Anne Arundel.blkgrps(streets).pdf` reveals that the `tract.BGs` 740602.1 and 740603.1 are on a military reservation or in a wildlife preserve in the western part of the county. If the eligible universe covers only the noninstitutional household population, it might be safe to drop these BGs. Instead, we combined them with BG 2 in their respective tracts for this exercise. The other deficient BGs were combined with other tracts.BGs as shown in the table.

There is also one tract that has a relatively small selection probability based on the initial calculations. Tract 741100 has a selection probability of 0.005; the next smallest is 0.022. This smallest tract contains a single BG, which, as shown in Table 11.2, was combined with `tract.BG` 740603.2. Tract 741100 is, thus, combined with tract 740603.

Table 11.2: Block groups where the expected workload exceeds the population count.

Tract	BG		18–24	25–44	45–54	55–64	65+	Action: combine with `tract.BG`
701400	3	Population	0	16	0	0	7	701400.2
701400	3	Workload	0	16.4	0	0	23.6	
740602	1	Population	0	0	0	0	0	740602.2
740602	1	Workload	NA	NA	NA	NA	NA	
740603	1	Population	0	0	0	0	0	740603.2
740603	1	Workload	NA	NA	NA	NA	NA	
740603	3	Population	5	101	0	0	0	740603.2
740603	3	Workload	6.7	33.3	0	0	0	
741100	1	Population	10	42	16	12	0	740603.2
741100	1	Workload	10.1	10.4	8.9	10.5	0	
750600	1	Population	0	0	45	0	7	750600.2
750600	1	Workload	0	0	32.6	0	7.4	
750700	2	Population	0	4	0	0	0	750700.1
750700	2	Workload	0	40	0	0	0	
750801	5	Population	0	21	30	8	0	750801.4
750801	5	Workload	0	7.2	23.1	9.7	0	

NA not applicable

As this example shows, tracts that are geographically adjacent may not have consecutive identification numbers. Figure 11.1 is a schematic map of the tracts in the county. Consulting a map like this may be necessary to make reasonable combinations. Alternatively, longitude-latitude centroids for tracts are available from the Census Bureau. These can be used to calculate

the distance between the centers of the tracts to determine which are geographically near each other. This approach will permit tracts to be combined via a computer algorithm without manual intervention. This is particularly useful when the frame of tracts is large.

Table 11.3: Summaries of tract and BG selection probabilities and weights after combining small units.

Probability or weight	Min.	First quartile	Median	Mean	Third quartile	Max.
π_i	0.0225	0.1828	0.2673	0.2660	0.3329	0.5920
π_{ij}	0.0015	0.0489	0.0783	0.0828	0.1030	0.4385
$1/\pi_i$	1.69	3.00	3.75	5.25	5.48	44.44
$1/\pi_{ij}$	2.28	9.71	12.78	22.08	20.45	684.90

After these combinations are made, the selection probabilities for tracts and BGs are summarized in Table 11.3. The range of selection probabilities for BGs is 0.0015–0.4385 while the range of weights for BGs is 2.28–684.90. Although the range of BG probabilities is substantial, a self-weighting sample of persons can still be selected from each domain since there are no deficient BGs after combining. The Sampford method was used to select a sample of 25 tracts and then 1 BG per sample tract. The code for combining BGs and tracts is in the file `Anne Arundel.MD.analysis.R`.

The selected sample tracts and BGs are listed in Table 11.4 and shaded in Fig. 11.2. The expected workloads in each BG are also shown in the table. The workloads are not integers. This means that when the samples of persons within a sample BGs are selected, the sampling will be done using fixed rates not fixed sample sizes. For example, `tract.BG` 701102.2 has a population of 76 in age group 18–24 and the sample size is 6.3 in Table 11.4. Persons in that age group and BG would be sampled at rate $6.3/76 \doteq 0.08289$.

Quality Control Checks

Checking the correctness of your work is always important. In this case, there are some simple assessments that will help determine whether computations and sample selections are correct. The weight for a sample BG is $1/\pi_{ij}$. These can be used to make population estimates which we can compare to frame numbers. There are two conditions that should hold exactly for any sample that has been selected. First, define

$$y_{ij}(d) = f_d Q_{ij}(d).$$

Table 11.4: Sample tracts and block groups within tracts with expected workloads in each BG.

Tract	Block group	18–24	25–44	45–54	55–64	65+	Total workload
				Workloads			
701102	701102.2	6.3	6.1	7.5	7.8	12.3	40
701200	701200.3	5.6	9.3	6.5	8.3	10.3	40
701300	701300.2	5.8	6.1	10.2	8.6	9.3	40
702100	702100.4	7.7	7.8	12	7.2	5.2	40
702300	702300.4	4.3	4.8	8.1	15	7.8	40
702401	702401.2	1.3	0.6	2	5.4	30.6	40
702700	702700.3	5.7	10.2	6.9	9.2	8	40
706300	706300.2	2.8	6.6	7.6	8.5	14.4	40
706600	706600.5	19.3	7.8	6.2	4.6	2.2	40
708000	708000.1	7.2	7	6.6	10	9.2	40
730100	730100.3	6.8	10.3	5.2	8	9.7	40
730402	730402.2	7.4	6.8	6.2	8.5	11.1	40
730502	730502.2	11.9	7.7	7.3	6.3	6.8	40
730601	730601.4	6.4	5.8	12.3	10.3	5.2	40
730800	730800.2	2	3.9	9.4	12.2	12.5	40
731204	731204.1	8.2	7.8	6.4	7	10.5	40
740102	740102.1	8.3	5.3	8.7	9.7	8	40
740201	740201.4	8.8	12.6	9.1	5.5	4	40
740301	740301.2	10.1	15.4	8	4.7	1.8	40
740500	740500.1	9.1	14.6	8.6	5.4	2.3	40
740601	740601.3	17.9	21.4	0.7	0	0	40
740700	740700.2	9.3	15.6	6.4	4.2	4.4	40
750804	750804.1	8	10.4	4.8	6	10.8	40
751000	751000.1	6.9	5.9	6.7	8.2	12.3	40
751103	751103.2	5.6	5.9	4.7	16.5	7.3	40

The π-estimator of the total for this variable across the domains in a BG is

$$\hat{t}_1 = \sum_d y_{ij}(d) / \pi_{ij}$$

since $S_{++} = mn\bar{\bar{q}}$. In this case $\bar{\bar{q}} = 40$. Thus, the estimator of the total of $y_{ij}(d)$ in any block group is the same constant, 40. The estimator of the population total of $y_{ij}(d)$ is

$$\hat{t}_2 = \sum_d \sum_{i \in s} \sum_{j \in s_i} y_{ij}(d) / \pi_{ij}, \tag{11.1}$$

which can be verified using the fact that $\pi_{ij} = mnS_{ij}/S_{i+}$ and the definition of S_{ij}. Since S_{++} is the total sample size, this π-estimator must be 1,000.

The population totals of the numbers of persons in each domain and across all domains can also be computed as

$$\hat{t}_3 \left(d \right) = \sum_{i \in s} \sum_{j \in s_i} Q_{ij} \left(d \right) / \pi_{ij}, d = 1, \ldots, 5$$

$$\hat{t}_4 = \sum_{d} \sum_{i \in s} \sum_{j \in s_i} Q_{ij} \left(d \right) / \pi_{ij}.$$

These do not necessarily equal the population counts but serve more as a reasonableness check. If the estimates are far from the frame counts, then further checking is warranted to decide whether errors have occurred. For this sample, we have $\hat{t}_1 = 40$ for each domain, $\hat{t}_2 = 1,000$, $\hat{t}_3 \left(d \right) = (38,011.38, 173,593.95, 63,811.75, 45,011.18, 52,714.43)$, and $\hat{t}_4 = 373,142.7$. The estimates $\hat{t}_3$ and $\hat{t}_4$ are reasonably near the population counts in Table 11.1. These checks can also be found in `Arundel.MD.analysis.R`.

Additional Considerations

We spent some time above worrying about the effects of tracts and block groups with small composite MOS. One of the paradoxes of designing samples is that a significant amount of time is spent considering events that may not happen. We may not select one of the BGs with an extremely small MOS, but if we do, its size may not support the desired sample sizes for domains. In addition, its weight will be large and can unnecessarily increase variances. This issue will be addressed again in Chap. 14.

The fact that only 1 BG is selected per tract might raise the question of whether variances can be estimated with this design. We can still estimate design variances because the number of first-stage units is 25, the number of sample tracts. However, there are alternative designs that might be worth considering. If the residents of different areas of the county were known to have different characteristics, it would be advisable to stratify by subcounty geography in some way. The BG numbers assigned by the Census Bureau can be used to sort the BGs in a more or less geographic order and strata created from the sorted list. A BG map should be consulted to be sure whether numeric sorting will achieve your stratification goals. A BG map for Anne Arundel County is in the file, `Anne Arundel.blkgrps(streets).pdf`, on the web site for this book.
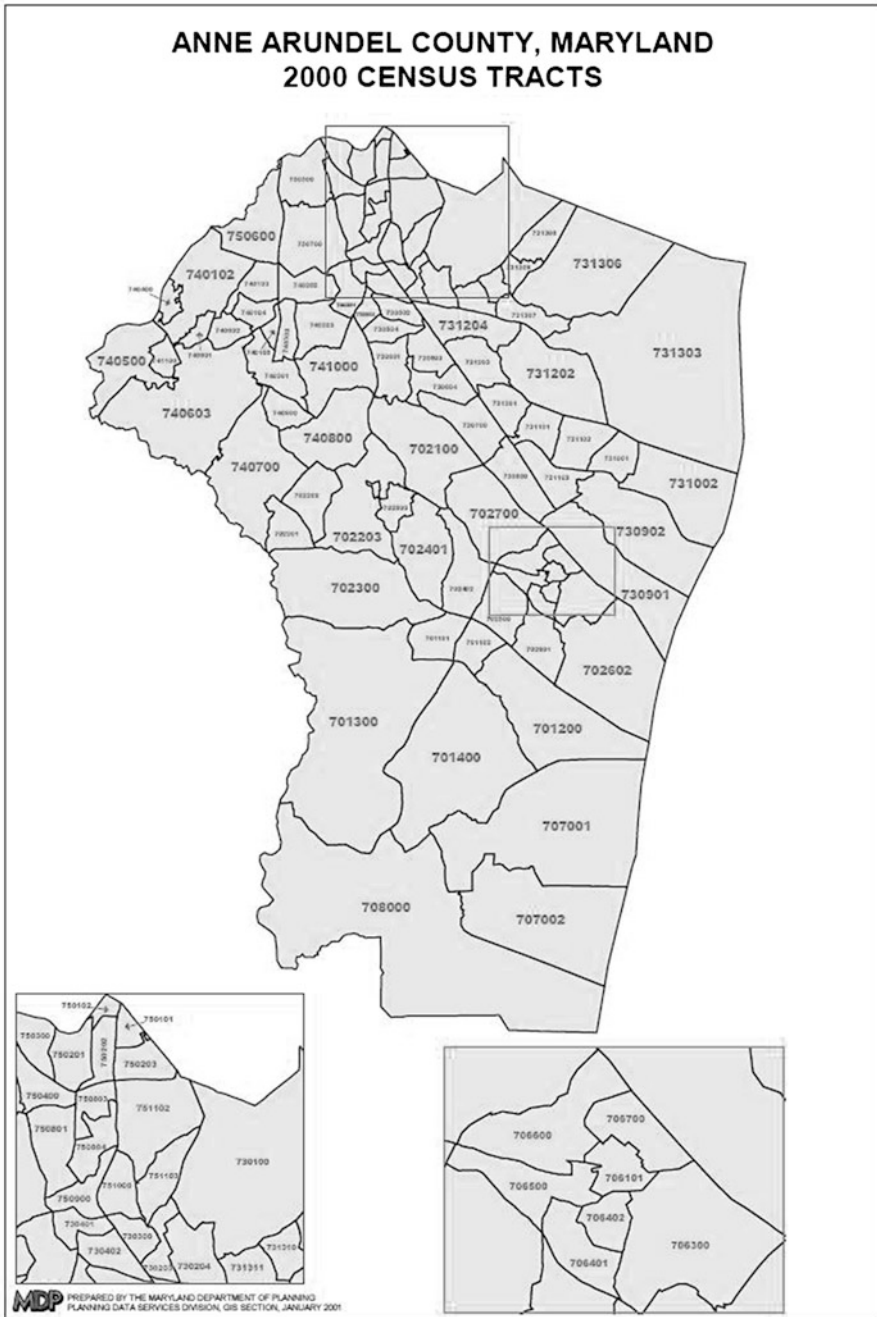
Fig. 11.1: Tract map for Anne Arundel County, Maryland. Source: Maryland Department of Planning, Planning Data Services Division, January 2001.
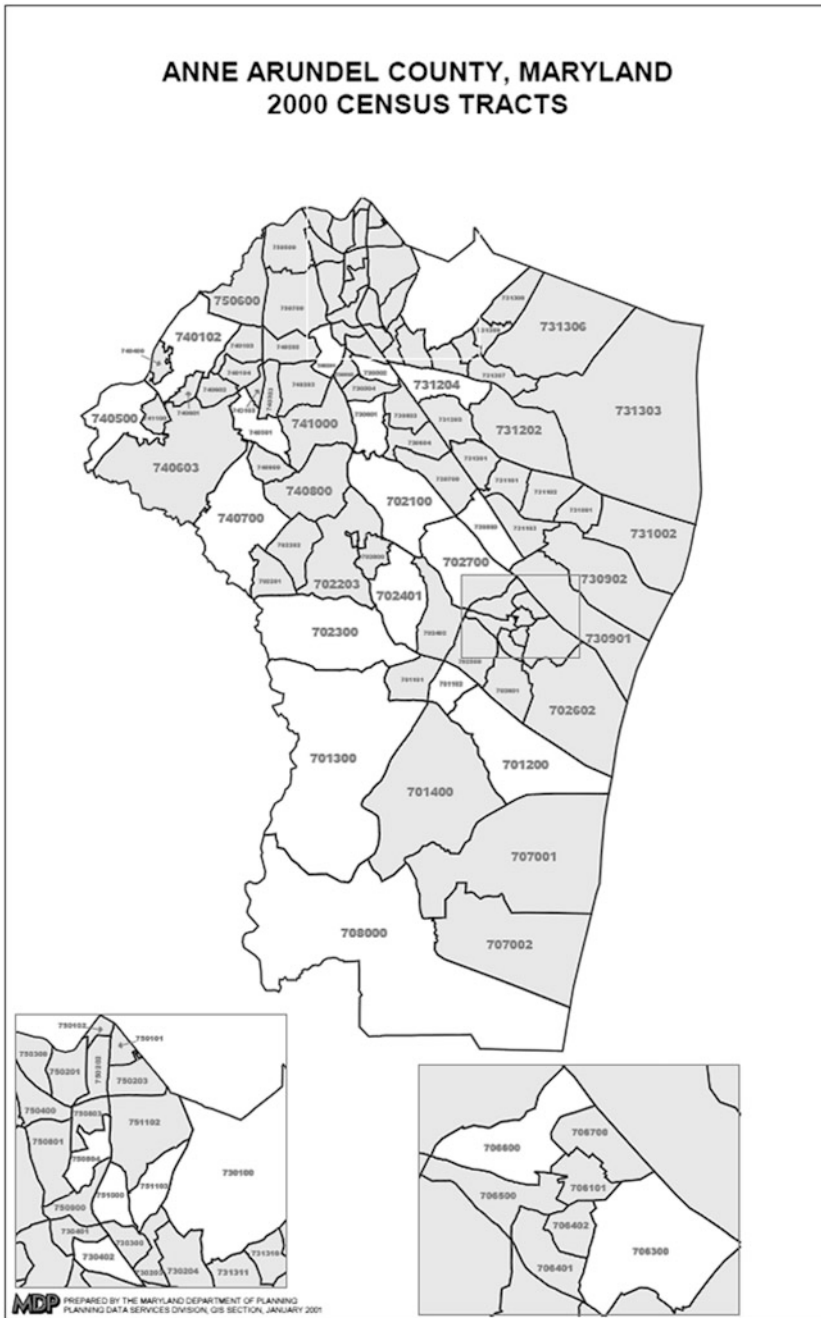
Fig. 11.2: Selected tracts in Anne Arundel County.

Part III
Survey Weights and Analyses

Chapter 12
Project 3: Weighting a Personnel Survey

In this project you will develop survey weights and deliver an analysis file for a survey of military personnel. Members of the military reserves were asked a variety of questions about job satisfaction. Some examples of the questions are:

- Suppose that you have to decide whether to continue to participate in the National Guard/Reserve. Assuming you could stay, how likely is it that you would choose to do so?
- Overall, how would you rate the current level of stress in your personal life?
- Taking all things into consideration, how satisfied are you, in general, with each of the following aspects of being in the National Guard/Reserve?
    The type of work you do in your military job
    Your total compensation (i.e., base pay, allowances, and bonuses)

The data file includes records for all persons who were in the initial sample—respondents, nonrespondents, and ineligibles. There are also several demographic variables from administrative record files for each sample person. The files to be used are listed at the end of this chapter.

The following tasks have yet to be completed and are assigned to your team. Each task should be documented in the project final report; be sure to justify the decisions your team has made.

(1) Develop the design weights (inverse of the selection probabilities) for this single-stage stratified simple random sampling design and verify your calculations. The field STRATUM defines the sample design strata. Each record contains counts of the number of persons in the population (NSTRAT) and the sample (NSAMP) in the design stratum to which the record belongs. The field V_STRAT identifies design strata that were collapsed together for variance estimation. Note that, if a population count is needed for a variance stratum, the values of NSTRAT will need to be summed for the design strata that are combined into a V_STRAT.

(2) Specify how you will classify the various response status codes (RESPSTAT) into the general categories—eligible respondent, eligible nonrespondent, ineligible, or unknown—described in Chap. 6. The variable values and value labels for RESPSTAT are provided in the section below, *Data Files and Other Information*.

(3) Apply weight adjustments to the design weights and verify your calculations. You should include the adjustment methods we have discussed in class—unknown eligibility, nonresponse, and calibration. In the case of either unknown eligibility or nonresponse adjustments, compare weighting cell and propensity adjustments by actually carrying out your own implementation of each method. You may encounter some cases, in either the data file for respondents or the file for population counts, that have missing data for fields that you would like to use in weighting. If so, you need to explain how you handled those in the various steps used in weighting.

(4) Prepare an analysis file (in SAS, Stata, or text format) containing the variables from the original data file (*SOFR.sas7bdat*), the base weights, the final analysis weights (you may choose only one set from task 3 above), and any adjustments applied to the design weights to create the final weights. Additionally, create any necessary indicators you would need to analyze the questionnaire responses and eliminate any unnecessary data records. All variables must have a descriptive label. For any newly created categorical variable, provide a description of the variable values in the report.

(5) Using your final analysis weights, tabulate the proportions of personnel who are

   (a) Dissatisfied or very dissatisfied with their total compensation (RA006A)
   (b) Very unlikely or unlikely to stay in the Reserves (RA008)

   Make these tabulations separately for each service and for enlisted personnel and officers. Include the point estimates of proportions and standard errors. Describe the method you use for standard error estimation and any limitations that the method may have.

(6) Include a description for data users of which cases and weights should be used for various types of data analyses. Provide some brief examples of software code that would be used to estimate means or proportions associated with a typical questionnaire item. Examples should be given for at least two software packages. Your report should describe how the software must be used in order to account for weights and design features like strata.

Contents of the Weighting Report

Below is a list of topic areas that should be included in your weighting report. Questions and suggestions are included in each section to assist with the development of the text. The order of the sections in your report does not have to be the same as that given below. You should construct your report in a way that presents topics in an order that seems logical to your team.

The report should be written to a client whose staff includes managers and technical personnel. Managers will be more interested in understanding the broad outline of the steps used in weighting. Technical personnel will be interested in understanding the details of weight computation, including appropriate formulas, and in being able to appropriately analyze the data. You should consider how to structure your report to serve these audiences.

Topic Areas for Weighting Report:

- Title Page (project title, date of submission, and name of project contact person)
- Introduction (overview of the document)
- Study Weights:
    Brief discussion of sampling design
    Methods to calculate design weights
    Types of weight adjustments and why they were used. Comparison of adjustments
    Evaluation of weights and methods used to check or compare calculations
- Analysis File:
    Summary of analysis file contents (include `PROC CONTENTS` or the equivalent in an appendix)
    Variables of interest
- References
- Appendix
    `PROC CONTENTS` or codebook of data file

Data Files and Other Information

- *SOFR codebook.pdf*—code values for each variable in the *SOFR.sas7bdat* data file.
- *RCCPDS57 codebook.pdf*—code values for each variable in the *RCCPDS57.sas7bdat* data file.
- *Annotated questionnaire.pdf*—the survey questionnaire with annotations showing variable names and code values for all questions. Note that the data file for this project contains only a subset of the questions in the survey. Also, some questions have been recoded to have different names and fewer values in the data file than are on the questionnaire.

- *SOFR.sas7bdat*—edited data file from the survey in SAS version 9 format. The same data is in the SAS transport file, *SOFR.xpt*.
- *RCCPDS57.sas7bdat*—file of population counts. The same data are in the SAS transport file, *RCCPDS57.xpt*.

  This file is the result of matching the sampling frame to the most current personnel file available as of the start of the data collection period. The personnel file consists of all persons on the payroll as of the date the file was constructed. Thus, these counts should cover only <u>eligible</u> cases. The labels for the field names contain the name of the variable in *sofr.sas7bdat* to which the counts correspond.
- *formats.sas7bcat*—format library for both SAS data files.

  To access this library in a SAS program include the following type of libname statement:

  `LIBNAME library ''C:\PracTools'';`

  To be sure that SAS searches that library for formats use

  `options fmtsearch=(library)`

  The folder name `PracTools` should be changed to the location where you save the format file. This format library will give access to the variable and value labels for the fields in *sofr.sas7bdat*.

Variable values and value labels for the RESPSTAT variable:

1	Questionnaire Returned—Completed
2	Questionnaire Returned—(Sufficient) Partial Complete
3	Questionnaire Returned—(Insufficient) Partial Complete
4	Questionnaire Returned—Ineligible
5	Questionnaire Returned—Blank
18	No Return—Deceased
19	No Return—Incarcerated
22	No Return—Separated/Retired
23	No Return—Active Refusal
25	No Return—Other
26	No Return—Eligible based on administrative records
27	Postal Nondelivery
29	Not Locatable
35	Ineligible—No Questionnaire Sent

Chapter 13
Basic Steps in Weighting

Survey weights are a key component to producing population estimates. For example, an estimated total has the form $\hat{t} = \sum_s w_i y_i$ where y_i is a response provided by the ith sample member and w_i is the corresponding analysis weight. Without their use, estimates may reflect only nuances of a particular sample and may contain significant levels of bias. This is the first of two chapters that address techniques for calculating analysis weights currently used in survey research. Articles detailing new research on survey weighting surface in the literature constantly. Therefore, we encourage survey researchers to use these chapters as a basis of understanding and to rely on journal articles for cutting-edge techniques.

There are a series of steps in weighting that are carried out in most, if not all, surveys. These include computation of base weights (also known as design weights), adjustments for unknown eligibility, nonresponse adjustments, and use of auxiliary data to reduce variances and, in some cases, correct for frame deficiencies. We cover the first three of these steps in this chapter. Chapter 14 will address the use of auxiliary data. Sections 13.1 and 13.2 give an overview of weighting and describe general theoretical approaches that are used to justify the use of weights in estimation.

In probability samples, the base weights are inverses of selection probabilities. Examples of base weight calculation are presented in Sect. 13.3 for various designs. These can be used for weighting a sample to the full finite population if the frame is perfect and all sample units respond. In some applications, a complete frame of units is available for sampling and frame problems are not a concern. In others, the frame may contain some units that are ineligible and may omit units that are. Having ineligible units in a frame is a type of overcoverage. A way of adjusting for ineligible units is presented in Sect. 13.4; the problem of frame undercoverage is dealt with in Chap. 14.

The failure of some units to respond is a worry in most surveys. Without adjusting for nonresponse, estimators can have significant levels of bias. There are different methods of adjustment, which we present in Sect. 13.5. Before

covering specific tools used in weighting, some general comments are needed about methods of inference and how they affect weight calculation.

13.1 Overview of Weighting

The general goal in weighting is to find a set of weights, w_i, that can be used in virtually all analyses to produce estimates for the target population under study. For example, an estimated total has the form $\hat{t} = \sum_s w_i y_i$ and a mean can be computed as $\hat{\bar{y}} = \sum_s w_i y_i / \sum_s w_i$ for a set of units in sample s (i.e., $i \in s$). Other statistics that can be written as combinations of estimated totals would use the same set of weights. Regression model analyses, for example, often begin with a type of estimated total that is used to derive parameter estimates. Estimates of medians and other quantiles depend on the same weights used to estimate totals. Properly constructed, a set of weights can provide approximately unbiased and consistent estimates[1] of many different population quantities. As a result, one set of weights can serve many purposes, which is a major practical advantage.

Figure 13.1 shows the general set of steps that are used in weighting in many surveys. The full sample (box A in the figure) can be split into the units whose eligibility is determined (A1) and those for which the eligibility is unknown (A2). The unknowns have their weight distributed in step W1 among the known sample cases (A1a, A1b, and A1c). Note that if the eligibility is known for all sample cases (say, through administrative records), then step W1 is not required. The next step, W2, is to make an adjustment for nonresponse. There are different ways of doing both the unknown eligibility and nonresponse adjustments as discussed in Sects. 13.4 and 13.5. One way is to put respondents and nonrespondents into classes and make a common adjustment to all respondents within each class. Classes can be formed based on estimated response propensities or classification algorithms.

In some surveys, no further steps are used and the final weights are the nonresponse-adjusted weights. In other cases, calibration to population values (step W3) can be used to correct for frame deficiencies and to reduce the variances of estimators, as described in Chap. 14. The auxiliary data used in calibration may come from an updated frame or from an independent source like a population census. Both the eligible respondents (A1a) and the known ineligibles (A1c) may enter into this step, depending on the source of the auxiliary data. There are a variety of methods for using auxiliary data that all fall under the heading of *calibration*. Among them are poststratification, general regression estimation, and raking, all of which are discussed in Chap. 14. Additionally, we discuss in Chap. 14 research associated with combining steps W2 and W3 into a single weighting procedure.

---

[1] See Sect. 4.1 for a discussion of unbiased and consistent estimates.
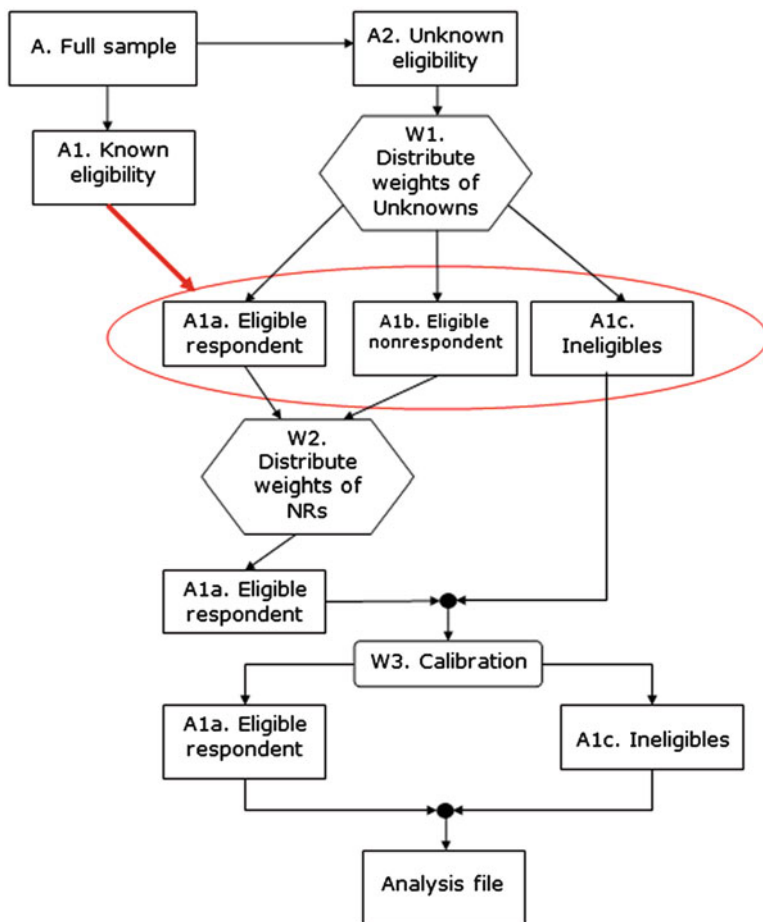
Fig. 13.1: General steps used in weighting.

13.2 Theory of Weighting and Estimation

Weights are used in constructing estimators. The key goal in weight construction should, thus, be to construct good estimators. To know whether an estimator is good or not, we have to evaluate its properties, like bias and variance, with respect to some statistical distribution. There are three methods of generating the distribution used for inference that we will emphasize in this and later chapters:

1. Design- or randomization-based
2. Model-based
3. Model-assisted

There are other approaches, most notably Bayesian [see Gelman et al. (1995)] that have some merit, but we will not cover them in this book. It is important to have at least an intuitive understanding of the thinking behind the three approaches above in order to understand why certain estimators work well or poorly in difference circumstances. In practice, the model-assisted method is the one most often used, as we will explain below.

In the design-based approach to sampling, the properties of estimators like bias and variance are evaluated with respect to repeated sampling. A probability sample must be selected to use this approach, i.e., a random mechanism is used to select units and, in principle, every unit has a known probability of selection, as described in Chap. 3. Nonetheless, it is not unusual for statisticians to "stretch the envelope" by applying repeated sampling analysis to samples that are not really selected with probability mechanisms. To compute, say, the expectation of an estimator, one thinks of a conceptual experiment where samples are repeatedly selected using the same plan. The estimate is computed for each sample. If these values average out to the full finite population value of the quantity being estimated, then the estimator is *design unbiased*. Other properties, like the design variance, are computed similarly.

There are a number of good reasons for using probability sampling. If a random mechanism is used in selection, conscious and unconscious biases are eliminated in selecting the sample. Random sampling is perceived as objective by the public and data users. It also provides a mathematical foundation for computing properties of estimates. However, most samples that start out as probability samples do not end up that way because of nonresponse (NR) and other problems that result in the loss of some sample units. Thus, strictly design-based inference is usually not feasible. Models for nonresponse, undercoverage, and other nonsampling errors are needed to completely reflect the processes that produce a sample. However, computation of base weights (i.e., inverses of selection probabilities) is usually the first step in weight computation in surveys that use probability samples.

Having good design-based properties is comforting. Surely, it is reasonable for a practitioner to be able to say that, if he or she selects random samples over the course of a career, then the methods used will produce correct answers on average. However, the design-based approach does not provide us with a systematic way of constructing good estimators. The relationship of response variables to predictors is not formally considered in design-based inference. Thinking about models that describe the variables in a population provides some structure that can be used as a guide.

By contrast, a strictly model-based approach ignores the sample design and considers only the population structure (i.e., a model) in deciding on an estimator and the corresponding weights. This approach can be applied to either probability or non-probability samples. For example, courses in mathematical statistics use estimators with the assumption that units are drawn from an infinite population. The resulting estimators are unbiased under the

model used to construct the estimators but can be biased if the model is misspecified or the model that fits the sample is different from the one that describes the population as a whole. In some cases, model-based estimation is the only choice. In an Internet survey of volunteer participants, there is no probability sample design, and estimators must be constructed using models. Whether the volunteers are so unlike the full population that estimation is impossible becomes a serious concern.

However, models inevitably need to be considered when developing weights, even in probability samples. Any sample with some degree of nonresponse requires assumptions about the nature of the analysis variables for the non-respondents and about the response mechanism. When computing weights for a volunteer survey, assumptions may be made about the mechanism that describes how likely a person is to participate. These assumptions, whether explicit or implicit, are models.

There are good, though fairly technical, arguments for why the randomization distribution itself should not be the basis for inference [e.g., see Valliant et al. (2000)], even in the absence of nonresponse. The general line of reasoning is that averaging over a randomization distribution involves averaging over samples that can be much different from the one actually selected. That is, design-based inference requires us to consider events that did not actually happen and are, therefore, irrelevant. These arguments do not necessarily have to be considered to develop a set of weights that give reasonable estimators. An interested reader can consult the references above along with Royall (1976) and Smith (1976, 1984, 1994) for discussion of the fundamental issues.

A hybrid approach uses both model-based and design-based thinking and is called model-assisted. A probability sample is selected, weights are calculated, and a model(s) guides the choice of the estimator. Inferences are made using the distribution generated by the probability sampling plan—not a model. Research suggests that the weights provide some level of protection against model misspecification. This is the approach that Särndal et al. (1992) espouse.

13.3 Base Weights

Base weights (or design weights) are computed when the sample is a *probability sample* drawn from a finite population. As defined in (Särndal et al. 1992 Chap. 1) and in Chap. 3 of this book, a probability sample is one realized under four conditions:

1. The set of all samples $\mathbf{S} = \{s_1, s_2, \ldots, s_M\}$ that can be selected from a finite population $\boldsymbol{U}$ can be defined given a specified sampling procedure.
2. A known probability of selection $p(s)$ is associated with each possible sample s in $\mathbf{S}$.

3. Every element in the target population has a nonzero probability of selection with the specified random sampling procedure.
4. One sample s^* is selected by a random mechanism under which each s in **S** receives the probability $p(s)$.

The function $p(s)$ defines a probability distribution on **S**, the set of all possible samples. The value for $p(s)$ is associated with each sample s and differs from the selection probability of an individual unit within the sample. To compute base weights, it is not necessary that we be able to compute $p(s)$. We only need the selection probabilities of the individual elements:

π_i = Selection (or inclusion) probability of element i.

The base weights, $d_{0i} = 1/\pi_i$, are the inverses of the selection probabilities. The selection probabilities may be computed as the product of conditional probabilities at different stages of selection, as illustrated in some of the examples below. Note that the size of the sample is not necessarily a fixed value and is also associated with the sampling procedure (see, e.g., the discussion on Poisson sampling in Chap. 3).

Base weights should be created as soon as the sample is selected if possible. This facilitates preliminary analyses, like performance rate calculations, and insures that the items required for computation of base weights do not get lost. Quality control checks need to be done on the computed weights. We cover these in detail in Chap. 18, but here are some things to note:

- Selection probabilities are all within the range (0,1].
- Base weights should sum to the total number of elements in the population or to a reasonable estimate of the population size. Similar checks on the sums of weights should be made for major subgroups (gender, race/ethnicity, establishments in retail trade, etc.)

Base Weights: An Exception. Using the inverses of selection probabilities as the base weights is usually the first step in weighting. An exception to this is a sampling method where some units are allowed to be selected more than once. These methods are sometimes used in the first stage of a multistage sample. For example, consider a sample of schools where school districts are selected at the first stage with probabilities proportional to number of students in each district. Very large districts may be selected more than once, in which case a larger subsample of schools within the district might be selected. When some units are allowed to be selected more than once, the expected number of selections or "hits" should be tracked; these can be greater than 1. The base weight would then be the inverse of the expected number of selections.

In the remainder of this section, we show the calculation of base weights for some specific designs.

Example 13.1 (Simple random sampling without replacement (srswor)). When n (fixed) units are selected from a population of size N, the selection probability of each unit is the same—$\pi_i = n/N$. The base weight is

$d_{0i} = \pi_i^{-1} = N/n$ and is also the same for all units. An *srswor* is called *self-weighting* or *epsem* (equal probability sampling and estimation method)—see Kish (1965). ∎

Example 13.2 (Stratified simple random sampling without replacement (stsrswor)). The population is divided into $h = 1, \ldots, H$ mutually exclusive strata that cover the whole population. An *srswor* of size n_h is selected in each stratum from a population of size N_h. The selection probability of unit i in stratum h is $\pi_{hi} = n_h/N_h$ and the base weight is $d_{0hi} = \pi_{hi}^{-1} = N_h/n_h$. This is the same for each sample unit in stratum h, but the sampling rates may be different from one stratum to another. ∎

Example 13.3 (Two-stage sampling leading to epsem). Suppose that a sample of students is selected in two stages—schools at the first stage and students at the second stage. In this case the primary (or first-stage) sampling units (PSUs) are schools. Assume that m PSUs are selected with probabilities proportional to size (*pps*) of the student body and that an equal probability sample of $\bar{n}$ students is selected in each PSU. Schools are selected in such a way that the inclusion probabilities are:

$\pi_i = mN_i/N$ for school i
N_i = number of students in PSU i
$N = \sum_{i \in U} N_i$ = total number of students in the population

If an equal probability sample of $\bar{n}$ students is selected in each sample school, then the probability of selecting a student within a school is $\pi_{j|i} = \bar{n}/N_i$ for student j within school i. The overall probability of selection is

$$\pi_{ij} = \pi_i \pi_{j|i} = \frac{mN_i}{N} \frac{\bar{n}}{N_i} = \frac{m\bar{n}}{N}$$

and the base weight for student j in school i is $d_{0ij} = \pi_{ij}^{-1} = N/m\bar{n}$. This particular type of sample is self-weighting since each student has the same base weight. ∎

When the frame contains little or no useful auxiliary information but target sample sizes are desired for some domains, two-phase or multiphase sampling can be used, as described in Chap. 17. The base weights can be computed as the product of weights associated with each phase.

Example 13.4 (Two-stage sampling for domains). A sample of m PSUs is selected and n_i secondary sampling units (SSUs) are selected within PSU i with probabilities proportional to the number of persons in each SSU. For convenience, let PSUs be defined in terms of small geographic segments of a country and SSUs as households within the segments for an area household survey. Each household can contain one or more persons. Suppose that

persons in each sample SSU are listed and classified into $G = 4$ age groups: less than 18, 18–25, 26–64, and 65 and above. Each person within a given SSU and age group is selected at the same rate. Suppose that the selection probability of PSU i is the same as in Example 13.3 and that SSU j in PSU i is selected with probability $n_i Q_{ij}/Q_i$ where Q_{ij} is the population size of SSU ij. The selection probability of each person in SSU ij and group g is

$$\pi_{ij}\left(g\right) = \frac{mn_i Q_{ij}}{Q} f_{ij}\left(g\right),$$

where Q is the total population count and $f_{ij}\left(g\right)$ is the rate at which age group g is sampled in SSU ij. The rates for the age domains will often be set in such a way that a self-weighting sample is obtained in each age group. The base weight for person k in age group g is then $d_{0ij}\left(g\right) = \pi_{ij}^{-1}\left(g\right)$. ∎

13.4 Adjustments for Unknown Eligibility

Frames and samples may contain units whose eligibility cannot be determined. Among the eligible units, most surveys will have some that do not respond. Chapter 6 discussed these problems in the context of determining initial sample sizes. Weight adjustments for unknown eligibility and nonresponse are also usually made to allow the respondents to weight up to the full *eligible* population. For use below, define these sets of sample units:

$s =$ Initial set of all sample units
$s_{IN} =$ Set of units in s that are known to be ineligible
$s_{ER} =$ Set of units that are eligible respondents
$s_{ENR} =$ Set of units that are eligible nonrespondents
$s_{KN} =$ Set of units whose eligibility is known ($s_{IN} \cup s_{ER} \cup s_{ENR}$, where $\cup$ denotes the union of one or more sets)
$s_{UNK} =$ Set whose eligibility is unknown

Some members of the sampling frame may be ineligible despite our best efforts to clean the frame in advance. In a survey of current military members, the frame may be the personnel file as of June of the current year, with a plan to collect data in August. By the time the survey is fielded in August, some people will have left the military. These "leavers" are ineligible, assuming the target population is all members at the time of data collection. Another example would be a telephone survey of households in which some telephone numbers turn out to be for businesses. In a household survey of childhood immunizations, households that do not have children are ineligible.

For a variety of reasons, it may not be possible to determine eligibility for all sample units. Some cases whose eligibility for a household survey may remain unknown after data collection is finished are:

- Ring/no answers in a telephone survey
- Undeliverable addresses in a mail survey
- Never at home in personal visit survey

As in Fig. 13.1, suppose that the final classification of sample units is:

- Known eligibility status:
  Eligible respondents
  Eligible nonrespondents
  Ineligibles
- Unknown eligibility status

If there are units known to be ineligible in the sample, this is evidence that there are other ineligibles in the unknown eligibility part of the sample and also in the nonsample. However, different decisions may be made in different surveys about how the unknowns are handled. For example, in an establishment survey done by mail, the unknowns may all be undeliverable addresses, in which case they all might be coded as out of business and, thus, ineligible.

The mechanics for adjusting for unknown eligibility are usually kept fairly simple. One method of handling the unknowns is to distribute their total sample weight among those whose eligibility status is known. Simple methods are usually used to do this, partly because little may be known about the cases with unknown eligibility and partly because nonresponse is considered to be a more serious problem that should receive more attention. A class-based approach, described below, can be used for unknown eligibility adjustment. The same approach can be used for nonresponse adjustment. We cover ways of forming classes in Sect. 13.5.1. The general idea is to make adjustments to the weights of cases with known status:

1. Form $b = 1, \ldots, B$ classes based on frame information known for all cases. Classes may cut across design strata. In practice, eligibility adjustment and nonresponse adjustment classes may be the same.
2. Let s_b be the set of sample units in class b, regardless of eligibility or response status, and $s_{b,KN} = s_b \cap s_{KN}$ be the set with known eligibility in class b. The symbol $\cap$ identifies the set of units in s_b *and* in s_{KN} (i.e., the intersection).
3. The unknown eligibility adjustment for sample units in class b is $a_{1b} = \frac{\sum_{i \in s_b} d_{0i}}{\sum_{i \in s_{b,KN}} d_{0i}}$, where d_{0i} is base weight.
4. The adjusted weight for unit i in $s_{b,KN}$ is $d_{1i} = a_{1b} d_{0i}$. The factor $1/a_{1b}$ functions as an estimate of the probability of having known status. The weights for the remaining units in class b, those with unknown eligibility, are set to zero, i.e., $a_{1b} = 0$ for $s_{b,UNK} = s_b \cup s_{UNK}$.

Example 13.5 (Ineligibles in a telephone survey). A telephone survey of the members of a campus student organization is conducted. The list is somewhat out-of-date so that some phone numbers are incorrect. Some persons on the list may have dropped out of school and are, therefore, ineligible. A portion of these ineligibles can be identified; 9.1 % of the sample is never contacted so that their eligibility is uncertain. A single adjustment class is used. The sums of the weights for different categories of cases are shown below. The sum of all sample weights is 110 and is 100 for the persons with known eligibility.

Category	$\sum_{i \in s_b} d_{0i}$	Percent dist.	Adjustment	Adj. sum of weights
Eligible respondents (R)	50	45.5	1.1	55
Eligible nonrespondents (NR)	40	36.4	1.1	44
Ineligible (IN)	10	9.1	1.1	11
Unknown eligibility (UNK)	10	9.1		
Total	110			110

The weight of the unknowns ($UNKs$) is assigned to $R{:}NR{:}IN$ in ratios of 5:4:1. Each individual base weight for the cases with known eligibility is increased by the factor $110/100 = 1.1$. Here $B = 1$ and $a_1 = 1.1$. ∎

13.5 Adjustments for Nonresponse

Adjusting for nonresponse can be either simple or elaborate, depending on how much is known about the nonrespondents. Sections 13.5.1–13.5.2 discuss weighting class and propensity scoring methods, along with some approaches to forming classes. First, we sketch some of the thinking needed to select a nonresponse adjustment method. Response can be thought of as either deterministic or stochastic (Kalton and Maligalig 1991):

1. **Deterministic**—Each eligible unit in the population will either respond or not if asked to participate. The choice is not random so that units could be sorted a priori into respondents and nonrespondents.
2. **Stochastic**—Each unit has some nonzero probability of responding. When asked to participate, a unit makes a random choice to cooperate or not.

The bias of a simple mean when there is deterministic response is

$$NR.Bias\left(\hat{\bar{y}}_r\right) = M\left(\bar{Y}_r - \bar{Y}_m\right)/N, \tag{13.1}$$

where $\hat{\bar{y}}_r$ is the estimated respondent mean, $\bar{Y}_r$ is the true mean for the respondent population, $\bar{Y}_m$ is the true mean for the nonrespondents population, and

M/N is the population nonresponse rate calculated as the ratio of the nonrespondent population size, M, over the population size, N. In the deterministic situation, there is a bias if the population mean for the respondents is different from that of the nonrespondents. The idea behind the weighting class adjustment method is to try and group units together in such a way that the class means for respondents and nonrespondents are equal, i.e., $\bar{Y}_r = \bar{Y}_m$.

Conditioning on the response pattern exhibited in the sample, the nonresponse bias in (13.1) can be estimated using base weights (or base weights adjusted for unknown study eligibility), denoted below as d_i, as

$$nr.bias\left(\bar{y}_r\right) = \bar{m}\left(\bar{y}_r - \bar{y}_m\right), \tag{13.2}$$

where $\bar{y}_r = \sum_{i \in s} r_i d_i y_i / \sum_{i \in s} r_i d_i$, the estimated mean of y for the respondent population with $r_i = 1$ if the ith sample member is a respondent ($r_i = 0$ otherwise); $\bar{y}_m = \sum_{i \in s} \left(1 - r_i\right) d_i y_i / \sum_{i \in s} \left(1 - r_i\right) d_i$, the estimated mean within the nonrespondent population; and $\bar{m} = \sum_{i \in s} \left(1 - r_i\right) d_i / \sum_{i \in s} d_i$, the weighted nonresponse rate. Note that the y-values are needed for both respondents and nonrespondents to evaluate (13.2). This usually means that frame variables available for all units must be used.

This kind of thinking carries over to stochastic response, but the algebra is more involved. Despite the added complexity, the stochastic approach underlies most of the nonresponse adjustment techniques used in practice. Define two indicators for being in the sample and responding:

$$I_i = \begin{cases} 1 & \text{if unit } i \text{ selected for sample} \\ 0 & \text{if not} \end{cases}$$

$$R_i = \begin{cases} 1 & \text{if unit } i \text{ responds given that it is in the sample} \\ 0 & \text{if unit } i \text{ does not respond} \end{cases}$$

The probability of being in the sample is $\Pr\left(I_i = 1\right) = \pi_i$ while the probability of responding given that unit i is in the sample is $\Pr\left(R_i = 1 \mid I_i = 1\right) = \phi_i$. Rosenbaum and Rubin (1983) call ϕ_i the propensity score for unit i. If $\phi_i = 0$ for some units, i.e., some units are "hard-core" nonrespondents who would never participate in a survey, this could cause bias. If all units have some nonzero probability of responding, then it may be possible to produce estimates that are, in some statistical sense, unbiased.

Suppose $d_{0i} = 1/\pi_i$ is the base weight we assign to unit i and consider this simple estimator of a mean: $\hat{\bar{y}}_\pi = \sum_{i \in s_R} d_{0i} y_i / \sum_{i \in s_R} d_{0i}$. Under the "quasi-randomization" setup, where sampling and responding are both considered to be random, Kalton and Maligalig (1991) showed the bias of $\hat{\bar{y}}_\pi$ is

$$B\left(\hat{\bar{y}}_\pi\right) \doteq \frac{1}{N\bar{\phi}} \sum \left(y_i - \bar{Y}_U\right)\left(\phi_i - \bar{\phi}\right), \tag{13.3}$$

where $\bar{\phi}$ is the average population probability of responding. In words, the bias depends on the covariance of the response variable and its response propensity. If y_i and ϕ_i are unrelated, there is no bias and nonresponse does not need to be corrected, at least when estimating a mean.

Generally, though, we need to do something to reduce or eliminate bias. One type of unbiasedness that we could strive for is design/response-mechanism unbiasedness. Suppose w_i^* is the weight we assign to unit i after nonresponse adjustment and consider this simple estimator of a total:

$$\hat{T} = \sum_{i \in s_R} w_i^* y_i, \; s_R = \text{set of respondents.}$$

The average of this estimator over sampling and response is

$$\begin{aligned} E_R E_I \left(\hat{T} \right) &= E_R E_I \left(\sum_{i \in U} R_i I_i w_i^* y_i \right) \\ &= \sum_{i \in U} w_i^* y_i E_R E_I \left(R_i I_i \right). \end{aligned}$$

If we can make $w_i^* = 1/E_R E_I \left(I_i R_i \right)$, this reduces to the population total, $\sum_{i \in U} y_i$. Since $E_R E_I \left(I_i R_i \right) = E_I \left[I_i E_R \left(R_i \mid I_i \right) \right] = \pi_i \phi_i$, the weight would be $w_i^* = \left(\pi_i \phi_i \right)^{-1}$. This, of course, requires that both π_i and ϕ_i be nonzero.

Before discussing the techniques for adjusting for unknown eligibility and nonresponse, we need to understand the ideas of *missing completely at random* (MCAR), *missing at random* (MAR), and *nonignorable nonresponse* (NINR). This terminology was introduced by Little and Rubin (2002). Lohr (1999, Sect. 8.4) gives a clear discussion of these ideas; we give a simplified sketch of them here. The definition of each term requires us to think of yet a third distribution—one for an analysis variable Y (in other words, a model for Y). In fact, if K analysis variables are collected on each unit, $y_i = (y_{i1}, y_{i2}, \ldots, y_{iK})$ must be considered. Suppose, also, that there is a set of auxiliary variables $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ available for each sample unit whether it responds or not. These can be items like age, race, and sex in a household survey or type of business and number of employees in a business establishment survey. The x's can also include information used in the sample design, like region of the country and type of area (urban, suburban, or rural) or observations reported by interviewers about the condition of a neighborhood. These observational data are referred to as paradata and are discussed in Kreuter et al. (2010).

However, some caution is required when using some kinds of paradata for nonresponse adjustment. Kreuter and Olson (2011) illustrate that if, say, trash on the streets of a neighborhood or difficulty in finding someone at home, are unrelated to the analysis variables collected in a survey, using those paradata in nonresponse adjustment may do more harm than good. Using irrelevant data may just inject pointless variability into estimates without correcting any bias.

Missing Completely at Random. If the probability of response ϕ_i does not depend on y_i or x_i, then the missing data are MCAR. In our personnel survey in Project 1 (Chaps. 2 and 7), nonresponse would be MCAR if whether a person responded or not did not depend on business unit, salary grade, tenure, or any of the job satisfaction measures collected in the survey. If everyone has the same probability of responding, ϕ, then any nonrespondents are MCAR.

Missing at Random. If the probability of response does not depend on y_i but does depend on some or all of the auxiliaries x_i, then the missing data are MAR. In this case, a model for response can be formed that depends on x_i since we know the auxiliaries for both respondents and nonrespondents. In the personnel survey, response could depend on salary grade—lower-paid workers might want to sound off about their complaints and respond at higher rates than others. This, as Lohr (1999) notes, is sometimes called *ignorable nonresponse*, meaning that if the response mechanism is modeled correctly and adjustments for nonresponse are made, then inferences to the population are possible.

Nonignorable Nonresponse. If the chances of responding depend on one or more analysis variables (i.e., the y's), and this dependence cannot be eliminated by modeling response based on x's that are known for both respondents and nonrespondents, then we have NINR. Suppose, in the personnel survey, we were able to model response as a function of business unit, pay grade, etc. plus an analysis variable that rates whether employees thought there was a clear link between performance rating and pay. If the coefficient on the rating variable was significant, this would be evidence of NINR. The practical problem with fitting this kind of model is that the rating for the nonrespondents will not be available. Consequently, NINR is difficult or impossible to detect except through a nonresponse follow-up study.

13.5.1 Weighting Class Adjustments

If we can create groups or classes where either all units have about the same probability of response or about the same y-values, then the nonresponse bias in (13.1) will be approximately eliminated. Thus, the ideal set of classes will be related to both the y's and the response probabilities as recommended in Kalton and Maligalig (1991) and Little and Vartivarian (2003, 2005). The practical difficulty with this is that the values of the response variables are not available for nonrespondents. Plus, a given set of classes will not be equally effective for all y's. Consequently, a set of classes is usually identified based

on response probabilities. If the covariates used to form the classes are also predictors of y variables, this is a bonus.

In this section, we cover the mechanics of using classes to make nonresponse adjustments. There are different ways of forming classes, which we describe in Sects. 13.5.2 and 13.5.3. We index the classes by $c = 1, \ldots, C$. The goal in forming classes is to put units together that have the same response propensity. As noted above, it is also desirable to have an association between the means of analysis variables and the way the classes are formed. If all units in a class have the same covariate values, x_c, and response propensity is a function of x_c, then $\phi_i = \phi(x_c)$ for all units in c. Denote the set of sample cases in class c as s_c, the set of eligible respondents as s_{ER}, and the set of eligible nonrespondents by s_{ENR}, as in Sect. 13.4. The cases that are known to be eligible in class c are $s_{c,E} = s_c \cap (s_{ER} \cup s_{ENR})$ and the set of eligible respondents in class c is $s_{c,ER} = s_c \cap s_{ER}$. The nonresponse adjustment for units in class c is computed using the unknown-eligibility adjusted weights discussed in Sect. 13.4:

$$a_{2c} = \frac{\sum_{i \in s_{c,E}} d_{1i}}{\sum_{i \in s_{c,ER}} d_{1i}},$$

that is, the ratio of the sum of the input weights for all eligible cases in the class to the sum of the input weights for the eligible respondents in that class. The resulting adjustment a_{2c} is applied only to the respondents in class c. The adjustment is set to zero for the unknowns or known eligible nonrespondents, $s_{UNK} \cup s_{ENR}$, and to one for the cases known to be ineligible, s_{IN}. The weight for unit i in the initial sample, after the adjustments for unknown eligibility and nonresponse, is then

$$d_{2i} = \begin{cases} d_{1i}a_{2c} & i \in s_{c,ER}, \\ d_{1i} & i \in s_{IN}, \\ 0 & i \in s_{UNK} \cup s_{ENR}, \\ d_{0i}a_{1b}a_{2c} & i \in s_{b,KN} \cap s_{c,ER}, \\ d_{0i}a_{1b} & i \in s_{b,KN} \cap s_{IN}, \\ 0 & i \in s_{UNK} \cup s_{ENR}. \end{cases}$$

Thus, eligible respondents get both the adjustment for unknown eligibility and the nonresponse adjustment. Known ineligibles ($s_{KN} \cap s_{IN}$) get only the unknown eligibility adjustment. Unknowns (s_{UNK}) and eligible nonrespondents (s_{ENR}) drop out.

The a_{2c} adjustment does not necessarily have to use the d_{1i} weights. Little and Vartivarian (2003) note that if all units in a nonresponse adjustment class have the same response probability, then an unweighted adjustment, $a_{2c} = n_{c,E}/n_{c,ER}$, will be unbiased with respect to the response model and can give more stable NR adjustments. This will be true even if the d_{1i}'s vary within each class.

The nonresponse adjustment classes may be formed by simply tabulating response rates among the known eligibles in different ways and trying to create classes with different rates. More formal and effective ways of creating classes are to use propensity models or classification algorithms, as described in the next two sections.

13.5.2 Propensity Score Adjustments

As noted in the previous section, an estimator of a total that is unbiased over the combined sampling/response process will result if the weight is $d_{2i} = 1/\pi_i \phi_i$. If $\phi_i = \phi(x_i)$, we can try to model the response probabilities as long as we measure the covariates on all initial sample cases. There are problems when units are not MAR or MCAR. For example, if $\phi_i = \phi(y_i)$, we do not have y's for nonrespondents ($R = 0$). If the nonrespondents follow a different model from the respondents, we will not know it. Another problem case would be $\phi_i = \phi(U_i)$ where U_i contains the unmeasured covariates or measured covariates incorrectly omitted from model. For example, it might be the case that response depends on age, race/ethnicity, and sex, but we omit race/ethnicity. A common situation would be that response depends on a covariate that is not measured on either the respondents or the nonrespondents.

We may fear that we are operating with inadequate information, but, in practice, model parameters must be estimated based on what is known for both respondents and nonrespondents. One approach is to fit a binary regression model for the response indicators R_i. The expected value of the indicator is

$$E_R\left(R_i \mid I_i = 1\right) = \Pr\left(R_i = 1 \mid I_i = 1\right) = \phi\left(x_i\right).$$

This is the conditional probability of response given that a unit is selected for the sample. This also has a bearing on whether to use base weights or not in fitting the model, as discussed later in this chapter.

Response as a Latent Variable Process

An interesting feature of this problem is that responding to a survey can be modeled as a realization of a latent variable process. This line of thought provides some motivation for the binary regression models that are often used to model response to a survey. The indicator R_i is the *manifest* variable (the one we see). Suppose that there is a *latent* variable R_i^* that is continuous but unobserved. If the value of R_i^* exceeds some threshold (say, bigger than some θ), unit i responds; otherwise, it does not. The latent variable is a unit's "motivation" to participate.

Other examples that can be modeled as latent variable processes are the decision to reenlist in the military and the decision to vote for some candidate for political office. In the former case, we see whether a person reenlists or not. Why or why not may require consideration of job satisfaction, family situation, potential future income after leaving the military, job skills, age, time in service, etc. Voting for a candidate may depend on the voter's perception of the candidate's honesty and the candidate's promises to improve schools or decrease crime. In the end what is observed is which candidate gets a person's vote.

To frame this mathematically, suppose that R_i^* is symmetrically distributed. Figure 13.2 illustrates the situation. If the unobservable R_i^* exceeds a threshold, then the unit responds; otherwise, it is a nonrespondent.
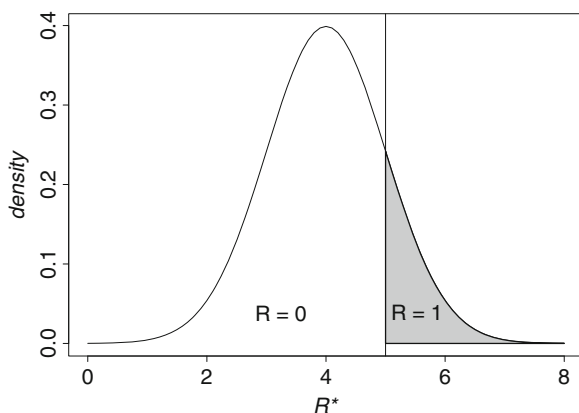
Fig. 13.2: Density of the latent variable for survey response.

Suppose the latent variable follows a linear model, $R_i^* = x_i^T \beta + u_i$, where u_i has distribution function F (not necessarily normal). Then, the probability of response, given selection for the sample, is

$$\phi(x_i) = \Pr(R_i = 1 | I_i = 1)$$
$$= \Pr(R_i^* > \theta).$$

Location of the R_i^* distribution is arbitrary, so we can set $\theta = 0$ or think about $R_i^* - \theta$, which has the same variance as R_i^*. The response probability can then be written as

$$\phi(x_i) = \Pr(R_i^* > 0) = \Pr(x_i^T \beta + u_i > 0)$$
$$= \Pr(u_i > -x_i^T \beta)$$
$$= 1 - F(-x_i^T \beta) = F(x_i^T \beta)$$

assuming a symmetric distribution F for u_i. Using different F distributions, leads to different binary regression models.

The *link function* is a transformation that will turn the probability into a linear function of the covariates, x_i. The link is determined by $F^{-1}[\phi(x_i)] = x_i^T\beta$. Thus, the link gives a quantity, $F^{-1}[\phi(x_i)]$, that is modeled as a linear combination of covariates, $x_i^T\beta$. The equation

$$F^{-1}[\phi(x_i)] = F^{-1}[E_R(R_i | I_i = 1)]$$
$$= x_i^T\beta$$

is called a *generalized linear model*. Some examples are the logistic, probit, and complementary log-log models.

Probit Model

In probit, the probability is modeled as being equal to the value of the cumulative normal distribution function, $\phi(x_i) = \Phi(x_i^T\beta)$. Here, $\Phi = F$ is standard normal distribution function, i.e., $u_i \sim N(0, 1)$. The probit link is $\Phi^{-1}[\phi(x_i)] = x_i^T\beta$, i.e., the inverse Gaussian cumulative distribution function or Gaussian quantile function. The link values have a range of $(-\infty, \infty)$ because they are quantiles of the standard normal distribution.

Logistic Regression

In a logistic regression model, $\phi(x_i) = \frac{\exp(x_i^T\beta)}{1+\exp(x_i^T\beta)}$ and the F^{-1} link is the *logit*, defined as

$$\log\left(\frac{\phi(x_i)}{1 - \phi(x_i)}\right) = x_i^T\beta.$$

Logits have a range of $(-\infty, \infty)$. The shape of the logistic distribution, $F(u) = \exp(u)/[1 + \exp(u)]$, is similar to normal distribution but with heavier tails as u ranges over $(-\infty, \infty)$. The logistic distribution has mean 0 and variance $\pi^2/3$.

Complementary Log-Log (c-log-log)

The probability of response in a complementary log-log model is $\phi(x_i) = 1 - \exp[-\exp(x_i^T\beta)]$. This is also called a log-Weibull distribution. The complementary log-log link is

$$\log\{-\log[1 - \phi(x_i)]\} = x_i^T\beta.$$

Using this model is equivalent to assuming that the error term in the latent variable model has what is called an "extreme value" distribution:

$$F(u_i) = e^{-e^{-u_i}}.$$

The extreme value distribution has mean of about -0.577 (known as Euler's constant) and variance $\pi^2/6$ (Weisstein 2010).

There are some differences in these distributions, but they are not extreme. Figure 13.3 shows probabilities on the vertical axis graphed versus standardized links on the horizontal axis. The standardized link for each distribution is defined as $[u - E(u)]/\sigma_u$. Probit and logit are almost identical while c-log-log has more probability at lower values of the link function.

Example 13.6 illustrates how to estimate response probabilities in R. A choice must be made whether to use the survey base weights when estimating the model parameters. Since probabilities *conditional on being selected for the sample* are desired, this implies that unweighted regressions should be fit. If the base weights were used, then the parameters estimated would be for the census-fit model, i.e., those that would be estimated if the entire population was in hand. If $\Pr(R_i = 1 | I_i = 1) = \Pr(R_i = 1)$, then the unweighted and weighted estimators would aim at the same quantities. However, even in that case, using variable base weights can give estimators with higher variances— a point illustrated by Little and Vartivarian (2003) in the context of class nonresponse adjustments.
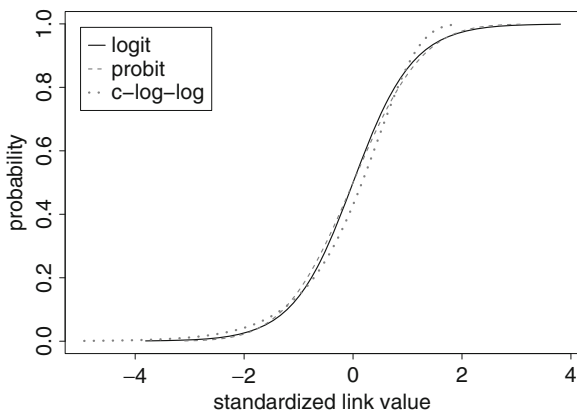
Fig. 13.3: Graph of probabilities versus standardized links for logit, probit, and c-log-log models.

Example 13.6 (Unweighted models). The 2003 NHIS (`nhis.RData`) data consists of 3,911 cases. We identified nonrespondents as persons who answered the question on personal income as Refused, Not Ascertained, and Don't Know or who reported their income only as above or below $20K. The `resp`
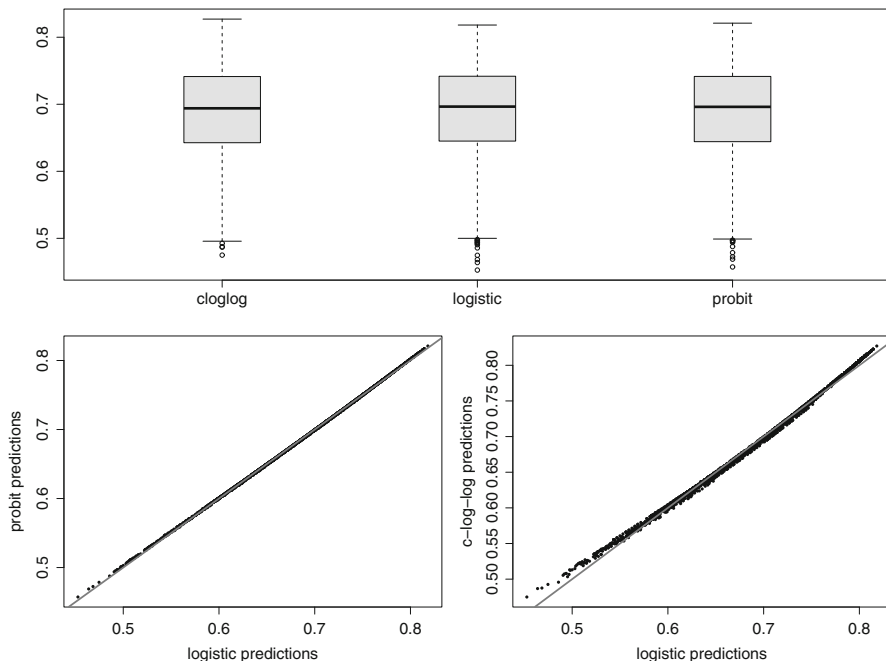
Fig. 13.4: Comparisons of predicted probabilities from logistic, probit, and complementary log-log models for response. A 45° line is drawn in the second row where the probabilities would be equal.

variable has values of 0 for nonrespondents and 1 for respondents. About 31 % are nonrespondents by this criterion. We fit logit, probit, and c-log-log models using the following covariates:

age	Age (continuous)
educ_r	Education recode (1 = High school, general education development degree (GED), or less, 2 = Some college 3 = Bachelor's or associate's degree 4 = Master's and higher)
hisp	Hispanic ethnicity (1 = Hispanic, 2 = non-Hispanic)
parents_r	Parent(s) of sample person present in the family (1 = Yes, 2 = No)
race	Race (1 = White, 2 = Black, 3 = Other)

The code for fitting an (unweighted) logistic model in R follows. Some of the output is in Table 13.1. The variables hisp, parents, and race are treated as R factor variables (class variables in SAS). R automatically creates dummy variables and omits the first level of each (reference level) to compute parameter solutions:

```
   # logistic regression
glm.logit <- glm(resp ~ age +
            as.factor(hisp) +
            as.factor(race) +
            as.factor(parents_r) +
            as.factor(educ_r),
            family=binomial(link = "logit"),
            data = nhis)
 summary(glm.logit)

   # extract link values
 L.hat <- glm.logit$linear.predictors
   # transform link values to probability scale
 pred.logit <- exp(L.hat) / (1 + exp(L.hat) )
```

To fit probit and c-log-log models, use

```
 family=binomial(link = "probit")
 family=binomial(link = "cloglog")
```

in the call to glm. Suppose the resulting models are stored in the objects glm.probit and glm.cloglog. To link values to predicted probabilities:

```
 L.hat <- glm.probit$linear.predictors
 pred.probit <- pnorm(L.hat)
```

or

```
 L.hat <- glm.cloglog$linear.predictors
 pred.cloglog <- 1 - exp(-exp(L.hat) )
```

The AIC values for the three models are: logistic, 4777.2; probit, 4777.1; and c-log-log, 4777.1—implying that all three fit equally well, at least by the AIC measure. Figure 13.4 shows boxplots of the predicted probabilities from the three models and scatterplots of the probit and c-log-log predictions versus the ones from the logistic model. These plots also confirm that the three models produce very similar results in this example.

The same unweighted models can also be fitted in SAS using proc genmod, the procedure designed to analyze data through a generalized linear model with a specified link function:

```
proc genmod data = nhis;
class hisp race parents_r educ_r;

model resp = age hisp race parents educ_r
            / dist = binomial link = logit
                  /* or probit or cloglog */ ; run;
```

■

The same models can be run with the base weights using `svyglm` in the R `survey` package (Lumley 2012). Because base weights are not available in the NHIS public use file, we have used the final survey weights (`svywt`) for illustration.

Example 13.7 (Weighted models). The R code for fitting the logistic model is shown below. First, a survey design object is created with `svydesign`:

```
require(survey)
nhis.dsgn <- svydesign(ids = ~psu, strata  = ~stratum,
                data    = nhis,
                nest    = TRUE,
                weights = ~svywt) # Note: base wt should
                                  # be used if available

glm.logit <- svyglm(resp ~ age + hisp + race +
                parents_r + educ_r,
            family = binomial(link = "l
            design = nhis.dsgn)
```

The weighted parameter estimates are shown in Table 13.1 along with the unweighted values from Example 13.6. The same parameters are significant in both the weighted and unweighted, albeit at different levels. ∎

Figure 13.5 plots the predicted response probabilities from the weighted models versus those from the unweighted models in each of the three models. The overall response rates for this data set are 69.0 % (unweighted) and 70.4 % (weighted). The survey-weighted predictions in Fig. 13.5 are mainly somewhat higher than the unweighted predictions, consistent with higher overall estimated response rate.

In SAS, `proc surveylogistic` can be used to compute weighted estimates of probabilities. SAS does not have procedures for fitting probit and c-log-log models with survey data, although this is probably no real limitation since logistic is used most often. One could use `proc genmod` with weights

Table 13.1: Unweighted and weighted parameter estimates from logistic models.

Parameter	Unweighted Estimate	z value	Pr(>|z|)		Survey weighted Estimate	z value	Pr(>|z|)	
(Intercept)	0.583	4.63	0.000	***	0.667	4.00	0.000	***
Age	−0.013	−5.74	0.000	***	−0.013	−5.74	0.000	***
as.factor(hisp)2	0.306	3.36	0.001	***	0.220	1.76	0.083	.
as.factor(race)2	−0.160	−1.61	0.109		−0.214	−1.61	0.111	
as.factor(race)3	−0.374	−2.31	0.021	*	−0.449	−2.23	0.028	*
as.factor(parents_r)2	0.522	4.74	0.000	***	0.547	4.84	0.000	***
as.factor(educ_r)2	0.249	2.54	0.011	*	0.341	3.07	0.003	**
as.factor(educ_r)3	0.346	3.79	0.000	***	0.383	3.99	0.000	***
as.factor(educ_r)4	0.276	1.94	0.052	.	0.310	2.15	0.035	*
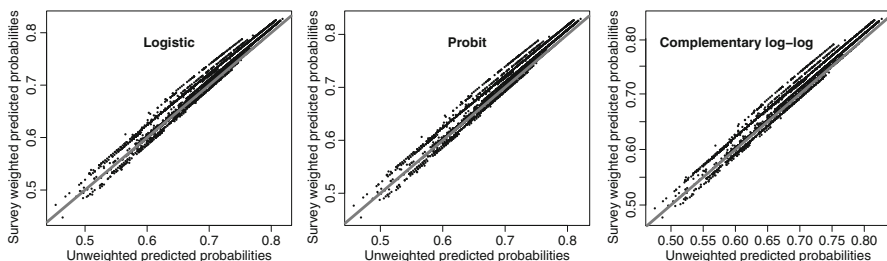
Fig. 13.5: Comparison of unweighted and weighted predicted probabilities from logistic, probit, and complementary log-log models. A 45° line is drawn in each panel.

to get point estimates for probit and c-log-log; the standard errors (and thus the test for significance) will not account for design features like clustering.

Use of Estimated Propensities for Nonresponse Adjustment

Response propensities can be used for nonresponse adjustments either individually or by grouping units into classes. The options are:

1. *Propensity weighting*—Adjust the weight for an individual responding unit by $1\big/\hat{\phi}_i$ with $\hat{\phi}_i$ computed from a binary regression.
2. *Propensity stratification*—Use the $\hat{\phi}_i$'s to create classes and make a common adjustment within each class to all respondents.

Propensity weighting was discussed in the previous sections. The use of propensity stratification was introduced by Rosenbaum and Rubin (1983) for observational studies. Propensity scores have found many uses, particularly in causal inference (Stuart 2010). In an observational study there may be a "treatment" and a "control" group, but no randomization to groups is used. With this kind of "found data" there can be many differences between the compositions of the groups that make inference difficult. For example, we may collect data on smokers and nonsmokers and measure the outcome variable lung cancer. Smokers and nonsmokers may differ on many covariates other than just whether or not they smoke. An observed difference in the rates of lung cancer in the two groups may be due to something other than smoking unless the effect of covariates can be "adjusted out" some way.

One way of doing the adjustment is to create classes. The general goal in class creation is to group units that have the same or very similar propensities of being in the "treatment" group (e.g., smokers). Units in a class will have the same configuration of covariates, or, at least, about the same $\phi(x)$, which summarizes the effect of the covariates. In theory, the difference between the estimated treatment and control means is unbiased for sets of units with same propensity score. Within each class, units are treated as if they were randomized to treatment or control since each has the same probability $\phi(x)$

of treatment. For example, we could group persons who have similar propensities for smoking. Then, compute proportions of smokers and nonsmokers with lung cancer in each class. The idea is that any differences in covariates like age, race/ethnicity, and social class have been adjusted out within each group because $\phi(x_i)$ summarizes the effects of the covariates (at least the ones in the model) and the $\phi(x_i)$'s are close to each other for all units in a group.

In the response/nonresponse case, the respondents are equivalent to the treated and the nonrespondents to the controls. The probability of treatment is the probability of responding. We create classes so that each unit in a class has same or similar probability of responding and make the same nonresponse adjustment to each respondent in a given class. Little (1986) was the first to suggest this for nonresponse adjustment; Czajka et al. (1992) give an example using tax returns.

Propensity Stratification: Creating Classes

First, a binary regression model is fit using covariates available for both respondents and nonrespondents. Ideally, these covariates are related to both the propensity to respond and the y's being measured. In practice, the set of available x's can either be extensive or quite limited, depending on the type of survey. In an employee satisfaction survey, like the one in Project 1, quite a bit may be known about all units in the sample. Panel surveys may also have data on nonrespondents in later waves if units responded in an early wave and provided some data. In a telephone survey, almost nothing may be known for the nonrespondents, other than, possibly, the geographic location of the phone number. Even this is changing in the USA, where mobile phone users can retain the same phone number wherever they move.

The general steps in class formation are:

1. Calculate $\hat{\phi}(x_i)$ for each unit in the sample used for modeling.
2. Sort the file by $\hat{\phi}(x_i)$—low to high.
3. Form classes with about the same number of initial (respondents + nonrespondents) sample units in each.

Five classes are usually recommended based on some analyses in Cochran (1968). With a large sample, there is no reason not to create more classes. This can help make each more homogeneous on covariates and propensity scores. More classes may decrease bias due to nonresponse but may increase variances by creating bigger spread in the weights. We will address the question of whether variation in weights increases variances of estimators in more detail in Chap. 14.

If the range of $\hat{\phi}(x_i)$ in each class is small, then using a single propensity value for each class is reasonable. In some data sets, there may be clumping of estimated probabilities, i.e., groups of units that have about the same

$\hat{\phi}(x_i)$. If there is separation among the groups, then creating classes with the same number of units may be a bad idea since you would mix units with different response propensities. There are several options for computing a single adjustment in each class c:

1. $\hat{\phi}_c = \sum_{i \in s_c} \hat{\phi}(x_i) \big/ n_c$, unweighted average estimated propensity where n_c is the unweighted number of cases in class c
2. $\hat{\phi}_c = \sum_{i \in s_c} d_i \hat{\phi}(x_i) \big/ \sum_{i \in s_c} d_i$, weighted average estimated propensity; where d_i is the input weight to the NR step and $\sum_{i \in s_c} d_i = \hat{N}_c$, the estimated number of population units in class c
3. $\hat{\phi}_c = n_{cR}/n_c$, unweighted response rate where n_{cR} is the unweighted number of respondents in class c
4. $\hat{\phi}_c = \sum_{i \in s_{cR}} d_i \big/ \sum_{i \in s_c} d_i$, weighted estimate of response rate
5. $\hat{\phi}_c = median \left[\hat{\phi}(x_i) \right]_{i \in s_c}$, unweighted median estimated propensity

If every unit in a class has the same probability of responding, i.e., the grouping is very effective, then (3) $\hat{\phi}_c = n_{cR}/n_c$ is best [see Little and Vartivarian (2003)]. If $\hat{\phi}(x_i)$'s vary within a class, (1) or (2) can be used. The fourth choice, $\hat{\phi}_c = \sum_{i \in s_{cR}} d_i \big/ \sum_{i \in s_c} d_i$, is an estimate of the population response rate in class c assuming MAR. This estimate is approximately unbiased with respect to the compound sampling/response mechanism or with respect to a model with a common response probability within each class. The fourth choice can be inefficient if weights vary much within class and units have a common ϕ_i. Choice (5), the median, might be considered if the response probabilities do vary quite a lot within a class or the distribution of the estimated probabilities is skewed. We will compare these options in an example below. In many applications, the options will give very similar answers.

Checking Balance on Covariates

D'Agostino (1998) gives a simple method for checking covariate balance within the classes formed in propensity stratification. After classes are formed, the idea is to make a check on the extent of differences in the covariate means. The covariate means should be different between the classes, but within a class, the means of the covariates should be the same for respondents and non-respondents. The latter condition is consistent with the response propensity being the same for all units within a class. Suppose classes are formed based on quintiles of $\hat{\phi}(x_i)$ giving five classes. Define a variable p.class with five values and treat it as a factor, i.e., `p.class <- as.factor(seq(1:5))`. Also, define the indicator variable `resp` = 1 if a unit is a respondent and 0 if an NR. Next, fit models for the mean of each covariate using `p.class` and `resp` as predictors:

- For quantitative x's, fit an analysis of variance (ANOVA) model, `x = p.class resp p.class*resp`.
- For dichotomous x's, fit a logistic model, `logit(x) = p.class resp p.class*resp`

The coefficients on resp and the interaction term `p.class*resp` should be nonsignificant if covariate means do not differ for Rs and NRs within quintile class. The coefficients on `p.class` should be nonzero and different from each other since units with different values of the propensities, and, consequently, the covariates, go into the different classes. Another simple, descriptive step is to look at the covariate means in a `p.class*resp table`. Balance checking is also relevant to other types of studies. For example, Harder et al. (2010) discuss balancing in causal inference in psychological studies.

Example 13.8 (Form classes from propensities). Continuing with the NHIS analyses given in Example 13.6, the object `pred.logit` holds predicted response probabilities from the unweighted logistic model. The R code below divides the persons into five classes and checks the count of persons per class:

```
    # Determine quintiles of response propensities
 quintiles <- quantile(pred.logit, probs = seq(0,1,0.2))
          # Create a factor to hold the class IDs
          # include.lowest=TRUE makes sure the smallest
          #    propensity is assigned to a class
 p.class <- cut(round(pred.logit,3), breaks = quintiles,
                include.lowest=TRUE)
 table(p.class)
 (0.453,0.631] (0.631,0.677] (0.677,0.714] (0.714,0.752]
           778           773           788           786
 (0.752,0.818]
           786
```

Next, we compare the five ways of estimating the class response propensity:

```
        # (1) Unweighted avg response propensity
by(data = pred.logit, p.class, mean)
        # (2 )Weighted response propensity
by(data = data.frame(pred.logit,
   wt = nhis[,"svywt"]),
   p.class,
   function(x) {weighted.mean(x$pred.logit, x$wt)})
       # (3) Unweighted response rate
 by(as.numeric(nhis[, "resp"]), p.class, mean)

       # (4) Weighted response rate
 by(data = data.frame(resp = as.numeric(nhis[,"resp"]),
   wt = nhis[,"svywt"]), p.class,
   function(x) {weighted.mean(x$resp, x$wt)})

       # (5) Unweighted Median response propensity
 by(pred.logit, p.class, median)
```

Table 13.2 lists the propensity values that would be used for each class based on the five methods above. In the NHIS data, all methods give similar results. Although all five methods give monotonically increasing propensity values across the classes, this does not have to be true. The unweighted and weighted response rates, in particular, do not have to increase from class 1 to 5, even though the estimated model propensities do.

Table 13.2: Five methods of estimating response propensities within classes based on fitting a logistic model to the NHIS data.

			(1)	(2)	(3)	(4)	(5)
Class	Boundaries	Count	Unweighted	Weighted	Unweighted	Weighted	Median
		of	avg.	avg.	RR	RR	
		persons	propensity	propensity			
1	[0.453,0.631]	778	0.588	0.591	0.589	0.591	0.595
2	(0.631,0.677]	773	0.655	0.655	0.662	0.679	0.657
3	(0.677,0.714]	788	0.696	0.696	0.694	0.702	0.696
4	(0.714,0.752]	786	0.732	0.732	0.707	0.717	0.733
5	(0.752,0.818]	786	0.777	0.778	0.796	0.804	0.775
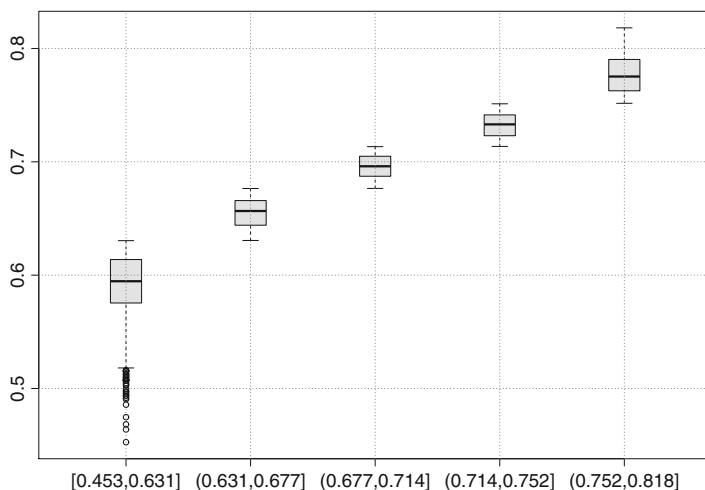
Fig. 13.6: Boxplots of predicted probabilities based on logistic regression after sorting into five propensity classes.

Figure 13.6 shows boxplots of the logistic regression probabilities within each of the five propensity classes. The horizontal line in each box is the unweighted mean of the propensities in the class. The class with the smallest propensities has a larger range than the others, which is typical in

these applications. The range of propensities in the last four classes is much shorter. Using the mean or another single value to adjust for nonresponse will eliminate the more extreme adjustments. For example, in the first class, the smallest estimated propensity is 0.453 whose inverse is 2.21. The unweighted mean in that class from Table 13.2 is 0.588 with an inverse of 1.70. Thus, using the mean would reduce the adjustment by about 23 %.

We illustrate a check on covariate balance by fitting an ANOVA model to age, which is continuous. We do not use the survey weights below since the interest is in whether balance has been achieved in the sample that was selected. Checks could be made using the weights, in which case the check would be on whether the census-fit model shows evidence of balance:

```
chk1 <- glm(age ~ p.class + resp + p.class*resp,
            data = nhis)
summary(chk1)
Coefficients:
                            Estimate  t value  Pr(>t)
(Intercept)                    56.12   63.06   < 2e-16 ***
p.class(0.631,0.677]           -7.84   -5.91   3.79E-09 ***
p.class(0.677,0.714]          -11.36   -8.37   < 2e-16 ***
p.class(0.714,0.752]          -12.98   -9.43   < 2e-16 ***
p.class(0.752,0.818]          -23.12  -15.00   < 2e-16 ***
resp                           -0.06   -0.05     0.957
p.class(0.631,0.677]:resp      -0.01   -0.01     0.994
p.class(0.677,0.714]:resp      -1.24   -0.73     0.464
p.class(0.714,0.752]:resp       0.22    0.13     0.896
p.class(0.752,0.818]:resp       1.57    0.86     0.390
Signif. codes:
  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In this case, the `p.class` factors all have coefficients that are significant while the `p.class*resp` interactions are not—the desired outcomes if mean age differs between classes but is the same for respondents and nonrespondents within a class. Another check is to fit a second model that includes only `p.class` and to test whether the models are equivalent:

```
chk2 <- glm(age ~ p.class, data = nhis)
anova(chk2, chk1, test="F")
```

The F-statistic is 0.457 with 3,906 and 3,901 degrees of freedom and has a p-value of 0.8084. Thus, the model without a factor for responding is judged to be adequate.

Balance on Hispanic ethnicity can be checked with a logistic regression after recoding `hisp = 1, 2` to a new binary variable `new.hisp=0,1`:

```
new.hisp <- abs(nhis$hisp-2)
chk1 <- glm(new.hisp ~ p.class + resp + p.class*resp,
            family=binomial(link = "logit"),
            data = nhis)
summary(chk1)
Coefficients:
```

```
                         Estimate    z    Pr(>z)
(Intercept)                 -0.51 -4.4    9.69E-06 ***
p.class(0.631,0.677]        -0.17 -0.9    0.344
p.class(0.677,0.714]        -0.76 -3.9    9.64E-05 ***
p.class(0.714,0.752]        -1.84 -7.1    1.75E-12 ***
p.class(0.752,0.818]        -2.92 -6.2    4.51E-10 ***
resp                         0.08  0.5    0.612
p.class(0.631,0.677]:resp   -0.25 -1.1   0.253
p.class(0.677,0.714]:resp   -0.05 -0.2   0.842
p.class(0.714,0.752]:resp   -0.04 -0.1   0.901
p.class(0.752,0.818]:resp   -0.99 -1.7   0.097
```

Three of the four coefficients for p.class are significant while the
p.class*resp interactions are not (at least at the 0.05 level). Outcomes
for race, parents_r, and educ_r are similar. Note that race and educ_r
need to be recoded to binary variables to use logistic. We can also fit a second
model with only p.class to compare to the one above:

```
chk2 <- glm(new.hisp ~ p.class,
            family=binomial(link = "logit"),
        data = nhis)
anova(chk2, chk1, test="Chisq")
```

The ANOVA statement tests whether the two models defined by chk1 and
chk2 are equivalent in the sense have having the same value of -2 log-
likelihood. The chi-square statistic is 3.7566 with five degrees of freedom and
has a p-value of 0.585. Consequently, balance was obtained on Hispanic also.

Numerical Note

The logistic model for checking covariate balance on parents has anomalous
results that seem to occur fairly often in practice and are worth a comment.
The model

```
 new.par <- abs(nhis$parents_r-2)
 chk <- glm(new.par ~ p.class + resp + p.class*resp,
            family = binomial(link = "logit"),
            data = nhis) summary(chk)
```

leads to an extremely large standard error on p.class(0.752,0.818]
and the interaction term p.class(0.752,0.818]:resp. The output from
this model is

```
Coefficients:
                         Estimate     SE     z Pr(>z)
(Intercept)                 -0.54   0.12 -4.64 3.49E-06   ***
p.class(0.631,0.677]        -1.11   0.20 -5.45 5.10E-08   ***
p.class(0.677,0.714]        -1.81   0.26 -7.07 1.51E-12   ***
p.class(0.714,0.752]        -3.27   0.47 -7.00 2.49E-12   ***
p.class(0.752,0.818]       -19.03 850.18 -0.02 0.982
resp                        -0.22   0.15 -1.45 0.148
p.class(0.631,0.677]:resp   0.47   0.25  1.85 0.064
```

```
p.class(0.677,0.714]:resp  0.38    0.31  1.23 0.220
p.class(0.714,0.752]:resp -1.19    0.75 -1.58 0.113
p.class(0.752,0.818]:resp  0.22 952.65  0.00 1.000
```

This is a symptom of "quasi-complete" separation in the data set, i.e., there
are one or more observations with a predicted probability equal to or near 1.
In this situation, a parameter estimate will diverge to infinity. In the NHIS
example, there are no cases in class 5 where the parents of the sample person
live in the home (`parents_r=1`). This does not have any bad effects on the
formation of propensity classes themselves. Since the value of the `parents_r`
covariate is the same for everyone assigned to that class, the goal of equalizing
the covariate values was accomplished. (Of course, if this problem cropped
up in logistic analysis where the goal is to find covariates related to the
occurrence of some characteristic, it would have to be addressed.)

The SAS code to create classes and check balance on the `age` and `hisp`
covariates is shown below. Here the logistic procedure is used, but `proc
genmod` is also an option as illustrated earlier in this chapter. The output
is not listed, although SAS users are encouraged to test the code using the
NHIS data:

```
proc import out= work.nhis
            datafile= "C:\nhis.csv"
            dbms=csv replace;
     getnames=yes;
     datarow=2;
run;
  * Model probability of response ;
proc logistic data = nhis;
   class hisp (ref = '1')
     race (ref = '1')
     parents_r (ref = '1')
     educ_r (ref = '1') / param=ref;
    model resp (event = '1') =
       age hisp race parents_r educ_r;
    output out = preds pred = pr;
 run;

  * Create quint1les based on the estimated propensity score;
proc rank groups = 5 out = r;
   ranks rnks;
   var pr;
run;

data a;
   set r;
   pclass = rnks + 1;
run;

* Show the breakdown of units by propensity class and response;
proc freq data = a;
   tables pclass * RESP;
run;
```

```
* Perform 2-way ANOVA and logistic regressions to determine   ;
* whether difference in covariate means was removed ;
* by creating classes.    ;
proc glm data = a;
   class pclass;
   model age = pclass resp pclass * resp;
run;

proc logistic data = a;
   class pclass (ref = '1')
         resp   (ref = '0');
   model hisp (event = '2') = pclass resp pclass * resp;
run;
```

Special Cases of Response Propensity Models

Seeing what a propensity model reduces to in some special cases is instructive. First, consider a model with main effects and interactions. For example, suppose we have gender (male, female) and race/ethnicity (Hispanic, non-Hispanic White, non-Hispanic Black, non-Hispanic other race/ethnicity). This gives $2 \times 4 = 8$ levels or classes in the model. The latent variable model defined without an intercept term for unit i in level j of gender and level k of race/ethnicity is

$$R_i^* = \alpha_j + \beta_k + (\alpha\beta)_{jk} + u_i,$$

where
   α_j is effect of jth level of gender (male or female)
   β_k is effect of kth level of ethnicity (Hispanic, non-Hispanic White, non-Hispanic Black, non-Hispanic other)
   $(\alpha\beta)_{jk}$ is an effect for the interaction of gender and ethnicity
   u_i is an error term assumed to have a standard normal distribution, i.e.,
      $u_i \sim N(0, 1)$
(If u_i has variance $\sigma^2 \neq 1$, then the model can be recast in terms of R_i^*/σ. Since the latent variable is unobservable, this is not an assumption that can be checked anyway.) This model would be fitted by (implicitly) creating dummy variables for every level of the interaction of gender with race/ethnicity (gender $\times$ race/ethnicity).

This is equivalent to a gender $\times$ race/ethnicity class adjustment model, as we now show. Suppose logistic regression is used to predict the response probability for a person with gender j and race/ethnicity k:

$$\phi(x_i) = \exp\left(\alpha_j + \beta_k + (\alpha\beta)_{jk}\right) \Big/ \left[1 + \exp\left(\alpha_j + \beta_k + (\alpha\beta)_{jk}\right)\right].$$

This probability is same for every unit i in each jk combination. This leads to the estimate of $\phi(x_i)$ being $n_{jk}^R \big/ n_{jk}$ in the unweighted case. That is, the estimated probability for units jk is simply the proportion of the sample in the class that are respondents, i.e., the unweighted response rate.

If survey weights are used, we get $\sum_{i \in s_{jk}^R} d_i \big/ \sum_{i \in s_{jk}} d_i$, the weighted response rate, with s_{jk} the set of sample units in level jk and s_{jk}^R the set of responders. If the appropriate model includes only main effects, so that $\phi(x_i) = \exp(\alpha_j + \beta_k) / [1 + \exp(\alpha_j + \beta_k)]$, this does not reduce to the class model, and the logistic predictions must still be used instead of a weighting class adjustment.

If no x's are significant in the model and an intercept-only is the best model, then this is evidence that the nonresponse is MCAR. If so, then one overall NR adjustment is appropriate.

Pros and Cons: Class Adjustment Versus Propensity Modeling

An obvious question is which method is better: class adjustment with classes defined by crosses of categorical variables or propensity modeling with adjustments based on either individual response propensities or propensity classes. Propensity modeling can be more flexible than class adjustment for several reasons, including:

- Categorical variables do not have to be completely interacted.
- Continuous x's can be used either by themselves or in combination with categorical variables.
- Explicit modeling can be done to decide what variables should be included.

The modeling may, of course, lead to choosing class adjustment if the model includes only main effects for categorical variables and all interactions. If little is known about nonrespondents, propensity modeling will not gain much over class adjustment and may be equivalent. In household surveys, for example, few person or household-specific items may be available. Neighborhood data may be more common.

At the other extreme, if there are a number of variables available for respondents and nonrespondents, propensity models may give a fairly wide range of estimated probabilities. In that case, grouping propensities into classes, as described above, will lead to less spread in the weight adjustments. On the other hand, if the model fits well, weight adjustments that are inverses of the individual propensities will eliminate bias while class adjustments may not. Class adjustment will eliminate bias if all units in the class have the same response probability, but if each unit has a separate response probability, a class adjustment may be too coarse to eliminate bias. (The bias referred to here is over the sampling and response distributions.) Since the models are usually not trusted completely, using propensity stratification is more common in practice.

A final computational point is this: a factor and its levels can be simple or elaborate. For example, a simple factor is gender (male, female). The equivalent to crossing two simple categorical variables would be to create a single variable that could take all values in the cross. For example, gender $\times$

race/ethnicity with levels (male, female) × (Hispanic, non-Hispanic White, non-Hispanic Black, non-Hispanic other) could be coded as a single variable with eight levels. This allows some flexibility in using a set of variables that are partially interacted in a model. As an illustration, we might have gender × race/ethnicity and gender × education as two factors in a propensity model.

13.5.3 Classification Algorithms

Another method of forming classes for nonresponse adjustments is via a classification algorithm. The idea of mathematically classifying units based on characteristics was introduced by Morgan and Sonquist (1963). Many algorithms are now available including classification and regression trees (CART; Breiman et al. (1993)), support vector machines (Vapnik 1995), and chi-square automatic interaction detection [CHAID, (Kass 1980)]. We will cover the CART algorithm which is available in the R package rpart (Therneau et al. 2012). The goal will be to classify units as respondents or nonrespondents based on covariates available for all sample cases. Thus, the input data are the same as for propensity modeling. Some of the primary applications of classification algorithms are in constructing decision trees. One of the more well-known examples is whether the space shuttle pilot should use the autolander or land manually (Michie 1989; Venables and Ripley 2002) based on wind direction and speed, visibility, and other factors.

In the nonrespondent application, the decision tree will classify cases using available covariates into classes that are related to their likelihood of being respondents. Advantages of CART compared to propensity modeling are that:

1. Interactions of covariates are handled automatically.
2. The way in which covariates enter the model does not have to be made explicit.
3. Selection of which covariates and associated interactions should be included is done automatically.
4. Variable values, whether categorical or continuous, are combined (grouped) automatically.

Judkins et al. (2005) and Rizzo et al. (1996) are two papers that compare propensity modeling and tree algorithms for nonresponse adjustment. As in the previous section, we want to form classes so that we can claim that we have MAR, i.e., given the x's that define classes, all units have the same response probability.

The following R code uses the package rpart to identify a tree using the NHIS data set based on the same variables as in the propensity model:

```
require(rpart)
set.seed(15097)
nhis <- data.frame(nhis)
t1 <- rpart(resp ~ age + hisp + race + parents_r + educ_r,
       method = "class",
       control = rpart.control(minbucket = 50, cp=0),
       data = nhis)

print(t1, digits=2)
plot(t1, uniform=TRUE, compress=TRUE, margin = 0.1)
text(t1, use.n=TRUE, all=TRUE,
     digits=4,
     cex=1.2,
     pretty=1.2,
     fancy=TRUE,
     xpd = TRUE,
     font = 3)
```

There is some randomness in how the algorithm determines the tree. The set.seed statement forces the internal random number generator to start in a particular place, which permits results to be reproduced in different runs of the same code.

The parameter minbucket = 50 in rpart.control requires that there be at least 50 cases (respondents + nonrespondents) in each final grouping of variable values known as a terminal node of the tree. The parameter cp=0 is a complexity parameter that prevents splits from being made unless the measure of fit improves by at least cp units. Venables and Ripley (2002) explain in some detail the criteria used by rpart to fit the tree. In our application, at each step, a split is found, based on one covariate from the available set of covariates that maximizes the log-likelihood of being a respondent. Although the default value of cp=0.1 seems small, we have found that rpart will often not construct a tree at all with the default. Setting cp=0, i.e., no penalty for complexity, may be necessary to construct a useful set of classes. The function print gives a fairly compact listing of the details of the tree:

```
 print(t1, digits=2)

 node), split, n, loss, yval, (yprob)
       * denotes terminal node
  1) root 3911 1200 1 (0.31 0.69)
    2) educ_r< 1.5 1964  690 1 (0.35 0.65)
      4) age>=56 588  240 1 (0.41 0.59) *
      5) age< 56 1376  450 1 (0.33 0.67)
       10) parents_r< 1.5 277  110 1 (0.40 0.60)
         20) age>=32 67   31 0 (0.54 0.46) *
         21) age< 32 210   75 1 (0.36 0.64) *
       11) parents_r>=1.5 1099  340 1 0.31 0.69) *
    3) educ_r>=1.5 1947  520 1 (0.27 0.73) *
```

The NHIS tree has five terminal nodes (or leaves) marked by *. Each row in the list shows the number of the node, the split, which is the combina-

tion of variable values for cases in the node, the total number of cases in the node (labeled n), the number of cases that are misclassified (labeled loss), the category of the majority of the cases in the node (0=nonrespondent, 1=respondent, labeled yval), and the proportion of cases that are 0s and 1s (labeled yprob). For example, the node labeled 4 in the output from the print statement above contains persons who have a high school education or less (educ_r<1.5) and who are aged 56 or more. There are 588 persons in that node, the majority of whom are respondents (yval=1). If all persons in the node were classified as respondents, 240 would be misclassified (loss=240). The proportion of cases that are respondents is 0.59.
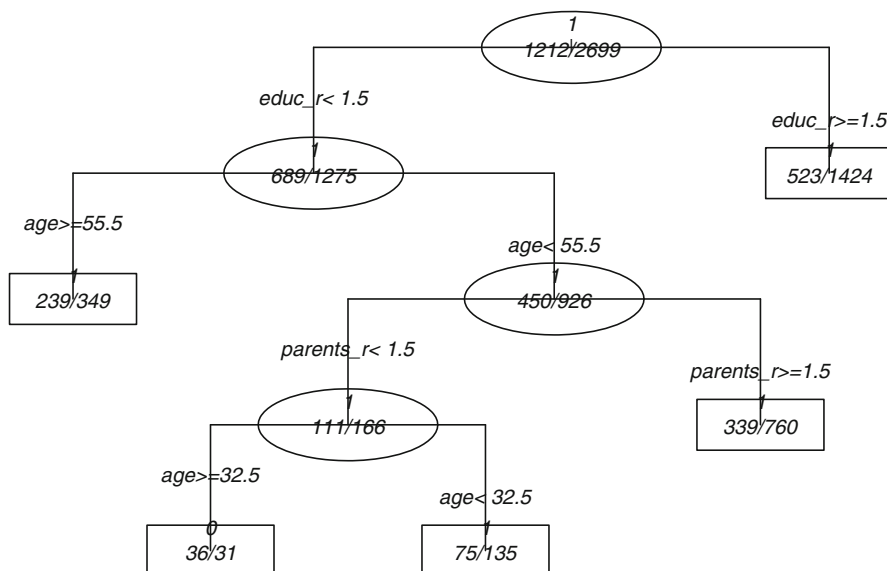
Fig. 13.7: Classification tree for nonresponse adjustment classes in the NHIS data.

Figure 13.7 is a picture of the tree, which may make clearer the combination of factors in each node. For example, the node labeled 11 in the output from print is defined by cases with a high school education or less (educ_r< 1.5), age < 56, and neither one of the sample person's parents present in the home (parents_r>=1.5). Node 11 has a 69 % response rate. Notice that the definitions of the nodes imply that there is some interaction among the variables that is being accounted for when modeling response. Hispanic and race were not used in constructing the tree although these factors were significant in the logistic model in Table 13.1. However, using more detailed categories for the presence of parents and education will lead to Hispanic being included, as illustrated in one of the exercises. Notice that the regression tree has identified a three-way interaction of education, age, and parents

as being important. This interaction was not included in the earlier logistic model, and probably could have been identified only by a lengthy trial-and-error process if we limited ourselves to logistic modeling.

A practical consideration in forming nonresponse adjustment classes is to assure that the sample size in each class is not too small. Deciding what is "too small" is subjective—50 cases (respondents + nonrespondents) is sometimes used. The sample size in a terminal node is controlled by setting `minbucket`. Notice that this does not set a constraint on the variance of the estimated response rate in a class because the variance would depend on the response rate itself. In some cases, a node will not be split even though the number of cases in a class is much larger than the value of `minbucket`. For example, terminal node 3 in our example has 1,947 cases but is not split because no improvement could be made in the log-likelihood that is being maximized.

There are a number of parameters that can be used in text to control the printing of the tree (see the help file for `text.rpart` in the `rpart` package). You may want to experiment with these to obtain a picture of the tree that you prefer. For example, setting `fancy=TRUE` represented the intermediate nodes by ellipses and the terminal nodes by rectangles. The edges connecting the nodes are labeled by left and right splits. If the default value of `fancy=FALSE` is used, the ellipses and rectangles are omitted and the edges are not labeled. When a tree has many branches and nodes, `fancy=FALSE` will produce a less cluttered looking picture.

In this example, the five terminal nodes are numbers 4, 20, 21, 11, and 3, with proportions of respondents equal to 0.59, 0.46, 0.64, 0.69, and 0.73. The range is 0.46–0.73. Recall that when five classes were created from propensity scores, the range in response probabilities was 0.59–0.78 using the unweighted average propensity in each class (see Table 13.2). Although we have five classes in both the propensity class analysis and the CART model, the cases assigned to each class are not necessarily the same. The assignment of each case to a node is given in component `t1$where`, which is a vector of length 3,911. The count of cases in each terminal node is given by

```
table(t1$where)
    3     6     7     8     9
  588    67   210  1099  1947
```

The labels (3, 6, 7, 8, 9) are not the same as the labels shown by `print(t1, digits=2)` above. The label 3, for example, means the third node produced by print:

```
  4) age>=56 588   240 1 (0.41 0.59} *
```

The nonresponse adjustment for units in terminal node c can be computed as either the inverse of the unweighted response rate, $1\big/\hat{\phi}_c = n_c/n_{cR}$, or the inverse of weighted response rate, $1\big/\hat{\phi}_c = \sum_{i\in s_c} d_i\big/\sum_{i\in s_{cR}} d_i$. The unweighted and weighted response rates are These rates can be computed

Adjustment class	Unweighted RR	Weighted RR
3	0.5935	0.6089
6	0.4627	0.4527
7	0.6429	0.6446
8	0.6915	0.7026
9	0.7314	0.7466

and merged onto the `nhis` data file to make the nonresponse adjustments using the following R code:

```
      # compute NR adjustments based on classes formed
      # by tree
      # Unweighted response rate
 unwt.rr <- by(as.numeric(nhis[, "resp"]), t1$where, mean)
      # Weighted response rate
 wt.rr <- by(data = data.frame(resp = as.numeric(nhis[,"resp"]),
            wt = nhis[,"svywt"]),
            t1$where,
            function(x) {weighted.mean(x$resp, x$wt)} )
      # merge NR class and response rates onto nhis file
 nhis.NR <- cbind(nhis, NR.class=t1$where)
 tmp1 <- cbind(NR.class=as.numeric(names(wt.rr)), unwt.rr, wt.rr)
 nhis.NR <- merge(nhis.NR, data.frame(tmp1), by="NR.class")
 nhis.NR <- nhis.NR[order(nhis.NR$ID),]
```

The merge uses the common field `NR.class` that is in both the `nhis.NR` and `tmp1` objects. Although we created a field that had the same name in the two objects, the `merge` statement is flexible enough to permit merging using fields that have different names.

13.6 Collapsing Predefined Classes

Designers of surveys often have a lengthy list of nonresponse adjustment classes, of the type described in Sect. 13.5.1, in mind when they develop weighting systems. However, using classes with a small number of sample cases will lead to imprecise estimates of response propensities. If the sample size in a class is small, the class may be collapsed with an adjacent one. The conventional justification for collapsing is that the possibility of creating extreme weights is reduced as are variances of estimates. However, a poor choice of the method for collapsing may lead to estimates that are quite biased.

Kalton and Maligalig (1991) and Kim et al. (2007) give some guidance on how the collapsing should be done, which we summarize here. Collapsing leads to bias when response rates and class means of the initial classes are

correlated within a collapsed class. The bias can be either positive or negative, depending on the correlation. Classes should be collapsed based on similarity of response rates, population class means, or both in order to avoid bias. This method of collapsing can be much different from procedures that only collapse "adjacent" classes, e.g., by combining contiguous age groups. If the adjacency coincides with classes that have similar response rates or means, no bias results.

There are at least two practical issues with collapsing based on class means. One is that, while the theory in the two papers above directs us to collapse based on population means, in a particular sample, we will only have estimates for the responding sample. If nonresponse is substantial, the means of the responding and nonresponding parts of the sample may be considerably different, even within the initial classes. This would be a case of NINR. In that case, adjustment based only on the initial set of classes or combinations of them cannot correct nonresponse bias. A second practical issue is that data on many items are collected in most surveys. Collapsing based on the class means for one variable may not work well for other variables. In that case, the compromise, suggested by Little and Vartivarian (2005) for nonresponse adjustment, of collapsing based on some weighted average of the means of an important set of variables might be a good solution. However, we will only have means for the respondents. Whether and how much the means of the nonrespondents differ cannot be checked in most surveys.

Regression tree software, such as `rpart` in R, has automated collapsing schemes that are informed by optimizing some criterion (like maximizing a log-likelihood or minimizing an error sum of squares) associated with the method chosen to partition the data. As shown in Fig. 13.7, the R function combined the continuous age variable into groups 55.5 years of age and older, and less than 55.5 years.

13.7 Weighting for Multistage Designs

The previous sections addressed weighting adjustments within a single stage of the survey design. These same basic techniques can be used within each stage of a multistage design and should sequentially reflect any appropriate adjustments from the previous level. We provide a few descriptive examples below.

Consider a stratified two-stage establishment survey where a stratified sample of businesses is randomly selected from a list in the first stage, and employees are randomly selected in second-stage strata from the sampled businesses. In this example, an establishment is a PSU, and an employee is an SSU. Establishments might be stratified by type of business (retail, manufacturing, etc.). Employees within an establishment might be stratified by occupational class (professional, clerical, etc.). The base weight for business

i in first-stage stratum h ($h = 1, \ldots, H \geq 2$) is calculated as the inverse probability of selection, $d_{0hi} = \pi_{hi}^{-1}$, as described in Sect. 13.3. The particular sampling mechanism is not important to the example and is left to the imagination of the reader. The corresponding (unconditional) base weight for employee k within stratum j is defined as defined as

$$d_{hijk} = \pi_{hi}^{-1} \pi_{jk|hi}^{-1},$$

where $\pi_{jk|hi}^{-1}$ is the base weight for employee jk within business hi, i.e., given that business hi was selected in the first stage of the design. If study eligibility cannot be confirmed for the business, such as its operational status or whether the company still manufactures a particular product, then the PSU base weight should be adjusted for unknown eligibility, $w_{hi} = d_{hi}a_{1hi}$. The resulting business-level weight would be used to create the final employee analysis weight such as

$$w_{hijk} = d_{hi} \, a_{1hi} \, a_{2hijk}$$

with a_{2hijk} signifying, for example, an SSU-level nonresponse adjustment.

A second example is a survey of teachers who instruct students between the age of 14 and 16. Schools (SSU) are randomly chosen from sampled geographic clusters (PSU) such as school districts or counties; teachers are then selected from rosters provided by a school administrator. If the status of school i in PSU h cannot be ascertained and the school administrator declines to participate in the study for some schools, then at least one adjustment should be applied to the SSU weight:

$$w_{ij} = \pi_i^{-1} \, \pi_{j|i}^{-1} \, a_{2ij}^{(1)} \, a_{2ij}^{(2)},$$

where π_i^{-1} is the PSU base weight, $\pi_{j|i}^{-1}$ is the conditional SSU base weight, $a_{2ij}^{(1)}$ is the (unconditional) unknown eligibility adjustment for SSU hi that was calculated with input weight $d_{ij} = \pi_i^{-1} \, \pi_{j|i}^{-1}$, and $a_{2ij}^{(2)}$ is the corresponding (unconditional) nonresponse adjustment calculated with input weight $w_{2ij}^{(2)} = \pi_i^{-1} \, \pi_{j|i}^{-1} \, a_{2ij}^{(1)}$. Note that neither an unknown eligibility adjustment nor a nonresponse adjustment would be required for stages of a survey design that involve geographic clusters as the sampling unit since, presumably, we will know whether each geographic unit is eligible and all will respond. However, if permission from school-district officials to contact the sample schools was required, then a PSU-level nonresponse adjustment would also be warranted. The adjusted school-level weight, w_{ij} above, would then be used to construct a nonresponse-adjusted analysis weight for teacher k, $w_{ijk} = w_{ij} \, a_{3ijk}$.

13.8 Next Steps in Weighting

The previous sections dealt with the development of base weights, the inverse selection probabilities, as well as adjustments to address problems with unknown study eligibility and nonresponse bias. The next chapter completes the picture by focusing on the use of auxiliary data (or covariates) whose totals are known for the target population. Use of auxiliary data can reduce variances of estimators and can adjust for incomplete sampling frames, a problem also known as undercoverage. These many weight adjustments, especially for multistage surveys, can unnecessarily inflate the variation in the analysis weights which in turn decreases the precision in the study estimates. Techniques used to manage this inflation are also discussed.

Exercises

13.1. Consider stratified simple random sampling without replacement (*stsrswor*). An *srswor* of size n_h is selected in each stratum from a population of size N_h, $h = 1, \ldots, H$. The selection probability of unit i in stratum h is $\pi_{hi} = n_h/N_h$ and the base weight is $d_{0hi} = \pi_{hi}^{-1} = N_h/n_h$. Show that the sum of the base weights across all units in the sample equals the population size N and that the sum within each stratum equals the stratum population size, N_h.

13.2. A two-stage sample of PSUs and persons within PSUs is needed for a pilot study on public transportation. A sample of three geographic PSUs has been selected with probabilities proportional to their total population counts, N_i, based on administrative records. Sampling was done in such a way that the probability of selecting PSU i is mN_i/N, using the notation from earlier in this chapter. Persons will be classified into two race/ethnicity groups for sampling—non-Hispanic Whites and others. You would like to select subsamples of non-Hispanic Whites and others so that the sample from each of these two groups is self-weighting. The desired sampling rates are 0.01 for non-Hispanic Whites and 0.04 for others.

PSU	N_i	Non-Hispanic White N_{Wi}	Others $N_{other,i}$
1	1,000	800	200
2	850	400	450
3	150	110	40
Pop. total N	10,000		

Find the following:

(a) Selection probabilities for the three sample PSUs
(b) Within-PSU sampling rates needed to achieve the desired overall sampling rates
(c) Base weights for each unit
(d) Expected number of sample persons in each PSU by race/ethnicity group and in total

13.3. Repeat Exercise 13.2 assuming that the target sampling rates are 0.02 for non-Hispanic Whites and 0.06 for others. Do you see any problems with this design? If so, what remedy would you suggest?

13.4. The following table gives sums of weights for samples of establishments in three cities that were classified as being in retail trade based on yellow page listings:

City	Eligible resp.	Eligible nonresp.	Known ineligible	Unknown eligibility	Total
1	50	46	11	17	124
2	77	89	19	12	197
3	44	31	8	23	106
Total	171	166	38	52	427

(a) Adjust the weights separately in each city first for unknown eligibility and then for nonresponse. Show your calculations in each step.
(b) What is the estimated total number of eligible units in each city and across all cities?
(c) What is the estimated number of ineligible establishments on the sampling frame?
(d) In what circumstance would it be reasonable to combine all three cities together to make the adjustments for unknown eligibility and nonresponse? Do those circumstances hold here?

13.5. A telephone survey of a sample of 500 members of a professional organization of podiatrists is conducted. The 500 are a simple random sample from the list of 2000 current members. Four hundred sample persons are definitely determined to be eligible. Among those, 320 respond to the survey and 80 refuse. The list is somewhat out-of-date so that some phone numbers are incorrect. Seventy sample people cannot be successfully contacted. Of the 70, there are 45 whose answering machine picks up, but a person is never contacted directly; 16 persons pick up the phone but immediately hang up when they hear that a survey is being done; for nine of the phone numbers neither a person nor an answering machine ever picks up. Some persons on the list may have dropped out of the organization and are, therefore, ineligible. You are able to identify 30 sample persons who have dropped their membership:

(a) Into what eligibility status would you classify the 70 people (45 answering machine, 16 hang-ups, 9 no answer): unknown, ineligible, or eligible refusal? Why?
(b) Given your decision in (a), use a single adjustment class to adjust for unknown eligibility. Which cases receive the adjustment? What is the adjustment value for each?
(c) After the adjustment for unknown eligibility, what are the estimated numbers in the population of eligibles and ineligibles?

13.6. Fit unweighted logistic, probit, and c-log-log models to the `resp` variable in the NHIS data set, `nhis.RData`:

(a) Use the covariates age, sex, hisp, and race.
(b) Which variables are significant predictors in each of the models?

(c) Compare the predicted probabilities from the three models.

13.7. Continuing Exercise 13.6, use the predicted response probabilities from the logistic regression that used all covariates and create two versions of propensity classes:

(a) Five classes with an equal number of respondents plus nonrespondents in each and
(b) Ten classes. Report the breaks used for the five and ten classes and the number of cases assigned to each class. (Check to see that all cases were assigned a non-missing class value. Use the parameter useNA="always" in table if you use R in order to see whether NAs were created.)
(c) Calculate the five alternative values of NR weight adjustment shown in Example 13.8. For the weighted adjustments, use the svywt variable. Discuss how the five alternative values of adjustments compare within (a) and (b) and how the adjustments using five and ten classes compare to each other.
(d) How do the class adjustments compare to using the inverses of individual propensity estimates as adjustments?
(e) Which set of adjustment values would you recommend and why?

13.8. Using the sets of five and ten propensity classes you created in Exercise 13.7, make the checks suggested by D'Agostino (1998) to see whether the propensity classification succeeded in balancing on the covariates. If balancing was not achieved, discuss what the implications of this might be for using the classes for nonresponse adjustment.

13.9. Using the NHIS data set, fit a classification tree for the response (resp) variable using the covariates age, sex, hisp, race, parents, and educ. Require that a minimum of 50 cases be assigned to each node. Describe the composition of each node in words and draw a picture of the tree. Compute the unweighted response rates in each of the nodes that are formed.

13.10. Calculate the unweighted and weighted values of NR weight adjustment (alternatives 3 and 4) shown in Example 13.8 for the classes identified in Exercise 13.9. For the weighted adjustments, use the svywt variable. How do these sets of values compare? Which would you recommend and why?

Chapter 14
Calibration and Other Uses of Auxiliary Data in Weighting

The previous chapter described the first few steps used in weight calculation: base weights, adjustments of unknown eligibility, and nonresponse adjustments. The last step, which is extremely important in many surveys, is to use auxiliary data to correct coverage problems and to reduce standard errors. By auxiliary data, we mean information that is available for the entire frame or target population, either for each individual population unit or in aggregate form. These may be obtainable because a frame of all units in the population was used to select the sample and each listing on the frame contains some data. Surveys of business establishments or institutions may have such frames. Population totals for some variables may be available from a source separate from the survey, like a census. In a business survey, the frame might have the number of employees from an earlier time period for each establishment. In a household survey, counts of persons in groups defined by age, race/ethnicity, and gender may be published from a census or from population projections that are treated as highly accurate.

Figure 14.1 shows two populations where a survey variable y is related to an auxiliary variable x. Regardless of the type of sample design used, exploiting the relationship between y and x can give more precise estimators than ignoring it. Using x will reduce variances more for the structure in the left-hand panel than in the right because of the (linear) association between the two variables is stronger. In either case, estimators can be used that take advantage of the relationship. We use a single auxiliary variable for ease of demonstration. Statisticians generally use a set of p $(p > 1)$ auxiliary variables, denoted in transposed vector form as $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T$ for the ith sample unit, to adjust the weights.

Another use of auxiliary data is to correct for coverage errors in a frame. For example, suppose that a household survey estimates that the number of African-American males, aged 18–24, is only 75 % of the latest census count or population projection for that group even after nonresponse adjustment. By creating weights that reproduce the census counts or population projections, we can "correct" for the undercoverage. To be effective, the responding
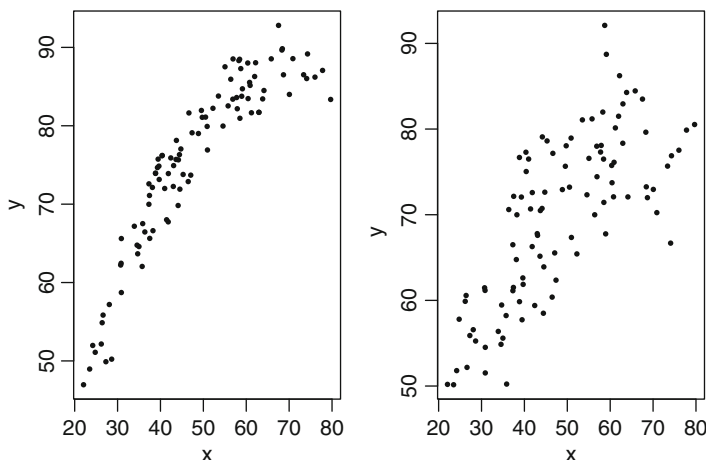
Fig. 14.1: Scatterplots of two hypothetical relationships between a survey variable y and an auxiliary x.

sample cases do have to be a good representation of the full population. This means that either (a) the adjustments to weights, covered in Chap. 13, correct any potential bias due to nonresponse, or (b) the analysis variables for the respondents follow the same model as exhibited in the full population and that model can be approximated by the calibration techniques in this chapter.

This chapter will cover some of the tools that are available for employing auxiliary data in estimation and the weights that are implied. Section 14.1 describes the general method of calibration estimation along with some examples. Two of the most common uses of auxiliary data are poststratification and raking, covered in Sect. 14.2. General regression estimation (GREG) and some examples of the broader class of calibration estimators are discussed in Sect. 14.3.

The steps of computing base weights, nonresponse adjustments, and calibration may result in weights whose sizes vary quite a bit. Practitioners usually are leery of having weights that have a large range since a few extremely large weights can destabilize estimates by increasing the associated standard errors (SEs). Having variable weights may or may not be something to worry about. Subgroups that respond at greatly different rates will lead to differing sizes of nonresponse adjustments. Weights that vary considerably may also be statistically efficient, as in the case of optimal allocation to strata that we studied in Chap. 3. However, if the weights vary for none of these reasons, this can be inefficient. Section 14.4 describes quadratic programming and weight-trimming methods that allow the weights themselves to be directly bounded. We also discuss two types of design effects that are sometimes useful when assessing weight variability.

14.1 Weight Calibration

The term *calibration estimator* was introduced for survey estimation by Deville and Särndal (1992). Kott (2009) gives a good review of the mathematics of the technique. The general idea is to use auxiliary variables to improve the efficiency of estimators. The auxiliaries may come from the frame, administrative records, published statistics, or other sources. Among the potential benefits of calibration are:

- Decrease in variances
- Bias correction for frame coverage and other survey errors
- Adjustment for nonresponse

Multiple auxiliaries can be used, but to illustrate the method, we begin with one of the simplest cases of calibration—the ratio estimator.

Example 14.1 (Ratio estimator). The ratio estimator of a mean for a simple random sample (*srswor*), introduced at the end of Sect. 3.2.2, is $\bar{y}_R = \bar{y}_s \bar{x}_U / \bar{x}_s$ where $\bar{y}_s = \sum_s y_i / n$ and $\bar{x}_s = \sum_s x_i / n$ are unweighted means of an analysis variable y and an auxiliary variable x and $\bar{x}_U$ is the population mean of x. The response variable might be the number of full-time employees in an establishment at the current time and x the corresponding number from a year ago. To compute the ratio estimate from a sample, we need the values of x for the individual sample units so that $\bar{x}_s$ can be calculated, and we need their population mean, $\bar{x}_U$. Notice that the individual values of x for the nonsample units are not required to compute $\bar{y}_R$, although we might have them from the frame. One property of the ratio estimator is that, if we treat y as the auxiliary variable, then $\bar{y}_s = \bar{x}_s$ and the ratio estimate reduces to $\bar{x}_U$. Thus, the estimate is calibrated (or benchmarked) in the sense that it reproduces the known population value when we substitute the auxiliary variable for the analysis variable. ∎

The ratio estimator is a member of a more general class that covers many of the estimators used in practice. Suppose that the weights used in the calibration step are denoted by d_i for the ith unit in the sample, like a person or a business establishment, for which data are collected. In Chap. 13 the product of a base weight, an unknown eligibility adjustment, and a nonresponse adjustment was called d_{2i}. We drop the subscript 2 here to reduce the notation. The goal of calibration is to find a new set of weights, $w = \{w_i\}_{i \in s}$ using set notation, that are near the input weights, $d = \{d_i\}_{i \in s}$, but when used to estimate totals of the auxiliaries, reproduce the population totals exactly. The thought in keeping the weights close in value is that the output weights can "borrow" any good estimation properties inherent in the input weights. For example, if the base weights are associated with a weighted mean that is design unbiased, then the same estimate calculated with the output weights should be (approximately) design unbiased as well. On the other hand, if the input weights produce high variance estimates, creating new weights that are

close to the old ones is no improvement. However, many efficient estimates are in the calibration class, making the class worth studying.

Formally, the following problem is solved with weight calibration:
Find the set of weights $\{w_i\}_{i \in s}$ that:

- Minimize a measure of the distance, $L\left(w, d\right)$, between the incoming weights and the calibrated weights.
- Subject to constraints:

$$\sum_{i \in s} w_i \mathbf{x}_i = \sum_{k \in U} \mathbf{x}_k, \tag{14.1}$$

where $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T$ is the set of p auxiliary variables for unit i and $w_i = d_i g_i$, a function of the input weight and an adjustment (the g-weight) that satisfies the constraints.

We will refer to the w_i's as the *final weights*. To determine the weights, we need the x-values for the individual sample units and the population totals for those x's. Typically, auxiliary information is not needed for the nonrespondents. Recall that the methods of nonresponse adjustment we studied in Chap. 13, like propensity adjustments, did require individual covariate information to be available for both respondents and nonrespondents. The auxiliaries can be either quantitative, such as the total number of students in a school, or qualitative, such as an indicator for gender=male.

One choice of L is the least-squares distance function,

$$L\left(w, d\right) = \sum_s \left(w_i - d_i\right)^2 \Big/ d_i. \tag{14.2}$$

Minimizing this, subject to the constraint in Eq. (14.1), leads to the *general regression estimator* or GREG. The GREGs include many of the estimators used in practice and studied in standard sampling books. We cover these in more detail in Sect. 14.3. Another distance function is

$$L\left(w, d\right) = \sum_s \left[w_i \log\left(\frac{w_i}{d_i}\right) - w_i + d_i\right]. \tag{14.3}$$

This leads to a type of raking estimator, which we discuss in Sect. 14.2.

There is usually a model under which a particular calibration estimator will be especially efficient in terms of the repeated sampling variance. We discuss the estimator-associated models in the Sects. 14.2 and 14.3. If the model correctly describes the dependence of an analysis variable on a set of auxiliaries, then the calibration estimator will be model unbiased also. In selecting a set of auxiliaries, good policy is to do some modeling using the auxiliaries as covariates. This will help in deciding which auxiliaries to use and whether any of the auxiliaries should be transformed by, say, taking the square

or logarithm. A few model-checking techniques are examined in Sect. 14.3.2; interested readers are referred to standard texts such as Cook and Weisberg (1982) and Weisberg (2005) for additional information. Diagnostics that are specialized to be appropriate for complex survey data can be found in Li and Valliant (2009, 2011) and Liao and Valliant (2012a,b).

14.2 Poststratified and Raking Estimators

Poststratified and raking estimators are two of the most commonly used calibration estimators. They are especially prevalent in household surveys of persons where the auxiliary variables are indicators for demographic groups. For example, persons may be classified by age group, gender, and race/ethnicity. Poststratification is implemented within calibration weighting classes formed by crossing *all* categories of the qualitative variables and constructing weights that reproduce the class-specific population counts in the weighted estimates. Poststratification can also be done using a single variable like age group. Formally, the poststratified estimator of a total is defined as

$$\hat{T}_{yPS} = \sum_{\gamma=1}^{G} N_\gamma \left(\hat{t}_{y\gamma} \Big/ \hat{N}_\gamma \right),$$

where $\hat{t}_{y\gamma} = \sum_{s_\gamma} d_i y_i$ is the estimated total of y in weighting class (or poststratum) γ based on the input weights, s_γ is the set of sample units in poststratum γ, $\hat{N}_\gamma = \sum_{s_\gamma} d_k$ is the estimated population size of poststratum γ based on the input weights, N_γ is the population count (also known as a control or control total) for the poststratum γ, and G is the total number of poststrata. The implied final weight for unit i in poststratum γ is

$$w_i = d_i \frac{N_\gamma}{\hat{N}_\gamma}, \tag{14.4}$$

where $g_i = N_\gamma \big/ \hat{N}_\gamma$ is the poststratification adjustment (factor). This is the g-weight in generic equation $w_i = d_i g_i$. With that definition of the weight, we can write the estimator as $\hat{T}_{yPS} = \sum_{i \in s} w_i y_i$, i.e., a weighted sum of the data values.

The weighting classes are referred to as poststrata because they are applied after the sample is selected and data are collected. They are not necessarily used at the design stage to select the sample. In fact, poststratification is a good way to use auxiliaries that you think are effective predictors of important variables collected in the survey but cannot be easily used for sample selection. For example, in a household survey, many countries do not have a frame of persons that includes race/ethnicity and educational attainment.

We can use those as poststrata as long as the population counts of persons in the cross of those two variables are available from a census or some other external source like projected population counts. In this example, poststrata would be defined as a combination of race/ethnicity and categorized education. Suppose that race/ethnicity is coded into three categories (1=White, 2=African-American, 3=other) while education, defined as the highest level of school completed, is coded into four categories: 1=less than high school); 2=high school graduate; 3=college or some college; and 4=graduate degree (master's, doctorate, professional degree beyond bachelor). Then, the cross of these variables leads to $G=12$ classes that could be used as poststrata. One note of caution before we proceed: using many important auxiliary variables for poststratification can reduce bias but may result in empty weighting classes or ones with a small number of respondent cases. This results in unstable estimates $\hat{N}_\gamma$ of the population controls and adds unnecessarily to the variability of the final weights—both instances should be avoided.

The poststratified estimator is a special case of the GREG with the distance function in Eq. (14.2). The model that naturally accompanies $\hat{T}_{yPS}$ is one where units have a common mean and variance within a poststratum:

$$E_M(y_i) = \beta_\gamma, Var_M(y_i) = \sigma_\gamma^2. \tag{14.5}$$

The **x** vector for a unit has G components, each containing a 0–1 indicator for whether or not a unit is in a particular poststratum. The `postStratify` function in the R `survey` package can be used to compute the estimate, as shown below in Example 14.2.

Example 14.2 (Poststratified estimator). To illustrate poststratification, we select a simple random sample of size 250 from the large NHIS population supplied with this book. Poststrata are defined by age group crossed with Hispanicity. First, we calculate the proportion of persons covered by Medicaid (a type of US governmental assistance for medical care provided to the poor) in domains defined by age group and Hispanicity. The `hisp` variable on the file is recoded to the 3-category variable and named `hisp.r`. The categories of age group and Hispanicity are shown in Table 14.1 along with

Table 14.1: Percentages of persons in the large NHIS population who reported receiving Medicaid.

	Age group (years)				
Hispanicity	under 18	18–24	25–44	45–64	65+
Hispanic	32.2	10.7	7.6	11.0	27.2
Non-Hispanic White	12.6	6.6	3.8	3.1	3.7
Non-Hispanic Black and other race/ethnicity	31.3	12.7	8.8	6.4	16.5

the percentages of persons receiving Medicaid. Hispanics and non-Hispanic Blacks under 18 have much higher percentages than other age groups; Hispanics who are 65 years and older also have a high rate of Medicaid. Of course, we could fit a model to predict whether people receive Medicaid, but the crosstable is sufficient to show that there is an interaction between age group and Hispanicity.

This is the code used to produce the percentages:

```
attach("nhis.large.RData", pos=2)
   # collapse hisp = 3,4
hisp.r <- nhis.large$hisp
hisp.r[nhis.large$hisp == 4] <- 3
nhis.large1 <- data.frame(nhis.large, hisp.r)
t1 <- table(nhis.large1$medicaid, nhis.large1$age.grp,
            nhis.large1$hisp.r)
100 * round(prop.table(t1[,,1],2),3)
100 * round(prop.table(t1[,,2],2),3)
100 * round(prop.table(t1[,,3],2),3)
```

Next, we select the sample and then create an *srswor* design object and a poststratified design object:

```
   # create single variable to identify age.grp x
   # hisp.r poststrata
m <- max(nhis.large1$hisp.r)
nhis.large1$PS <- (nhis.large1$age.grp - 1)*m + nhis.large1
$hisp.r
N.PS <- table(PS = nhis.large1$PS)

       # select  srswor of size n
set.seed(-1570723087)
n <- 250
N <- nrow(nhis.large1)
sam <- sample(1:N, n)
samdat <- nhis.large1[sam, ]

       # compute srs weights and sampling fraction
d <- rep(N/n, n)
f1 <- rep(n/N, n)

       # srswor design object
nhis.dsgn <- svydesign(ids = ~0,         # no clusters
                       strata = NULL,    # no strata
                       fpc = ~f1,
                       data = data.frame(samdat),
                       weights = ~d)

       # poststratified design object
ps.dsgn <- postStratify(design = nhis.dsgn,
                        strata = ~PS,
                        population = N.PS)
```

The postStratify function takes three main parameters: design (survey design object), strata (formula or data frame of poststratifying variables), and population (table or data frame with population frequencies).

There are several steps above where particular syntax requirements must be observed. A single variable was created to denote the poststrata. The statements

```
m <- max(nhis.large1$hisp.r)
nhis.large1$PS <- (nhis.large1$age.grp - 1)*m +
                  nhis.large1$hisp.r
```

number the age group $\times$ Hispanicity poststrata from 1 to 15 (=5*3). The `poststratify` function accepts a single variable (called `PS` in this case) to define the poststrata.

The `svydesign` function creates an object, `nhis.dsgn`, that contains the *srswor* design information. The weights are in `d` and are all equal to N/n. In order to include a finite population correction (*fpc*) factor in variance estimates, the `fpc` parameter must be specified when creating the design object. The parameter must be a vector whose length is equal to the sample size. Using a scalar will generate an error, even when the finite correction is a single value. Rather than specifying $1 - n/N$, which is the textbook definition of an *fpc*, the `survey` package requires the `fpc` parameter to be either the population total N or the sampling fraction, n/N. This may seem idiosyncratic, but the same convention is used by both Stata and SAS.

The population control counts, N_γ, are contained in `N.PS`. The name `PS` given to the table in the statement

```
    N.PS <- table(PS = nhis.large1$PS)
```

must match the name of the variable in the data set that holds the poststrata. Next, we can verify that the poststratified weights do sum to the population counts using the `svytotal` function below. Only the first four of 15 poststrata are shown. The estimated counts do match the population counts (which the reader can verify). The SEs of the estimates are zero since there is no variation from sample to sample in the estimates—they will always equal the population counts. This issue is revisited in Chap. 18 where we cover quality control tabulations in more detail:

```
        # Check that weights are calibrated for x's
 svytotal(~as.factor(PS), ps.dsgn)

               total    SE
as.factor(PS)1   1952    0
as.factor(PS)2   2870    0
as.factor(PS)3   1169    0
as.factor(PS)4    581    0
```

Note that the weights of the individual sample cases can be examined with the command, `weights(ps.dsgn)`. The estimated proportion of persons receiving Medicaid, their SEs, and coefficients of variation (CV) are produced by

```
        # PS standard errors and cv's
svytotal(~ as.factor(medicaid), ps.dsgn, na.rm=TRUE)
cv(svytotal(~ as.factor(medicaid), ps.dsgn, na.rm=TRUE))
        # srs standard error and cv's
svytotal(~ as.factor(medicaid), nhis.dsgn, na.rm=TRUE)
cv(svytotal(~ as.factor(medicaid), nhis.dsgn, na.rm=TRUE))
```

The parameter, `na.rm=TRUE`, is used because some cases have missing values for Medicaid; without it, results will all be `NA` (i.e., missing). To force Medicaid to be treated as a class (factor) variable, `as.factor` is used. The poststratified and *srswor* estimates for the total number of persons receiving Medicaid are

	Total	SE	*CV*
Poststratified	1870.8	344.5	0.184
srswor	1899.7	385.3	0.203

In this sample, the *srswor* and poststratified estimated totals are similar and the latter has slightly smaller SE and *CV*. ∎

Example 14.3 (Poststratified estimator as a way of correcting for undercoverage). Suppose that the sample frame only covers 75 % of the population subgroups of Hispanic and non-Hispanic Blacks and other race/ethnicities:

```
        # create frame with undercoverage
        # 75% coverage of Hispanics and non-Hispanic Black &
        # Other. These correspond to poststrata
        #      1,3,4,6,7,9,10,12,13, and 15.
PS.prob <- rep(c(0.75, 1, 0.75), 5)
cov.prob <- PS.prob[nhis.large1$PS]
N <- nrow(nhis.large1)
rn <- runif(N)
nhis.cov <- nhis.large1[rn <= cov.prob, ]
```

The code above generates a uniform random variable in the interval [0,1] for each person in the population. This random number is compared to the coverage rate (0.75 or 1) for the poststratum containing each person and a "covered" population is created with the statement

```
  nhis.cov <- nhis.large1[rn <= cov.prob, ]
```

This treats coverage as a random phenomenon—every person has some chance of being in the frame. This may or may not be a realistic assumption but is typical in the literature that analyzes the effects of undercoverage. Some linkage is needed between the units in the frame, the sample selected from it, and the rest of the universe in order to make inferences for the entire target

population. Modeling coverage as a random occurrence is one way of doing this. We then selected an *srswor* sample of n=500 from `nhis.cov` using the seed, `set.seed(610376119)`, and computed poststratified weights (code not shown).

The estimated totals of Medicaid recipients, their SEs, and CVs are shown in Table 14.2. The totals for Hispanics and non-Hispanics can be computed with the statement

```
svyby(~as.factor(medicaid), ~hisp.r, ps.dsgn, svytotal,
         na.rm=TRUE)
```

The proportions are found by substituting `svymean` for `svytotal`.

The estimates calculated with (unadjusted) base weights are labeled as π-estimates in this example and in subsequent text to distinguish them from estimates calculated with (adjusted) final weights. As shown in Table 14.2, the π-estimated totals are too small due to the undercoverage, but the poststratified estimates are much closer to the population totals. The π-estimates do have smaller SEs, but 95 % confidence intervals would not contain the population totals. (The method of variance estimation used here is called linearization. We will cover methods of SE estimation in more detail in Chap. 15.) Of course, this is just a single sample. In other samples, it is possible for the poststratified estimates to be too large and the π-estimates closer to the truth. However, poststratification will, on average, reduce bias due to undercoverage in practical applications (e.g., see Kim et al. 2007), making it one of the standard techniques for correcting undercoverage. ∎

Table 14.2 also shows the estimated proportions for both types of weights. In this example, poststratification makes less difference in either the point estimates or the SEs. This is also typical—estimates that are ratios are often less affected by coverage problems than are estimated totals.

A popular alternative to poststratification is *raking*, which can also use more than one auxiliary variable. In the example above with age group and Hispanicity, all weighting classes formed by the cross-classification are used as poststrata. A population control value is needed for each weighting class. In addition, minimum sample size requirements are usually imposed; otherwise $\hat{N}_\gamma$ can be unstable. In raking only the marginal age group and Hispanicity control counts are needed. This is especially relevant when only marginal counts are available in published sources.

As with the poststratified estimator, the raked estimator is also associated with a linear model. For example, the model in a two-variable raking problem is

$$E_M(y_i) = \mu + \alpha_j + \beta_k, Var_M(y_i) = \sigma^2 \qquad (14.6)$$

for i having level j of the first variable and level k of the second. The parameters α and β are fixed effects. The poststratified model for the mean that naturally goes with the cross of two variables is $E_M(y_i) = \mu + \alpha_j + \beta_k + (\alpha\beta)_{jk}$ where $(\alpha\beta)_{jk}$ is an interaction term. This model is equivalent to

Table 14.2: Comparison of π-estimates and poststratified estimates in Example 14.3 of totals and proportions of persons receiving Medicaid when the frame has undercoverage.

Statistic	Estimate	SE	CV
Estimated totals			
Full population			
Actual population total	2,281		
π-estimate	1,770	246	0.139
PS estimate	2,381	322	0.135
Hispanic			
Actual population total	935		
π-estimate	616	150	0.243
PS estimate	954	209	0.219
Estimated proportions			
Full population			
Actual population proportion	0.107		
π-estimate	0.093	0.013	0.139
PS estimate	0.112	0.015	0.135
Hispanic			
Actual population proportion	0.189		
π-estimate	0.184	0.041	0.223
PS estimate	0.190	0.042	0.219

expression (14.5). Thus, the raking model has main effects only and fewer parameters.

Even when the main effects only model seems inadequate, raking is often a way to use more variables that may be important predictors of analysis variables or of frame coverage rates. In poststratification, crossing several variables may quickly create more classes than the sample can support.

Example 14.4 (Raking by age group and Hispanicity). To illustrate the procedure, we rework Example 14.3 by raking to the age group and Hispanicity margins. The same *srswor* sample of 500 from the covered population is used, and a survey design object called nhis.dsgn is created. The code below uses the calibrate function to do the raking. An alternative is the function rake, which will give the same answer (see Lumley 2010, Sect. 7.3):

```
        # create marginal pop totals
 N.age <- table(nhis.large1$age.grp)
 N.hisp <- table(nhis.large1$hisp.r)
 pop.totals <- c('(Intercept)' = N, N.age[-1], N.hisp[-1])

        # create raked weights
rake.dsgn <- calibrate(design = nhis.dsgn,
         formula = ~as.factor(age.grp) + as.factor(hisp.r),
         calfun = "raking",
         population = pop.totals)
```

The `calibrate` function accepts a number of parameters:

design   survey design object
formula   model formula for calibration model
population   vectors of population column totals for the model
             matrix in the calibration model or list of such vec-
             tors for each cluster
calfun   calibration   function.   Allowable   values   are
             calfun=c("linear","raking","logit","rrz").
             The function is flexible enough to accept a user-
             defined distance function also.

Notice how the vector of population totals is defined. The first position is for the population total of an intercept μ in Eq. (14.6), which is just the number of units in the population. When categorical variables are in a model, the R convention is to drop the first level so that the system of estimating equations for the parameters can be solved. That is, the first category is treated as the reference level. The first levels of `age.grp` and `hisp.r` are omitted in forming the pop.totals vector by "subtracting" the first position from the vector, e.g., `N.age[-1]`. For those versed in matrix algebra, this ensures that the calibration equations can be solved by creating an auxiliary matrix with full column rank. As in poststratification, we can check that the calibration has succeeded by estimating the totals of the two auxiliary variables:

```
       # Check that weights are calibrated for x's
 svytotal(~as.factor(age.grp), rake.dsgn)

                      total  SE
as.factor(age.grp)1   5991   0
as.factor(age.grp)2   2014   0
as.factor(age.grp)3   6124   0
as.factor(age.grp)4   5011   0
as.factor(age.grp)5   2448   0

 svytotal(~as.factor(hisp.r), rake.dsgn)
                    total  SE
as.factor(hisp.r)1   5031   0
as.factor(hisp.r)2  12637   0
as.factor(hisp.r)3   3920   0
```

The totals, proportions, and their SEs and CVs can be estimated using `svytotal`, `svymean`, and `cv` as in Example 14.3. The results are in Table 14.3. The estimates are very close to those for poststratification in Table 14.2. Note, however, that only the marginal control totals are guaranteed to be satisfied with raking, not the control totals for the cross-classification.

Table 14.3: Raking estimates in Example 14.3 of totals and proportions of persons receiving Medicaid when the frame has undercoverage.

	Total	SE	CV
Estimated totals			
Full population	2,360	316	0.134
Hispanic	943	210	0.223
Estimated proportions			
Full population	0.111	0.015	0.134
Hispanic	0.187	0.042	0.223

14.3 GREG and Calibration Estimation

To define the GREG, we need some vector and matrix notation that is more elaborate than used in other parts of this book. Särndal (2007) gives a good general discussion of GREGs. Understanding the notation is not essential to follow the examples later in this chapter, and some readers may wish to skip to the illustrations of using software to compute GREGs. To discuss the GREG, it is easier to begin with totals rather than means. Suppose there are n sample units. The GREG estimator of the population total of y can be written as

$$\hat{T}_{yGREG} = \hat{t}_y + \left(\mathbf{t}_x - \hat{\mathbf{t}}_x\right)^T \hat{\mathbf{B}}$$
$$= \sum_{i \in s} \left[1 + \left(\mathbf{t}_x - \hat{\mathbf{t}}_x\right)^T \left(\mathbf{X}^T \mathbf{D} \mathbf{V}^{-1} \mathbf{X}\right)^{-1} \mathbf{x}_i / v_i\right] d_i y_i$$

where $\hat{t}_y = \sum_s d_i y_i$ is the estimator of the total based on the input weights, the superscript T represents the transpose of the specified vector, $\mathbf{t}_x = (t_{x1}, \ldots, t_{xp})^T$ is the $p \times 1$ vector of population (or control) totals of the p auxiliaries using the number of rows by the number of columns matrix notation, $\hat{\mathbf{t}}_x = \sum_s d_i \mathbf{x}_i$ is the estimate of totals of the x's based on the d_i weights, $\mathbf{x}_i$ is the $p \times 1$ vector of auxiliary values for the ith sample unit,

$\mathbf{D} = diag(d_i)$ is the $n \times n$ diagonal matrix of input weights,

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{pmatrix}$$ is the $n \times p$ matrix of auxiliaries for the n sample units,

$$\hat{\mathbf{B}} = \left(\mathbf{X}^T \mathbf{D} \mathbf{V}^{-1} \mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{D} \mathbf{V}^{-1} \mathbf{y}$$

with $\mathbf{y} = (y_1, \ldots, y_n)^T$ being the vector of y's for the sample units, and $\mathbf{V} = diag(v_i)$ is an $n \times n$ diagonal matrix of values associated with the variance parameters in an underlying linear model. It is possible to formulate

the GREG using a block-diagonal or some other non-diagonal covariance matrix, but this is seldom done in practice.

The $p \times 1$ vector, $\hat{\mathbf{B}}$, is an estimator of the slope in the model $y_i = \mathbf{x}_i^T \beta + \varepsilon_i$ where the ε_i have mean 0 and variance v_i. Note that in the case of *srswor* design and base weights, $\hat{\mathbf{B}}$ reduces to $\left(\mathbf{X}^T \mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{y}$, familiar to many from a regression modeling course. If the model errors were all 0, then $\hat{\mathbf{B}} = \beta$ and the GREG reduces to $\mathbf{t}_x^T \beta$, which is also the population sum of the y's, T_y. In that case, the y for each unit in the population can be predicted without error as $\mathbf{x}_i^T \beta$, and the GREG would be exactly equal to T_y in every sample. As a result, the better the predictor that x is of y, the smaller the variance of the GREG.

An estimated total for a y is calculated as $\hat{T}_{yGREG} = \sum_s w_i y_i$, a function of the weights resulting from the calibration procedure of the form:

$$w_i = d_i g_i$$

The term in brackets is called the g-weight or a calibration adjustment (factor) in this text and many other references on calibration such as, for example, Särndal et al. (1992) and Stukel et al. (1996). Notice that the final w_i weights do not depend on any analysis variables (y's). As a result, the same set of weights can be used for any estimated total. As we pointed out at the beginning of Chap. 13, estimates of many quantities, like means, model parameters, and quantiles, depend on estimating totals. For example, a mean would be estimated as $\hat{\bar{y}}_{GREG} = \sum_s w_i y_i / \sum_s w_i$.

A GREG is approximately unbiased in repeated sampling if the frame provides full coverage of the target population, and $\hat{\mathbf{t}}_x$ is an unbiased (or, at least consistent) estimator of the population total, $\mathbf{t}_x$. Roughly speaking, the unbiasedness follows if the input weights, d_i, lead to $\hat{t}_y = \sum_s d_i y_i$ being an unbiased estimator of the population total and the difference $\mathbf{t}_x - \hat{\mathbf{t}}_x$ estimates 0. In the case of frame undercoverage, $\hat{t}_y$ will be too small on average but so will $\hat{\mathbf{t}}_x$. Thus, $\mathbf{t}_x - \hat{\mathbf{t}}_x$ will be positive and provide a correction for the undercoverage. Some of the other practical considerations in using GREGs are:

(1) The population totals of the auxiliaries, $\mathbf{t}_x$, which are also called calibration controls, should ideally be true values and known without error. If the population x totals are incorrect, then either $\mathbf{t}_x - \hat{\mathbf{t}}_x$ will not estimate 0 when it should, or $\mathbf{t}_x - \hat{\mathbf{t}}_x$ will not give the correct coverage adjustment. In some surveys, however, it may be desirable to use estimates of the $\mathbf{t}_x$ controls from a larger or higher quality survey than the one you are weighting (see, e.g., Dever and Valliant 2010). This may be true if there are x's that are felt to be very predictive of analysis variables, but only population estimates from another survey are available.

(2) The estimated auxiliary totals, $\hat{\mathbf{t}}_x = \sum_s d_i \mathbf{x}_i$, should be measured in the same way in the population as in the survey. For example, suppose one of the x's is household annual income. If a census and the survey collect

income using different question wordings, this noncomparability could bias $\mathbf{t}_x - \hat{\mathbf{t}}_x$.

(3) As alluded to earlier in this chapter, an association exists between outcome and auxiliary variables and is "effectively" represented by a linear model. Although a close association between x's and the analysis variables is not necessary for the GREG to be approximately unbiased, a model that fits well will yield lower variances. Consequently, some formal model fitting is an important step in weighting.

(4) The fact that the calibration adjustments $\{g_i\}_{i=1}^{n}$ are sample-dependent needs to be accounted for in variance estimation. We will cover methods of doing this in Chap. 15.

14.3.1 Links Between Models, Sample Designs, and Estimators—Special Cases

Although some practitioners prefer to think of GREGs as model-free, we feel that this is obscurantist at best. The motivation for choosing a particular form of GREG is much easier to understand when an underlying model is considered. Many sample-design/estimator combinations used in practice are special cases of GREG. Some examples of estimator/sample-design/model combinations are given in Table 14.4. These estimators are described in various texts like Cochran (1977) and Särndal et al. (1992).[1]

GREGs flow from various kinds of linear models, as noted above. Nonetheless, they are often used to estimate totals of binary 0–1 variables even though this implies that a linear model describes the association with a dichotomous variable. Although fitting a linear model to a binary variable would probably be considered a gaffe by most data analysts, it is commonplace in survey estimation. This is an offshoot of using estimators of the form $\hat{T} = \sum_s w_i y_i$. In some cases, like the poststratification model (14.5) where every unit in a weighting class has the same mean, a linear model is fine for a binary variable. But, in others where quantitative auxiliaries are used, the implicit predictions for 0–1 variables may be outside of the range [0,1] for some units. A limited amount of work has been done on using traditional binary regression models to estimate totals of 0–1 variables in surveys (e.g., Lehtonen and Veijanen 1998; Valliant 1985). However, these methods can result in g-weights that are a function of the analysis variables and are not in common use in surveys. We will not discuss these techniques further here.

---

[1] The form of the combined regression estimator shown in Table 14.4 is from Särndal et al. (1992) and differs from the one in Cochran (1977).

Table 14.4: GREG estimators of population total by design and model assumptions.

Estimator	Design	Estimated total	Model mean	Model variance
Expansion	srswor	$\hat{T}_0 = N\bar{y}_s$	$E_M(y) = \mu$	$Var_M(y) = \sigma^2$
Ratio	srswor	$\hat{T}_R = N\bar{y}_s \frac{\bar{x}_U}{\bar{x}_s}$	$E_M(y) = \beta x$	$Var_M(y) = \sigma^2 x$
Poststratified	srswor	$\hat{T}_{yPS} = \sum_{g=1}^G N_g \left(\hat{t}_{yg}\big/\hat{N}_g\right)$	$E_M(y_i) = \beta_g$	$Var_M(y_i) = \sigma_g^2$
Stratified expansion	stsrswor	$\hat{T}_{y,st} = \sum_h N_h \bar{y}_{hs}$	$E_M(y_{hi}) = \mu_h$	$Var_M(y_{hi}) = \sigma_h^2$
Stratified combined ratio	stsrswor	$\hat{T}_{CR} = \hat{T}_{y,st} \frac{N\bar{x}_U}{\sum_h N_h \bar{x}_{hs}}$	$E_M(y_{hi}) = \beta x_{hi}$	$Var_M(y_{hi}) = \sigma^2 x_{hi}$
Stratified separate ratio	stsrswor	$\hat{T}_{SR} = \sum_h N_h \bar{y}_{hs} \frac{\bar{x}_{Uh}}{\bar{x}_{hs}}$	$E_M(y_{hi}) = \beta_h x_{hi}$	$Var_M(y_{hi}) = \sigma_h^2 x_{hi}$
Stratified combined regression	stsrswor	$\hat{T}_{CL} = \hat{T}_{y,st} + N\hat{B}(\bar{x}_U - \bar{x}_{st})$ with $\bar{x}_{st} = \sum_h N_h \bar{x}_{hs}/N, \hat{B} = \tilde{s}_{xy}/\tilde{s}_x^2$ $\tilde{s}_{xy} = \sum_h \frac{N_h}{n_h} \sum_{s_h} (x_{hi} - \bar{x}_{hs})(y_{hi} - \bar{y}_{hs})$ $\tilde{s}_x^2 = \sum_h \frac{N_h}{n_h} \sum_{s_h} (x_{hi} - \bar{x}_{hs})^2$	$E_M(y_{hi}) = \alpha + \beta x_{hi}$	$Var_M(y_{hi}) = \sigma^2$
Stratified separate regression	stsrswor	$\hat{T}_{SL} = \sum_h N_h \left[\bar{y}_{hs} + \hat{B}_h(\bar{x}_{Uh} - \bar{x}_{hs})\right]$	$E_M(y_{hi}) = \alpha_h + \beta_h x_{hi}$	$Var_M(y_{hi}) = \sigma_h^2$

14.3.2 More General Examples

To illustrate a GREG that uses both quantitative and qualitative auxiliaries, we use the Survey of Mental Health Organizations population. The file, `smho.N874`, contains 874 hospitals and is a subset of the `smho98` file introduced in Chap. 3. The variables on the file are:

`hosp.type`	hospital type (1=psychiatric, 2=residential/veterans, 3=general, 4=outpatient/partial case, and 5=multi-service/substance abuse)
`EXPTOTAL`	total expenditures in 1998
`BEDS`	total inpatient beds
`SEENCNT`	unduplicated client/patient count seen during year
`FINDIRCT`	hospital receives money from the state mental health agency (1=yes, 2=no)

The following code will load the data set into R:

```
smho.N874 <- read.csv("smho.N874.csv", row.names = 1)
```

Suppose the goal is to estimate the total of expenditures in some year after 1998, but we use the 1998 file to explore whether any of the covariates, `BEDS`, `SEENCNT`, `EOYCNT`, and `FINDIRCT`, would be useful predictors. For this illustration, we drop the cases with hospital type = 4. Many of these are outpatient units that do not have inpatient beds; beds will obviously not be related to expenditures for them. The 725 organizations other than type 4 can be retained with the following R code. Note that the exclamation point instructs the software to keep only records in `smho.N874` that are not in the delete vector:

```
delete <- smho.N874$hosp.type == 4
smho <- smho.N874[!delete, ]
```

A useful first step is to make a scatterplot matrix of the quantitative variables in the problem, as shown in Fig. 14.2. The correlation of expenditures (`EXPTOTAL`) with number of beds (`BEDS`) is reasonably high at 0.70 but is less for patient count (`SEENCNT`) and end-of-year patient count (`EOYCNT`), 0.35 and 0.30, respectively. Nevertheless, the two count variables may be useful predictors. To explore relationships further, we drew Fig. 14.3 which plots expenditures versus beds separately for each hospital type. The gray line in each panel is a nonparametric smoother designed to reflect the relationship of two variables without specifying any particular model. There is some evidence that the slope for beds depends on hospital type. The same may be true for the slopes for `SEENCNT` and `EOYCNT`, but, for this example, we will not pursue this possibility.

Next, we can do some more formal modeling. The R code below fits a model with common slopes for SEENCNT and EOYCNT but a different slope for BEDS in each hospital type:
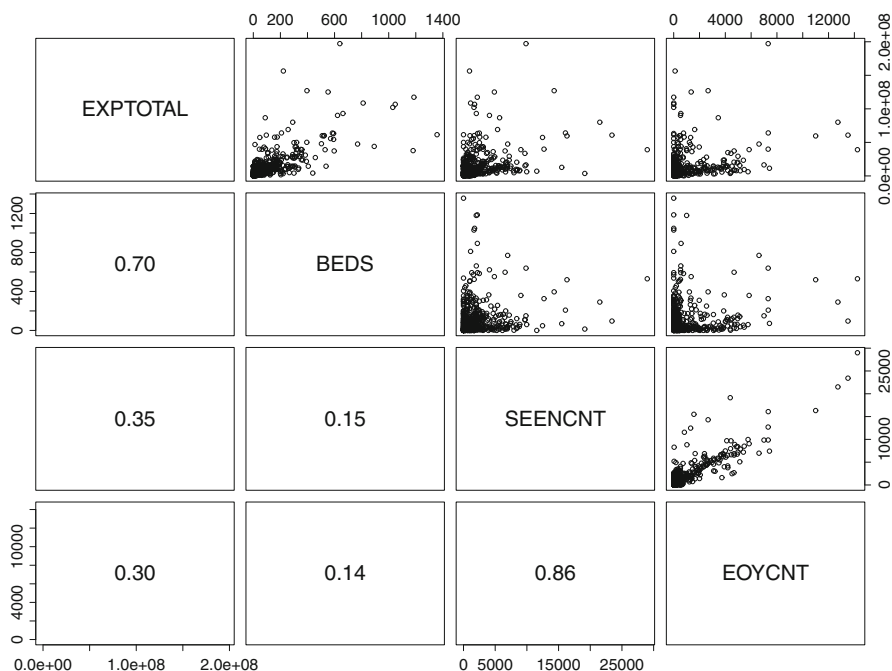
Fig. 14.2: Scatterplot matrix of variables in the `smho.N874` data set.

```
                # Separate slope on beds in each hosp type
m2 <- glm(EXPTOTAL ~SEENCNT + EOYCNT +
                    as.factor(FINDIRCT) +
                    as.factor(hosp.type):BEDS,
            data = smho)

summary(m2)
Coefficients:
                            Estimate Std. Error t value Pr(>t)
(Intercept)                1318589.1   912432.2   1.445 0.148856
SEENCNT                       1033.9      310.6   3.329 0.000918 ***
EOYCNT                        2036.2      603.6   3.373 0.000782 ***
as.factor(FINDIRCT)2         78026.1   965237.6   0.081 0.935595
as.factor(hosp.type)1:BEDS   98139.3     3318.8  29.570 < 2e-16  ***
as.factor(hosp.type)2:BEDS   39489.4     5644.5   6.996 6.05e-12 ***
as.factor(hosp.type)3:BEDS   77578.4    15082.2   5.144 3.48e-07 ***
as.factor(hosp.type)5:BEDS   36855.8     8650.5   4.261 2.31e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

All parameters estimates are significantly different from 0 except for
FINDIRCT and the intercept. We show one final diagnostic plot in Fig. 14.4,
although there are many that could be done. This figure plots the studentized
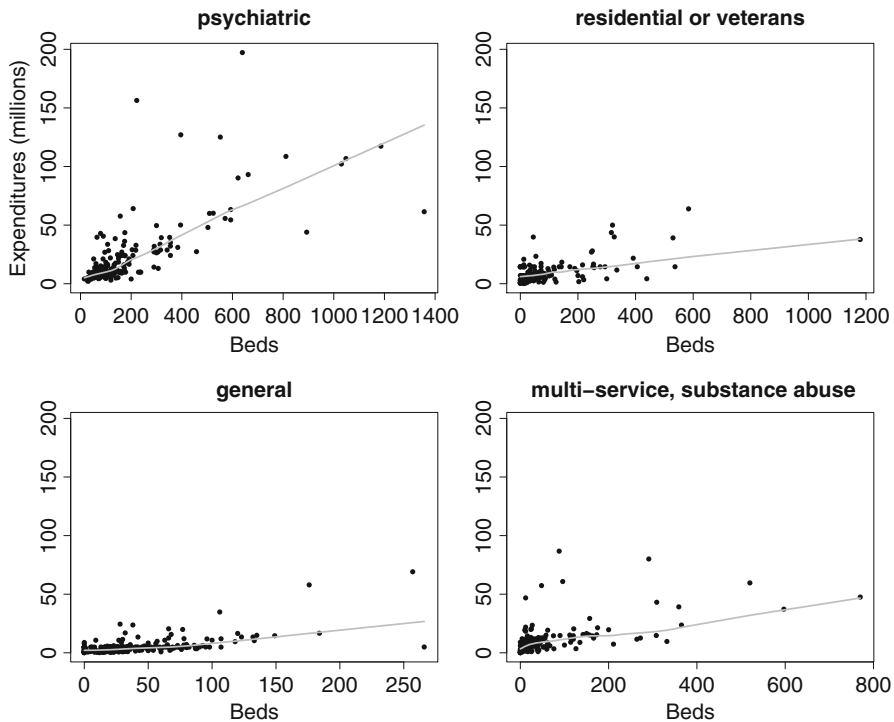
Fig. 14.3: Plots of expenditures versus beds for the four hospital types. The line in each panel is a nonparametric smoother.
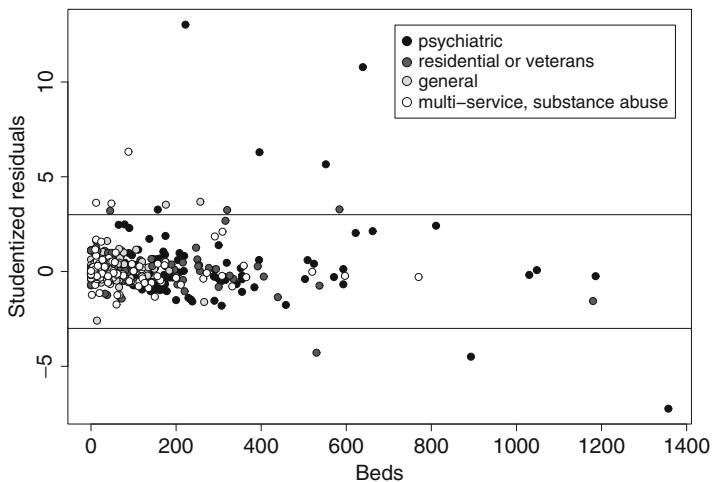
Fig. 14.4: Studentized residuals plotted versus beds for the `smho.N874.sub` data. Reference lines are drawn at ± 3.

residuals from the model versus beds. Hospital types are shown in different shades. Most points fall within the bands drawn at ± 3, but there are a number outside the bands. Part of this may be due to nonhomogeneous variance, which is visible in Fig. 14.3. Using a weighted regression with weights proportional to beds to some power might help, but some of the smaller hospitals have large standardized residuals. Some of the most extreme are psychiatric hospitals that have either a large number of beds or large value of expenditures.

In the first panel of Fig. 14.3, we saw that the plot of expenditures versus beds was extremely diffuse for psychiatric hospitals. If these large organizations could be recognized in advance of sampling, they might be selected with certainty, as described in Chap. 3. After the sample is selected, it might be prudent to exclude such units, whether they are certainties or not, from the process of computing calibration weights. They can have a harmful effect on weights and resulting estimates since the implied slope in a GREG estimate can be affected by extreme points. Of course, residuals for variables other than total expenditures may not be extreme. As a result, the decision about whether to exclude particular units from computation of weights is not clear-cut.

To illustrate calibration, we will select a sample from the subset of `smho98` that excludes hospital type 4 and use the same model as above. The code below uses the `sampling` package to select a sample with probability proportional to the square root of (recoded) beds. The method of selection is to randomize the order of the population and then select a systematic sample (see Hartley and Rao 1962). First, the value of beds is recoded to have a minimum of 5; otherwise, any hospital with 0 beds cannot be selected. The base weights are in the `d` vector:

```
require(sampling)
x <- smho[,"BEDS"]
        # recode small hospitals to have a minimum MOS
x[x <= 5] <- 5
x <- sqrt(x)
n <- 80
set.seed(428274453)
pk <- n*x/sum(x)
sam <- UPrandomsystematic(pk)
sam <- sam==1
sam.dat <- smho[sam, ]
d <- 1/pk[sam]
```

The counts of sample units by hospital type are 33, 15, 17, and 15 so that all types are represented. Next, the `survey` package is used to create a design object, `smho.dsgn`, that is then used in the calibrate function to compute GREG weights. This function accepts a number of parameters as discussed in Example 14.4:

```
smho.dsgn <- svydesign(ids = ~0,                # no clusters
                       strata = NULL,           # no strata
                       data = data.frame(sam.dat),
                       weights = ~d)

        # Compute pop totals of auxiliaries
        # Note these are the original not the recoded x's
x.beds <- by(smho$BEDS, smho$hosp.type, sum)
x.seen <- sum(smho$SEENCNT)
x.eoy <- sum(smho$EOYCNT)
N <- nrow(smho)

pop.tots <- c('(Intercept)' = N,
              SEENCNT = x.seen,
              EOYCNT = x.eoy,
              x.beds = x.beds)

sam.lin <- calibrate(design = smho.dsgn,
                     formula = ~SEENCNT + EOYCNT +
                         as.factor(hosp.type):BEDS,
                     population = pop.tots,
                     calfun="linear")
```

The parameter setting `calfun=c("linear")` results in GREG weights being computed. As in poststratification and raking, we can check whether the calibration constraints were satisfied:

```
svyby(~BEDS, by=~as.factor(hosp.type), design=sam.lin,
      FUN=svytotal)
  as.factor(hosp.type)  BEDS      se.BEDS
1                    1 37978 1.826570e-12
2                    2 13066 6.289865e-13
3                    3  9573 6.799993e-13
5                    5 10077 1.398118e-12

svytotal(~SEENCNT, sam.lin)
          total        SE
SEENCNT 1349241 5.755e-11

svytotal(~EOYCNT, sam.lin)
       total        SE
EOYCNT 505345 1.911e-11
```

Since the SEs are essentially 0, a set of weights has been obtained that satisfy $\sum_s w_i \mathbf{x}_i = \mathbf{t}_x$. The calibrate function will also issue an error message if the calibration fails for any reason. Examining summary statistics for the weights is always wise. When this is done, we see that at least one GREG weight (-0.3983) is negative even though the smallest base weight was 2.714:

```
summary(weights(smho.dsgn))
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.714   5.693   8.150   8.763  10.090  33.680
summary(weights(sam.lin))
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.     Max.
-0.3983  5.7470  8.8320  9.0630 10.9300 33.8300
```

There is nothing in the GREG algorithm that prevents negative weights, although in a sample where all selection probabilities are small and the resulting input weights are large, this is unlikely to happen. In theory, even with negative weights, the GREG will be approximately design unbiased and, if the model is specified correctly, model unbiased for population totals. However, negative weights could have a serious effect on some domain estimates, and users are generally uncomfortable with weights that are negative. In fact, some software packages will not allow negative weights. To help remedy this potential problem, the calibrate function has a bounds parameter which gives the relative amount that the final weights can differ from the input weights. To do the bounding, this restriction is added to the calibration problem described in Sect. 14.1:

$$L \leq \frac{w_i}{d_i} \leq U \text{ for all } i \in s.$$

In words, the calibrated weight for each sample case must be larger than a lower bound L times the input weight and less than an upper bound U times the input weight. (This is synonymous with bounding the g-weights because $w_i = d_i g_i$.) Thus, the bound is on the relative change in the initial weight— not on the final weight itself. The bounds are arbitrary. For example, you might want to require that a final weight be somewhere between $1/2$ and 3 times the initial weight. If the input weights are positive (which they will be if they are inverses of selection probabilities or nonresponse-adjusted, inverse probabilities), then the bounded weights will be positive. It is easy to make the bounds so tight that the calibration will fail, and some trial and error may be needed to arrive at values that will work in a particular problem.

In Sect. 14.4 the issue of weight variability will be covered in more detail. Here, we illustrate how to bound the weight changes using either the GREG or raking distance functions. In this case, the final weights are required to be within 0.4 and 3 times the input weights. When bounds are set, an iterative procedure is used to arrive at a final set of weights. Three parameters that may be useful are:

maxit	Number of iterations allowed before the procedure stops. Default value 50
epsilon	Tolerance in matching population total. Default value 10^{-7}
force	Return an answer even if the specified accuracy was not achieved. Default value is FALSE

If convergence is not obtained with the default settings, increasing the number of iterations allowed and loosening the tolerance may help. If force=TRUE, the approximately calibrated design object will still be returned. Checking

how closely the constraints are satisfied may help determine why the calibration failed. The bounds parameter can be used with either the least squares or raking distance functions as shown below (the parameter, `bounds`, is not available if `calfun = "logit"`):

```
            # Linear calibration with bounds
sam.linBD <- calibrate(design = smho.dsgn,
                       formula = ~SEENCNT + EOYCNT +
                           as.factor(hosp.type):BEDS,
                       population = pop.tots,
                       bounds = c(0.4, 3),
                       calfun = "linear")

            # Check controls
svyby(~BEDS, by=~as.factor(hosp.type), design=sam.linBD,
      FUN=svytotal)
svytotal(~BEDS, sam.linBD)
svytotal(~SEENCNT, sam.linBD)
svytotal(~EOYCNT, sam.linBD)

            # raking
sam.rake <- calibrate(design = smho.dsgn,
      formula = ~ SEENCNT + EOYCNT + as.factor(hosp.type):BEDS,
      population = pop.tots,
      bounds = c(0.4, 3),
      calfun = "raking",
      maxit = 100, epsilon = 1e-4)
```

In the raking code above, the settings `maxit = 100, epsilon = 1e-4` were used in order to obtain convergence. With the default settings, `calibrate` will report that convergence was not achieved, although for practical purposes it has been.

The results are generated with the `svytotal` and `cv` R functions (as in Example 14.2) and are summarized in Table 14.5. The estimates are greater than the population total, but in this case, a 95 % normal-approximation confidence interval will contain the actual population total of \$8.774 billion in all cases. For example, the confidence interval based on the GREG estimate can be found with

Table 14.5: Estimated totals of expenditures, standard errors, and coefficients of variation for the π-estimate, GREG, bounded GREG, and bounded raking estimates in a *pps* sample from a subset of the Survey of Mental Health Organizations population.

Estimator (design object)	Estimated total (000s)	SE (000s)	$CV(\%)$
Population	8,774,651		
π-estimate (`smho.dsgn`)[1]	9,322,854	915,126	9.81
GREG (`sam.lin`)	9,563,683	748,596	7.82
Bounded GREG (`sam.linBD`)	9,612,035	744,746	7.75
Bounded raking (`sam.rake`)	9,529,511	732,273	7.68

```
confint(svytotal(~ EXPTOTAL, sam.lin))
```

As Table 14.5 shows, the GREG, bounded GREG, and bounded raking esti-
mates have smaller estimated SEs and CVs than the π-estimate in this sam-
ple. Each of the CVs for the calibrated estimates is about 79 % of that of the
π-estimate. Thus, calibrating gives a substantial increase in precision.

Exploring how the weights for these different methods compare is worth-
while. Figure 14.5 shows a plot of the weights for the three calibration meth-
ods versus the base weights in the left-hand panel. A 45° line is drawn where
the weights would equal the base weights. Most weights are increased slightly
to hit the control totals, but a few are noticeably decreased. In the right-
hand panel, the ratios, w_i/d_i, are plotted versus the base weights. The upper
bound of 3 clearly has no effect. The unit with the negative weight is marked
by an arrow. Using a lower bound of 0.4 causes several weights, including
the negative one, to move to the boundary. Comparing the points from the
unbounded linear GREG and the two bounded methods, it is apparent that
bounding would not affect most units much but would eliminate the objec-
tionable negative weight.
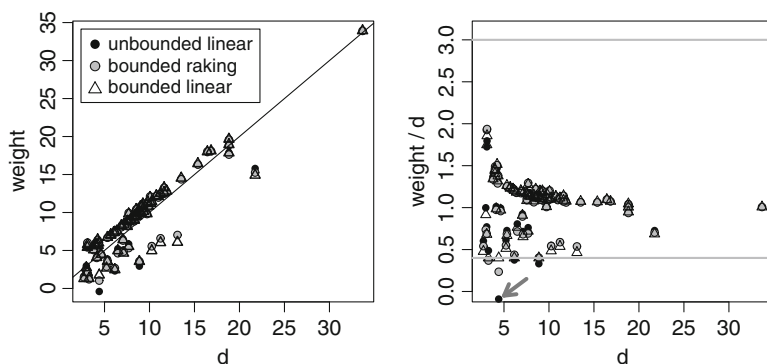
Fig. 14.5: Plots of weights for the different methods of calibration in a *pps sample*.
A 45° line is drawn in the left-hand panel. Reference lines are drawn at the weight
bounds, 0.4 and 3, in the right-hand panel.

Selecting covariates to use in calibration is, in some ways, even more dif-
ficult than in a typical modeling problem because the weights can be used
for many response variables. To illustrate how much difference the choice
of covariates can make, we recompute GREG weights using a model that
has parameters for SEENCNT, EOYCNT, a common slope for BEDS, and
main effects for hospital type. This differs from the model above which had a
separate slope for each hospital type and did not include controls on the
number of hospitals. The code for computing unbounded GREG weights
follows:

```
N.hosp <- table(smho$hosp.type)
x.beds <- sum(smho$BEDS)
pop.tots <- c(BEDS = x.beds,
              SEENCNT = x.seen,
              EOYCNT = x.eoy,
              HOSP = N.hosp)

sam.lin2 <- calibrate(design = smho.dsgn,
                         formula = ~0 + BEDS + SEENCNT + EOYCNT
                           + as.factor(hosp.type),
                         population = pop.tots,
                         bounds=c(-Inf,Inf),
                         calfun="linear")
```

Next, we estimate the total of expenditures and the proportion of hospitals receiving financing from state mental health agencies (FINDIRCT). Results are in Table 14.6. The CV of total expenditures, the quantitative total, is 7.77 % for the new model, labeled GREG 2 in the table. This is about the same as for the model with a separate slope for beds in each hospital type and no controls on hospital counts by type, labeled GREG 1. But, for FINDIRCT, the GREG-2 estimate has a CV of 9.91 % compared to 16.92 % for GREG 1. This gain is consistent with the fact that the population proportions in the four hospital types are much different—0.67, 0.80, 0.94, and 0. That is, the means differ by hospital type, implying that a hospital type factor should be in a model predicting FINDIRCT.

Based on these two examples, a model with BEDS + SEENCNT + EOYCNT + as.factor(hosp.type) might be preferred. However, there may be other analysis variables for which another set of auxiliaries could be more efficient. When computing weights in most surveys, using different auxiliaries for different analysis variables is cumbersome and impractical. The goal is to find a general-purpose set of weights that will be reasonably efficient for most estimators. Considering a broad set of analysis variables may be necessary to make a good decision about which ones to select.

In some respects, considering covariates seems to have made the estimation problem much harder because of the uncertainty in which ones to use. Using the π-estimator, in contrast, is simple since we just compute selection probabilities and invert them and we have a set of weights. However, this simplicity is misleading because a set of good covariates will reduce SEs appreciably as illustrated in Tables 14.5 and 14.6.

The unfortunate thing that often happens in practice is that weights are computed without examining any analysis variables at all. This may be because the time schedule is so tight that weighting and editing of data must occur in parallel, so that the analysis variables are unavailable to the staff constructing the weights. Or, it may be that an organization has always done the weighting without benefit of the analysis variables (whether they are available or not). Divorcing weighting from analysis is common in one-time surveys that will not be repeated. In that case, general rules of thumb may be used to select covariates or a simple procedure like poststratification

Table 14.6: Estimated totals of expenditures and proportion of hospitals with direct state financing, standard errors, and coefficients of variation for the π-estimate and two choices of GREG in a *pps* sample from a subset of the Survey of Mental Health Organizations population.

	Estimate or population value	SE	CV (%)
Total expenditures (000s)			
Population	8,774,651		
π-estimate	9,322,854	915,126	9.82
GREG 1	9,563,683	748,596	7.83
GREG 2	9,161,491	711,633	7.77
Proportion with financing from state mental health agency			
Population	0.336		
π-estimate	0.323	0.059	18.16
GREG 1	0.303	0.051	16.92
GREG 2	0.340	0.034	9.91

may be used. In continuing surveys that are periodically repeated, there is an opportunity to use prior data to guide weight creation. Regardless of the circumstances, looking at how a proposed implementation of poststratification, raking, GREG, etc., performs for some important estimates is always a good practice.

14.4 Weight Variability

Having survey weights that vary is common. Reasons for variability include:

(1) Varying selection probabilities as would occur in *pps* sampling or stratified sampling with different sampling rates in the strata
(2) Over or undersampling groups of units in two-phase sampling based on domain membership
(3) Unequal response rates (and/or rates of unknown eligibility) in different subgroups leading to unequal weight adjustments
(4) Calibration to auxiliaries to reduce variances or correct for frame coverage errors

In some cases, varying weights may be designed into the sample, as in (1) and (2) above. In other applications, varying weights are needed to correct for potential nonresponse bias or differential undercoverage as in (3) and (4). However, highly differential weights can increase the variances of estimates even if they decrease bias.

Practitioners will often worry about having unequal weights, particularly in household surveys. Whether this is a genuine concern depends on the

situation. This section reviews some measures of weight variability, how they are derived, and how they should be interpreted. We also show how to use quadratic programming and more arbitrary weight-trimming methods to bound weights.

14.4.1 Quantifying the Variability

Kish (1965, 1992) introduced a "design effect due to weighting" which is equal to one plus the relvariance of the sample weights:

$$deff_w = 1 + relvar(w)$$
$$= 1 + n^{-1}\sum_s (w_i - \bar{w})^2 \Big/ \bar{w}^2,$$

where $\bar{w} = n^{-1}\sum_s w_i$. The term $deff_w$ is also known as an unequal weighting effect (e.g., Liu et al. 2002). This is a widely used, and possibly over-used, measure that is interpreted as the increase in variance of an estimator due to having weights that are not all the same. For example, Kish also writes $deff_w$ as $1 + L$ with L being the inflation above the variance that would be obtained with a self-weighting sample. Practitioners often compute $deff_w$ while developing the final weights and use it to make a judgment about whether any weights should be modified because they are "too variable." There appear to be no universally accepted rules of thumb to gauge when $deff_w$ is "large." For better or worse, $deff_w$ values of 1.5 or larger frequently lead to some action being taken.

To understand whether this measure is applicable to a specific survey, we need to understand how it is derived. Consider an *stsrswor* with n_h sample units allocated to stratum h. The number of units in the stratum population is N_h, and the proportion of the population in stratum h is $W_h = N_h/N$. As shown in Kish (1965), $deff_w$ is the ratio of the variance of the stratified expansion estimator, $\bar{y}_{st} = \sum_{h=1}^{H} W_h \bar{y}_{sh}$, with unequal weighting to the variance of the same estimator with proportional allocation, assuming stratum variances are *equal*:

$$deff_w = \frac{V(\bar{y}_{st}) \text{ with non} - \text{proportional allocation}}{V(\bar{y}_{st}) \text{ with proportional allocation}}.$$

The key assumption is that a proportional allocation is optimal for the study. This allocation, as discussed in Sect. 3.1.2, is used only when the assumption that the stratum population standard deviations are all equal $(S_h = S)$ is reasonable. In this special case, $deff_w$ measures the change in the variance associated with the deviation from the presumed optimal design. However, variation in the weights is appropriate if the $S_h = S$ assumption is not reasonable or if any of the conditions discussed at the beginning of

Sect. 14.4 exist. Although $deff_w$ is motivated by stratified sampling, it is commonly applied to any type of sample where weights vary.

The measure $deff_w$ may be useful if equal weighting is appropriate, i.e., stratum variances are equal or, at least, not expected to be extremely different. This may be true in household surveys. However, $deff_w$ is largely irrelevant in many applications. Among them are:

- Establishment or institution surveys where variances are known to differ among strata
- Household surveys where different subgroups are intentionally sampled at different rates to obtain desired sample sizes
- Surveys where different groups respond at substantially different rates so that nonresponse adjustments, which are needed to reduce bias, create different size weights even though the initial sample may be self-weighting.

In these cases, as noted in Kish (1992), differential weights can be much more efficient than equal weights.

The best use of $deff_w$ may be as a diagnostic after weights are calculated. Large values may signal that the results of different steps should be checked to see whether any errors have occurred or whether a particular step injects a lot of unjustified variability in the weights. The nonresponse adjustment step is usually one that can be quite sensitive to how weighting classes are formed or propensities are estimated. If it is felt that extreme adjustments are untrustworthy and are not really correcting bias, this is a good reason to modify the procedure in some way.

To get a feel for the values that $deff_w$ can take, consider the case of two strata and an *stsrswor* design. Suppose that the proportion of the sample in stratum h is $p_h = n_h/n$ and the weight of each unit in stratum h is w_h ($h = 1, 2$). When the sampling fractions are negligible in each stratum, the value of Kish's $1 + L$ can be shown to be (see the exercises):

$$deff_w = \frac{p_1 w_1^2 + p_2 w_2^2}{(p_1 w_1 + p_2 w_2)^2}. \tag{14.7}$$

This is evaluated for a few values of w_1 and w_2 for $f_1 = f_2 = 0.5$ in Table 14.7. If the ratio of weights in strata 1 and 2 is 3:1, then $SE(\bar{y}_{st})$ is inflated by only 12 %. If the ratio is 50:1, $\sqrt{deff_w} = 1.39$. Ratios of the maximum to minimum weight can be much larger than 50:1 in some surveys. For example, in the 1998 US Survey of Mental Health Organizations (SMHO) that we are using as an example in this book, this ratio was about 160:1 (Li and Valliant, 2009). The exercises give an example using smho.N874 where, in a *pps* sample, $deff_w$ is almost 20. However, to intelligently interpret these ratios, always be aware of the caveat that unequal weights may be needed for efficient estimation. You need to consider the characteristics of particular survey variables to decide whether weight variability is a problem.

Kish (1987b) also suggested a measure similar to $deff_w$ for cluster samples. A formal justification of the measure using a model was given by

Table 14.7: Kish's $deff_w$ measure for variance inflation due to unequal weighting for a case of two strata with equal allocations ($f_1 = f_2 = 0.5$) to the strata.

w_1	3	5	10	15	20	50
w_2	1	1	1	1	1	1
$deff_w$	1.25	1.44	1.67	1.77	1.82	1.92
$\sqrt{deff_w}$	1.12	1.20	1.29	1.33	1.35	1.39

Gabler et al. (1999). Suppose that a cluster sample is selected and each sample unit is assigned to one of $\gamma = 1, \ldots, G$ weighting classes. The number of sample units in class γ is $n_\gamma \left(n = \sum_\gamma n_\gamma \right)$. Suppose the following simple variance model holds for an analysis variable y_{ij} associated with unit j in cluster i $(i = 1, \ldots, I)$:

$$Cov_M \left(y_{ij}, y_{i'j'} \right) = \begin{cases} \sigma^2 & i = i', j = j', \\ \rho\sigma^2 & i = i', j \neq j', \\ 0 & \text{otherwise.} \end{cases} \tag{14.8}$$

In words, units all have a common variance σ^2, different units in the same cluster have correlation ρ, and units in different clusters are uncorrelated. Gabler et al. (1999) considered the weighted sample mean, $\bar{y}_w = \sum_{i \in s} \sum_{j \in s_i} w_{ij} y_{ij} \big/ \sum_{i \in s} \sum_{j \in s_i} w_{ij}$. They showed that, in the case where all units in a weighting class have the same weight ($w_i = w_\gamma$, $i \in s_\gamma$), the ratio of the model variance of $\bar{y}_w$ under (14.8) to the variance of the unweighted mean under a model where all y's are uncorrelated is

$$deff_{2w} = n \frac{\sum_\gamma n_\gamma w_\gamma^2}{\left(\sum_\gamma n_\gamma w_\gamma \right)^2} \left[1 + \rho \left(b^* - 1 \right) \right],$$

where $b^* = \sum_s \left(\sum_\gamma w_\gamma n_{i\gamma} \right)^2 \big/ \sum_\gamma w_\gamma^2 n_\gamma$ with $n_{i\gamma}$ being the number of sample units in weighting class γ that are in cluster i. If the sample size in each cluster is the same, $\bar{b}$, then $deff_{2w}$ is bounded above:

$$deff_{2w} \leq n \frac{\sum_\gamma n_\gamma w_\gamma^2}{\left(\sum_\gamma n_\gamma w_\gamma \right)^2} \left[1 + \rho \left(\bar{b} - 1 \right) \right].$$

The value of the bound was the suggestion in Kish (1987b).

SUDAAN® (RTI International, 2012), software designed to analyze survey and other correlated data mentioned in Chap. 3, allows the calculation of four different design effects. The various versions are calculated through

different assumptions used for the denominator variance estimator. The formulas estimate the amount of variance inflation associated with combinations of the design features (i.e., clustering, stratification, differential sampling rates, and unequal weights). Specific SUDAAN procedures, accessible within a SAS program through its "SAS-callable" version, are discussed in Chap. 15.

A measure that comes closer to accounting for the possibility that variable weights may be efficient is one derived by Spencer (2000). Suppose that p_i is the 1-draw selection probability of unit i and that p_i and an analysis variable y_i are correlated. For example, this would be reasonable in the hospitals population if a probability proportional to number of beds (x_i) sample were selected and the analysis variable was number of patients discharged. In that case, $p_i \propto x_i$, and as we saw in Chap. 3, the number of discharges is related to number of beds. Suppose this linear model holds for Y:

$$y_i = \alpha + \beta p_i + \varepsilon_i. \tag{14.9}$$

The finite population ordinary least-squares estimates of α and β are $\alpha_U = \bar{y}_U - \beta_U \bar{p}_U$ and

$$\beta_U = \sum_U (y_i - \bar{y}_U)(p_i - \bar{p}_U) \left/ \sum_U (p_i - \bar{p}_U)^2, \right.$$

where $\bar{y}_U$ and $\bar{p}_U$ are finite population means. These equations can be rewritten using the fact that $\bar{p}_U = \sum_U p_i/N = 1/N$. The finite population variance of the errors, $\varepsilon_i = y_i - (\alpha_U + \beta_U \bar{p}_U)$, is $\sigma_\varepsilon^2 = \left(1 - \rho_{yp}^2\right)\sigma_y^2$ with $\sigma_y^2 = N^{-1}\sum_U (y_i - \bar{y}_U)^2$ and ρ_{yp} being the population correlation between y and p. The weight for unit i is $w_i = 1/(np_i)$. If the sample is selected with replacement, the *pwr*-estimator from Sect. 3.2.1 of the population total is $\hat{T}_{pwr} = \sum_s w_i y_i$. Its design variance is $V\left(\hat{T}_{pwr}\right) = n^{-1}\sum_U (y_i/p_i - T)^2$. Substituting values from model (14.9) in this variance formula and taking the ratio of the result to the variance of the estimated total under *srs* with replacement, Spencer obtained the following approximate expression for a design effect due to unequal weighting:

$$deff_{Sw} = \left(1 - \rho_{yp}^2\right)\frac{n}{N}\bar{w}_U + \left(\frac{\alpha_U}{\sigma_y}\right)^2\left(\frac{n}{N}\bar{w}_U - 1\right), \tag{14.10}$$

where $\bar{w}_U = \sum_{i \in U} w_i/N$ is average weight in population. Spencer's $deff_{Sw}$ can be estimated by

$$\widehat{deff_{Sw}} = \left(1 - \hat{\rho}_{yp}^2\right)[1 + relvar(w)] + \left(\frac{\hat{\alpha}}{\hat{\sigma}_y}\right)^2 relvar(w),$$

where $\hat{\rho}_{yp}^2$ and $\hat{\alpha}$ are the R-squared and estimated intercept values calculated from fitting model (14.9) by survey-weighted least squares.

The estimated population unit variance is $\hat{\sigma}_y^2 = \sum_s w_i (y_i - \bar{y}_w)^2 \big/ \sum_s w_i$ and $[1 + relvar(w)] = n \sum_s w_k^2 \big/ \left(\sum_s w_k\right)^2$ as in Kish's $deff_w$. When σ_y is large relative to α and $\rho_{yp} = 0$, Spencer's and Kish's measures are about the same. Note that, in general, Spencer's $deff_{Sw}$ depends on y and will, thus, be different depending on the analysis variable considered.

A deficiency of Spencer's formula is that it applies only to a *pwr*-estimator. In practice, in cases where auxiliary variables are used in sampling, they are also used in estimation. Henry (2011) filled this gap by extending Spencer's result to GREG estimators.

The next example evaluates Kish's and Spencer's design effects for a sample from a population where there is a clear relationship between y and an auxiliary variable used in sample selection. For illustration we use an artificial "HMT" population generated in the same way as the one in Hansen et al. (1983), which is a famous paper published by three of the important, historical figures in applied sampling. The generating model was $y_i = \alpha + \beta x_i + \varepsilon_i$ where x and y both have gamma distributions and the errors have a variance that increases in proportion to $x^{3/2}$. The R function, HMT.fcn, on the book web site was used to create a population of 5,000 units. Figure 14.6 is a plot of a sample of 500 units from the population.

Example 14.5 (Comparison of Spencer's and Kish's deff's). Using the R sampling package, one sample of $n = 80$ was selected from the HMT population with probabilities proportionate to x. Kish's and Spencer's *deff*'s were computed using the following code:

```
            # load sampling package
require(sampling)
            # Load function to generate HMT pop
source("HMT.fcn.R")
            #Random seed for sample selection
set.seed(-500398777)

    # Generate HMT population
pop.dat <- as.data.frame(HMT.fcn())
            #Population size
N <- nrow(pop.dat)

# Calculate 1-draw selection probabilities - pps
            #MOS = x
mos <- pop.dat$x
            #Calculate 1-draw selection probabilities
pop.dat$prbs.1d <- mos / sum(mos)

# Select sample - pps
            #Define size of sample
n <- 80
            # probabilities for selecting a sample of n
pk <- n * pop.dat$prbs.1d
            # PPS sample
sam <- UPrandomsystematic(pk)
```

```
sam <- sam==1
sam.dat <- pop.dat[sam, ]
            # Base weights
dsgn.wts <- 1/pk[sam]

# Spencer's Deff
            # Calculate WLS values
sam.wls <- lm(y ~ prbs.1d, data = sam.dat, weights = dsgn.wts)
            # DEFF component - var of y
sam.mean.y <- sum(sam.dat$y * dsgn.wts) / sum(dsgn.wts)
sam.var.y  <- sum(dsgn.wts * (sam.dat$y - sam.mean.y)^2) /
                            sum(dsgn.wts)
            # DEFF component - alpha squared
sam.alpha2 <- sam.wls$coefficients[1] \^{}2
            # DEFF component - squared correlation
sam.rho2.yP <- summary(sam.wls)$r.squared
            # DEFF component - Kish
kish.deff <- n*sum(dsgn.wts^2) / (sum(dsgn.wts)^2)
            #Spencer's DEFF
spencers.deff <- as.numeric((1 - sam.rho2.yP) * kish.deff +
                (sam.alpha2 / sam.var.y) * (kish.deff - 1))
```

The resulting values of the two design effect formulas are

```
kish.deff
[1] 1.882999
spencers.deff
[1] 0.6468291
```

Kish's *deff* claims that the variance of the π-estimator is 88 % larger than it would be with an equal probability sample. On the other hand, Spencer's *deff* of 0.65 says that $pp(x)$ sampling and the resulting unequal weighting will be more efficient than equal probability sampling. Based on the plot in Fig. 14.6, *pps* sampling is obviously better in this population. ∎
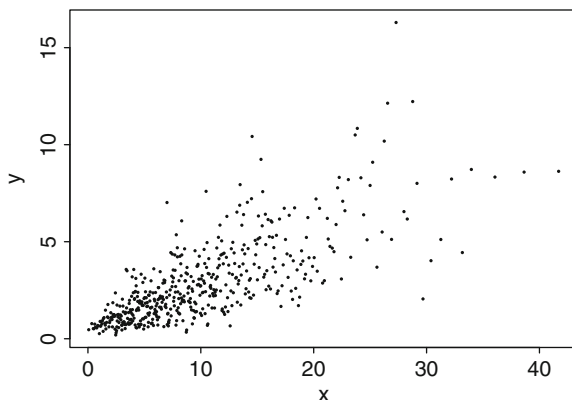
Fig. 14.6: Plot of a subsample of 500 points from the Hansen, Madow, and Tepping (1983) population.

Despite the incorrect impression conveyed by Kish's *deff* in the above example, having extremely variable weights is likely to be inefficient for at least some variables collected in a survey. In the next section, we cover several ways of limiting weight variation.

14.4.2 Methods to Limit Variability

Procedures are often used to trim extreme weights, especially large ones. The methods used in practice are mainly *ad hoc* but may improve the efficiency of estimators for some variables. We have explored some techniques earlier that are geared toward limiting extreme weights. For example, in Sect. 13.5.1, weighting classes were created for nonresponse adjustment based on response propensities. Using classes rather than individual propensities can be a way of eliminating a few large nonresponse adjustments. Constrained calibration, discussed in Sect. 13.5.1, is another way of attempting to avoid excessive weight adjustments. There are also off-the-cuff procedures that can be used to limit the range of base weights. For example, the number of phone lines or number of household residents can be top-coded when computing selection probabilities within a household.

The first method we cover is quadratic programming (QP) with constraints. Much like GREG with weight bounds, QP allows a final set of weights to be found that is calibrated to population totals for some auxiliary variables. The second method is less formal but is probably more common in practice. Large weights are arbitrarily trimmed back to an upper bound. The total weight trimmed away is then spread among the other sample units.

Quadratic Programming

One option for limiting the range of weights is quadratic programming as described in Isaki et al. (2004). A QP problem with constraints has the following general form:

Find the vector $\mathbf{k}$ to minimize $\Phi = \frac{1}{2}\mathbf{k}^T\boldsymbol{\Sigma}\mathbf{k} - \mathbf{z}^T\mathbf{k}$
Subject to the constraints $\mathbf{C}^T\mathbf{k} \geq \mathbf{c}_0$

where $\boldsymbol{\Sigma}$ is a symmetric matrix of constants and $\mathbf{z}$ is a vector of constants. The task of finding constrained weights can be phrased as a QP problem. Suppose that the input weights, which could be base weights or nonresponse-adjusted weights, are d_k ($k \in s$). The final weights to be computed are $\{w_k\}_{k \in s}$. If we require the final weights to be calibrated to population totals of some auxiliaries $\mathbf{x}$, then one QP formulation is

Find the set of weights $\{w_k\}_{k \in s}$ that minimizes $\sum_s (w_k - d_k)^2 \big/ d_k$
Subject to $\sum_s w_k \mathbf{x}_k = \mathbf{t}_x$ and $L \leq w_k \leq U$.

To see that this fits into the general QP mold, first note that

$$\sum_s (w_k - d_k)^2 \big/ d_k = \mathbf{w}^T \mathbf{D}^{-1} \mathbf{w} - 2\mathbf{d}\mathbf{D}^{-1}\mathbf{w} + \mathbf{d}\mathbf{D}^{-1}\mathbf{d}$$
$$= \mathbf{w}^T \mathbf{D}^{-1} \mathbf{w} - 2\mathbf{1}_n^T \mathbf{w} + \sum_s d_k$$

with $\mathbf{D} = diag\,(d_k)$, $\mathbf{w} = (w_1, \ldots, w_n)^T$, and $\mathbf{1}_n$ representing an $n \times 1$ vector of ones. The formulation above then corresponds to the general problem with $\mathbf{k} = \mathbf{w}$, $\mathbf{z} = 2 * \mathbf{1}_n$, and $\mathbf{\Sigma} = \mathbf{D}^{-1}$. The sum of input weights, $\sum_s d_k$, is a constant, given the sample. So, solving the weight calibration problem is equivalent to minimizing

$$\Phi = \mathbf{w}^T \mathbf{D}^{-1} \mathbf{w} - 2\mathbf{1}_n^T \mathbf{w}.$$

The bounds on the weights fit the general form, $\mathbf{C}^T \mathbf{k} \geq \mathbf{c}_0$, with

$$\mathbf{C} = \begin{pmatrix} \mathbf{X_s} \\ \mathbf{I_n} \\ -\mathbf{I_n} \end{pmatrix} \text{ and } c_0 = \begin{pmatrix} \mathbf{t_x} \\ L\mathbf{1_n} \\ -U\mathbf{1_n} \end{pmatrix},$$

where $\mathbf{X}_s$ is the $n \times p$ matrix of auxiliaries for the sample units, $\mathbf{I_n}$ is an $n \times n$ identity matrix, and $\mathbf{k} = \mathbf{w}$ as before. Note that the bounds $L \leq w_k \leq U$ are different from the bounds, $L \leq w_k/d_k \leq U$, used for the GREG weights in Sect. 14.3.2. Using $L \leq w_k \leq U$ is in some ways preferable to the earlier GREG constraint because it directly bounds the sizes of the final weights. In contrast, the GREG constraint bounds only the relative change from the initial weights, i.e., the size of the g-weights. If the initial weights are extreme, then the final GREG weights are likely to be also.

The R package `quadprog` (Turlach and Weingessel 2011) can solve the QP problem. To illustrate this, we work a variation of the example in Sect. 14.3.2 using the `smho.N874` data set. Auxiliaries are BEDS, SEENCNT, EOYCNT, and hosp.type.

Example 14.6 (Constrain weights using quadratic programming). This example uses the same sample of 80 hospitals as in Sect. 14.3.2, which was selected with seed 428274453, after recoding all units to have a minimum of 5 beds and after eliminating type-4 hospitals. The range of base weights in the sample that was selected with probabilities proportional to the square root of recoded beds was 2.71 to 33.68. Suppose that we want to constrain the weights to be in the range $[L, U] = [2, 18]$. The model is the same as the one in the earlier section and includes the auxiliaries, SEENCNT, EOYCNT, and `as.factor(hosp.type):BEDS`. The complete listing of code to select

the sample and compute the QP weights is in the file `Example 14.6 quadprog.wts.R` on the web site; excerpts are shown below.

The `quadprog` package contains a function, `solve.QP`, that solves general quadratic programming problems. As in the earlier section, the sample data are in the object `sam.dat` and the base weights are in `d`. The object names in the code below match the vectors and matrices above fairly closely. One thing to note is the `model.matrix` function which creates the matrix of auxiliary values for the sample units that is implied by a model having a separate slope on beds for each hospital type (`X.hosp`). The transposed version of this matrix is combined with the quantitative auxiliaries—BEDS, SEENCNT, EOYCNT—to create a matrix called `X`. A requirement of solve.QP is that any equality constraints be listed first in $\mathbf{C}^T\mathbf{k}$. This is accomplished in the statement that creates `Cmat` by placing the vector of ones for the overall total number of hospitals (`one`) and `X` first. Then, in the call to `solve.QP`, the parameter `meq=7` tells the function that the first seven constraints in the `bvec` vector are equalities. The resulting sample weights are calibrated to the desired population totals. The fact that this succeeded is checked in `Example 14.6 quadprog.wts.R`. The weights, $\mathbf{w}$, are contained in `fs.wts$solution`:

```
library(quadprog)
        # Tabulate pop totals for constraints
x.beds <- by(smho$BEDS, smho$hosp.type, sum)
x.seen <- sum(smho[,"SEENCNT"])
x.eoy <- sum(smho[,"EOYCNT"])
N <- nrow(smho)
X.hosp <- model.matrix(~ 0 + as.factor(hosp.type):BEDS,
        data = sam.dat)

X <- rbind(sam.dat[, "SEENCNT"],
     sam.dat[, "EOYCNT"],
     t(X.hosp))
c0a <- c(N, x.seen, x.eoy, x.beds)

        # Lower and upper weight bounds
L <- 2
U <- 18
        # Compute full sample weights via QP
In <- diag(nrow = n)
one <- rep(1, n)
c0b <- c(L * one,
        -U * one)
Cmat <- rbind(one, X, In, -In)
fs.wts <- solve.QP(Dmat = diag(1/d),
                dvec = 2 * one,
                Amat = t(Cmat),
                bvec = c(c0a, c0b),
                meq = 7
                # first 7 constraints are equality
                   constraints
        )
```

∎

One caveat in the use of `solve.QP` is that it can run out of memory when the sample size is large. The function requires that `Dmat` be input as an $n \times n$ matrix even though it is diagonal and more compact methods of storage are available. Increasing the amount of memory available to R may help. In Windows, do this with the command `memory.limit(size=4095)`.

The extent to which the constraints affect the input weights depends on which units are randomly sampled. This variation will not be reflected by standard variance formulas, but one option is to use a replication variance estimator. We will cover these in more detail in Chap. 15, but will take this opportunity to illustrate one version of replication—the jackknife. The idea behind the jackknife is to delete one unit from the sample, adjust the input weights for the jackknife subsampling, compute weights in whatever way is being used, and then use the resulting weights to compute an estimate. The process is repeated until n replicate estimates have been computed. The variation of the replicate estimates is then computed around the full sample estimate. For many types of estimates, e.g., totals, means, and combinations of them, a theory has been developed to justify the use of the jackknife. However, it does not work for all types of estimates, and there is no theory to say that a consistent or unbiased variance estimator is produced when the weights are quadratically constrained. In this example, the jackknife does produce reasonable answers, and we use it for illustration.

The code below loops through all units in the sample, deleting one at a time and resolving the quadratic program to give a set of $n=80$ jackknife weights. We then use the `survey` package to compute estimates and SEs. The results are listed in the "Bounded QP" rows of Table 14.8. For comparison, we include estimates generated from the other calibration methods with `svytotal` and `cv` in the previous examples (they are also computed in Example 14.6 `quadprog.wts.R`.):

```
        # Compute jackknife version of weights
        # Matrix to hold jackknife weights
 rep.wts <- matrix(0, nrow = n, ncol = n)

 for (k in 1:n){
     fill <- (1:n)[-k]
     In <- diag(nrow = n-1)
     one <- rep(1, n-1)
     c0b <- c( L * one,
               -U * one)
     Cmat <- rbind(rep(1,n-1), X[,-k], In, -In)

     wts <- solve.QP(Dmat = diag(1/d[-k]),
             dvec = 2 * one,
             Amat = t(Cmat),
             bvec = c(c0a, c0b),
             meq = 7)
     rep.wts[k, fill] <- wts$solution
 }
```

```
        # make jackknife design object
library(survey)
qp.dsgn <- svrepdesign(weights = fs.wts$solution,
           repweights = t(rep.wts),
           type = "JK1",
           scale = (n-1)/n,     #JK subsampling adjustment
           data = data.frame(sam.dat),
           combined = TRUE)
```

As Table 14.8 shows, the QP weights yield a CV for estimated total expenditures that is somewhat smaller than for the π-estimate (9.3 % vs. 8.4 %). The GREG and bounded GREG estimates have smaller CVs. But, the gain from the QP weights is substantial for mean expenditures. For the mean, the CVs for QP, GREG, and bounded GREG are the same as for the estimated totals because the total number of hospitals is a constraint, implying that the denominator of the mean is a constant, N. This is not the case for the π-estimate. QP, GREG, and bounded GREG are only slightly more efficient than the π-estimator for the proportion of hospitals receiving financing from state mental health agencies. As noted earlier, a model that has hosp.type as a factor would be more efficient for this statistic.

Table 14.8: Estimated total expenditures and proportions of hospitals receiving direct financing, standard errors, and coefficients of variation for the π-estimate, GREG estimates, and bounded quadratic program weights in a *pps* sample from a subset of the Survey of Mental Health Organizations population.

Estimator (design object)	Estimate or population value	SE	CV (%)
Total expenditures (000s)			
Population	8,774,651		
π-estimate (smho.dsgn)	9,322,854	915,126	9.82
Unbounded GREG (sam.lin)	9,563,683	748,596	7.83
Bounded GREG (sam.linBD)	9,612,035	744,746	7.75
Bounded QP (qp.dsgn)	9,509,333	800,769	8.42
Mean expenditures (000s)			
Population	12,103		
π-estimate (smho.dsgn)	13,299	1,712	12.88
Unbounded GREG (sam.lin)	13,191	1,033	7.83
Bounded GREG (sam.linBD)	13,258	1,027	7.75
Bounded QP (qp.dsgn)	13,116	1,105	8.42
Proportion with financing from state mental health agency			
Population	0.336		
π-estimate (smho.dsgn)	0.323	0.059	18.16
Unbounded GREG (sam.lin)	0.303	0.051	16.92
Bounded GREG (sam.linBD)	0.302	0.051	16.87
Bounded QP (qp.dsgn)	0.306	0.053	17.39

Simultaneous Adjustment for Nonresponse and Calibration

An obvious question is whether calibration to population totals alone will be enough to correct for nonresponse and any coverage errors. Bounded calibration adjustments, as well as bounded nonresponse adjustments, can be produced with the WTADJUST procedure beginning in SUDAAN v.10. This SUDAAN procedure implements the methods discussed in Folsom and Singh (2000) where either weight adjustment is calculated by way of a generalized exponential model. Both adjustments are calculated sequentially from this type of model by noting that both can be viewed as a calibration problem—input weights can be calibrated either to the sum of the input weights (in a nonresponse adjustment step) or to the population control totals (in a calibration step). The different adjustments are generated by the following specifications:

- *Nonresponse adjustment.* The model is specified with a dependent variable equal to the response indicator (1=respondent, 0=nonrespondent). The recommended lower bound on the weight adjustment is 1.0 to ensure that every sample member at least represents itself in the target population estimates.
- *Calibration adjustment.* The model is specified with a dependent variable equal to a calibration indicator (1=units included in the calibration, 0=otherwise). The recommended lower bound on the weight adjustment is 0 so that input weights may be reduced to meet the control totals.

Example 14.7 (Constrain weights using WTADJUST). The SAS-callable SUD-AAN syntax for PROC WTADJUST used to recompute Example 14.6 is provided below. For comparison, an SAS transport file was first created from the data frame containing the sample of 80 hospitals (sam.dat in Sect. 14.3.2) and with appended design weights (called dwt) using the following R code:

```
require(SASxport)
smho_80 <- cbind(sam.dat, dwt=d)
write.xport(smho_80, file="C:\\SMHO\\DATA\\Ex14_7.dat")
```

Note that SAS does not support "periods" either in the variable names or in transport file names, hence the use of the underline in the augmented data frame named smho_80.

The SAS transport file is loaded into the SAS program using PROC COPY, verified (code not shown) and submitted to the SUDAAN procedure to produce calibrated weights with constrained weight adjustments. Currently, the procedure only allows hard-coded control totals (POSTWGT) instead of sourcing the information from a data file. The calibration adjustment and final calibrated weights are called ADJFACTOR and WTFINAL in the SAS data file CAL_WTS. The output information has been renamed to have more descriptive

variable titles. The interested reader can verify the weight sums by the calibration variables with the DESCRIPT procedure below. The resulting summary statistics are provided below the program:

```
 options nocenter;

LIBNAME in "C:\\SMHO\\DATA";
LIBNAME tmp "C:\\";

   * Load SAS transport file and create unique IDs *;
LIBNAME smho_xpt XPORT "C:\SMHO\DATA\Ex15_7.dat";
PROC COPY in=smho_xpt OUT=tmp;  RUN;
DATA SMHO_80;
  LENGTH ID 3;
  SET tmp.SMHO_80;
  ID = _n_;
RUN;

   * Constrained calibration *;
PROC WTADJUST DATA=CAL_WTS DESIGN=WR ADJUST=POST;
  NEST _one_;          * No stratification or clustering;
  WEIGHT DWT;
  LOWERBD 0.4;
  UPPERBD 3.0;
  CLASS HOSP_TYP;
  MODEL _one_ = SEENCNT EOYCNT HOSP_TYP*BEDS;

                         * Corresponds to pop.tots in R program;
  POSTWGT 725 1349241 505345 37978 13066 9573 10077;
  IDVAR ID SEENCNT EOYCNT HOSP_TYP BEDS;
  OUTPUT / PREDICTED=ALL FILENAME=BCAL_WTS REPLACE;
RUN;

   * Rename Constrained calibration *;
DATA BCAL_WTS;
  SET BCAL_WTS(DROP=_one_
             RENAME=(WTFINAL=BCAL_WT ADJFACTOR=BCAL_ADJ));
  LABEL BCAL_WT  = "Calibrated weights w/bounded adjustments"
        BCAL_ADJ = "Bounded calibration adjustments";
RUN;

PROC DESCRIPT DATA=BCAL_WTS DESIGN=WR;
  NEST _one_;
  WEIGHT BCAL_WT;
  CLASS HOSP_TYP;
  VAR _one_ SEENCNT EOYCNT BEDS;
  TABLES HOSP_TYP;
  PRINT TOTAL SETOTAL / STYLE=NCHS;
RUN;

PROC MEANS DATA=BCAL_WTS NOLABELS MIN P25 P50 MEAN P75 MAX;
  VAR DWT BCAL_WT;
RUN;
```

Weight	Min	25th %-tile	Median	Mean	75th %-tile	Max	Sum
Base	2.71	5.69	8.15	8.76	10.11	33.68	701.00
WTADJ with bounds	1.41	5.93	8.91	9.06	11.03	33.75	725.00
GREG	−0.39	5.75	8.83	9.06	10.93	33.83	725.00
GREG with bounds	1.30	5.78	8.91	9.06	10.92	33.92	725.00

For comparison we also show the summaries for the unbounded GREG weights and the bounded GREG weights from Sect. 14.3.2. The quantiles of the weights from WTADJ and bounded GREG are very similar in this case. ∎

Weight Trimming and Redistribution

Potter (1990, 1993) describes several other methods of weight trimming. Some try to identify a method of trimming that will minimize mean square error; others look only at the distribution of the weights when deciding how to trim. These methods are *ad hoc* and largely theoretical. The form of weight trimming that may be most common can be summarized as follows:

(1) *Set upper and lower bounds on weights.* Methods for setting the bounds are generally arbitrary and a matter of agency preference or historical precedence. For example, one method used in the National Assessment of Educational Progress (National Center For Education Statistics 2008) is to trim any weight greater than 3.5 times the median weight ($3.5w_{med}$, say) back to $3.5w_{med}$.

(2) Any weight greater than upper bound (less than lower bound) is reset to the bound. That is,

$$w_{k,trim} = \begin{cases} U & w_k \geq U, \\ w_k & L < w_k < U, \\ L & w_k \leq L. \end{cases}$$

Define $\{w_{k,trim}\}_{k \in s}$ to be the set of trimmed weights.

(3) Determine the sum $K = \sum_{k \in s} |w_k - w_{k,trim}|$, i.e., the net amount of weight lost by trimming.

(4) Distribute K evenly among the units whose weights were not trimmed.

(5) Repeat steps (2)–(4) until no weights fail the bounds check.

If the input weights respect a set of control totals, the trimmed weights typically will not. One could then recalibrate the weights after trimming and iterate through the trimming and calibrating steps until a set of weights is obtained that respect the weight bounds and the controls. Since the same thing is achieved by the quadratic programming method, it is doubtful that this would be worthwhile.

The `trimWeights` function in the R `survey` package will trim weights to a specified bound and redistribute the trimmed-off amount to the other sample units. By using the parameter, `strict = TRUE`, the function calls itself recursively until the bounds are satisfied.

Example 14.8 (Trim and redistribute weights). We repeat Example 14.6 in which a *pps* sample of hospitals is selected from smho.N874 after dropping type-4 facilities. The design object is smho.dsgn. We then calibrate with the model SEENCNT + EOYCNT + as.factor(hosp.type):BEDS, as in the example in Sect. 14.3.2, to create the object `sam.lin`. The full R code for this example is in `trim.wts.R` shown in Appendix A and on the book web site. The code to trim the weights to the range [2, 18] and to summarize the results is:

```
sam.lin.tr1 <- trimWeights(design = sam.lin,
                           lower = 2,
                           upper = 18,
                           strict = TRUE)

summary(weights(sam.lin.tr1))
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.002   5.957   9.043   9.062  11.140  18.000
```

As the summary shows, the range restriction on the weights is satisfied. However, the weights are no longer calibrated. For example, the population total of SEENCNT is 1,349,241, but the estimate with the trimmed weights is

```
 svytotal(~SEENCNT, sam.lin.tr)
          total      SE
SEENCNT 1426878 240798
```

Note that the SE is nonzero using the variance formulas for a with-replacement design.

Figure 14.7 is a plot of the trimmed weights versus the base weights and the GREG weights. The GREG weights are shown in black. Nine points have been trimmed to the [2, 18] boundaries. The changes from the GREG weights to the trimmed weights are minimal for the other points. The base weights (inverses of selection probabilities) are plotted in gray. There is a considerable amount of change between the base weights and both the GREG weights and trimmed GREG weights.

The SUDAAN WTADJUST procedure can also be used to trim weights. Continuing with Example 14.6, bounds are placed on the input weights by adding WTMIN and WTMAX statements as shown in the code below. Note that the WTFINAL variable has been renamed to CAL_WTinit in the CAL_WTS data set created in the previous calibration step:

```
 PROC WTADJUST DATA=CAL_WTS DESIGN=WR ADJUST=POST;
   NEST _one_;     * No stratification or clustering;
   WEIGHT CAL_WTinit;
```
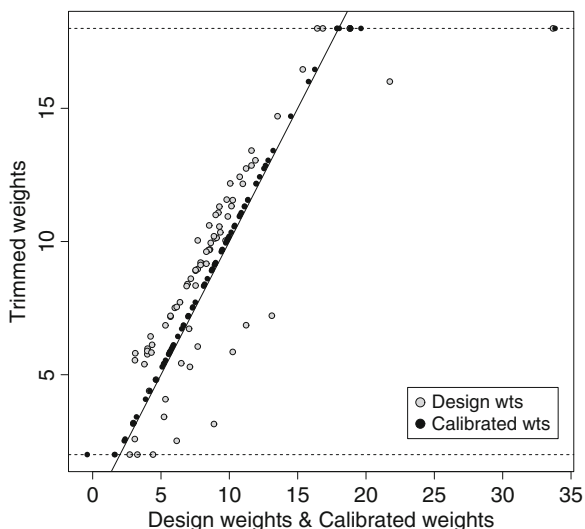
Fig. 14.7: Trimmed weights plotted versus base weights and GREG weights in a sample from the `smho.N874` population. The diagonal line is a 45° reference line. Horizontal lines are drawn at 2 and 18.

```
   CLASS HOSP_TYPE;
   LOWERBD 0;
   UPPERBD 18;
   WTMIN 2;
   WTMAX 18;
   MODEL _one_ = N SEENCNT EOYCNT HOSP_TYPE*BEDS;

                   * Corresponds to pop.tots in R program;
   POSTWGT 725  1349241  505345  37978  13066  9573  10077;
   OUTPUT ADJFACTOR WTFINAL
     / FILENAME=CAL_WTS2 FILETYPE=SAS REPLACE;
 RUN;
```

The POSTWGT statement above is used to force the trimmed and bounded weights to also satisfy the control totals used in Example 14.6. Notice that the control totals must be entered as constants in the POSTWGT statement; they cannot be read from a file.

Exercises

14.1. Use the `smho.N874` data set to complete this exercise on poststratification. Select a simple random sample of size $n = 80$ without replacement. If you use R, set the random number seed to -530049348 with the set.seed command.

(a) What are the means of expenditures in the five hospital types in the population? What should you look for in order for poststratification to be worth considering?
(b) Compute the population counts of facilities by hospital type, treating the smho98 data set as the full population. Compute the unweighted sample counts by hospital type to verify that each type is represented in the sample. If one of the hospital types was not represented in the sample, what would be the practical and theoretical implications? Discuss this in the context of design-based and model-based inference.
(c) Calculate the set of poststratified weights for the sample using hospital type as the poststratification variable. What do the weights sum to before and after poststratification? Is this what you expect?
(d) Verify that the calibration controls are met by the set of poststratified weights.
(e) Estimate the population total of expenditures and its standard error for the expansion estimator under the *srswor* design and for the poststratified estimator. Be sure and incorporate a finite population correction factor into the variance estimates. Discuss any similarities or differences in the estimated totals and SEs.

14.2. Repeat the exercise above after selecting a probability proportional to size sample.

(a) If you are using R, use the function `UPrandomsystematic` in the `sampling` package to select a probability proportional to size sample. Define the measure of size (mos) as a recoded version of the square root of beds. After taking the square root of beds, recode any mos ≤ 5 to 5. If you use R, set the random number seed to -530049348 and select a sample of size $n = 80$.
(b) Compute the unweighted sample counts by hospital type to verify that each type is represented in the sample.
(c) Calculate the set of poststratified weights for the sample using hospital type as the poststratification variable. What do the weights sum to before and after poststratification? Is this what you expect?
(d) Verify that the calibration controls are met by the set of poststratified weights.
(e) Estimate the population total of expenditures and its standard error for the π-estimator under the *pps* design and for the poststratified estimator. Discuss any similarities or differences in the estimated totals and SEs.

14.3. Use the model `BEDS + SEENCNT + EOYCNT + as.factor(hosp.type)` and the sample described in Sect. 14.3.2 to compute GREG weights. That is, select a sample with probabilities proportional to recoded square root of beds (using the random number seed 428274453 if you are using R). Restrict the population to facilities other than type 4.

(a) Verify that the weights are calibrated, i.e., $\sum_s w_i x_i = t_x$, for the auxiliary variables in the calibration model.
(b) What are the ranges of the base weights and the calibrated weights?
(c) Experiment with bounding the weight adjustments using lower and upper bounds of $[L, U] = [0.01, 3]$. Use FORCE=TRUE in the calibrate function if convergence is not obtained. Are these weights fully calibrated? Plot the GREG weights with no bounds and the bounded-adjustment weights versus the base weights. Use different symbols or colors to distinguish the sets of weights. What do these results tell you about numerical problems that may occur with bounded calibration?

14.4. Consider a stratified simple random sample in which n_h units are selected from N_h units in stratum h. The unit variance in stratum h is S_h^2. The proportional allocation to the strata has $n_h/n = N_h/N$ with $n = \sum_h n_h$ and $N = \sum_h N_h$. The weight for each unit i in stratum h is $w_{hi} \equiv k_h = N_h/n_h$. Define the relvariance of the weights as

$$relvar\,(w) = n^{-1} \sum_h \sum_{i=1}^{n_h} (w_{hi} - \bar{w})^2 \bigg/ \bar{w}^2$$

with $\bar{w} = n^{-1} \sum_h \sum_{i=1}^{n_h} w_{hi}$. Derive the three versions, (a), (b), and (c), below of Kish's $1 + L$ formula. That is, in the case with $S_h^2 = S^2$ in each stratum, show that

$$
\begin{aligned}
1 + L &= \frac{V(\bar{y}_{st}|\text{general allocation})}{V(\bar{y}_{st}|\text{proportional allocation})} \\
&= \left(\sum_h W_h k_h\right)\left(\sum_h W_h/k_h\right) & (a) \\
&= 1 + relvar\,(w) & (b) \\
&= \frac{n \sum_h \sum_{i=1}^{n_h} w_{hi}^2}{\left(\sum_h \sum_{i=1}^{n_h} w_{hi}\right)^2} & (c)
\end{aligned}
$$

14.5. Show that, in the case of $H = 2$ strata with an *srswor* selected in each stratum, Kish's $1 + L$ measure is

$$deff_w = \frac{p_1 w_1^2 + p_2 w_2^2}{(p_1 w_1 + p_2 w_2)^2}.$$

where $p_h = n_h/n$ and the weight of each unit in stratum h is w_h $(h = 1, 2)$. Assume that sampling fractions are negligible in each stratum.
Use this formula to verify the calculations in Table 14.7.

14.6. Using the random seed value of 15097 in R, select a sample of $n=50$ hospitals from the data file Hospital_pop.txt with probabilities proportional to the square root of the number of BEDS, i.e., *pps* $(x^{1/2})$. The hospital file has 393 records.

(a) Calculate the estimated design effects using Spencer's formula and Kish's approximation.

(b) Describe the estimators of the population total to which the Kish and Spencer *deff*'s refer. Why do the computed values differ? Which do you think is the most relevant here? Why?

(c) Estimate the total of discharges (y) in the population using the π-estimator along with its SE and *CV*. How does this compare to the estimate of the variance of the total from a simple random sample of $n=50$. Estimate the *srswor* variance from the sample of 50 selected for this problem. (Hint: you need to use the methods in Chap. 3 to estimate a population variance.)

14.7. Use the data set `nhispart.xpt` and the R `survey calibrate` function to compute some sets of calibration weights. The weights will be based on the categorical x variables, `SEX`, `R_AGE1`, and `RACRECI2`, codes for which are given below along with the population control counts for each category. The function `read.xport` in the `foreign` package (R Core Team and contributors worldwide 2012a) can be used to read the data into R.

SEX	Code	Pop totals	R_AGE1 Age range	Code	Pop totals	RACRECI2 Race	Code	Pop totals
Male	1	2,000	18–24	3	500	White	1	3,350
Female	2	2100	25–44	4	1,800	Black	2	650
			45–64	5	1,000	All other	3	100
			65–69	6	250			
			70–74	7	250			
			75+	8	300			

The counts on the file `nhispart.xpt` are below. You should use these to verify that you have read the file correctly.

SEX	Code	File totals	R_AGE1 Age range	Code	File totals	RACRECI2 Race	Code	Pop totals
Male	1	1805	18–24	3	512	White	1	3,138
Female	2	2119	25–44	4	1,555	Black	2	601
			45–64	5	1,255	All other	3	185
			65–69	6	164			
			70–74	7	150			
			75+	8	288			

(a) Select a simple random sample without replacement of size $n = 200$, setting the random seed to 15097. List the indexes of the sample you selected sorted in order from low to high. (Hint: use the `sample` function.)

(b) Create a new variable equal to 1 if the family income is less than 1.5 times the poverty threshold and 2 otherwise. The ratio of family income to poverty threshold is `RAT_CAT` and has the values below. Keep the unknowns as a separate category. Show a table with the sample counts of your new variable. Also create versions of the variables `R_AGE1` and `RACRECI2` that have a minimum of 10 cases per category. Do this by

collapsing R_AGE1=6 and 7 together and RACRECI2=2 and 3 together. Tabulate the numbers of sample cases in the recoded versions of RAT_CAT, R_AGE1, and RACRECI2.

Code	Ratio of family income to poverty threshold
01	Under 0.50
02	0.50 to 0.74
03	0.75 to 0.99
04	1.00 to 1.24
05	1.25 to 1.49
06	1.50 to 1.74
07	1.75 to 1.99
08	2.00 to 2.49
09	2.50 to 2.99
10	3.00 to 3.49
11	3.50 to 3.99
12	4.00 to 4.49
13	4.50 to 4.99
14	5.00 and over
99	Unknown

(c) Create a set of calibrated weights using the linear distance function and no bounds on the weight adjustments. Verify that your weights are calibrated. Show the minimum, maximum, and the three quartiles of the weights. (Hint: use the `weights` extractor and the `summary` function.)

(d) Create a set of weights using the linear distance function with lower and upper bounds on the weight adjustments of 0.5 and 1.6. Verify that your weights are calibrated. Show the minimum, maximum, and the three quartiles of the weights.

(e) Create a set of weights with the raking distance function with no bounds on the weight adjustments. Verify that your weights are calibrated. Show the minimum, maximum, and the three quartiles of the weights.

(f) Using the three sets of weights (linear with no bounds, linear with bounds, and raking with no bounds), compare the individual unit weights with a pairs plot. Comment on the comparisons.

(g) Using the four sets of weights—*srs*, linear with no bounds, linear with bounds, and raking with no bounds—estimate the proportions of the population with family incomes less than 1.5 and greater than or equal to 1.5 times the poverty income ratio and their estimated standard errors. (Hint use `svymean`.) Comment on the estimates.

14.8. Using the data file `smho.N874`, answer the following:

(a) Calculate the probabilities for all population units in a sample of 50 selected with probabilities proportional to the following measure of size (MOS): `EXPTOTAL`. Identify certainties, if any, i.e., units with selection

probability greater than or equal to 1. If there are certainties, assign
them probability 1, and recalculate the selection probabilities for the non-
certainty part of the population, keeping the total sample at 50.
(b) Select a sample of size 50 using the probabilities computed in (a). If you
use R, set the random number seed to 429336912.
(c) Compute Kish's $1 + L$ and Spencer's *deff* for this sample. In the case of
Spencer's *deff*, use the variable SEENCNT as y.
(d) Explain in words the meaning of the value you obtained in (c) for $1 + L$.
What should be considered in determining whether the value is excessively
large or not? How do Kish's and Spencer's measures compare in this
problem?
(e) Repeat parts (a)–(d) using BEDS as the MOS. Set the MOS for any unit
with BEDS $= 0$ to the minimum value of BEDS for those with non-zero
BEDS. Use EXPTOTAL as the y for Spencer's *deff* and 429336912 as the
random number seed. You may find it useful to examine the individual
weights when discussing the Kish and Spencer measures.

14.9. Show that when a probability proportional to x sample is selected, the
weights are calibrated to the total of x in the population. That is, $\sum_s w_i = t_x$
where w_i is the inverse of the selection probability of unit i and t_x is the total
of x across all units in the frame. Do you think that the π-estimator is the
most efficient estimator, i.e., smallest variance, in any population where $pp(\mathrm{x})$
sampling is reasonable? Why or why not?

14.10. Using the data file smho.N874, select a sample of $n{=}50$ units with
probabilities proportional to recoded BEDS as the measure of size. Set the
MOS for any unit with BEDS $= 0$ to the minimum value of BEDS for those
with nonzero BEDS. If you use R, set the random number seed to 429336912.

(a) Report the summary for the resulting weights, i.e., the min, max, quar-
tiles, and the mean. Do any units have weights that seem to be of concern?
(b) Use quadratic programming to bound the weights in the range $[1, 50]$.
Plot the resulting weights versus the base weights. What was the effect
of the bounding? Is quadratic programming an effective way of bounding
the weights here?
(c) Re-do parts (a) and (b) but recode any unit with BEDS $= 0$ to BEDS$=10$.
Discuss your results. Are the weight adjustments as extreme as in (b)?

Chapter 15
Variance Estimation

In previous chapters we considered the variance of estimators in order to determine the sample size and allocation to the design strata. After the sample data are collected, estimates are made and their variances and standard errors (SEs) must be computed. An SE (square root of the estimated variance) is a basic measure of precision that can be used as a descriptive statistic, e.g., as part of a coefficient of variation (CV), or for making inferences about population parameters via confidence intervals. Estimating SEs that faithfully reflect all sources of (or a significant portion of the) variability in a sample design and an estimator is our goal, but this can be complicated. This is especially true when several (random) weight adjustments described in Chaps. 13 and 14 are used. For example, when an adjustment for nonresponse is applied and then weights are raked to population controls, both procedures contribute to the variance of an estimator in addition to the randomness due to selecting the initial sample itself.

Many analysts, however, often estimate SEs in ways that do not account for all sources of variability. This may be due to inadequate information about how the data were collected and estimates made, use of inappropriate software, ignorance of proper procedures, or some combination of these. Also, published analysis files may only contain the final set of analysis weights instead of providing users with the individual weight adjustments. As discussed in this chapter, this problem is remedied in many public-use data files through multiple (replicate) weights. The importance of capturing the various random components is demonstrated in this chapter along with methods used to fulfill this objective that are specific to the sample design and point estimator.

There are several alternative methods of variance estimation that will be covered in this chapter—exact formulas, linearization, and replication variance estimators. We summarize the methods along with some of their strengths and weaknesses, including how easily each can account for different sources of variability. Exact methods are covered in Sect. 15.1 and apply to a limited number of sample designs and estimators. However, one of the

exact methods in multistage sampling, called the ultimate cluster estimator, is the basis for some of the theory that supports linearization and replication estimators. Strictly speaking, the ultimate cluster estimator is exact only for sampling designs where the primary sampling units (PSUs) are selected with replacement, but it is a useful approximation in other designs when the sampling fraction of PSUs is small.

Exact methods do not apply when an estimator is nonlinear; Sect. 15.2 describes the circumstances that make an estimator nonlinear. In Sect. 15.3, we cover linearization variance estimators, which apply to many estimators for which exact formulas are not available. Section 15.4 contains a discussion of three replicate variance estimation methods—jackknife, balanced repeated replication, and bootstrap—that are applicable to most public-use analysis files that have been treated to minimize identification of the survey participants. The linearization and replication variance estimation techniques, the methods most applicable to design-based estimation, are built around doing something with the PSUs or first-stage units. For example, one linearization method computes a variance based on differences among the weighted PSU totals. In the replication methods, subsamples called replicates are formed by designating subsets of the PSUs. The entire sample of units within a PSU is retained if a PSU is in a replicate.

The last two sections of this chapter discuss some specialized topics—combining PSUs or strata for variance estimation and ways of handling certainty PSUs when estimating variances.

15.1 Exact Methods

In a few simple cases, theoretical variances and their estimators have exact formulas. We first encountered these situations in Chap. 3 where the notation that we use below was defined. There are three designs—simple random samples, stratified simple random samples, and varying probability sampling with replacement—that we have dealt with most often that admit exact variance formulas. For example, if a stratified simple random sample without replacement (*stsrswor*; discussed in Sect. 3.1.1) of size $n = \sum_{h=1}^{H} n_h$ is selected and the population mean is estimated with $\bar{y}_{st} = \sum_{h=1}^{H} W_h \bar{y}_{sh}$, then its variance is estimated with

$$v\left(\bar{y}_{st}\right) = \sum_{h=1}^{H} W_h^2 \frac{1 - f_h}{n_h} \hat{S}_h^2,$$

where $\hat{S}_h^2 = (n_h - 1)^{-1} \sum_{i \in s_h} (y_{hi} - \bar{y}_{sh})^2$ and $W_h = N_h/N$.

Another common design is to select units with varying probabilities and without replacement. If n units are selected and π_i is the selection probability of unit i, the π-estimator is $\hat{y}_\pi = \sum_{i=1}^{n} y_i/\pi_i$. Defining π_{ij} as the probability

that units i and j are both selected for a sample, one of the variance estimators recommended for $\hat{y}_\pi$ is the Yates-Grundy estimator:

$$\text{var}_{YG}\left(\hat{y}\right) = -\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}}\right) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2. \tag{15.1}$$

One difficulty with this estimator is that samples are sometimes selected using systematic sampling so that some of the π_{ij}'s are zero. In that case, no design-unbiased estimator of the variance exists. Särndal et al. (1992, Chap. 3) provide the technical details. Even if a design is used where the YG estimator might be feasible, the π_{ij}'s may not be available. This is especially true when doing secondary data analysis using a file prepared by someone else, like a government agency.

If a sample is selected with varying probabilities and with replacement (*ppswr*) and the *pwr*-estimator, $\hat{\bar{y}}_{pwr} = \frac{1}{Nn} \sum_s \frac{y_i}{p_i}$, is used, its variance is estimated with

$$v\left(\hat{\bar{y}}_{pwr}\right) = \frac{1}{N^2} \frac{1}{n(n-1)} \sum_{i \in s} \left(\frac{y_i}{p_i} - \hat{t}_{pwr}\right)^2 \tag{15.2}$$

where s is the set of sample units, $\hat{t}_{pwr} = N\hat{\bar{y}}_{pwr}$, and p_i is the 1-draw selection probability (i.e., the selection probability if only one unit were selected). This has the obvious advantage of not requiring any π_{ij}'s.

Another important case in which the *pwr* formula applies is a multistage design in which the first-stage units are selected with replacement. In that design, formula (15.2) can be used with y_i defined as the estimated total for the units in first-stage unit i. The technical requirement is that y_i must be an unbiased estimator of the PSU total of y. If the PSUs are stratified, then the *pwr*-estimator of a mean is $\hat{\bar{y}}_{pwr} = N^{-1} \sum_h n_h^{-1} \sum_{i \in s_h} y'_{hi}/p_{hi}$, where p_{hi} is the 1-draw probability of selection of PSU i in stratum h and $y'_{hi} = \sum_{k \in s_{hi}} d_{k|hi} y_{hik}$ is the estimated total just for units in PSU hi. The set of sample units in PSU hi is s_{hi} while $d_{k|hi}$ is the weight for unit k in PSU hi that expands the PSU sample to only the population of that PSU. The full weight for unit k in s_{hi} is $d_k = d_{k|hi}/p_{hi}$ where $d_{k|hi}$ is sometimes referred to as the conditional within-PSU weight for unit k (conditional on PSU hi being selected) and d_k as the unconditional weight. The *pwr* variance formula is then

$$v\left(\hat{\bar{y}}_{pwr}\right) = \frac{1}{N^2} \sum_h \frac{1}{n_h(n_h-1)} \sum_{i \in s_h} \left(\frac{y'_{hi}}{p_{hi}} - \hat{t}_{pwr,h}\right)^2, \tag{15.3}$$

where $\hat{t}_{pwr,h} = n_h^{-1} \sum_{s_h} y'_{hi}/p_{hi}$. This formula is also often written as

$$v\left(\hat{\bar{y}}_{pwr}\right) = \frac{1}{N^2} \sum_h \frac{n_h}{(n_h-1)} \sum_{i \in s_h} \left(\hat{Y}_{hi} - \hat{\bar{Y}}_h\right)^2, \tag{15.4}$$

where $\hat{Y}_{hi} = \sum_{k \in s_{hi}} d_k y_k$ and $\bar{\hat{Y}}_h = n_h^{-1} \sum_{i \in s_h} \hat{Y}_{hi}$. The form in Eq. (15.4) is convenient because it uses the full-sample weight, d_k, rather than both the 1-draw weight, $1/p_{hi}$, and the conditional within-PSU weight, $d_{k|hi}$. An analyst will typically not have $1/p_{hi}$ and $d_{k|hi}$ separately. The formula in Eq. (15.3) or Eq. (15.4) is called the *ultimate cluster* variance estimator (Hansen et al. 1953a).

The PSU terminology can potentially be confusing in area probability samples. As discussed in Chaps. 9 and 10, the term PSU usually denotes a geographic area that is one or more local government jurisdictions, like a county. Some PSUs may be selected with probability 1 (the certainties) while others have selection probabilities less than 1. The certainty PSUs are not the first-stage units, although practitioners habitually call them PSUs. A certainty PSU is really a stratum composed of lower-level units. In a certainty, the first-stage units are actually census tracts, block groups, or some other subcounty units. For example, Washington DC might be a certainty PSU in a US area sample, but 20 block groups might be sampled from it. The 20 block groups are the PSUs for purposes of variance calculation. In this chapter, when we refer to PSUs for variance estimation, we really mean "first-stage units." You need to be cognizant of this when setting up a data file for variance estimation.

Many variance estimators shown in sampling textbooks assume with-replacement sampling. However, most designs do not use without-replacement sampling at the first stage. Consequently, Eq. (15.2) or Eq. (15.4) is not strictly appropriate for most designs used in practice. The real utility of the with-replacement formulae lies in the fact that they are good approximations to the variance of estimators in many situations where without-replacement sampling is used. Practitioners often make use of this kind of thinking. In Chap. 3, the with-replacement variance formula was a handy vehicle for computing sample sizes when a sample was selected with varying probabilities. Similarly, when analyzing data that has already been collected, Eq. (15.2) or Eq. (15.4) is easier to compute than most exact formulas that account for without-replacement sampling. Because of its convenience, expression (15.4), in particular, is the building block for many of the variance estimates that software packages provide. We will cover this idea in more detail in Sect. 15.3.

15.2 Linear Versus Nonlinear Estimators

Being able to use an exact variance formula depends not only on the sample design but also on using what is known as a *linear* estimator, which has a particular meaning in the design-based world. Knowing what a linear estimator is (and is not) will be important since the linearization and replication variance estimators covered in later sections are designed to handle *nonlinear* estimators.

In model-based or mathematical statistics, a linear estimator is usually defined to have the form $\hat{\theta} = \sum_{i \in s} \alpha_i y_i$ where the α's are constants in the random sample s ($i \in s$) and the y variables are treated as random variables under some model, e.g., $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$. In design-based sampling, the randomness comes from how the sample is selected. A random variable is defined for whether a unit is in the sample or not:

$$\delta_i = \begin{cases} 1 & \text{if unit } i \text{ is in the sample } (i \in s) \\ 0 & \text{if not } (i \notin s). \end{cases},$$

The probability that $\delta_i = 1$ is the selection probability of unit i, i.e., π_i using the established notation. A detailed discussion on this conceptual difference is found in, e.g., Valliant et al. (2000). There are several definitions of linear that have been proposed for design-based sampling [see (Wolter 2007, chap. 1)]. We will use a slightly simplified version that is precise enough for our purposes. A *linear* estimator is one that can be written as $\hat{\theta} = \sum_{i \in U} \delta_i \alpha_i y_i$ where U is the set of all units in the finite population, and the value of α_i is the same regardless of the set of sample units that are selected.

A *nonlinear* estimator is one where the δ_i's are combined in a way that is more complicated than just a weighted summation. For example, an estimator defined as $\hat{\theta} = \sum_{i \in U} \delta_i \alpha_i y_i / \sum_{i \in U} \delta_i \alpha_i x_i$ is nonlinear since it is the ratio of two linear estimators. The poststratified estimator, $\hat{T}_{yPS} = \sum_{\gamma=1}^{G} N_\gamma \left(\hat{t}_{y\gamma} / \hat{N}_\gamma \right)$ from Sect. 14.2, is nonlinear. The weight for each sample unit is $w_i = d_i N_\gamma / \hat{N}_\gamma$ where d_i represents the base weight; N_γ the population count in poststratum γ; and $\hat{N}_\gamma = \sum_{i \in U_\gamma} \delta_i d_i = \sum_{i \in s_\gamma} d_i$, the estimate of N_γ defined for U_γ, the set of all population units, and for s_γ, the set of sample units that are in poststratum γ. The fact that $\hat{N}_\gamma$ is in the denominator makes $\hat{T}_{yPS}$ nonlinear.

When weighting class adjustments for nonresponse are used, as in Sect. 13.5.1, a nonlinear estimator is created. The adjusted weights involve terms like

$$a_{2c} = \frac{\sum_{i \in s_{c,E}} d_{1i}}{\sum_{i \in s_{c,ER}} d_{1i}},$$

where c is a weighting class, d_{1i} is a base weight adjusted for unknown eligibility, $s_{c,E}$ is the set of eligible sample units in c, and $s_{c,ER}$ is the set of responding eligible sample units in c. An estimated total using this type of nonresponse-adjusted weight can be written as $\hat{T} = \sum_c \sum_{i \in s_{c,ER}} a_{2c} d_{1i} y_i$. Both the numerator and denominator of a_{2c} are random with respect to the sample design, making the nonresponse-adjusted estimator nonlinear.

Another example is a GREG estimator in Sect. 14.3, which involves the inverse of a sample matrix, among other complications, that make it highly nonlinear. If the GREG calibration is preceded by a nonresponse adjustment, then even more nonlinearity is injected into the estimator.

Estimating the variance of a nonlinear estimator is somewhat more difficult than for a linear estimator. However, the linearization method, described in the next section, is a solution to this problem (at least in principle).

15.3 Linearization Variance Estimation

This section sketches how linearization variance estimation works. We also cover some more specialized matters that naturally accompany variance estimation, including confidence interval construction, degrees of freedom for variance estimators, accounting for sampling fractions, domain estimation, and the effects of multiple steps in weighting on variances.

15.3.1 Estimation Method

Linearization is a method of approximating variances. The technique is also known as the *Taylor series* or *delta* method. The general idea is to approximate a complicated estimator like a ratio, an odds ratio, or a regression coefficient by a linear function. The theoretical, designed-based variance is calculated for the linear approximation and then the theoretical variance is estimated based on whatever design was used to select the sample. Although understanding the details of the method is not essential for the presentation here, understanding the general approach is worthwhile. Suppose that an estimator can be written as a function f of estimated totals:

$$\hat{\theta} = f\left(\hat{t}_1, \ldots, \hat{t}_p\right).$$

Each estimated total must be a linear estimator of the form, $\hat{t}_j = \sum_{i \in s} \alpha_i y_{ji}$. The standard choice for α_i is the inverse of the selection probability so that $\hat{t}_j$ is a π-estimator. For example, in the case of a ratio, we might have $\hat{\theta} = \hat{t}_1/\hat{t}_2$ where $\hat{t}_1 = \sum_s d_i y_i$ and $\hat{t}_2 = \sum_s d_i x_i$. The first step is to form a linear approximation to the nonlinear function $\hat{\theta}$:

$$\hat{\theta} - \theta \doteq \sum_{j=1}^{p} \frac{\partial f\left(\hat{t}\right)}{\partial \hat{t}_j} \left(\hat{t}_j - t_j\right) \tag{15.5}$$

where $\hat{\theta}$ is the estimate of the population parameter θ; $\hat{t} = \left(\hat{t}_1, \ldots, \hat{t}_p\right)^T$, the vector of estimated totals; $\partial f\left(\hat{t}\right)/\partial \hat{t}_j$ is the partial derivative of f with respect to the j-th estimated total in $\hat{t}$; and t_j is the population total for the j-th variable. The theory behind the approximation requires that the partial derivatives can be derived and are evaluated at the population values

(which, of course, we do not know). Sample estimates are substituted for the population quantities in order to calculate an estimated variance as shown below.

The apparently complicated problem of estimating the variance of the non-linear $\hat{\theta}$ is, thus, reduced to the simpler problem of estimating the variance of a weighted combination of the $\hat{t}_j$'s. We then compute the variance of the right-hand side of Eq. (15.5) by squaring both sides of the equation and evaluating the expectation with respect to the sample design thereby obtaining

$$V\left(\hat{\theta}\right) \doteq \sum_{j=1}^{p} \left[\frac{\partial f\left(\hat{t}\right)}{\partial \hat{t}_j}\right]^2 V\left(\hat{t}_j\right) + \sum_{j=1}^{p}\sum_{k \neq j}^{p} \frac{\partial f\left(\hat{t}\right)}{\partial \hat{t}_j}\frac{\partial f\left(\hat{t}\right)}{\partial \hat{t}_k} cov\left(\hat{t}_j, \hat{t}_k\right). \quad (15.6)$$

The terms, θ and t_j, in Eq. (15.5) do not have to be considered in the variance approximation since they are population values that are treated as constants.

For many different sample designs, we know how to compute the variances and covariances in Eq. (15.6). For instance, if the design is *stsrs*, $V\left(\hat{t}_j\right)$ has the form $V\left(\hat{t}_j\right) = \sum_{h=1}^{H} N_h^2 \left(1 - f_h\right) S_h^2/n_h$. The covariances, $cov\left(\hat{t}_j, \hat{t}_k\right)$, under this design are similar with S_h^2 replaced by a population covariance of the jth and k^{th} variables. To compute the resulting sample variance estimator, here denoted as $v_L\left(\hat{\theta}\right)$, the derivatives, variances, and covariances in Eq. (15.6) are evaluated using their corresponding sample estimates.

Example 15.1 (Linearization variance estimator for the ratio of two totals). Consider a point estimator defined as the ratio of two estimated totals: $\hat{\theta} = \hat{t}_1/\hat{t}_2 \equiv f\left(\hat{t}_1, \hat{t}_2\right)$ with $\hat{t}_j = \sum_{k \in s} d_i y_{jk}$ $(j = 1, 2)$. Using the notation above, we say that $\hat{\theta} = f\left(\hat{t}_1, \hat{t}_2\right)$, a function of two unique estimators. This quantity estimates the population parameter $\theta = t_1/t_2$, where $\theta = f\left(t_1, t_2\right)$ and $t_j = \sum_{k \in U} y_{jk}$ $(j=1,2)$. To compute the linearization variance estimator, we begin with a Taylor expansion as shown in Eq. (15.5):

$$\hat{\theta} - \theta \doteq \frac{\partial f\left(t\right)}{\partial t_1}\left(\hat{t}_1 - t_1\right) + \frac{\partial f\left(t\right)}{\partial t_2}\left(\hat{t}_2 - t_2\right)$$

so that

$$\left(\hat{\theta} - \theta\right)^2 \doteq \left(\frac{\partial f\left(t\right)}{\partial t_1}\right)^2\left(\hat{t}_1 - t_1\right)^2 + \left(\frac{\partial f\left(t\right)}{\partial t_2}\right)^2\left(\hat{t}_2 - t_2\right)^2$$
$$+ 2\frac{\partial f\left(t\right)}{\partial t_1}\frac{\partial f\left(t\right)}{\partial t_2}\left(\hat{t}_1 - t_1\right)\left(\hat{t}_2 - t_2\right).$$

Taking the expectation of both sides of the equal sign with respect to the particular sample design in use, we obtain

$$V\left(\hat{\theta}\right) = E_\pi\left[\left(\hat{\theta} - \theta\right)^2\right]$$
$$\doteq \left(\frac{\partial f(t)}{\partial t_1}\right)^2 V\left(\hat{t}_1\right) + \left(\frac{\partial f(t)}{\partial t_2}\right)^2 V\left(\hat{t}_2\right) + 2\frac{\partial f(t)}{\partial t_1}\frac{\partial f(t)}{\partial t_2}Cov\left(\hat{t}_1, \hat{t}_2\right)$$

where $\frac{\partial f(t)}{\partial t_1} = \frac{1}{t_2}$ and $\frac{\partial f(t)}{\partial t_2} = -t_1\left(\frac{1}{t_2}\right)^2$. Estimate values for $V\left(\hat{t}_1\right)$, $V\left(\hat{t}_2\right)$, $Cov\left(\hat{t}_1, \hat{t}_2\right)$, and the derivatives are generated using the sample design and data and plugged into this formula to obtain $v_L\left(\hat{\theta}\right)$, the estimated sample variance of $\hat{\theta}$. ∎

An alternative method that avoids computing the individual variances and covariances in Eq. (15.6) is called the *linear substitute* method (Wolter, 2007, Sect. 6.5). The idea is to substitute the formula for $\hat{t}_j$ into Eq. (15.5) and reverse the summation over variables and units before calculating the variance. Suppose that a multistage design is used and $\hat{t}_j = \sum_{i \in s}\sum_{k \in s_i} d_k y_{jk}$ is the statistic of interest, where d_k is the base weight for unit k in PSU i and y_{jk} is the value of the j-th analysis variable for unit k in PSU i. Then, the reversal of Eq. (15.5) leads to

$$\hat{\theta} - \theta \doteq \sum_{i \in s}\sum_{k \in s_i} d_k z_k + \text{constants} \tag{15.7}$$

with $z_k = \sum_{j=1}^p \frac{\partial f(\hat{t})}{\partial \hat{t}_j} y_{jk}$ $(k \in s_i)$. The "constants" in Eq. (15.7) depend on the population totals and derivatives and neither contribute to the design variance. The sum $\hat{z} = \sum_{i \in s}\sum_{k \in s_i} d_k z_k$ is the estimated total of the z_k, which are called the linear substitutes. The variance estimation problem is then reduced to estimating the variance of a single estimated total. Often, the ultimate cluster variance estimator in Eq. (15.4) is used. If the design was a stratified cluster sample, then, using the linear substitutes, the ultimate cluster formula would be

$$v_L\left(\hat{\theta}\right) = \sum_h \frac{n_h}{(n_h - 1)}\sum_{i \in s_h}\left(\hat{z}_{hi} - \hat{\bar{z}}_h\right)^2,$$

where $\hat{z}_{hi} = \sum_{k \in s_{hi}} d_k z_k$ and $\hat{\bar{z}}_h = n_h^{-1}\sum_{i \in s_h}\hat{z}_{hi}$.

Example 15.2 (Continuation of Example 15.1, ratio of two totals). Take the estimator of a ratio defined in Example 15.1, $\hat{\theta} = \hat{t}_1/\hat{t}_2$ with $\hat{t}_j = \sum_{k \in s} d_i y_{jk}$ $(j=1,2)$. The linear substitute is $z_k = t_2^{-1}\left(y_{1k} - \theta y_{2k}\right)$. The approximate variance is $V\left(\sum_s d_k z_k\right)$. How this is estimated depends on the sample design. If the design is *srswor*, then the estimated variance is

$$v\left(\hat{\theta}\right) = \frac{N^2}{n}\left(1 - \frac{n}{N}\right)\frac{\sum_s\left(z_k - \bar{z}_s\right)^2}{n - 1}$$

with $\bar{z}_s$ being the unweighted sample mean of the z_k's. If the design is *ppswr*, then

$$v\left(\hat{\theta}\right) = \frac{1}{n\left(n-1\right)} \sum_{k\in s} \left(\frac{z_k}{p_k} - \hat{t}_{pwr,z}\right)^2$$

with $\hat{t}_{pwr,z} = n^{-1} \sum_s z_k/p_k$. If a two-stage (or more) design was used, then a variance formula appropriate for that design would be used. ∎

Example 15.3 (Log-odds in a 2×2 table). Suppose that the following table gives the estimated counts of persons who have diabetes classified by gender. Suppose that a multistage, stratified sample is used and that each estimated

	Has diabetes	Does not have diabetes
Male	$\hat{t}_1$	$\hat{t}_2$
Female	$\hat{t}_3$	$\hat{t}_4$

total has the form $\hat{t}_j = \sum_h \sum_{i \in s_h} \sum_{k \in s_{hi}} d_k y_{jk}$. Notice that each cell in the table is a domain so that y_{jk} is 1 ($k \in s_{hi}$) if unit k is in cell j ($j = 1, 2, 3, 4$) and 0 if not. The log of the ratio of the odds of males having diabetes to the odds ratio for females is

$$\hat{\theta} = \log\left(\frac{\hat{t}_1\hat{t}_4}{\hat{t}_2\hat{t}_3}\right) = \log\left(\hat{t}_1\right) - \log\left(\hat{t}_2\right) - \log\left(\hat{t}_3\right) + \log\left(\hat{t}_4\right).$$

The linear substitute is $z_k = \frac{y_{1k}}{t_1} - \frac{y_{2k}}{t_2} - \frac{y_{3k}}{t_3} + \frac{y_{4k}}{t_4}$, and the log-odds is approximately $\hat{\theta} \doteq \sum_h \sum_{i \in s_h} \sum_{k \in s_{hi}} d_k z_k$. The ultimate cluster variance estimator, in this case, is

$$v\left(\hat{\theta}\right) = \sum_h \frac{n_h}{\left(n_h - 1\right)} \sum_{i \in s_h} \left(\hat{z}_{hi} - \hat{\bar{z}}_h\right)^2,$$

where $\hat{z}_{hi} = \sum_{k \in s_{hi}} d_k z_k$ and $\hat{\bar{z}}_h = n_h^{-1} \sum_{s_h} \hat{z}_{hi}$. To evaluate $v\left(\hat{\theta}\right)$, we replace each t_j in the linear substitute z_k with its sample estimate. ∎

Software packages have certain special cases of the linear substitute formula programmed. The user specifies the sample design and the type of estimator, and the software evaluates the appropriate formula. R, Stata, SUDAAN, and SAS all use the linear substitute method as one of their options. The user is limited to statistics for which the linear substitute has been programmed. For customized statistics, the statistician may need to construct his/her own specialized program.

15.3.2 Confidence Intervals and Degrees of Freedom

Confidence intervals are usually computed using either the normal or t-approximation. A $100\,(1-\alpha)\,\%$ confidence interval for θ is either

$$\hat{\theta} \pm z_{1-\alpha/2}\sqrt{v_L\left(\hat{\theta}\right)} \text{ or } \hat{\theta} \pm t_{1-\alpha/2}\,(df)\,\sqrt{v_L\left(\hat{\theta}\right)}$$

where $z_{1-\alpha/2}$ is the point in a standard normal distribution with $1-\alpha/2$ of the area to its left and $t_{1-\alpha/2}\,(df)$ is the corresponding point in a central t-distribution with df degrees of freedom. Some of the rules of thumb used for setting degrees of freedom are described below.

The degrees of freedom are a characteristic of a variance estimator as well as the sample design. If the data, $y_1,\ldots,y_n$, were each independently generated by a normal distribution with mean μ and variance σ^2, then $\sum_{i=1}^{n}\left(y_i-\bar{y}_s\right)^2\big/\sigma^2$ has a chi-square distribution with n-1 degrees of freedom. In design-based theory, no assumptions are made about an underlying model distribution. As a result, large sample theory is used to assign approximate degrees of freedom to variance estimators (e.g., see Rust (1984, 1985)). The standard rule of thumb is

$$df = \text{(number of sample PSUs)-(number of design strata)}. \tag{15.8}$$

If there are $n = \sum_{h=1}^{H} n_h$ sample PSUs and H strata, the rule says that $df = n - H$. In other words, $n_h - 1$ degrees of freedom are picked up from each stratum.

How accurate this rule is depends on the variability and kurtosis of the analysis variable, which in this section is the linear substitute z_k. Kurtosis is a measure of how "peaked" the distribution of y_k or z_k is in comparison to a standard normal distribution. In many cases, the rule of thumb may be poor as illustrated in Valliant and Rust (2010). Among the things that will taint its accuracy are:

(i) Non-normality of the z_k which can be caused by a small number of sample PSUs.

(ii) The z_k having heavier tails than a normal distribution.

(iii) The underlying variances of the analysis variables being different among strata.

(iv) The statistic is the proportion of the population that has a rare characteristic. This can result in a heavy-tailed distribution of the z_k.

(v) PSUs and/or strata are collapsed together to reduce computational burden. This is common when using the replication variance estimators discussed in subsequent sections. Collapsing is described in Sect. 15.5.

Example 15.4 (Evaluating the partial derivatives: ratio estimator of a mean). To construct a linearization variance estimator, alternatives are sometimes

available for how to evaluate the partial derivatives in Eq. (15.5). The ratio estimator of a mean under *srswor*, $\bar{y}_R = \bar{y}_s \bar{x}_U / \bar{x}_s$, which we covered in Sect. 3.5.2, will illustrate the options. The linear approximation to $\bar{y}_R$ is

$$\bar{y}_R - \bar{y}_U \doteq \frac{\partial \bar{y}_R}{\partial \bar{y}_s} (\bar{y}_s - \bar{y}_U) + \frac{\partial \bar{y}_R}{\partial \bar{x}_s} (\bar{x}_s - \bar{x}_U).$$

The theorem that leads to the approximation says that the partials should be evaluated at population values. Dropping the terms in $\bar{y}_U$ and $\bar{x}_U$, the part of the approximation that depends on the sample quantities is

$$\frac{\partial \bar{y}_R}{\partial \bar{y}_s} \bar{y}_s + \frac{\partial \bar{y}_R}{\partial \bar{x}_s} \bar{x}_s = n^{-1} \sum_s \left(\frac{\partial \bar{y}_R}{\partial \bar{y}_s} y_k + \frac{\partial \bar{y}_R}{\partial \bar{x}_s} x_k \right).$$

The partial derivative of $\bar{y}_R$ with respect to $\bar{y}_s$ is $\bar{x}_U / \bar{x}_s$. If evaluated at population quantities, then the partial derivative is equal to one. Otherwise, if the partial derivative is evaluated at sample quantities, we have $\bar{x}_U / \bar{x}_s$. The partial with respect to $\bar{x}_s$ is $-\bar{y}_s \bar{x}_U / \bar{x}_s^2$. When evaluated at population and sample quantities, this partial derivative is $\bar{y}_U / \bar{x}_U$ and $-\bar{y}_s \bar{x}_U / \bar{x}_s^2$, respectively. Thus, two choices for linear approximations are

$\bar{y}_R - \bar{y}_U \doteq n^{-1} \sum_s \left(y_k - \frac{\bar{y}_U}{\bar{x}_U} x_k \right)$ derivatives evaluated at population values

$\bar{y}_R - \bar{y}_U \doteq n^{-1} \frac{\bar{x}_U}{\bar{x}_s} \sum_s \left(y_k - \frac{\bar{y}_U}{\bar{x}_U} x_k \right)$ derivatives evaluated at sample estimates.

The first approximation leads to the *srswor* variance estimator

$$v_0 = \frac{N^2}{n} \left(1 - \frac{n}{N} \right) \sum_s \left(y_k - \frac{\bar{y}_s}{\bar{x}_s} x_k \right)^2.$$

We use $\bar{y}_s / \bar{x}_s$ instead of $\bar{y}_U / \bar{x}_U$ in the squared term because the population mean of the y's is unknown. The second approximation leads to

$$v_2 = \frac{N^2}{n} \left(1 - \frac{n}{N} \right) \left(\frac{\bar{x}_U}{\bar{x}_s} \right)^2 \sum_s \left(y_k - \frac{\bar{y}_s}{\bar{x}_s} x_k \right)^2.$$

The estimator v_2 has better conditional performance than does v_0. By "conditional," we mean that v_2 tracks the variance of $\bar{y}_R$ better than v_0 for samples where $\bar{x}_s$ differs from $\bar{x}_U$. More formally, good conditional performance means that an estimator is unbiased (or approximately so) under a model that describes the dependence of y on x. In this case, the model that motivates the ratio estimator is $E_M(y_i) = \beta x_i$, $V_M(y_i) = \sigma^2 x_i$. The estimator v_2 is both model unbiased and approximately design unbiased under *srswor* giving it a kind of *double robustness*—a term used in the clinical trials literature (Kang and Schafer 2007). ∎

To arrive at v_2, a somewhat arbitrary choice is made to evaluate the partials in a way different than dictated by Taylor's theorem. An interesting

feature of replication estimators, discussed in Sect. 15.4 is that they automatically are approximately design unbiased and model unbiased. This is not quite as good as it sounds because the design-unbiasedness is under with-replacement sampling of PSUs, and the model-unbiasedness is under a model for which the point estimator itself is unbiased. The actual design may not be with-replacement, and the model under which the replication variance estimator is unbiased may not be the best one for the analysis variable. Nonetheless, replication conveys a kind of automatic, double robustness while this is not always true of linearization variance estimators.

15.3.3 Accounting for Non-negligible Sampling Fractions

The general, large sample theory for linearization is built around the assumption that PSUs are selected with replacement. As noted earlier, this is not much of a limitation if the size of the sample of first-stage units is small compared to the population size of first-stage units. Accounting for large sampling fractions of PSUs or selections of PSUs that cannot be treated as approximately independent is difficult except in some simple designs.

If the selections of PSUs cannot be realistically treated as being independent, then the basic question for an analyst is whether (a) the software you are using has a variance formula that matches the design or (b) you can program the correct formula yourself. If (b) is within your grasp, then an elaborate formula can be programmed that fully accounts for the complexity of the design and estimator. For most analysts, though, (a) is probably more realistic. This is especially true if your analysis requires many different domain estimates. Programming these correctly is not a trivial exercise. SUDAAN accommodates more types of sample designs than the other packages we cover (see RTI International 2012, chap. 3). For example, it covers designs in which (15.1) is the right formula, but the user must input the values of π_i and π_{ij}. The π_{ij} values, in particular, may not be available.

An option offered by R, SAS, Stata, and SUDAAN is to include an *ad hoc* finite population correction (*fpc*) factor into formula (15.4) as applied to $\hat{z}$ rather than $\hat{\bar{y}}_{pwr}$. This is theoretically correct if the PSUs are selected by *srswor* or *stsrswor* where the *fpc* is either $1 - n/N$ or $1 - n_h/N_h$, and there is no subsampling within each PSU. If the PSUs are selected with varying probabilities without replacement, then this kind of *fpc* may be crude. In the R survey package the option fpc is included in the svydesign statement; in Stata the *fpc* is included in the svyset statement; in each SAS procedure (like surveyfreq) the statement is rate. In R and Stata the value of the sampling rate, n/N, is the value of the *fpc* variable—not $1 - n/N$, which is the textbook definition of the *fpc*. In R the *fpc* must be a vector of the same length as the number of records in the sample file; it could be a column in

the object that holds the sample data. We gave an example of the R syntax in Sect. 15.2. In Stata and SAS, `fpc` or `rate` should be a field in the sample data file.

Additionally, *fpc*'s for different stages of sampling can be included in R, Stata, and SUDAAN. These are appropriate only when each stage is a simple random sample selected without replacement from the units at each stage. Multistage designs with *srswor* at each stage are fairly unusual, but the option to include several *fpc*'s is available. You should consult the manual for the software package you are using to learn how the data file needs to be set up to use this option.

Example 15.5 (Accounting for fpc's). We illustrate the effect of using *fpc*'s by selecting an *stsrswor* from the `smho.N874` population. A sample of $n_h = 50$ is selected in each of $H = 5$ strata defined by hospital type. The strata function in the `sampling` package selects the sample. The stratum-specific sampling fractions, n_h/N_h (0.23, 0.43, 0.20, 0.34, and 0.35 for strata 1 through 5, respectively), are stored in `sam$Prob`, whose length is that of the full sample, 250 (5*50) because this rate is the same for every sample unit in a given stratum. The full set of R code for this example is in "`Example 15.5 FPCs.R`."

```
require(survey)
require(sampling)
        # Population stratum counts
Nh <- table(smho.N874[, "hosp.type"])

        # Select a stratified simple random sample within
        # hospital type strata
set.seed(428274453)
n <- 50
H <- length(Nh)
sam <- strata(data = smho.N874, stratanames = "hosp.type",
              size = rep(n,H), method=c("srswor"),
              description = TRUE)
sam.dat <- smho.N874[sam\$ID\_unit,]
d <- 1/sam$Prob
sam.rates <- sam$Prob
        # Create a design object with fpc's
smho.dsgn <- svydesign(ids = ~0,           # no clusters
                       strata = ~hosp.type,
                       fpc = ~sam.rates,
                       data = data.frame(sam.dat),
                       weights = ~d)

cv(svyby(~EXPTOTAL, by=~as.factor(hosp.type), design=smho.dsgn,
         FUN=svytotal))
cv(svytotal(~EXPTOTAL, design=smho.dsgn))

        # Create a design object without fpc's
smho.nofpc.dsgn <- svydesign(ids = ~0,
                       strata = ~hosp.type,
```

```
                        data = data.frame(sam.dat),
                        weights = ~d)
 cv(svyby(~EXPTOTAL, by=~as.factor(hosp.type),
              design=smho.nofpc.dsgn, FUN=svytotal))
 cv(svytotal(~EXPTOTAL, design=smho.nofpc.dsgn))
```

Two design objects are created: smho.dsgn, which uses *fpc*'s, and smho.nofpc.dsgn, which does not. The results for *CVs* of the estimated total of expenditures by stratum and overall are given below. Omitting the *fpc*'s leads to SEs and *CVs* being overestimated from 12 to 33 %. The increased SE size

Stratum	1	2	3	4	5	Full pop.
CV (%) with *fpc*	17.6	11.3	9.5	17.1	13.1	8.7
CV (%) without *fpc*	20.1	15.0	10.6	21.0	16.3	10.1
Ratio of *CVs*	1.14	1.33	1.12	1.23	1.24	1.16

could result in, for example, failing to reject the null hypothesis specified for a statistical test when the hypothesis could have been rejected or suppression of survey estimates if they exceed some specified relative standard error. ∎

15.3.4 Domain Estimation

Estimates for domains (i.e., subpopulations or domains) are important in the analysis of data from most surveys. The estimates for cells in a crosstab are examples of domain estimates. One way of characterizing domains (also referred to as subpopulations or subgroups) is by whether the sample size from the domain is fixed by the design or not. If the domain sample size is fixed, then analysis of the domain can be done by creating a subfile that contains only the units in the domain. For instance, if the employees of a company are stratified by division in which they work (data processing, field operations, statistical, human resources, etc.), then each division can be analyzed separately. If the sample sizes are not fixed, then the randomness of the domain sample size should be incorporated in the variance estimates. In the employee survey, we might be interested in the domain of persons who feel that they are underemployed considering their education levels. Assuming that we do not know who those people are prior to doing the survey, their sample size will be random.

   The technique used for estimating the design variance of a domain estimate for which the sample size is random is to code a unit as having a value of 0 if they are not in the domain and as having its observed value if it is in the domain:

$$y_k(d) = \begin{cases} y_k & \text{if } k \text{ is in domain } d, \\ 0 & \text{if not.} \end{cases}$$

Some texts use an indicator variable, $\Delta_i(d) = 1$ if unit i is in domain d and 0 if not. Then $y_k(d) = y_k \Delta_i(d)$. The recoded $y_k(d)$ is then used in whatever variance formula is appropriate for the design. For a linearization variance estimator, $y_k(d)$ is used in the linear substitute. As we will see in Sect. 15.4, this zero-coding trick is unnecessary in replication variance estimation—another advantage of the replication approach.

15.3.5 Assumptions and Limitations

Theory is available for linearization variance estimators to show when they are approximately unbiased and consistent. Krewski and Rao (1981) provide the fundamental theory, which is summarized also by Wolter (2007). The type of sample design does have to be considered—in particular whether the PSUs were sampled with or without replacement. In an easy case like *stsrs* the linearization approach can be applied to without-replacement designs, as illustrated in Examples 15.1 and 15.2. In multistage samples, much of the theory has been developed for designs in which PSUs can be selected with varying probabilities but *with replacement*. In that case, the ultimate cluster variance estimator can be applied to the linear substitutes. When the PSU sampling is without replacement, a with-replacement variance estimator is usually conservative, but this is a compromise that most practitioners can accept.

There are also some mathematical assumptions needed to derive the theory that applies to nonlinear estimators. Three of the key mathematical requirements are that (i) the number of sample PSUs is large, (ii) the variables being analyzed (the y's) cannot be highly variable or affected by any extreme outliers, and (iii) the nonlinear function, $\hat{\theta} = f(\hat{t}_1, \ldots, \hat{t}_p)$, must be differentiable with respect to its $\hat{t}_j$ ingredients. Different types of designs can satisfy requirement (i). In a stratified design with a limited number of strata, there must be a large number of PSUs in each stratum. A stratified design with a small number of units per stratum can satisfy (i) if the number of strata is large.

As noted earlier, linearization variance estimators are used to compute confidence intervals (CIs) of the form $\hat{\theta} \pm z_{1-\alpha/2} \sqrt{v_L(\hat{\theta})}$. There are two assumptions needed to say that this interval (or the version using a t multiplier) covers the desired population quantity in $100(1-\alpha)\%$ of samples:

(i) The distribution of $\hat{\theta}$ is approximately normal when the sample is large.

(ii) $v_L(\hat{\theta})$ is a consistent estimator of the theoretical variance, $V(\hat{\theta})$, in the sense that $v_L(\hat{\theta}) / V(\hat{\theta}) \xrightarrow{p} 1$ as the sample becomes large.

By "large sample" we mean that the number of sample PSUs is "large." This, of course, raises the question: how big does the PSU sample have to be to be considered large? Naturally, this is a question without a clear-cut answer. Highly skewed continuous variables will need a bigger sample size than more symmetric ones. Rare or prevalent characteristics will require a larger sample than the ones whose proportion is nearer.

Some practitioners will say that 30 PSUs are enough for CIs to perform as advertised. We prefer a much larger number—60 or more. Even if 30 would be sufficient to treat $\hat{\theta}$ as normal, a variance estimator based on 30 PSUs can be quite unstable. This will seriously foul the performance of confidence intervals. Having at least 60 PSUs offers a modicum of protection against unstable variance estimation. We return to this point later in the chapter in the discussion of replication variance estimators.

The method outlined above does not work for estimating the variance of a quantile, like the median or the first and third quartiles. However, a linearization-like method was developed for quantiles by Francisco and Fuller (1991). Their method is available in the R survey package and in SUDAAN. Example 15.8 illustrates its calculation along with another method due to Woodruff (1952).

15.3.6 Special Cases: Poststratification and Quantiles

In Chap. 14, several examples showed standard errors estimated via linearization. Examples 14.2, 14.3, and 14.4 covered poststratified and raked estimators and their standard errors. The next example shows a simpler case of linearization that would often be used for public-use data sets provided by federal governments.

Example 15.6 (Linearization variance estimation). The nhis.large data set can be treated as a stratified, cluster design with 2 PSUs selected per stratum. Since this was extracted from a public-use data set published by the US government, no design information was provided other than identifiers (IDs) for the design strata and PSUs and a survey weight. The R code below uses the RData version of this file to estimate the proportions of the population in five age groups. The with-replacement variance estimator is used since we have no information to do anything more elaborate. The svydesign function defines a design object that specifies the variable that holds the PSU IDs (ids), the strata (strata), the survey weights (weights), and the data set used to create the object. Notice that the ids, strata, and weights variables have to be entered as a formula within the front. After the design object is created, tabulations of different kinds can be done. In this case, svymean is used to compute proportions, their linearization standard errors, and a design effect for each cell estimate. The ftable function is used to format

the table in a slightly nicer way than the default from `svymean`. The full set of code is in the file `Example 15.6 lin var`.

```
attach("C:\\nhis.large.RData", pos=2)
require(sampling)
            # create a design object
nhis.dsgn <- svydesign(ids = ~psu,
                       strata = ~stratum,
                       nest = TRUE,
                       data = nhis.large,
                       weights = ~svywt)

a <- svymean(~factor(age.grp), deff=TRUE, design=nhis.dsgn)
b <- ftable(a, rownames = list(age = c("< 18", "18-24", "25-44",
                   "45-64", "65+")))
round(b,3)

age
< 18   mean  0.253
       SE    0.004
       Deff  1.575
18-24 mean   0.101
       SE    0.004
       Deff  3.872
25-44 mean   0.285
       SE    0.004
       Deff  1.463
45-64 mean   0.240
       SE    0.004
       Deff  2.092
65+    mean  0.122
       SE    0.004
       Deff  3.268
```

The same tabulation can be done in Stata with this code after telling the package to use the `nhis.large` data set:

```
svyset psu [pweight=svywt], strata(stratum)
svy: tab agegrp, percent se deff
```

After reading the data into a file called `nhis_large`, the SAS code for the table is:

```
proc surveyfreq data=nhis_large;
 tables agegrp / deff;
 strata stratum;
 cluster psu;
 weight svywt;
 run;
```

(The variable `agegrp` is used above because SAS and Stata do not support variable names containing certain characters such as a period, e.g., `age.grp`.)

■

Ignoring calibration to population controls is usually a serious error in variance estimation. How serious the mistake is depends on how well the model that underlies the type of calibration used fits the data. (We covered the links between calibration and models in Sect. 14.3.1.) The better the fit, the bigger the mistake. Not accounting for calibration in variance estimation is an especially easy mistake to make when analyzing data from public-use data sets. Sometimes no special guidance is provided for how to estimate variances with different software packages. Stratum and PSU codes may be on the file, and the documentation may say that weights were adjusted to hit certain population controls, like age/sex/race-ethnicity counts in a household survey. But, no population counts are provided to users, or the survey documentation does not give explicit definitions of the control categories. In such a case, you may be able to still do something that is roughly correct (or, at least, better than ignoring the fact that controls were used).

Take the case of poststratification for illustration. The sum of the final weights in each poststratum will satisfy

$$\hat{N}_\gamma \left(= \sum_{k \in s_\gamma} w_k \right) = N_\gamma,$$

where s_γ is the set of sample units in poststratum γ and N_γ is the population control. The variance of the estimates poststratum count $\hat{N}_\gamma$ is zero because of the forced equality with the population controls above. Thus, if you have a reasonable guess about what the poststrata definitions are, you can recover the control totals. You may need to use those controls to create a new set of poststratified weights, depending on the requirements of the software package you are using. To illustrate the possibility that ignoring poststratification would be an error, we return to Example 15.2.

Example 15.7 (Linearization with poststratification). The R code for this example is in the file `Example 15.7 poststrat.R`. Recall that in Example 15.2, an *srswor* of 250 cases was selected from the nhis.large population. Fifteen age groups x Hispanicity poststrata were used. A design object called nhis.dsgn was created and, in turn, used to create an object with poststratified weights, ps.dsgn.

```
        # collapse hisp = 3,4
hisp.r <- nhis.large$hisp
hisp.r[nhis.large$hisp ==4] <- 3
nhis.large1 <- data.frame(nhis.large, hisp.r)

        # create single variable to identify
        #            age.grp x hisp.r poststrata
 m <- max(nhis.large1$hisp.r)
 nhis.large1$PS <- (nhis.large1$age.grp - 1)*m + nhis.large1$hisp.r
 N.PS <- table(PS = nhis.large1$PS)
 ps.dsgn <- postStratify(design = nhis.dsgn,
                         strata = ~PS,
                         population = N.PS)
```

(In this example, in contrast to Example 15.2, we omit an *fpc.*) A critical requirement is that the name associated with the population total vector, N.PS, must be the same as the name of the variable used to identify post-strata. The statement, table(PS=nhis.large1$PS), insures that the name PS is used for the population totals. The poststratified estimated totals of persons receiving Medicaid (medicaid=1) or not (medicaid=2) are estimated with:

```
svytotal(~ as.factor(medicaid), ps.dsgn, na.rm=TRUE)
                       total      SE
as.factor(medicaid)1   1870.8 346.47
as.factor(medicaid)2  19467.6 372.59
```

On the other hand, we can use the poststratified weights to form a design object assuming that the sample was selected with varying probabilities and with replacement, and then estimate the same total.

```
wts <- weights(ps.dsgn)
         # design object ignoring PS
noPS.dsgn <- svydesign(ids = ~0,
          strata = NULL,
          data = data.frame(samdat),
          weights = ~wts)

svytotal(~ as.factor(medicaid), noPS.dsgn, na.rm=TRUE)
                       total      SE
as.factor(medicaid)1   1870.8 384.73
as.factor(medicaid)2  19467.6 470.38
```

The estimated totals are, of course, the same with these two alternatives. However, the SEs for the total number of persons receiving Medicaid are 346.47, accounting for poststratification, and 384.73, ignoring it. Consequently, we would overestimate the SE by about 11 % (384.73 vs. 346.47). For the estimated total not receiving Medicaid the SE would be overestimated by 26 % (470.38 vs. 372.59). The overestimation would also occur with the replication methods, considered subsequently, if poststratification is ignored. ∎

Estimating the SE of an estimated quantile requires different methods from those introduced earlier for linearization. The Francisco and Fuller (1991, FF) and Woodruff (1952) methods are available in R. Both methods first compute a confidence interval (CI) on a quantile. A standard error is then computed by dividing the CI length by $2z_{1-\alpha/2}$ where $100(1-\alpha)\%$ is the level of the confidence interval. FF uses what is called a *test inversion* method. For the median, for example, the CI consists of all potential population values that would be accepted in a hypothesis test that the value was equal to the median. The Woodruff method is simpler and consists, roughly, of putting a CI around the proportion associated with the quantile (like the median) and then translating the CI endpoints back to the data scale.

Example 15.8 (Quantiles). We use the `smho.N874` population to illustrate the computation of quantiles and the same sample described in Sect. 15.3.2. Type 4 hospitals are deleted and the variable, beds, is recoded to have a minimum value of 5. A sample of 80 hospitals is selected from the edited list frame with probabilities proportional to the square root of the number of (recoded) beds. The complete set of code is given in the file `Example 15.8 FF quantile.R`.

```
smho.dsgn <- svydesign(ids = ~0,
          strata = NULL,
          data = data.frame(sam.dat),
          weights = ~d)

      # population quantiles
popq <- quantile(smho$SEENCNT, c(0.25, 0.50, 0.75))

      # Compute quantiles and CIs
      # Francisco-Fuller method
FF <- svyquantile(~SEENCNT, design=smho.dsgn,
                  quantiles = c(0.25, 0.50, 0.75),
                  ci=TRUE, interval.type="score",
                  se = TRUE)

      # Woodruff method
wood <- svyquantile(~SEENCNT, design=smho.dsgn,
                    quantiles = c(0.25, 0.50, 0.75),
                    ci=TRUE, interval.type="Wald",
                    se = TRUE)
round(cbind(t(FF$quantiles), t(FF$CIs[,,1])), 0)
     SEENCNT (lower upper)
0.25     581     208    846
0.5     1458     846   1613
0.75    1932    1654   4182

round(cbind(t(wood$quantiles), t(wood$CIs[,,1])), 0)
     SEENCNT (lower upper)
0.25     581     184    753
0.5     1458     829   1622
0.75    1932    1663   4759

   # extract SEs
 round(SE(FF),1)
 0.25   0.5  0.75
162.8 195.7 644.9

 round(SE(wood),1)
 0.25   0.5  0.75
145.3 202.4 790.0
```

The object `sam.dat` holds the data for the 80 sample hospitals. The function `svyquantile` computes the first and third quartiles and the median via the parameter, `quantiles = c(0.25,0.50,0.75)`. The FF and Woodruff

methods are specified with `interval.type="score"` or `"Wald"`, respectively. The output is a list with components named `quantiles` and `CIs`. For FF we examine these by binding the point estimates and CI limits together with

```
round(cbind(t(FF$quantiles), t(FF$CIs[,,1])), 0).
```

A similar statement displays the results for Woodruff. The standard error estimates are extracted with `SE(FF)` and `SE(wood)`. ∎

Effect of Duplicate Values on a Quantile. A word of warning is appropriate here for variables that have many duplicated values. In physical measurements, for example, data for some items may be rounded to integers for inclusion in a data set even though, in principle, the underlying measurement is continuous. For example, the NHANES data files provided by the US National Center for Health Statistics have many ties in high-density lipoprotein (HDL) cholesterol. HDL is measured in milligrams per deciliter, which is an integer, but conceptually HDL could take on a continuum of values. CIs and SEs of quantiles are sensitive to ties in data values. Depending on how these are handled, point estimates will differ somewhat, but CIs and SEs can differ a lot. The R `svyquantile` function has two options: `ties=``discrete''` and `ties=``rounded''`. With the former, the data are treated as genuinely discrete so that the CDF is a step function. With rounded, interpolation is used to construct the CDF. If the discreteness of the data is an artifact of the measurement or reporting process, then using `ties=``rounded''` seems preferable.

15.3.7 Handling Multiple Weighting Steps with Linearization

The implementations of linearization in software packages typically do not account for the effects of multiple stages of weight adjustment. For example, if nonresponse adjustments are used, followed by poststratification to population control totals, the linearization formulas that are preprogrammed in R, Stata, SAS, and SUDAAN will account only for poststratification if properly specified (see Example 15.7).

The theory for the method can certainly be adapted to reflect both steps. However, as noted earlier, users do not often have all of the information that would be needed to properly compute a linearization variance. For example, suppose that nonresponse adjustment cells were formed, as described in Sect. 13.5.1, and the input weights adjusted by the ratio of sums of weights for the full sample and for the respondents. If a total of some y is estimated, the analyst would have to know which cell each respondent and nonrespondent fell into, along with the sum of the input weights for the full sample

and for the respondents in each cell. If a poststratified estimator is used on top of this, the poststratum of each unit must be known. The poststrata may be different from the nonresponse adjustment cells. Users may have the poststratum codes but not the nonresponse adjustment information. In some public-use data sets, users will have neither.

On the other hand, replication, discussed in the next section, makes it relatively easy to account for such multiple weighting steps. As long as the replicate weights are properly constructed, an analyst can use them to get correct SE estimates even if the analyst has no knowledge of whether nonresponse adjustments, poststratification, raking, GREG estimation, or something else was used.

15.4 Replication

The other, general method for estimating the variance of nonlinear estimators is replication. The idea is to create a series of subsamples, i.e., *replicates*, each of which can be used to estimate the same parameter as the full sample. The variance is then computed among the replicate estimates. There are three alternatives that we will cover—the jackknife, balanced repeated replication (BRR), and the bootstrap.

In each of the methods, subsamples of the PSUs are selected—not of the units within PSUs. A subsample is referred to as a *replicate*. If a PSU is selected for a replicate, every sample unit within the PSU is retained. The base weights for the units in a replicate are adjusted in a way that depends on the method of replication. Then, any additional weight adjustments such as calibration (if they are used) are carried out separately for each replicate. This leads to a set of weights for each replicate in addition to the full-sample weights generated for linear variance estimation (Sect. 15.3). These weights are appended to the record for each sample unit and are used to compute the replicate estimates.

Three types of replicate variance estimators are reviewed in the subsequent sections. For each, we provide an overview of the procedures to calculate the corresponding weights, references for the theoretical details, and the advantages and limitations.

15.4.1 Jackknife Replication

The basic jackknife method creates replicates by dropping one first-stage unit and reweighting the remaining units to produce a full population estimate. For example, if unit i out of n first-stage units is dropped, then the weight for unit k is

$$d_{k(i)} = \frac{n}{n-1} d_k,$$

where $(n-1)/n$ is the subsampling fraction or probability of retaining the $n-1$ first-stage units in the ith replicate sample. The estimated total for a variable y based on replicate i is

$$\hat{t}_{(i)} = \sum_{k \in s(i)} d_{k(i)} y_k,$$

where $s(i)$ denotes the set of sample units excluding unit i. The jackknife variance estimator is calculated across the n replicate estimates using

$$v_J = \frac{n-1}{n} \sum_{i=1}^{n} \left(\hat{t}_{(i)} - \hat{t} \right)^2, \tag{15.9}$$

where $\hat{t} = \sum_{k \in s} d_k y_k$, the full-sample estimate of $t = \sum_{k \in U} d_k y_k$. There are some other variations of the jackknife derived from centering the replicate estimates around the mean of the $\hat{t}_{(i)}$'s and some other options (e.g., see Krewski and Rao 1981). All of these are numerically about the same in large samples.

As an example, consider the estimator of a total from a simple random sample, $\hat{t} = N\bar{y}_s$. Expression (15.9) reduces to

$$\frac{N^2}{n(n-1)} \sum_s (y_k - \bar{y}_s)^2,$$

which is the standard formula for the variance of $\hat{t}$ in *srswr*. Since this variance estimator can be computed directly, the jackknife has no advantage for $\hat{t} = N\bar{y}_s$ nor for any other linear estimator.

The benefit of the jackknife is that it is approximately unbiased and consistent for the variance of *nonlinear* estimators. If the nonlinear estimator, $\hat{\theta} = f\left(\hat{t}_1, \ldots, \hat{t}_p\right)$, is being analyzed, the jackknife is constructed by deleting unit i and computing $\hat{\theta}_{(i)} = f\left(\hat{t}_{1(i)}, \ldots, \hat{t}_{p(i)}\right)$. Each replicate estimate $\hat{\theta}_{(1)}$, $\hat{\theta}_{(2)}, \ldots, \hat{\theta}_{(n)}$ is computed corresponding to dropping each of the n units in the sample. Each replicate estimate has the same form as the full-sample estimate. The results are then plugged into Eq. (15.9) to estimate the variance.

If a multistage sample is selected, "deleting a unit" means "delete a PSU." By deleting a PSU, we mean that all sample units in a PSU are dropped when the PSU is dropped. Dropping one unit at a time from within a PSU will give incorrect variance estimates. With a stratified multistage design, one PSU is omitted at a time to create a replicate, and the weight adjustment for a replicate applies only to the PSUs within the stratum where a PSU was dropped. Suppose that PSU i in stratum h $(h = 1, \ldots, H)$ is removed to form replicate (hi). Denote the adjusted base weight for unit k in replicate (hi) by $d_{k(hi)}$. Then, the base weights d_k are adjusted this way:

$$d_{k(hi)} = \begin{cases} 0 & \text{if unit } k \text{ is in PSU } i \text{ in stratum } h, \\ \frac{n_h}{n_h-1}d_k & \text{if unit } k \text{ is in stratum } h \text{ but not in PSU } i, \\ d_k & \text{if unit } k \text{ is not in stratum } h. \end{cases} \quad (15.10)$$

In other words, all units in the deleted PSU hi have their weights set to 0. All units in the other $(n_h - 1)$ PSUs within stratum h have their base weights multiplied by $n_h/(n_h - 1)$, the inverse of the within-stratum subsampling fraction. Units in the strata where no PSU was dropped retain their original weight. Thus, the weights for the retained units in stratum h are adjusted to represent the full stratum and the weights for units in other strata are left alone.

The adjusted weights are then used in the same way as they would be in a single-stage sample to compute a replicate estimate denoted by $\hat{\theta}_{(hi)}$. The stratified jackknife variance estimator is then

$$v_J\left(\hat{\theta}\right) = \sum_h \frac{n_h - 1}{n_h} \sum_{i \in s_h} \left(\hat{\theta}_{(hi)} - \hat{\theta}\right)^2, \quad (15.11)$$

where s_h denotes the sample of PSUs in stratum h and $\hat{\theta}$ is the full-sample estimate calculated with, for example, the linearization weight or as the weighted average of the replicate estimates. Expression (15.11) is sometimes referred to as the JKn formula. Since one PSU is dropped to form each replicate, the total number of replicates in JKn equals the number of sample PSUs. As in the unstratified case, there are some other options for how the variance estimate can be centered. But, as long as the PSU sample is large, these will be numerically similar.

Special Cases

There are two special cases of the jackknife that sometimes crop up in the literature that are worth a brief discussion. One is the unstratified jackknife, introduced at the beginning of this section, which is sometimes called JK1. This is really just a special case of JKn with one stratum. Two sample PSUs in each stratum lead to another special case. When $n_h = 2$, the JKn formula for an estimated total, $\hat{t} = \sum_h \sum_{i=1}^{2} \sum_{k \in s_{hi}} d_k y_k$, reduces to

$$v_J\left(\hat{t}\right) = \sum_h \left(\hat{Y}_{h1} - \hat{Y}_{h2}\right)^2, \quad (15.12)$$

where $\hat{Y}_{hi} = \sum_{k \in s_{hi}} d_k y_k$ as defined below (15.4). Expression (15.12) is known as JK2 and is available only in WesVar. Since JK2 only requires deleting the first PSU in each stratum, it is important to avoid numbering PSUs as 1 and 2 in some systematic way. For example, if PSU 1 is always the one with the smaller population size and size is related to the analysis variables, then JK2 can be biased. As a result, randomly numbering the PSUs as 1 or 2 within each stratum is a good idea.

The JK2 variance estimator has no particular theoretical support for nonlinear estimators but does lead to fewer replicates being used than in JKn. Within JKn, $2H$ replicates would be needed in a 2-per-stratum design where H is the number of strata. In JK2 only H replicates are needed. This could be quite a savings if the number of strata is large. However, the BRR method to be covered in Sect. 15.4.2 applies in the 2-units-per-stratum case and has been proven to work for nonlinear estimators and for quantiles like the median. Neither JKn nor JK2 converges to the correct variance for quantiles. BRR also requires only slightly more replicates than JK2. Thus, there seems to be no good reason to use JK2 in any application.

Domain Estimation and Replication

The jackknife, BRR, and the bootstrap all correctly handle domain estimation without doing the explicit zero-coding for non-domain members that was needed for linearization. Using the recoded variable, $y_k(d) = 0$ for units outside the domain and y_k for domain units, is still correct for replication. But, this is equivalent to dropping the zero-coded units when computing $\hat{\theta}_{(hi)}$ and $\hat{\theta}$ to use in Eq. (15.11). Dropping the non-domain units is the standard way of computing the jackknife (or BRR or bootstrap) variance estimates. Recall that deleting the non-domain units and computing a linearization variance estimate from that subset of the file would generally be a mistake.

Assumptions, Advantages, and Limitations

The assumptions for the jackknife to be approximately unbiased and consistent for the variance of a nonlinear estimator are the same as for linearization: the PSU sample must be large, the analysis variable y has no extreme outliers and is not highly variable, and it must be possible to take all first derivatives of the nonlinear function. Krewski and Rao (1981) give the full set of technical conditions. The theory for the jackknife basically says that it is equivalent to the linearization estimator in very large samples. Thus, anywhere linearization works, the jackknife should work.

The great advantage of the jackknife (and BRR and the bootstrap) is that it can also implicitly reflect the effects on variances of nonresponse and calibration adjustments. If a nonresponse adjustment procedure is used for the full sample, the same procedure should be done separately for each replicate weight. For example, if calibration, say via poststratification or GREG, is used, then that should also be done separately for every replicate. The reason that the jackknife reflects these adjustments is that even with a series of nonlinear adjustments, many estimators can still be written in the form $\hat{\theta} = f\left(\hat{t}_1, \ldots, \hat{t}_p\right)$. For example, the poststratified estimator, $\hat{T}_{yPS} = \sum_{\gamma=1}^{G} N_\gamma \left(\hat{t}_{y\gamma} \middle/ \hat{N}_\gamma\right)$, is a function of $2G$ estimated totals—

$\hat{t}_{y1}, \ldots, \hat{t}_{yG}, \hat{N}_1, \ldots, \hat{N}_G$. If, in addition, nonresponse adjustments within cells are used, this simply adds some more estimated totals to the nonlinear f.

The available software that does linearization variance estimation does not, typically, account for the effects of multiple stages of weight adjustment. The reasons for this are twofold: (i) the appropriate linear substitute that accounts for all stages of sampling is not programmed and (ii) users generally cannot provide all the information that would be necessary for the software to compute the appropriate linear substitute. With a replicate variance estimate like the jackknife all that is necessary is to recompute every adjustment separately for every replicate. The jackknife implicitly estimates the variance for the linear approximation for a complicated nonlinear estimator and, thus, implicitly accounts for all the adjustment steps. Consequently, if the database constructor has computed the replicate weights in this way, any analyst can use them and obtain correct variance estimates.

Software packages that will compute jackknife replicate base weights using the complete sample file include R survey, WesVar, and the svr package, which is an add-on to Stata (Winter 2002). These packages will also compute BRR base weights, to be covered later in this chapter. Any additional weight adjustments applied to the full-sample weights to address nonresponse or calibration would need to be applied as a second step to each replicate base weight. Other software packages require the finalized replicate weights as inputs to the procedures.

Example 15.9 (JKn variance estimation). In this example, we show the syntax needed in R to do the tabulation in Example 15.6 using JKn variance estimation. As in the earlier example, we create a design object called `nhis.dsgn`. Then JKn weights are calculated by calling the `as.svrepdesign` function.

```
        # create a design object
 nhis.dsgn <- svydesign(ids = ~psu,
                        strata = ~stratum,
                        nest = TRUE,
                        data = nhis.large,
                        weights = ~svywt)
        # JKn
 jkn.dsgn <- as.svrepdesign(design = nhis.dsgn, type = "JKn")
        # 1-way table
 a <- svymean(~factor(age.grp), deff=TRUE, design=jkn.dsgn)
 ftable(a, rownames = list(age = c("< 18", "18-24", "25-44",
           "45-64", "65+")))
```

The results for this table are exactly the same to three decimal places as those for linearization in Example 15.6 and are not shown. The weight adjustments used for the JKn variance estimate can be examined with the extractor function

```
 weights(jkn.dsgn)
```

The nhis.large sample is a 2-PSU-per-stratum design. The function `as.svrepdesign` makes a weight adjustment of 2 to the weight of each unit in the PSU that is retained in a particular stratum for a replicate. The weight of each unit in a deleted PSU is set to 0. The weight for PSUs in strata where no PSU is deleted is unchanged. The dimension of `weights(jkn.dsgn)` is 21588×150, i.e., the number of persons in the file by twice the number of strata. ∎

Example 15.10 (JKn with cell nonresponse adjustments). As discussed previously, nonresponse adjustments should be applied separately for replicates to reflect their effects on variances. In this example, we use the this data set and the nonresponse adjustment classes determined using `rpart` in Sect. 13.5.3. Some snippets of the R code are shown below. The full program is in the file `Example 15.10 JKn NR.R`. The code uses the packages `rpart` and `doBy` in addition to survey.

```
        # create a design object
 nhis.dsgn <- svydesign(ids = ~psu,
                        strata = ~stratum,
                        nest = TRUE,
                        data = nhis,
                        weights = ~svywt)
        # JKn
 jkn.dsgn <- as.svrepdesign(design = nhis.dsgn,
                            type = "JKn")
        #  Compute a tree using rpart; code not shown
        #  Store cells in t1$where
        # append NR classes to nhis object
 nhis.NR <- data.frame(nhis, NR.class=t1$where)
        # wt adjustments for JKn (values are 0, 1, or 2)
 JKwtadj <- weights(jkn.dsgn)
 nreps <- ncol(JKwtadj)
 fswts <- nhis$svywt
 rep.adjwt <- matrix(0, nrow=nrow(JKwtadj), ncol=nreps)

        # compute NR adjustments for full sample
 wt.rr <- by(data = data.frame(resp = as.numeric(nhis$resp),
             wt =fswts),
             nhis.NR$NR.class,
             function(x) {weighted.mean(x$resp, x$wt)})
 tmp1 <- cbind(NR.class=as.numeric(names(wt.rr)), wt.rr)

 sam.nr <- merge(nhis.NR, data.frame(tmp1), by = "NR.class")
 sam.nr$fs.adjwt <- sam.nr$svywt / sam.nr$wt.rr
 sam.nr <- data.frame(ID = sam.nr$ID, fs.adjwt = sam.nr$fs.adjwt,
                wt.rr = sam.nr$wt.rr)
 sam.nr <- orderBy(~ID, data=sam.nr)
 fs.adjwt <- sam.nr$fs.adjwt

        # compute NR adjustments for each replicate
 for (r in 1:nreps){
```

```
    adjwts <- fswts * JKwtadj[,r]
            # wtd RR; adjwts=0 for units not in replicate
    wt.rr <- by(data = data.frame(resp = as.numeric(nhis.NR$resp),
                wt = adjwts),
                nhis.NR$NR.class,
                function(x) {weighted.mean(x$resp, x$wt)})
    tmp1 <- cbind(NR.class=as.numeric(names(wt.rr)), wt.rr)

    sam.nr <- merge(nhis.NR, data.frame(tmp1), by = "NR.class")
    sam.nr <- data.frame(sam.nr, wt.rr = sam.nr$wt.rr)
    sam.nr <- orderBy(~ID, data=sam.nr)
            # adjust rep wts for NR
    rep.adjwt[,r] <- adjwts / sam.nr$wt.rr
}

       # assign names to rep.adjwt columns and
       # append NR-adjusted weights onto nhis data file
rname <- vector("character", length=nreps)
for (r in 1:nreps){
   rname[r] <- paste("repwt",r,sep="")
}
dimnames(rep.adjwt)[[2]] <- rname

R <- nhis$resp == 1
nhis.NR <- cbind(nhis[R==1, ], fs.adjwt=fs.adjwt[R==1],
                rep.adjwt[R==1,])
       # extract wts for respondents only
rep.adjwt <- rep.adjwt[R==1,]

       # JKn design object with NR-adjusted weights
jkn.NR.dsgn <- svrepdesign(data = nhis.NR[,1:16],
                           repweights = rep.adjwt,
                           type = "JKn",
                           weights = nhis.NR$fs.adjwt,
                           combined.weights = TRUE,
                           scale = 1,
                           rscales = rep(1/2,nreps))

svytotal(~factor(age_r), design=jkn.NR.dsgn)
a <- svymean(~factor(age_r), design=jkn.NR.dsgn)
b <- ftable(a, rownames = list(age_r = c("18-24 years",
            "25-44 years","45-64 years","65-69 years",
            "70-74 years","75 years and older")))
round(b,4)

18-24 years          mean  0.1281
                     SE    0.0070
25-44 years          mean  0.3984
                     SE    0.0097
45-64 years          mean  0.3153
                     SE    0.0096
65-69 years          mean  0.0434
                     SE    0.0038
```

```
 70-74 years         mean   0.0417
                     SE     0.0044
 75 years and older mean   0.0731
                     SE     0.0058
```

First, a survey design object, `jkn.dsgn`, is created with JKn replicate weights. The `nhis` data set has 87 strata and 2 sample PSUs per stratum. Consequently, there are 174 JKn replicates. The object `JKwtadj` holds the weight adjustments for each replicate—not the adjusted weights themselves. In a 2-PSU-per-stratum design the JKn adjustments are 0, 1, or 2, i.e., the special case of Eq. (15.10) with $n_h = 2$. In this example, the nonresponse adjustment in each class is the inverse of the weighted response rate, computed using the function, `weighted.mean`, for the full sample and for each of the `nreps=174` replicates. We use the `by` function to get the response rate in each class.

Another design object is then created using `svrepdesign`. The parameter, `combined.weights = TRUE`, means that the replicate weights include the full-sample weights and the adjustments used when forming replicates. The parameters, `scale` and `rscales`, relate to the way that R survey forms the replicate variance formula. As described in Lumley (2010, Sect. 2.3.1), the formula used is

$$var\left(\hat{\theta}\right) = a_{\bullet} \sum_{\alpha=1}^{M} a_{\alpha} \left(\theta_{\alpha}^{*} - \bar{\theta}^{*}\right)^{2},$$

where α denotes a replicate, $a_{\bullet}$ is the scale parameter, a_{α} is the `rscales` parameter, θ_{α}^{*} is a replicate estimate, and $\bar{\theta}^{*}$ is the mean of the replicate estimates. To make this correspond to Eq. (15.11), we set $A = 1$ and $a_r = (n_h - 1)/n_h = 1/2$. The replicate variance estimate is centered around the mean of the replicate estimates, unlike the variance in Eq. (15.11) which is centered around the full-sample estimate. The variance centered around the mean will be somewhat smaller than the one centered around the full-sample estimate. But, the difference will be slight when the PSU sample is large. In this case, the JKn SEs that account for the nonresponse adjustment are very little different from the linearization SEs that ignore it. You can verify this with the code in `Example 15.10 JKn NR.R`. ∎

Example 15.11 (JKn with poststratification). The effect of calibration can also be reflected with a jackknife variance estimator. The poststratification adjustment must be redone separately for each replicate. We illustrate the calculation in R with the same poststratification example as used in Examples 15.2 and 15.7. The full listing of the R code is in `Example 15.11 JKn poststrat.R`. A design object called `nhis.dsgn` is created as in Example 15.7; `as.svrepdesign` creates the unstratified, jackknife design, which is specified by `type="JK1"`. Poststratified jackknife weights are computed with `postStratify` using poststratum totals stored in `N.PS`.

```
jk1.dsgn <- as.svrepdesign(design = nhis.dsgn, type = "JK1")
        # poststratified design object
jk1.ps.dsgn <- postStratify(design = jk1.dsgn,
                        strata = ~PS,
                        population = N.PS)

        # Check that weights are calibrated for x's
svytotal(~ as.factor(PS), jk1.ps.dsgn)
        # PS standard errors and CVs
svytotal(~ as.factor(medicaid), jk1.ps.dsgn, na.rm=TRUE)
                        total      SE
as.factor(medicaid)1   1870.8 390.60
as.factor(medicaid)2  19467.6 416.89
```

The results for the line, svytotal(as.factor(PS), jk1.ps.dsgn),
which are not listed here, show that the SE of the estimated total number of
persons in each poststratum is 0, as it should be.

In this example, the jackknife SE for the total of persons receiving Med-
icaid is somewhat larger than the linearization estimate that ignores post-
stratification (390.60 above vs. 384.73 in Example 15.7). The jackknife SE
for persons not getting Medicaid is less than the linearization estimate that
ignores poststratification (416.89 vs. 470.38 in Example 15.7). This apparent
contradiction is a reflection of the fact that standard error estimates are just
that—estimates. Expected gains due to stratification are not necessarily man-
ifested in the SE for every estimate. Also, poststratification is not guaranteed
to reduce SEs for all estimates, only those for variables that are related to
the ones used to create poststrata. ∎

15.4.2 Balanced Repeated Replication

Balanced repeated replication (BRR) or balanced half-sampling is a method
devised by McCarthy (1969) for designs where two PSUs are selected in
each stratum. This type of design is common in area probability samples
where a goal is often to spread the PSUs geographically as much as possible.
The number of strata and the geographic dispersion can be maximized by
selecting only 1 PSU per stratum. However, a one-per-stratum design does
not permit a within-stratum variance component to be estimated while a
two-per-stratum design does. Generally when a 1-PSU-per-stratum sample is
selected, the strata each containing one PSU are then paired to form "analytic
strata" following the order in which the PSUs were selected. The design is
then treated as if it were 2 PSUs per stratum. In that case, BRR can be
applied to the combined strata.

When $n_h = 2$, it would be possible to form a replication variance estimate
by randomly selecting one of the two PSUs from each stratum and doing this
many times. However, there would be 2^H possible half-samples that could be
randomly selected. McCarthy devised an ingenious method that, for linear

estimators, produces the same variance estimate as would be obtained by selecting all 2^H half-samples, yet takes far fewer replicates. The replicates are designated in a prescribed way using something called a Hadamard matrix. The number of replicates, A, needed is the smallest multiple of 4 that is greater than or equal to the number of strata, i.e., $H \leq A \leq H + 3$. A set of replicates that follows this prescription is called an *orthogonal* set. The savings in the number of replicates compared to using all 2^H half-samples are substantial. The savings increase dramatically as H increases as evidenced by this table:

H	A	2^H
5	8	32
10	12	1,024
20	24	1,048,576

Hadamard matrices are usually represented by +1's and -1's. A 4×4 example is

$$H = \begin{pmatrix} +1 & +1 & +1 & +1 \\ +1 & -1 & +1 & -1 \\ +1 & +1 & -1 & -1 \\ +1 & -1 & -1 & +1 \end{pmatrix}.$$

Rows are for strata; columns are for replicates. The first column having all +1's means that the first PSU from each stratum should be selected for replicate 1. The second column of $(+1, -1, +1, -1)$ means that the second replicate contains PSU 1 from stratum 1, PSU 2 from stratum 2, PSU 1 from stratum 3, and PSU 2 from stratum 4. If $H = 4$, the number of replicates needed for an orthogonal set is 4.

There is also a concept called *full orthogonal balance* for which the number of half-samples must be divisible by 4 but must be strictly greater than H. Full orthogonal balance results in the average of the replicate estimates equaling the full-sample estimate for linear estimators (but not for nonlinear estimators). The R `survey` and WesVar packages both calculate orthogonal sets of BRR base weights and neither calculate the full-orthogonally balanced sets. As with jackknife weights, the final set of analytic BRR weights are calculated from the base weights after applying the adjustments used to generate the full-sample weight. To date, other software packages rely on the analyst to provide the final BRR weights.

Like the jackknife, deleting a PSU means that the entire sample within the PSU is dropped. The base weights for the units in the PSUs that are retained are multiplied by 2. Thus, the weights for units in replicate α are

$$d_{k(\alpha)} = \begin{cases} 0 & \text{if unit } k \text{ is in a PSU that is not in the half-sample,} \\ 2d_k & \text{if unit } k \text{ is in a PSU that is in the half-sample.} \end{cases}$$

The adjusted weights are then used to compute a replicate estimate denoted by $\hat{\theta}_\alpha$. If the full-sample estimate has the form $\hat{\theta} = f\left(\hat{t}_1, \ldots, \hat{t}_p\right)$, then a

half-sample estimate is $\hat{\theta}_\alpha = f\left(\hat{t}_{1\alpha}, \ldots, \hat{t}_{p\alpha}\right)$ where $\hat{t}_{j\alpha}$ is the estimated total for the j-th variable based on the units in half-sample α. The BRR variance estimator is then

$$v_{BRR}\left(\hat{\theta}\right) = A^{-1} \sum_{\alpha=1}^{A} \left(\hat{\theta}_\alpha - \hat{\theta}\right)^2. \tag{15.13}$$

The variance estimator can be centered around quantities other than the full-sample estimate, but these will all be similar when the number of strata is large. In fact, the R `survey` package uses the mean of the replicate estimates in Eq. (15.13) rather than $\hat{\theta}$. WesVar, on the other hand, uses the full-sample estimate as in Eq. (15.13).

Fay BRR

One potential problem with the standard BRR is that one-half of the sample is eliminated to form a replicate. This may lead to instability for domain estimates. If a domain occurs in only a subset of the PSUs, all sample units in the domain could be dropped in a particular replicate. Although this will not bias the variance estimator, it will make the variance estimator itself unstable, i.e., the variance of the variance estimator may be unnecessarily high.

A modification of BRR due to Robert Fay (Fay 1984; Dippo et al. 1984; Judkins 1990) addresses this problem. Rather than dropping a PSU entirely, the Fay BRR simply down-weights it. Half-samples are identified using a Hadamard matrix as above. The weights are then calculated as

$$d_{k(\alpha)} = \begin{cases} \rho d_k & \text{if unit } k \text{ is in a PSU that is not in the half-sample,} \\ (2 - \rho)\, d_k & \text{if unit } k \text{ is in a PSU that is in the half-sample,} \end{cases} \tag{15.14}$$

where $0 \leq \rho < 1$. If $\rho = 0$, this is the standard BRR. If, for example, $\rho = 0.5$, the weights for PSUs in a half-sample are multiplied by 1.5; the weights of units in the other PSUs are multiplied by 0.5. Another choice, which Judkins (1990) found to perform well, is $\rho = 0.3$.

The Fay BRR solves the small domain problem because no PSU is completely dropped from the sample. Even if the domain has few sample cases, it will always be in each replicate.

Assumptions, Advantages, and Limitations

The BRR variance estimator is approximately unbiased and consistent for the variance of nonlinear estimates, as is the jackknife. The assumptions are similar to those for the linearization and jackknife estimators. For large sample theory the requirement is that the number of strata H be large since each stratum must have $n_h = 2$ PSUs.

   An important characteristic of BRR and the Fay BRR is that both provide
legitimate estimates of the variance of a quantile, unlike the jackknife. Rao
and Wu (1985) and Rao and Shao (1999) provide the theoretical support.
A key property that BRR does share with the jackknife is that it can be used
to reflect multiple stages of weight adjustment—like nonresponse adjustment
and calibration. As long as the database constructor redoes the weighting
steps separately for each replicate, BRR will give correct variance estimators.

Example 15.12 (BRR variance estimation). The `nhis.large` design is 2
PSUs per stratum. Thus, BRR is appropriate. In R, BRR and Fay-BRR
design objects can be created from the `nhis.dsgn` object used in the previ-
ous examples. The code below creates an object for standard BRR by calling
`as.svrepdesign` with `type="BRR"` and another object for Fay BRR using
`type="Fay"`. The ρ parameter for Fay BRR is defined by `fay.rho = 0.3`.

```
brr.dsgn <- as.svrepdesign(design = nhis.dsgn, type = "BRR")
faybrr.dsgn <- as.svrepdesign(design = nhis.dsgn,
                 type = "Fay", fay.rho = 0.3)
```

The weight adjustments can be examined with
   `weights(brr.dsgn)` or `weights(fay.brr.dsgn)`.
In this case the dimension of the weight adjustment matrix is 21588×80
with 80 being the size of the Hadamard matrix that was used. Although the
NHIS design has 75 strata and the smallest multiple of 4 greater than or
equal to 75 is 76, a Hadamard matrix of dimension 80 is the one that R has
available. ∎

Example 15.13 (Quantile variance with BRR). BRR and Fay BRR can be
used to estimate the SEs of quantiles. The code below uses `smho.N874` and
stratifies the population by a measure of size based on `BEDS`. `BEDS` is recoded
to remove the zero values. Strata are then formed and an *stsrswor* selected.
The method of stratification is the one described in Sect. 3.2.1 where the
population is sorted by size and strata formed to each have about the same
total measure of size. The `cut` function is useful for this. Selecting an *stsrswor*
is then very similar to *pps* sampling. In this example, we formed 25 strata
and selected 2 sample hospitals per stratum for a total of 50.

```
x <- smho.N874$BEDS
x[x <= 10] <- 10
x <- sqrt(x)
smho.N874 <- smho.N874[order(x), ]
x <- sort(x)
N <- nrow(smho.N874)
n <- 50
H <- 25

cumx <- cumsum(x)
size <- cumx[N]/H
```

```
brks <- (0:H)*size
strat <- cut(cumx, breaks = brks, labels = 1:H)
pop <- data.frame(smho.N874, strat = strat)
set.seed(428274453)
sam <- strata(data = pop,
                stratanames = "strat",
                size = rep(2,H), method=c("srswor"))
sam.dat <- pop[sam$ID\_unit,]
d <- 1/sam$Prob
smho.dsgn <- svydesign(ids = ~0,
                        strata = ~strat,
                        data = sam.dat,
                        fpc = sam$Prob,
                        weights = ~d)
smho.BRR.dsgn <- as.svrepdesign(design = smho.dsgn,
                    type = "BRR")
smho.FayBRR.dsgn <- as.svrepdesign(design = smho.dsgn,
                    type = "Fay",
                    fay.rho = 0.3)

svyquantile(~EXPTOTAL, design = smho.BRR.dsgn, quantile=0.5,
        interval.type="quantile")
svyquantile(~EXPTOTAL, design = smho.FayBRR.dsgn, quantile=0.5,
        interval.type="quantile")
```

Two versions of BRR were used: standard BRR and Fay BRR with $\rho = 0.3$.
One thing to note is that replication objects in R will not use *fpc*'s. Although
stratum-level *fpc*'s are in the smho.dsgn object above, they are stripped away
when the BRR design objects are created. The survey package will warn
you that this is happening. The results of svyquantile are that the median
is 6,966,393 with estimated SEs of 1,015,020 with BRR and 1,009,630 with
Fay BRR. The estimated CVs with the two methods are 14.6 % and 14.5 %
—very close to each other. ∎

15.4.3 Bootstrap

The bootstrap, invented by Efron (1982), has become extremely popular in
non-survey statistics because it is easy to compute and seems to be good
for any and everything. The general idea is again to select subsamples from
the full sample, do this many times, and to summarize the properties of a
statistic across the subsamples. There are several variations that have been
proposed for the bootstrap for finite population estimation. One due to Rao
and Wu (1988) applies to a stratified, multistage design with $n = \sum_{h=1}^{H} n_h$
sampled PSUs and uses the following steps. There are some variations on the
bootstrap for finite population estimation (e.g., Saigo et al. 2001; Shao and
Sitter 1996; Sitter 1992), but only the Rao-Wu version is currently available
in any software package.

(1) In each stratum, draw an *srswr* of m_h PSUs from the n_h initial sample PSUs. Let m_{hi}^* denote the number of times that PSU i is selected from stratum h so that $\sum_{i=1}^{n_h} m_{hi}^* = m_h$. Note that $m_{hi}^* = 0$ for PSUs not selected for the bootstrap sample. Create a replicate weight for each sample unit k within the initial sample PSUs ($k \in s_{hi}$) as

$$d_k^* = d_k \left(\left\{ 1 - \sqrt{\frac{m_h}{(n_h-1)}} \right\} + \sqrt{\frac{m_h}{(n_h-1)}} \frac{n_h}{m_h} m_{hi}^* \right),$$
$$= d_k B_{hi}$$

where B_{hi} is defined by the last equality. This is computed for units in *all* sample PSUs, not just those in the bootstrap sample. Provided that $m_h \le (n_h - 1)$, all such weights will be nonnegative, but not otherwise.

(2) Calculate $\hat{\theta}$, the desired estimate, using weights d_k^* in place of d_k.

(3) Repeat this process $B > 1$ time. Denote the corresponding bootstrap sample estimates as $\hat{\theta}_{(1)}, \hat{\theta}_{(2)}, \ldots \hat{\theta}_{(B)}$.

We will refer to the process in steps 1–3 as the Rao-Wu bootstrap. The bootstrap variance estimator is

$$v_{boot}\left(\hat{\theta}\right) = \frac{1}{B} \sum_{b=1}^{B} \left(\hat{\theta}_{(b)} - \hat{\theta}\right)^2.$$

We can choose m_h to be any values greater than or equal to 1. The simplest choice is $m_h = n_h - 1$, in which case

$$d_k^* = d_k \frac{n_h}{(n_h - 1)} m_{hi}^*.$$

Hence, units not included in a given bootstrap replicate get weight 0, those included exactly once get weight

$$d_k \frac{n_h}{n_h - 1},$$

those in twice get

$$d_k \frac{2n_h}{(n_h - 1)},$$

and so on. If $m_h \ne n_h - 1$, then units not included in the bootstrap sample get nonzero weights, as in the Fay BRR. If $m_h = n_h$, then a bootstrap weight can even be negative since $1 - \sqrt{n_h/(n_h - 1)} < 0$.

The two-PSUs-per-stratum case is worth examining since we saw above that BRR gave us an especially efficient way of forming replicates. If $m_h = n_h - 1$ and $n_h = 2$, then $m_h = 1$. Thus, the bootstrap is like BRR in this case, but without the control over the number of replicates. For just estimating the variance of a statistic, BRR is a more economical choice in the 2-per-stratum case. But, the bootstrap has advantages even in that case, particularly for constructing confidence intervals, as we describe next.

Assumptions, Advantages, and Limitations

The Rao-Wu bootstrap provides a consistent and approximately unbiased estimator of the variance of nonlinear statistics and for the variance of a quantile. The assumptions to derive theory for the bootstrap are the same as for the jackknife and BRR. In particular, for multistage samples, the PSUs are assumed to be selected with replacement. Rao and Wu (1988) do give some specialized ways of constructing the bootstrap estimates that will account for some types of designs using without-replacement sampling. However, these are not available in the software that we cover.

   A major selling point of the bootstrap is that it can be used to approximate the full distribution of a statistic, $\hat{\theta}$, not just its variance. By drawing many bootstrap samples and computing an estimate from each, an empirical distribution of $\hat{\theta}$ can be formed. A confidence interval for θ can be constructed in one of two ways:

(i) **Bootstrap percentile method**. Order the bootstrap estimates from lowest to highest. The lower $100\,(\alpha/2)\,\%$ confidence limit for θ is the $100\,(\alpha/2)\,\%$ point of the empirical distribution of the bootstrap estimates. The upper $100\,(1-\alpha/2)\,\%$ point of the empirical distribution is the upper confidence limit.

(ii) **Studentized bootstrap method**. In each bootstrap sample compute $t_{(b)} = \left(\hat{\theta}_{(b)} - \hat{\theta}\right)\big/\sqrt{v_{(b)}}$ where $v_{(b)}$ is an estimate of the variance of $\hat{\theta}_{(b)}$ based on the b-th bootstrap sample *only*. The value for $v_{(b)}$ could be generated from any consistent estimator appropriate for the design, e.g., linearization, jackknife, or BRR. After determining the $100\,(\alpha/2)\,\%$ and $100\,(1-\alpha/2)\,\%$ points, as well as $t^*_{\alpha/2}$ and $t^*_{1-\alpha/2}$, of the distribution of $t_{(b)}$, the CI is then computed as

$$\left(\hat{\theta} - t^*_{1-\alpha/2}\sqrt{v_{boot}\left(\hat{\theta}\right)},\ \hat{\theta} - t^*_{\alpha/2}\sqrt{v_{boot}\left(\hat{\theta}\right)}\right).$$

   If the distribution of $\hat{\theta}_{(b)}$ is fairly symmetric, then choice (a) performs well in the sense of giving coverage probability near $100\,(1-\alpha)\,\%$. If the distribution is extremely asymmetric or the sample size is small, (a) may not work as well. In non-survey statistics, choice (b) appears to be the best in theory (Efron and Tibshirani 1998). Rao and Wu (1988) proved that the second method does give correct coverage rates in large samples when using the Rao-Wu bootstrap weights. Method (b) is more computationally demanding, especially if a replication variance estimator like the jackknife is used in each bootstrap sample. In addition, the studentized bootstrap method is not available in R survey. Special programming will be needed.

Example 15.14 (Bootstrap). This example uses the same type of sample from `smho.N874` as in Example 15.13. The population is stratified by a measure of size based on beds and a 2-per-stratum sample of size 50 is selected

(i.e., $H=25$ and $n_h=2$). For comparison, we also select an unstratified simple random sample of 50 (i.e., $H=1$ and $n_h=n=50$). We estimate the total of end-of-year count of patients (EOYCNT) and get 95 % CIs using the bootstrap percentile method and the t-approximation for comparison. Part of the code is shown below; the file Example 15.14 bootstrap.R contains all of the R code. After the sample is selected, the design object, smho.dsgn, is created and, in turn, used to create an object, smho.boot.a, that holds Rao-Wu bootstrap weights from the stratified sample. The bootstrap object for the simple random sample is smho.boot.b. To create the bootstrap object, as.svrepdesign is called with the parameter, type="subbootstrap". Five hundred replicates are used. Although smho.dsgn contains an *fpc*, this is not retained when the bootstrap design object is created. R survey will warn you that this is happening.

```
        # stsrswor from strata based on a measure of size
        # create design with boostrap wts.
        # Rao-Wu version used with mh = nh-1
 smho.boot.a <- as.svrepdesign(design = smho.dsgn,
                                type = "subbootstrap",
                                replicates = 500)
        # mean & CI for EOYCNT based on RW bootstrap
 a1 <- svytotal(~EOYCNT, design = smho.boot.a,
                na.rm=TRUE,
                return.replicates = TRUE)
            # Compute CI based on bootstrap percentile method.
 ta1 <- quantile(a1$replicates, c(0.025, 0.975))

         # t approximation with v.boot
 La <- a1$mean + qt(0.025,df=degf(smho.boot.a)*sd(a1$replicates)
 Ua <- a1$mean + qt(0.975,df=degf(smho.boot.a)*sd(a1$replicates)
 c(La[1], Ua[1])

         # srswor of same size as above
 sam <- sample(1:N, n)
 sam.dat <- pop[sam,]
 d <- rep(N/n,n)
 smho.dsgn <- svydesign(ids = ~0,
                          data = sam.dat,
                          weights = ~d)

 smho.boot.b <- as.svrepdesign(design = smho.dsgn,
                                type = "subbootstrap",
                                replicates = 500)
 b1 <- svytotal(~EOYCNT, design = smho.boot.b,
          na.rm=TRUE,
          return.replicates = TRUE)
             # Compute CI based on bootstrap percentile method.
 tb1 <- quantile(b1$replicates, c(0.025, 0.975))
          # t approximation with v.boot
 Lb <- b1$mean + qt(0.025,df=degf(smho.boot.b)*sd(b1$replicates)
 Ub <- b1$mean + qt(0.975,df=degf(smho.boot.b)*sd(b1$replicates)
 c(Lb[1], Ub[1])
```

Calling `svytotal` above with `return.replicates=TRUE` saves the replicate estimates. The code to summarize the results and plot histograms of the replicates estimates follows.

```
        # pop total
sum(pop$EOYCNT)
        # totals \& SEs
rbind(c(a1$mean, SE=SE(a1)),
      c(b1$mean, SE=SE(b1)))

        # CIs
rbind("stsrswor boot" = ta1,
      "stsrswor t CI" = c(La[1], Ua[1]),
      "srswor boot" = tb1,
      "srswor t CI" = c(Lb[1], Ub[1]))

par(mfrow = c(2,1),
    mar = c(3,3,1,1))
r <- range(a1$replicates/10^3, b1$replicates/10^3)
truehist(a1$replicates/10^3, nbins=25,
         xlim = r, col = "gray85")
abline(v = a1$mean/10^3, col="gray50")
title(paste("stsrswor, n =",n), cex.main = 1)
truehist(b1$replicates/10^3, nbins=25,
         xlim = r, col = "gray85")
title(paste("srswor, n =",n), cex.main = 1)
abline(v = b1$mean/10^3, col="gray50")
```

The population total of `EOYCNT` is 727,723. The estimated totals and SEs from the two samples are:

	Estimated total	SE
stsrswor	528,635	122,674
srswor	732,867	221,723

The stratified sample is clearly much more efficient, in the sense of having a smaller SE, but its estimated total is farther from the truth. The 95 % confidence intervals (in thousands) are:

	Lower Bound	Upper Bound
stsrswor bootstrap percentile CI	280	766
stsrswor t CI	276	781
srswor bootstrap percentile CI	388	1,227
srswor t CI	287	1,178

The *t*-intervals are computed with 25 *df* for the stratified sample (where $n - H$ equals 50–25) and 49 *df* for the simple random sample (for $n - 1$ is 50 – 1). All intervals do cover the population total in this sample, but the

CIs are not the same for the bootstrap and the t-intervals. The bootstrap percentile intervals are not symmetric around the point estimate of the total, while, of course, the t-intervals are.

Looking at the histograms of the bootstrap replicate estimates in Figure 15.1 makes it clearer why this is so. Neither of the histograms is symmetric with the *srswor* bootstrap distribution being noticeably skewed. In contrast, we could not get this information from other replication methods. For example, in the stratified sample, BRR would give us only 28 replicate estimates. In the simple random sample, the jackknife would give 50 replicate estimates. Neither 28 nor 50 replicate estimates are enough to draw much of a histogram. But, with 500 bootstrap estimates, we can get a good idea of the underlying distribution of the estimator of the total. ■

We did not compute studentized bootstrap CIs in the last example. This can be done, but the user must do some programming. As an example, suppose that the mean of `EOYCNT` in the `smho.N874` population is to be estimated. In the stratified design above, the bootstrap weights can be retrieved with `weights(smho.boot.a)`. For each set of replicate weights, form a design object that will use the desired method of variance estimation (linearization, jackknife, or BRR). Use the function `svyttest` to test that the mean of `EOYCNT` is 0. The t-statistic is

$$t^*_{(b)} = \hat{\theta}_{(b)} \Big/ \sqrt{v_{(b)}},$$

where $\hat{\theta}_{(b)}$ is the estimated mean from replicate b and $v(b)$ is its estimated variance. This t-statistic will be available in `svyttest$statistic`. Adjust the t-statistic to obtain $t_{(b)} = \left(\hat{\theta}_{(b)} - \hat{\theta}\right)\Big/\sqrt{v_{(b)}}$ by subtracting $\hat{\theta}\Big/\sqrt{v_{(b)}}$ where $\hat{\theta}$ is the full-sample estimated mean. A loop can then be used to compute and collect these adjusted t-statistics from all replicates. A similar approach was used in Example 15.10 where nonresponse adjustments were computed for JKn replicates. From the collection of t-statistics, locate the $100\,(\alpha/2)\,\%$ and $100\,(1-\alpha/2)\,\%$ points, $t^*_{\alpha/2}$ and $t^*_{1-\alpha/2}$, of the distribution of $t_{(b)}$. These would then be used to calculate the studentized bootstrap interval, $\left(\hat{\theta} - t^*_{1-\alpha/2}\sqrt{v_{boot}\left(\hat{\theta}\right)}, \ \hat{\theta} - t^*_{\alpha/2}\sqrt{v_{boot}\left(\hat{\theta}\right)}\right)$. The same algorithm could be used for more elaborate tabulations, like a 2-way table of means, proportions, or totals.

The bootstrap can also be used to estimate the SE of a quantile and to get a standard error for a quantile, as shown in the next example.

Example 15.15 (Bootstrap quantiles). Continuing Example 15.13 with the same sample, we create a bootstrap object with 500 replicates. The BRR and Fay BRR design objects, `smho.BRR` and `smho.FayBRR`, are also used to create 95 % CIs for the median of expenditures, `EXPTOTAL`. The full code is in `Example 15.15 bootstrap quantile.R`.
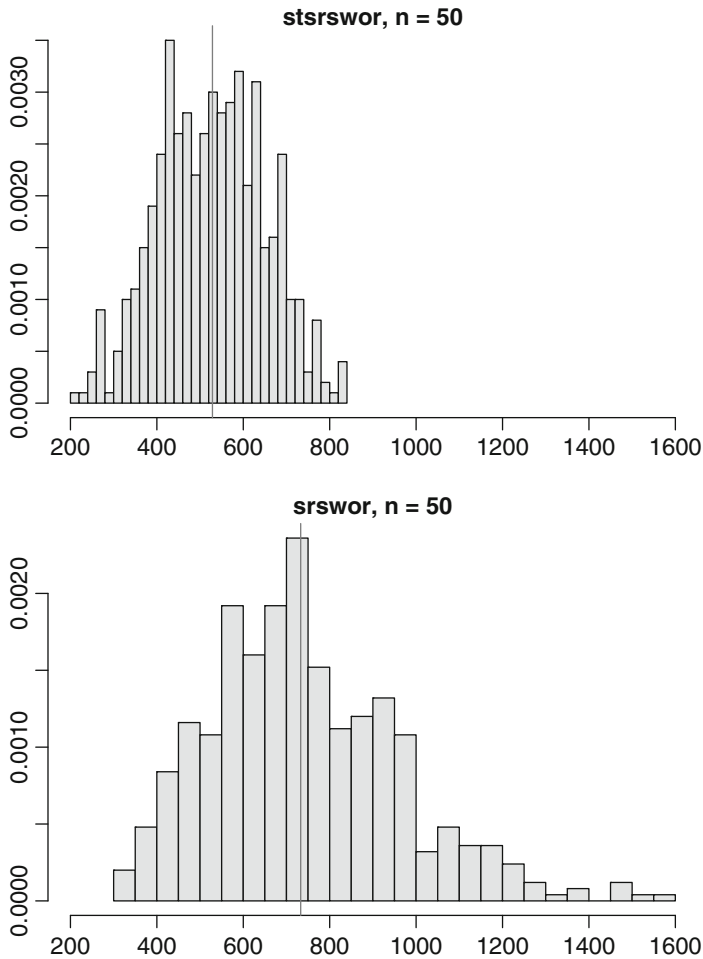
Fig. 15.1: Histograms of bootstrap estimates of total end-of-year count of patients in the SMHO population. Horizontal scales in thousands; a gray reference line is drawn at the full-sample estimate.

```
smho.boot <- as.svrepdesign(design = smho.dsgn,
                  type = "subbootstrap",
                  replicates = 500)

a1 <- svyquantile(~EXPTOTAL, design = smho.BRR, quantile=0.5,
        interval.type="quantile")
a2 <- svyquantile(~EXPTOTAL, design = smho.FayBRR, quantile=0.5,
        interval.type="quantile")
a3 <- svyquantile(~EXPTOTAL, design = smho.boot, quantile=0.5,
        interval.type="quantile",
        return.replicates = TRUE)

      # t approximation with BRR
```

```
La1 <- a1 + qt(0.025,df=degf(smho.BRR))*SE(a1)
Ua1 <- a1 + qt(0.975,df=degf(smho.BRR))*SE(a1)

        # t approximation with Fay.BRR
La2 <- a2 + qt(0.025,df=degf(smho.FayBRR))*SE(a2)
Ua2 <- a2 + qt(0.975,df=degf(smho.FayBRR))*SE(a2)

        # t approximation with v.boot
La3 <- a3 + qt(0.025,df=degf(smho.boot))*sd(a$replicates)
Ua3 <- a3 + qt(0.975,df=degf(smho.boot))*sd(a$replicates)
ta3 <- quantile(a3$replicates, c(0.025, 0.975))

rbind(c(La1[1], Ua1[1]), c(La2[1], Ua2[1]),
      c(La2[1], Ua2[1]), ta3)
```

The results for CIs on the median (in thousands) are:

	Lower	Upper
BRR	4,876	9,057
Fay BRR	4,887	9,046
Bootstrap t	4,723	9,209
Bootstrap percentile	4,750	8,272

The population median is 6,240 (also in thousands). The t-intervals for BRR and Fay BRR are almost identical while the bootstrap t-interval is wider. The bootstrap percentile interval is noticeably different. One reason for this is the irregular histogram of the bootstrap estimates shown in Fig. 15.2. A simulation study would be needed to decide whether this provides a better coverage rate than the symmetric intervals. ∎

15.5 Combining PSUs or Strata

There are two reasons to combine strata or PSUs: one is to reduce the number of replicates required when using the jackknife or BRR methods, and the other is to create pseudo- (or analytic) strata for variance estimation when one PSU has been selected per stratum or when only one PSU in a stratum participates. Both these cases are covered in this section.

15.5.1 Combining to Reduce the Number of Replicates

In some sample designs, the number of PSUs can be extremely large. This is especially true in education and establishment surveys where there can be thousands of first-stage units. Although replication variance estimators are attractive because of their ability to reflect multiple stages of weight adjustments, a strict application of, say, the jackknife can lead to far more replicates and far larger databases than most practitioners think are feasible.
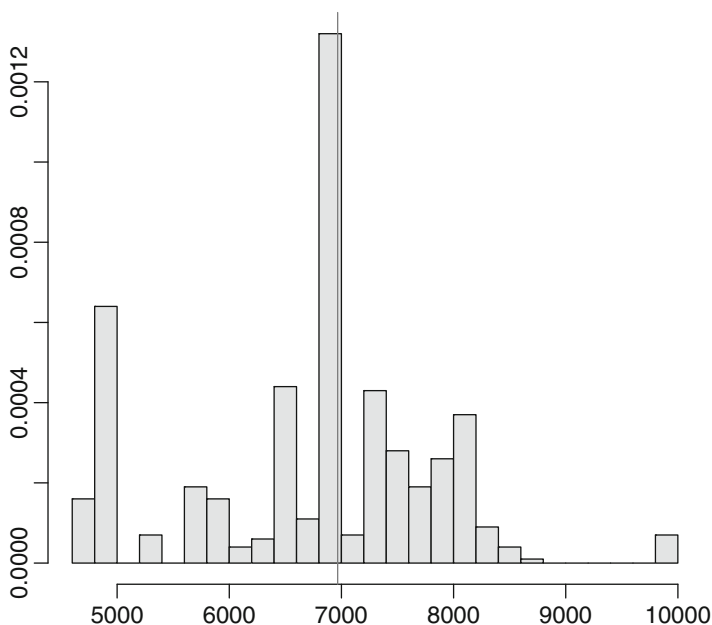
Fig. 15.2: Histogram of bootstrap estimates of median expenditure total in the SMHO population. Horizontal scales in thousands; a gray reference line is drawn at the full-sample estimate.

In such cases, PSUs or strata or both may be grouped together. Replication is then applied to the groups. Properly done, grouped replication estimators can still be consistent and approximately unbiased.

Appendix D of the WesVar manual (Westat 2007) describes how the groupings can be legitimately done for several types of sample designs. Rust (1984, 1985) also discusses some options. We summarize a few of the considerations here. Table 15.1 shows a simple case to illustrate the possibilities. There are 3 design strata and a total of 14 PSUs. If the JKn variance estimate were used, 14 replicates would be required. The third and fourth columns show the combinations of strata (labeled VarStrat) and of PSUs (labeled VarUnit) that could also be used for JKn variance estimation. The terms VarStrat and VarUnit are the ones used in WesVar and are apposite in conveying their use as the combined strata and PSUs used for variance estimation.

The three design strata are combined to form two VarStrat—design strata 1 and 2 are combined as VarStrat 1, and design stratum 3 is left alone as VarStrat 2. Design strata 1 and 2 both contain 4 PSUs. These are grouped into 2 VarUnits in each design stratum. PSUs 1 and 2 from each of design strata 1 and 2 are grouped into VarUnit 1. PSUs 3 and 4 from each of design strata 1 and 2 are grouped into VarUnit 2. If a grouped version of JKn is used, two replicates would be formed from VarStrat 1: one by dropping VarUnit 1

Table 15.1: Example of grouping strata and PSUs for variance estimation.

Design stratum	PSU	VarStrat	VarUnits for JKn	VarUnits for BRR
1	1	1	1	1
	2		1	1
	3		2	2
	4		2	2
2	1	1	1	1
	2		1	1
	3		2	2
	4		2	2
3	1	2	1	1
	2		1	1
	3		2	1
	4		2	2
	5		3	2
	6		3	2
Total	14		5	4

(PSUs 1 and 2 from each of design strata 1 and 2) and another by dropping VarUnit 2 (PSUs 3 and 4 from design strata 1 and 2). In VarStrat 2 (design stratum 3), three VarUnits are formed. PSUs 1 and 2 are VarUnit 1; PSUs 3 and 4 are VarUnit 2; and PSUs 5 and 6 are VarUnit 3. A total of 5 VarUnits is formed, which compares to the 14 original PSUs.

When creating VarUnits within a VarStrat, the size of the variance estimate can be affected by the groupings. For example, in VarStrat 2, if we sorted the six PSUs based on their size of weighted total of a y variable and then assigned the three VarUnits as shown in Table 15.1, this would tend to produce a large estimated variance of a total, at least for that particular y. If we randomly ordered the PSUs and assigned the VarUnits as (1, 1, 2, 2, 3, 3), as shown in the table, this would give a better reflection of the variance of an estimated total. Random ordering within a design stratum prior to numbering is the recommended procedure. An exception to this would be a case in which PSUs were implicitly substratified within a design stratum, say, by sorting the frame and using systematic selection. In that case, a better procedure would be treat the substrata as explicit design strata and then decide how to group strata or PSUs.

The grouped jackknife variance estimator is computed using a formula that is parallel to Eq. (15.11):

$$v_{GJ}\left(\hat{\theta}\right) = \sum_{\tilde{h}=1}^{\tilde{H}} \frac{G_{\tilde{h}} - 1}{G_{\tilde{h}}} \sum_{g=1}^{G_{\tilde{h}}} \left[\hat{\theta}_{(\tilde{h}g)} - \hat{\theta}\right]^2, \qquad (15.15)$$

where $G_{\tilde{h}}$ is the number of VarUnits in VarStrat $\tilde{h}$, $\tilde{H}$ is the total number of VarStrat, and $\hat{\theta}_{(\tilde{h}g)}$ is the estimate computed after dropping VarUnit g

in VarStrat $\tilde{h}$. To compute $\hat{\theta}_{(\tilde{h}g)}$, the weight for each unit retained in $\tilde{h}g$ is multiplied by $G_{\tilde{h}} / (G_{\tilde{h}} - 1)$, the inverse of the subsampling fraction. That is, the weights are increased to reflect the fact that there is one less group being used to make the replicate estimate when group $\tilde{h}g$ is dropped. In the example above $\tilde{H}$=2, G_1=2, and G_2=3. The total number of replicates created is $G = \sum_{\tilde{h}=1}^{\tilde{H}} G_{\tilde{h}}$. In the example, we have G=5. Without the grouping of strata and PSUs, G=14 replicates would be formed.

To declare the VarStrat and VarUnits for use with a software package, you would simply say that the stratum variable was the field that holds the VarStrat codes; the PSU variable would be the field for the VarUnit codes. The value of $(G_{\tilde{h}} - 1) / G_{\tilde{h}}$ is specified in the `rscales` parameter to R survey. Other software packages will count the values of $G_{\tilde{h}}$.

There are many references that explore the properties of the grouped jack-knife variance estimator (e.g., Kott 1999; Kott 2001; Lu et al. 2006 and (Wolter 2007, chap. 4). Although there is no unique way to create the groups, it is possible to create a biased variance estimator by doing the grouping badly. As a general rule each replicate estimate $\hat{\theta}_{(\tilde{h}g)}$ must be a legitimate estimate for the full population. In Table 15.1, if we had numbered all the VarUnits from design stratum 1 as "1" and all the VarUnits from design stratum 2 as "2," dropping VarUnit 1 in VarStrat 1 would result in dropping the entire sample from design stratum 1. Likewise, dropping VarUnit 2 would drop all of design stratum 2. As a result, neither $\hat{\theta}_{(11)}$ nor $\hat{\theta}_{(12)}$ would be estimates for the full population and $v_{GJ}\left(\hat{\theta}\right)$ would be biased.

Another type of problem is created if the number of PSUs in each group within a VarStrat is not the same. In Table 15.1, for example, if in VarStrata 2, we put 4 PSUs in VarUnit 1 and 2 in VarUnit 2, the jackknife variance estimator in Eq. (15.15) would again be biased. The reason is that the weight adjustment, $G_{\tilde{h}} / (G_{\tilde{h}} - 1)$, is too crude for each of the replicate estimates to be unbiased. If the number of PSUs per group varies, a more nearly unbiased choice is

$$v_{GJ2}\left(\hat{\theta}\right) = \sum_{\tilde{h}} \sum_{g=1}^{G_{\tilde{h}}} \frac{\left(n_{\tilde{h}} - n_{\tilde{h}g}\right)}{n_{\tilde{h}}} \left[\hat{\theta}_{(\tilde{h}g)} - \hat{\theta}\right]^2, \qquad (15.16)$$

where $n_{\tilde{h}}$ is the total number of PSUs in VarStrat $\tilde{h}$ and $n_{\tilde{h}g}$ is the number of PSUs in group $\tilde{h}g$. The weight adjustment applied to the retained cases when VarUnit $\tilde{h}g$ is deleted is $n_{\tilde{h}} / \left(n_{\tilde{h}} - n_{\tilde{h}g}\right)$. In other words, the weight adjustment depends on how many original PSUs are dropped. As illustrated in Valliant et al. (2008), even small differences in the numbers of PSUs per group can produce large biases if the JKn formula in Eq. (15.15) is used along with the $G_{\tilde{h}} / (G_{\tilde{h}} - 1)$ weight adjustments. This is more likely to be an issue in single-stage surveys where there are a large number of units in some strata, and the number does not divide evenly into the desired number of groups. Formula (15.16) is not available in standard software and must

be programmed by the user. Because of that, creating VarUnits within a VarStrat that all have the same number of original sample PSUs is the best, practical solution.

Grouping can also be used for BRR, but two VarUnits must be created in each VarStrat. The last column in Table 15.1 shows one way of doing this in our little example. The only change from the grouping used for JKn is in VarStrat 2 where PSUs 1–3 are grouped into VarUnit 1 and PSUs 4–6 into VarUnit 2. The weights for each unit retained for a half-sample would be multiplied by 2 for the standard BRR or by either ρ or $2 - \rho$ for Fay BRR as in Eq. (15.14). As for JKn, the variance estimator is biased if each VarUnit does not contain the same number of PSUs, and a weight adjustment of 2 for the standard BRR or 2 and $2 - \rho$ for the Fay BRR are used. Before assigning the PSUs in VarStrat 2 to VarUnits, they should be randomly ordered and then coded as (1, 1, 1, 2, 2, 2). This will avoid biasing the variance estimator by sorting on a characteristic that is related to the analysis variable.

15.5.2 How Many Groups and Which Strata and PSUs to Combine

If grouping is legitimate, the natural questions are how many groups to form and which strata and PSUs should we combine? The number of groups is related to the degrees of freedom (*df*) of the variance estimator. The more *df* a variance estimator has, the more stable the variance estimator tends to be. Consequently, the basic goal is to have a large number of *df*. Ideally, this would be done for full population estimates and for domain estimates. Domains that occur in all strata and PSUs will not need special consideration—they behave about the same as the full population. In a household survey where PSUs are geographically stratified, domains defined by gender (male, female) will be spread across all strata. For domains that occur in only a subset of the strata, achieving efficient creation of groups can be complicated. Regions of a country would be examples of domains that occur in only some of the strata. When region estimates are important, it may be possible to group strata in such a way that the estimate for each region retains almost the same number of *df* as for an ungrouped variance estimator, even though the grouping does lose *df* for full population estimates. We will illustrate this below with a simple example.

The rule of thumb that is often used for full population estimates is that *df* equals the number of sample PSUs minus the number of strata. That is, each stratum contributes the number of sample PSUs minus 1 to the overall *df*. When grouping of strata and/or PSUs is used, then the rule of thumb is applied to the number of VarStrat and VarUnits.

Table 15.2: Approximate coefficients of variation of variance estimators and standard error estimators based on different numbers of degrees of freedom.

df	CV of variance estimator (%)	CV of standard error estimator (%)
10	45	22
25	28	14
50	20	10
75	16	8
100	14	7
200	10	5
400	7	4

Suppose that v is a variance estimator and V is the theoretical variance of some estimator. If we treat $df * v/V$ as having a chi-square distribution with df degrees of freedom, then the CVs of v and of $\sqrt{v}$ can be approximated as shown in Table 15.2. In particular, $CV(v) = 2/\sqrt{df}$ and $CV(\sqrt{v}) \doteq CV(v)/2$. If we wanted $CV(v)$ to be 10 %, we need 200 df. If the criterion is $CV(\sqrt{v}) = 0.10$, then we need $df = 50$. If we use grouping, the rule of thumb is that $df = G - \tilde{H}$. For a domain estimate, the rule of thumb is to compute $df = G - \tilde{H}$ but only over the VarStrat in which the domain occurs. If we use the BRR method, we should create at least 50 VarStrat to have $CV(\sqrt{v}) = 0.10$. Considering that some domains may occur in only a subset of design strata, having $G - \tilde{H}$ equal to at least 100 seems prudent.

To answer the question of which design strata to combine, the possibility of making domain estimates should be considered. For example, suppose that a design has $H = 10$ strata and 2 sample PSUs per stratum as shown in Table 15.3. Strata 1–5 are in region 1 while strata 6–10 are in region 2. The rule of thumb says that there are 10 df for the full design and 5 for each region. The full BRR method requires 12 replicates—the smallest multiple of 4 greater than or equal to the number of strata. If we want to use 8 replicates rather than 12, $\tilde{H} = 8$ VarStrat can be created, each of which has 2 VarUnits. Table 15.3 lists two ways of creating the 8 VarStrat. In set 1, design strata 1 and 2 are combined as design strata 6 and 7. The degrees of freedom based on the rule of thumb are 8 for full-sample estimates and 4 for estimates for both regions 1 and 2. In set 2, design strata 1 and 6 are combined while 2 and 7 are combined. The number of VarStrat assigned to each of regions 1 and 2 is 5. Thus, set 2 has the same df for each region as does the original sample design. By judicious creation of groups, we reduce the full-sample df from 12 to 8 but retain the same df for regions as in the full sample.

On the other hand, other domains may occur in all strata. Degrees of freedom would be lost for their variance estimates. The reduction would be from 12 to 8, as is the case for full population estimates.

Table 15.3: Two options for combining design strata to reduce number of BRR replicates.

Design stratum	Region of design stratum	VarStrat	Set 1		Set 2	
			Design strata	Region in VarStrat	Design strata	Region in VarStrat
1	1	1	1, 2	1	1, 6	1,2
2	1	2	3	1	2, 7	1,2
3	1	3	4	1	3	1
4	1	4	5	1	4	1
5	1	5	6, 7	2	5	1
6	2	6	8	2	8	2
7	2	7	9	2	9	2
8	2	8	10	2	10	2
9	2					
10	2					

15.5.3 Combining Strata in One-PSU-per-Stratum Designs

As discussed in Chap. 10, area probability samples often are stratified to such a degree that only one PSU is selected from each non-self-representing stratum. This deep stratification allows more control over the geographic dispersion of PSUs than does selecting two per stratum or some larger number. The trouble with this method is that neither a design-unbiased nor consistent estimator of the variance is available, even for linear estimators. This is a long-standing problem in survey sampling and is studied in Hansen et al. (1953a, Sect. 9.15) and Wolter (2007, Sect. 2.5).

The usual procedure is to combine strata into pairs for variance estimation. Alternative terminology, used by Wolter, is to "collapse" strata. After pairing strata, BRR, Fay BRR, or the jackknife can be used. The resulting variance estimator will generally be an overestimate. As HHM emphasize, strata should be combined based on stratum-level characteristics—not those of the selected PSUs. For example, if population size and degree of urbanization were used to form strata, then two strata of urban and similar size PSUs could be combined. For a systematic sample, the frame is generally sorted by characteristics within a design stratum. These characteristics are sometimes referred to as implicit stratification in comparions with the explicit design strata. Thus, sample units are selected in a prespecified order defined by the implicit strata. This order should be maintained when forming the PSU pairs. A useful thought process is to consider which strata would have been put together if the plan had been to select two PSUs per stratum.

Combining strata based on sample characteristics could lead to negatively biased variance estimates, at least for some point estimates. To take an extreme case, suppose we want to estimate the total expenditures in the SMHO hospital population from a 1-hospital-per-stratum sample. Hospitals

are stratified by number of beds and one selected at random from each stratum. If we collect data and pair strata whose sample PSUs have expenditures that are near each other, this procedure, applied repeatedly to different samples, would give variance estimates for total expenditures that are artificially low. By peeking at the data to do the pairing, we depress the value of the estimated within-group variance contribution. The strata that are paired could vary from one sample to another based on what data are observed. On the other hand, if we pair adjacent strata because of similarity of the number of beds per hospital, these pairs would be set once and not vary depending on how the samples came out.

In a simple case, Wolter (2007, Sect. 2.5) shows that bias of the collapsed-stratum variance estimator (with collapsing set in advance of viewing the results) is positive and depends on

$$\sum_{\tilde{h}=1}^{\tilde{H}} \left(\mu_{\tilde{h}1} - \mu_{\tilde{h}2} \right)^2,$$

where $\tilde{h}$ is a collapsed stratum and $\mu_{\tilde{h}g}$ is the population total of the analysis variable for stratum g in collapsed stratum $\tilde{h}$. The pairing might still not give us full credit for the gains from stratification, but the possibility for negative bias would be removed.

One final note on collapsing strata and PSUs should be mentioned. The sample design is one source for generating a single-PSU-per-stratum situation. Another is linked to sample loss such as with nonresponse or ineligibility. For example, consider a design stratum containing 5 sample schools (PSUs) where students are selected from within the randomly sampled school. If two schools close prior to data collection (ineligible) and administrators from two other schools decline to participate because of funding and time restrictions, then data from only one sample school is available for analyses within this stratum. Variance strata and PSUs are then formed using the same criteria as discussed above.

15.6 Handling Certainty PSUs

There are two cases to consider when PSUs are selected with certainty (i.e., selection probability equals 1.0): (1) certainties in a single-stage sample and (2) certainty first-stage units in multistage surveys. In both cases, we need to determine how the certainties should be handled when using linearization variances or for creating replicates for replication variance estimates. As noted in earlier chapters, certainties are also called self-representing (SR) while other units are called non-self-representing (NSR). In a single-stage

sample, a certainty does not contribute to the repeated sampling variance. For linearization variance estimators, each certainty can be assigned its own stratum code and has a base weight equal to 1.0. In replication variance estimator, a certainty can be forced to be a member of every replicate. In both the full sample and a replicate, each certainty gets a weight of 1. Consequently, for linear estimators, the contribution from each certainty will subtract out when taking the difference in a replicate estimate and the full-sample estimate. For example, in JKn, for a linear estimator:

$$\hat{\theta}_{(hi)} = \text{(contribution from SRs)} + \text{(contribution from NSRs in the replicate)},$$
$$\hat{\theta} = \text{(contribution from SRs)} + \text{(contribution from NSRs in the full sample)}.$$

The contribution from the SRs subtracts out when computing $\hat{\theta}_{(hi)} - \hat{\theta}$.

In a multistage sample with SR PSUs, the SRs are really strata containing lower-level sample units. For example, suppose that an area probability sample is selected in the US and that Cook County, Illinois, which contains Chicago, is a certainty. In the Cook County stratum, the first-stage units might be block groups (BGs), as discussed in Chap. 10. The BGs are the PSUs in Cook County for the purposes of variance estimation.

A common approach in area samples is to use BRR for variances. The first-stage units within the SRs are often divided into only two VarUnits. There are two worries with this: (i) two VarUnits give 1 *df* in each SR which may be much less than the maximum number of *df* available, and (ii) if there are not an even number of first-stage units in an SR, the standard replicate weight adjustments and variance formulas may be biased, as discussed earlier for the jackknife. In the Chicago example, suppose that 20 BGs are selected. BRR can be used if 2 VarUnits of 10 BGs are created, resulting in 1 *df*. But, we can just as easily create 10 VarUnits of 2 BGs and get 10 *df*. Given the computational speed and storage capacity of modern computers, a miserly savings of a few replicates in the 2-VarUnit example is hardly worth it.

For the second worry mentioned above, the most prudent procedure is to select an even number of first-stage units within each SR. Sometimes that cannot be done due to workload restrictions for field personnel or supplementation of the sample to meet a target number of respondents (see Sect. 6.5.2 on the use of data collection replicates). If the final number of first-stage units (like BGs) in an SR is odd, a practical approach is to combine two BGs. This is an example of a VarUnit formed by combining first-stage units.

Example 15.16 (Handling certainties). Using the `smho98` population, which includes one extremely large hospital, as seen in Chap. 3, we select a single-stage sample of 80 from the 875 hospitals in the population. The sample is *pps*

with the measure of size based on number of beds. Any value of BEDS that
is less than 10 is recoded to be 10. With this plan, 9 units are certainties and
71 are non-certainty. The R code is in Example 15.16 certainties.R.
To notify the survey package that there are certainties, a variable called
stratum is created, having a value of 1 for non-certainties and 2 for cer-
tainties. The fpc=0 for all sample hospitals in stratum 1 and fpc=1 for
stratum 2.

```
pop <- smho98
        # recoded BEDS as MOS
set.seed(428274453)
n <- 80
N <- nrow(pop)
x <- pop$BEDS
x[x<10] <- 10
pik <- n*x/sum(x)

        # check for certainties & adjust selection probs of
        #       non-certainties
n.cert <- sum(pik >= 0.8)
certs <- (1:N)[pik >= 0.8]
x.nc <- x[-certs]
n.nc <- n - n.cert
pik <- n.nc*x.nc/sum(x.nc)
sam <- UPrandomsystematic(pik)
pop.nc <- pop[-certs,]

        # extract rows for non-certainties, then append rows
        #       for certainties
sam.dat <- pop.nc[sam==1,]
sam.dat <- rbind(sam.dat, pop[certs,])

        # append strata codes and fpc's
        # stratum = 1 for non-certs, 2 for certs
        # fpc = 0 for non-certs, 1 for certs
stratum <- c(rep(1,n.nc), rep(2,n.cert))
fpc <- c(rep(0,n.nc), rep(1,n.cert))
sam.dat <- cbind(sam.dat, stratum, fpc)
probs <- c(pik[sam==1], rep(1,n.cert))
d <- 1/probs

        # Create a design object with fpc's
smho.dsgn <- svydesign(ids = ~0,
                        strata = ~stratum,
                        fpc = ~fpc,
                        data = data.frame(sam.dat),
                        weights = ~d)

svytotal(~EXPTOTAL, design=smho.dsgn)
cv(svytotal(~EXPTOTAL, design=smho.dsgn))
svytotal(~SEENCNT, design=smho.dsgn)
cv(svytotal(~SEENCNT, design=smho.dsgn))
```

The design object, smho.dsgn, uses both the stratum and fpc variables. For comparison, a design object (not shown above but in the code file) was also created that did not include the *fpc*'s. The estimated totals of expenditures and patients seen, along with the SEs and *CV*s, are shown below. If the certainties are accounted for, the *CV*s are 8.6 % and 11.0 %. But, if the certainties are thrown in with the non-certainty selections for SE calculation, the *CV*s are 9.9 % and 16.5 %. Thus, ignoring the fact that there are certainties leads to a substantial overstatement of *CV*s and SEs.

Variable	Estimated total (millions)	With *fpc* SE	*CV*	No *fpc* SE	*CV*
EXPTOTAL	8711.50	748.78	8.6 %	864.17	9.9 %
SEENCNT	1.17	0.13	11.0 %	0.19	16.5 %

Accounting for the certainties can also be accomplished with the jackknife using the code below. First, a separate stratum code is assigned to each of the 9 certainties and stored in strat.rep. The statement

```
options(survey.lonely.psu="certainty")
```

results in single PSUs in a stratum being omitted from variance calculations (but not from estimates of means, totals, etc.).

```
strat.rep <- c(rep(1,n.nc), 2:(2 + (n.cert-1)))
options(survey.lonely.psu="certainty")
rep.dsgn <- svydesign(ids = ~0,
                      strata = ~strat.rep,
                      data = data.frame(sam.dat),
                      weights = ~d)

 jkn.dsgn <- as.svrepdesign(design = rep.dsgn, type = "JKn")
```

The estimated totals and SEs of EXPTOTAL and SEENCNT are the same as those above—a result that you can verify by running the code for this example. ∎

Exercises

15.1. Consider the situations described below. In each case, classify the
estimator as linear or nonlinear and explain your reasoning. Which of the
following methods of variance estimation would you use: exact formula, lin-
earization, or replication? Explain your choices. If more than one method
can be used, discuss the considerations that should be made when selecting
a particular variance estimator.

(a) Stratified simple random sample of business establishments selected with-
out replacement. The estimate is the ratio of the π-estimator of total
before-tax profits (across all establishments) to π-estimator of total rev-
enues (again across all establishments).

(b) Stratified simple random sample of business establishments selected with-
out replacement. Estimate is π-estimator of total expenditures on capital
improvements in 2001.

(c) Two-stage stratified sample design of households. At the first stage, a
sample of primary sampling units (PSUs) is selected with varying prob-
abilities without replacement. PSUs are geographic areas like counties
or groups of counties. The frame of PSUs is stratified by region of the
country. Four PSUs are selected from each stratum. At the second stage
an equal probability sample of households is selected within each PSU
selected in the first stage. The population quantity to be estimated is
the average household income for households whose head is classified as
Hispanic. The estimator is the π-estimator.

(d) A single-stage sample of schools is selected with probabilities proportional
to the square root of enrollment from a prior academic year. The frame is
sorted hierarchically based on the following variables: region of the coun-
try, location of the school (urban, suburban, rural), and the percentage
of students in the school who receive free or reduced price lunches. The
population quantity to be estimated is the proportion of students who
scored at or above a specified proficiency level on a standardized mathe-
matics test.

15.2. The following data were collected from a sample of two PSUs selected
from each of two strata.

h	PSU	Y_{hi}
1	1	5
1	2	6
2	1	10
2	2	4
Total		25

Y_{hi} is the weighted PSU total observed for PSU i in stratum h.

(a) Compute the balanced repeated replication (BRR) variance estimator for
the estimated total $\hat{y} = \sum_{h=1}^{2} \sum_{i=1}^{2} Y_{hi}$. Specify which form of the BRR

estimator you are using. Use the following orthogonal matrix where rows designate the strata and columns the replicates:

$$A = \begin{bmatrix} + & + & + & + \\ + & - & + & - \\ + & + & - & - \\ + & - & - & + \end{bmatrix}$$

(b) What is the variance formula for the estimated total $\hat{y}$ if PSUs are assumed to be selected with replacement? Evaluate this formula using the data in the table above. How does it compare with your answer in part (a)?

15.3. What are the "rules of thumb" values for the following combinations of sample design and variance estimators?

(a) 2 sample PSUs selected per stratum with replacement and with varying probabilities, balanced repeated replication (BRR) variance estimator.
(b) A design with H strata, n_h sample PSUs selected in stratum h, and the stratified delete-one jackknife estimator.
(c) 2 sample PSUs selected per stratum with replacement and with varying probabilities, Fay-BRR variance estimator with $\rho = 0.3$.
(d) A design with 100 strata and two sample PSUs selected per stratum with replacement. PSUs are randomly numbered 1 or 2 within each stratum. Strata are then combined into 25 superstrata with 8 PSUs per superstratum. The PSUs numbered 1 in a superstratum are treated as one group while the PSUs numbered 2 are treated as a second group. A BRR variance estimator is used treating the superstrata as variance estimation strata.

15.4. Suppose that $\hat{y}$ is an unbiased estimate of a finite population total, Y. You are interested in the estimator $g(\hat{y}) = \sqrt{\hat{y}}$.

(a) Write down the first-order Taylor series approximation to $g(\hat{y})$.
(b) Based on your answer to (a), what is the approximate design variance of $g(\hat{y})$? Write your answer in general terms that apply to any design.
(c) Specialize your answer in (b) to the following designs: simple random sampling without replacement, stratified simple random sampling without replacement, and a single-stage design where units are selected with varying probabilities with replacement.

15.5. Use the `nhis.large` data set which is a stratified, cluster design with 2 PSUs selected per stratum. Estimate the proportions of the population in each age group (`age.grp`) that had an overnight hospital stay (hosp.stay). Estimate the standard errors using linearization, BRR, Fay BRR with $\rho = 0.5$, and the JKn jackknife. How do the estimated SEs compare?

15.6. Use the nhis.large file as a population and select a simple random sample of size $n = 500$. If you are using R, use a random number seed of 428274453. Poststratify the sample to population counts for `age.grp`.

(a) Compute the estimated proportion of the population who reported a doctor visit (`doc.visit`) in the two weeks prior to the interview.
(b) Calculate the SEs using the linearization method and JKn. What would be the effect on estimated SEs of ignoring the poststratification?
(c) Estimate the proportions and SEs of the population who reported a doctor visit in a table defined by Hispanic ethnicity (`hisp`). Combine categories 3 and 4 of `hisp` together. What would be the effect of ignoring the poststratification for these estimates?

15.7. Use the sample from Example 15.8 from the `smho.N874` population. Estimate the quartiles (30th, 50th, and 75th) of the end-of-year count of patients (`EOYCNT`). Find the 95 % CIs and SEs estimated by the Woodruff and Francisco-Fuller methods. What are the SEs implied by these two methods? If you try to estimate the first quartile and its SE using the Francisco-Fuller method, an error will occur. What characteristics about the sample data do you think causes the error?

15.8. Repeat Exercise 15.6 on poststratification using the bootstrap method with 500 replicates. If you are using R, use a random number seed of -711384152. How do your estimates of standard errors and CVs compare to the linearization and jackknife estimates in exercise 15.6?

15.9. Repeat Example 15.8 using the Rao-Wu bootstrap method with 500 replicates. Delete type 4 hospitals and recode the variable, beds, to have a minimum value of 5. Estimate the 25th, 50th, and 75th quantiles of SEENCNT in smho.N874 and the 95% confidence intervals for each using the t-approximation. If you are using R, use a random number seed of -711384152. Draw histograms of the bootstrap replicate estimates for the 25th, 50th, and 75th quantiles of SEENCNT. How do the 95 % CIs from the bootstrap percentile method compare to those from the Woodruff and Francisco-Fuller methods?

15.10. Use the `smho.N874` population and select a sample that is stratified by hospital type.

(a) Determine the proportional allocation of 120 hospitals and the sampling fraction in each stratum.
(b) Select a stratified simple random sample without replacement using the sample sizes computed in (a). If you are using R, use a random number seed of -69716384.
(c) Compute the estimated average number of beds per hospital overall and for each hospital type. Use both the linearization and JKN methods of variance estimation and account for finite population corrections.

(d) How do the SEs in part (c) compare? Is there a reason to prefer one method of variance estimation over the other for this sample? Explain your answer.

15.11. Suppose that you know from the survey documentation that the `nhis.large` file was poststratified by age group (`age.grp`) and race (`race`).

(a) Describe how can you account for this when estimating standard errors?
(b) Using your method from (a), compute estimates of the proportions of persons who delayed medical care in the last 12 months (`delay.med`) for different income levels (`inc.grp`). Calculate their SEs using linearization.
(c) What are the SEs of the estimated totals and proportions of the population that are in each of the `age.grp` × `race` domains?

15.12. Repeat Exercise 15.11 using BRR and Fay BRR with $\rho = 0.5$.

15.13. Use the `nhis.large` file and compute SEs via linearization, BRR, and Fay BRR with $\rho = 0.5$ but ignore poststratification.

(a) Compute estimates of the proportions of persons who delayed medical care in the last 12 months (`delay.med`) for different income levels (`inc.grp`). Are the estimates of proportions the same or different from those in Exercises 15.11 and 15.12?
(b) Compare the SEs ignoring poststratification with those that account for it from Exercises 15.11 and 15.12. How serious an error would be made by ignoring poststratification?

15.14. The following table lists the PSUs in a national sample in the US Regions are numbered 1 to 4. The PSUs with county names (e.g., Kings County NY, Maricopa County AZ) are certainties (or non-self-representing). The non-certainty or non-self-representing PSUs are labeled as region.nsr.nn. For example, NE.nsr.1 is the first NSR PSU in the northeast region. Each of these PSUs is a sample of size 1 from a stratum of NSR PSUs. Strata have been formed within each region so that adjacent strata are similar in population size, i.e., consecutively numbered NSR PSUs within a region are from similar strata. Suppose that each PSU has 10 sample clusters of households.

(a) If you pair NSR PSUs within a region and randomly split each SR PSU into two groups of 5 clusters, how many degrees of freedom would a variance estimator have for national estimates and for regional estimates? Use the standard rule of thumb to count *df*.
(b) What is another method that you might use within each SR PSU to pick up more *df*? What would the resulting total *df* be with this method?

(c) Suppose that you begin with the pairs of NSR PSUs and splits of
    SR PSUs as used in part (a). If you plan to use the BRR method
    of variance estimation with only 20 replicates, how could you combine
    strata to accomplish this in a way that preserves the same number
    of degrees of freedom for the variance estimates for regional estimates
    as in (a)?

Region	Stratum	Region	Stratum	Region	Stratum	Region	Stratum
1	Kings County, NY	2	Cook County IL(1)	3	Miami-Dade County, FL	4	Maricopa County, AZ
1	Queens County, NY	2	Cook County IL(2)	3	Harris County, TX	4	Los Angeles CA(1)
1	NE.nsr.1	2	MW.nsr.1	3	Dallas County, TX	4	Los Angeles CA(2)
1	NE.nsr.2	2	MW.nsr.2	3	S.nsr.1	4	San Diego County, CA
1	NE.nsr.3	2	MW.nsr.3	3	S.nsr.2	4	Orange County, CA
1	NE.nsr.4	2	MW.nsr.4	3	S.nsr.3	4	W.nsr.1
1	NE.nsr.5	2	MW.nsr.5	3	S.nsr.4	4	W.nsr.2
1	NE.nsr.6	2	MW.nsr.6	3	S.nsr.5	4	W.nsr.3
1	NE.nsr.7			3	S.nsr.6	4	W.nsr.4
1	NE.nsr.8			3	S.nsr.7	4	W.nsr.5
				3	S.nsr.8	4	W.nsr.6
				3	S.nsr.9		
				3	S.nsr.10		

Chapter 16
Weighting the Personnel Survey: One Solution

The project assigned in Chap. 12 was to compute a set of weights for a survey of members of the military reserves. A stratified simple random sample of personnel was selected and queried about satisfaction with their jobs. The project provides an opportunity to put into practice the techniques covered in Chaps. 13–15. Completing the project requires calculation of base weights, an adjustment to account for cases whose eligibility status is unknown, an adjustment for nonresponse, and calibration to some finite population totals. There are several practical problems to be solved, including selecting a particular method of nonresponse adjustment, deciding how to use the population counts that are available, and determining how to handle missing values in both the sample cases and the population counts.

Although this chapter is not written as a formal report to be delivered to a client as was requested in the Chap. 12 assignment, we want to again emphasize the importance of good documentation. Clear documentation of all weighting steps is critical for several reasons. It may be necessary to repeat some or all steps at a later time. For example, errors may be discovered in some details of calculations, or problems may be found in one of the input data sets. If a survey will be repeated at a later date, a well-written weighting report can guide the work in the next survey. Very detailed specification memos, like the ones described in Chap. 18, will remove any doubt about what should be done and can lead to reduced costs if the survey is repeated at a later date.

The R code for the solution to this project is in the files,

```
16.1 Solution bwt-unknown adj.R
16.2 Solution NR adj.R
16.3 Solution calibration adj.R
16.4 Example tabulations.R
```

all of which are on the book's web site.

16.1 The Data Files

Two data files were provided for the project. One file (`SOFR.sas7bdat` or `SOFR.xpt`) contained records for all 71,701 sample members who were initially selected. The file includes the 19 variables shown in Table 16.1. The fields include identification number, final respondent status code, stratum identifier, stratum sample count and population count, frame variables

Table 16.1: Contents of data file SOFR.sas7bdat.

#	Variable	Label
1	REC_ID	Unique Record Identification Number
2	RESPSTAT	Final respondent status code
3	SRMARST	What is your marital status?
4	RA006A	Taking all things into consideration, how satisfied are you, in general, with each of the following aspects of being in the National Guard/Reserve? Your total compensation (i.e., base pay, allowances, and bonuses)
5	RA006B	Taking all things into consideration, how satisfied are you, in general, with each of the following aspects of being in the National Guard/Reserve? The type of work you do in your military job
6	RA008	Suppose that you have to decide whether to continue to participate in the National Guard/Reserve. Assuming you could stay, how likely is it that you would choose to do so?
7	RA115	Overall, how well prepared are you to perform your wartime job?
8	RA118	Overall, how would you rate the current level of stress in your personal life?
9	SRED	What is the highest degree or level of school that you have completed? Mark the one answer that describes the highest grade or degree that you have completed
10	RA112RA	In past 12 months, how many days did you spend in a compensated Reserve/Guard status?
11	XSRRCR	Service
12	XACT2R	Activated 30 days—3 level: In the last 24 months were you ever activated longer than 30 consecutive days?
13	XRETH4R	Imputed race/ethnicity—2 level
14	XSEXR	Recoded: imputed gender
15	XCPAY1R	Recoded: imputed pay grade group 1
16	NSAMP	Stratum sample count
17	NSTRAT	Stratum population count
18	V_STRAT	Variance estimation stratum
19	STRATUM	Design stratum

Table 16.2: Contents of data file RCCPDS57.sas7bdat.

#	Variable	Label
1	SERVICE	(XSRRCR) Branch of military service
2	GENDER	(XSEXR) Gender
3	PG_GROUP	(XCPAY1R) Pay grade group
4	RACETH	(XRETH4R) Race/ethnicity
5	EDUCCAT	(SRED) Highest degree/level of school completed
6	MARIT	(SRMARST) Current marital status
7	ACTIVATD	(XACT2R) Activated more than 30 consecutive days or less in last 24 months
8	COUNT	Person count

(gender, pay grade, race, etc.), and respondents' answers to key questions. The variable RESPSTAT for final respondent status code has information about the eligibility and the response status for each sample member.

The fields NSAMP and NSTRAT contain the number of cases in the sample and in the frame for the stratum to which a person belongs. The values are the same for all records for persons in a given stratum. Based on inspecting the file of sample persons, there were 404 strata, defined by combinations of branch of the service, race/ethnicity, gender, and pay grade.

As shown in Table 16.2, the other data file (RCCPDS57.sas7bdat or RCCPDS57.xpt) has population counts for seven frame variables (branch of the service, gender, pay grade, race/ethnicity, education, marital status, and whether a person had been called to active service more than 30 consecutive days in the last 24 months.). These frame variables have different names than in the sample data file, but the alternative names are indicated in the labels. Population counts are provided in variable COUNT.

16.2 Base Weights

Base weights can be computed as soon as the sample is selected. We do not need to know the dispositions of any of the sample cases because the base weights in this survey depend only on the frame counts and the sample sizes in each of the design strata. Since a stratified simple random sample was selected, the selection probability of each person i in stratum h was $\pi_{hi} = n_h/N_h$ where

$$n_h = \text{number of persons sampled from stratum } h,$$

$$N_h = \text{number of persons on the frame in stratum } h.$$

The base weight for person hi is the inverse of the selection probability: $w_{hi} = N_h/n_h$. This is computed as `NSTRAT/NSAMP`. The sum of the base weights is 870,373, which is exactly equal to the count of the persons on the frame since the sample is *stsrs*.

16.3 Disposition Codes and Mapping into Weighting Categories

Table 16.3 gives counts of persons by the disposition codes in the `RESPSTAT` field. These codes are specific to the survey of Reserve personnel, as is apparent from some of the categories. For example, code 22 (no return—separated/retired) would probably not be used in surveys of most other populations. Because there was a time lapse between the time the sample was selected and the data were collected, the status of some persons changed. This is the reason for having codes for retirees, deceased, incarcerated, etc. Addresses for some personnel are out-of-date leading to the inability of the postal service to deliver the survey (code 27). To compute weights, the disposition codes need to be mapped into the groups:

ER	Eligible respondents
ENR	Eligible nonrespondents
IN	Known ineligibles
UNK	Unknown eligibility

Table 16.3: Counts for each final respondent status code.

Respondent status (as stored in `RESPSTAT` variable)	Count
1 = questionnaire returned—completed	25,539
2 = questionnaire returned—(sufficient) partial complete	20
3 = questionnaire returned—(insufficient) Partial Complete	524
4 = questionnaire returned—ineligible	503
5 = questionnaire returned—blank	97
18 = no Return—deceased	9
19 = no Return—incarcerated	2
22 = no Return—separated/retired	35
23 = no Return—active refusal	193
25 = no Return—other	8
26 = no Return—eligible based on administrative records	39,872
27 = postal nondelivery	1,339
29 = not locatable	6
35 = ineligible—no questionnaire sent	3,554
Total	71,701

To compute the various AAPOR response rates described in Chap. 6, the disposition codes are mapped to a slightly different set of categories:

I	Complete interview
P	Partially complete interview
R	Refusal/break-off
NE	Not eligible
U	Unknown eligibility
O	Other eligible noninterview

The mappings we used for both the weighting and AAPOR categories are shown in Table 16.4. A number of decisions have to be made about how to map the dispositions. Some choices are obvious, like mapping code 1 (questionnaire returned—Completed) to ER and I. Others are less so, like codes 5 (questionnaire returned—blank), 25 (no return—other), 27 postal nondelivery), and 29 (not locatable). Unless more is known about such cases, a conservative approach would be to consider the eligibility of these persons as unknown, which we did in Table 16.4. Since there is disposition code 26 (no return—eligible based on administrative records), it is apparent that efforts were made to match the sample file against personnel records. Consequently, the alternative argument could be made that persons in codes 5, 25, 27, and 29 are ineligible. Clearly, there is some subjectivity in the mapping.

Table 16.5 shows the counts of cases in the weighting and AAPOR categories. Judging from the counts, unknown eligibility is a minor problem. On the other hand, the response rate is well under 50%. Thus, concentrating efforts on the nonresponse adjustment is prudent in this sample.

Chapter 6 reviewed various outcome rates that may be computed in a survey. As illustrations, we compute $RR1$ and $RR4$ which are defined as

$$RR1 = \frac{100I}{(I + P) + (R + O) + U},$$

$$RR4 = \frac{100\,(I + P)}{I + P + R + O + e * U},$$

where

$$e = \frac{I + P + R + O}{I + P + R + O + NE}$$

is the proportion of unknowns that are allocated to being eligible. In this sample, $e = 0.941$, $RR1 = 37.78\%$, and $RR4 = 37.83\%$. Since the number of unknowns is a small part of the full sample, the values of these two response rates are virtually the same.

Table 16.4: Mapping of disposition codes into collapsed respondent statuses.

Response status (as stored in RESPSTAT)	Weighting code	Weighting category	AAPOR code	AAPOR description
1 = questionnaire returned—completed	ER	Eligible respondent	I	Complete interview
2 = questionnaire returned—(sufficient) partial complete	ER	Eligible respondent	P	Partially complete interview
3 = questionnaire returned—(insufficient) partial complete	ENR	Eligible nonrespondent	R	Refusal/break-off
4 = questionnaire returned—ineligible	IN	Ineligible	NE	Not eligible
5 = questionnaire returned—blank	UNK	Unknown eligibility	U	Unknown eligibility
18 = no return—deceased	IN	Ineligible	NE	Not eligible
19 = no return—incarcerated	IN	Ineligible	NE	Not eligible
22 = no return—separated/retired	IN	Ineligible	NE	Not eligible
23 = no return—active refusal	ENR	Eligible nonrespondent	R	Refusal/break-off
25 = No Return—other	UNK	Unknown eligibility	U	Unknown eligibility
26 = No Return—eligible based on administrative records	ENR	Eligible nonrespondent	O	Other eligible noninterview
27 = postal nondelivery	UNK	Unknown eligibility	U	Unknown eligibility
29 = not locatable	UNK	Unknown eligibility	U	Unknown eligibility
35 = ineligible—no questionnaire sent	IN	Ineligible	NE	Not eligible

Table 16.5: Counts for each weighting and AAPOR category.

Disposition	Indicator	Count	Percent
Weighting category (disposition codes)			
Eligible respondent (1,2)	ER	25,559	35.6
Eligible nonrespondent (3,23,26)	NR	40,589	56.6
Known ineligible (4,18,19,22,35)	IN	4,103	5.7
Unknown eligibility (5,25,27,29)	UNK	1,450	2.0
Total		71,701	100.0
AAPOR category (disposition codes)			
Complete (1)	I	25,539	35.6
Partial (2)	P	20	0.03
Refusal/break-off (3,23)	R	717	1.0
Other eligible noninterview (26)	O	39,872	55.6
Not eligible (4, 18, 19, 22, 35)	NE	4,103	5.7
Unknown eligibility (5, 25, 27, 29)	U	1,450	2.0
Total		71,701	100.0

16.4 Adjustment for Unknown Eligibility

Using the base weights, we can estimate the numbers of persons on the frame that are in the weighting categories, ER, ENR, IN, and UNK:

Weighting category		Estimated count	Percentage (%)
Eligible respondent	ER	320,677	36.8
Eligible nonrespondent	NR	474,675	54.5
Known ineligible	IN	55,770	6.4
Unknown eligibility	UNK	19,251	2.2
Total		870,373	100.0

The estimated population counts are distributed in about the same way as the unweighted counts in Table 16.5. Since only 2.2% of the frame is estimated to be unknowns, we will make one overall adjustment, which, using the notation from Sect. 13.4, is equal to

$$a_1 = \frac{\sum_{i \in s} d_{0i}}{\sum_{i \in s_{KN}} d_{0i}} = \frac{320,677 + 474,675 + 55,770}{870,373} = 1.0226.$$

The adjustment is made in the file "16.1 Solution bwt-unknown adj.R".

16.5 Variables Available for Nonresponse Adjustment

There are four variables that have non-missing data for both the sample respondents and nonrespondents: branch of the service, race/ethnicity, gender, and pay grade. These are the same variables that were used in defining design strata. The other personal characteristics—education, marital status, and whether a person spent more than 30 consecutive days on active duty in the last 2 months—are each missing for almost all nonrespondents. Table 16.6 shows sample counts of responding and nonresponding persons for each of the variables that we can use for nonresponse adjustment; Table 16.7 shows

Table 16.6: Sample counts of respondents and nonrespondents and population counts for the four variables with no missing data for sample persons.

(Code value) variable	Nonrespondent		Respondent			Population controls (before imputation)	
	n	%	n	%	Total	N	%
Service							
(1) Army National Guard	10,060	65.0	5,424	35.0	15,484	322,053	40.2
(2) Army Reserve	8,398	61.9	5,179	38.1	13,577	190,235	23.7
(3) Naval Reserve	4,686	56.4	3,617	43.6	8,303	77,022	9.6
(4) Marine Corps Reserve	7,869	70.6	3,283	29.4	11,152	36,094	4.5
(5) Air National Guard	4,855	53.6	4,207	46.4	9,062	105,092	13.1
(6) Air Force Reserve	4,721	55.1	3,849	44.9	8,570	71,022	8.9
Missing	–	–	–	–	–	291	0.04
Race/ethnicity							
(1) Non-Hispanic White	20,625	55.1	16,833	44.9	37,458	540,473	67.4
(2) Total minority	19,964	69.6	8,726	30.4	28,690	260,734	32.5
Missing	–	–	–	–	–	602	0.1
Gender							
(1) Male	34,100	61.9	21,007	38.1	55,107	663,122	82.7
(2) Female	6,489	58.8	4,552	41.2	11,041	138,574	17.3
Missing	–	–	–	–	–	113	0.01
Pay group							
(1) E1–E3	7,026	82.5	1,494	17.5	8,520	112,244	14.0
(2) E4	12,936	75.8	4,125	24.2	17,061	198,048	24.7
(3) E5–E6	10,146	64.2	5,653	35.8	15,799	265,388	33.1
(4) E7–E9	2,810	47.1	3,162	52.9	5,972	110,397	13.8
(5) W1–W5	987	42.1	1,356	57.9	2,343	10,948	1.4
(6) O1–O3	3,185	45.7	3,783	54.3	6,968	41,176	5.1
(7) O4–O6	3,499	36.9	5,986	63.1	9,485	63,608	7.9
Missing	–	–	–	–	–	–	–
Grand totals	40,589	61.4	25,559	38.6	66,148	801,809	100.0

Table 16.7: Sample counts of respondents and nonrespondents and population counts for education level, marital status, and activation.

Variable (Code value)	Nonrespondent		Respondent			Population controls (before imputation)	
	n	%	n	%	Total	N	%
Education							
(1) 12 years or less of school (no diploma)	0	0.0	146	100.0	146	10,819	1.3
(2) High school graduate—high school diploma or equivalent	0	0.0	2,059	100.0	2,059	116,933	14.6
(3) Some college credit but less than 1 year	1	0.0	2,465	100.0	2,466	113,512	14.2
(4) One or more years of college, no degree	2	0.0	4,967	100.0	4,969	223,581	27.9
(5) Associate's degree	1	0.0	2,399	100.0	2,400	96,073	12.0
(6) Bachelor's degree	8	0.1	7,750	99.9	7,758	147,450	18.4
(7) Master's, doctoral, or professional school degree	1	0.0	4,912	100.0	4,913	66,614	8.3
Missing	40,576	97.9	861	2.1	41,437	26,827	3.3
Marital status							
(1) Married	233	1.4	16,934	98.6	17,167	455,603	56.8
(2) Separated	6	1.5	397	98.5	403	11,748	1.5
(3) Divorced	40	1.6	2,538	98.4	2,578	75,025	9.4
(4) Widowed	0	0.0	75	100.0	75	3,324	0.4
(5) Never married	157	2.7	5,577	97.3	5,734	254,468	31.7
Missing	40,153	99.9	38	0.1	40,191	1,641	0.2
Activated more than 30 days							
(1) Activated $\leq$ 30 days	7	1.1	611	98.9	618	37,171	4.6
(2) Activated 30 days	148	1.1	12,912	98.9	13,060	250,808	31.3
(3) Not activated	24	0.2	11,814	99.8	11,838	508,083	63.4
Missing	40,410	99.5	222	0.5	40,632	5,747	0.7
Grand totals	40,589	61.4	25,559	38.6	66,148	801,809	100.0

similar counts for the other three demographic variables that are available mainly for respondents. The two tables also show population counts from the RCCPDS57.XPT file.

The file from which population counts were made had some missing data for each variable, other than pay group. For example, branch of the service in Table 16.6 was missing for 291 persons; race/ethnicity was missing for 602 persons. Later, in Sect. 16.7, when we calibrate to the population counts, imputations will have to be made for those missing values.

16.6 Nonresponse Adjustments

Two options for nonresponse adjustment that we covered in Chap. 13 are to use estimated response propensities and cells formed with a regression tree. Both alternatives are examined in this section.

Propensity Models

First, we will examine the option of creating classes based on estimated response probabilities or propensities. A model with main effects and all two-way interactions, using the four available variables, was fitted without using survey weights. The R code is shown below and is in the file "16.2 Solution NR adj.R". The variable resp is 1 for respondents and 0 for nonrespondents:

```
glm.logit2  <-  glm(resp ~ as.factor(xsrrcr)*as.factor(xreth4r)
                     + as.factor(xsrrcr)*as.factor(xsexr)
                     + as.factor(xsrrcr)*as.factor(xcpay1r)
                     + as.factor(xreth4r)*as.factor(xsexr)
                     + as.factor(xreth4r)*as.factor(xcpay1r)
                     + as.factor(xsexr)*as.factor(xcpay1r),
                     family=binomial(link = "logit"),
                     data = sofr.d1.elig)
 anova(glm.logit2, test="Chisq")
```

The data set sofr.d1.elig is a subset of sofr.sas7bdat that contains only the 66,148 eligible respondents and nonrespondents. Part of the output from the anova command is shown below (all deviance degrees of freedom are over 66,000):

	Df	Deviance	P(> Chi)	
as.factor(xsrrcr)	5	951.3	< 2.2e-16	***
as.factor(xreth4r)	1	1376.9	< 2.2e-16	***
as.factor(xsexr)	1	13.2	0.0002764	***
as.factor(xcpay1r)	6	6081.2	< 2.2e-16	***
as.factor(xsrrcr):as.factor(xreth4r)	5	71.0	6.379e-14	***
as.factor(xsrrcr):as.factor(xsexr)	5	15.9	0.0070584	**
as.factor(xsrrcr):as.factor(xcpay1r)	30	209.9	< 2.2e-16	***
as.factor(xreth4r):as.factor(xsexr)	1	31.2	2.291e-08	***
as.factor(xreth4r):as.factor(xcpay1r)	6	7.2	0.3004589	
as.factor(xsexr):as.factor(xcpay1r)	6	36.8	1.937e-06	***

```
 ---
 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

All main effects and interactions are highly significant except for the xreth4r*xcpay1r interaction. In a survey-weighted regression the same factors and interactions were significant. With such a large sample size, we could probably find some significant three-way interactions also. But, for this project, we will not attempt to extend the model above.

Given predicted response probabilities from this model, we can create classes based on their quantiles. Table 16.8 shows the ranges of propensities and counts of persons in each class when 5 and 10 classes are created. Notice that the counts of persons in each class are not equal. Since the model uses only factors as predictors, there are many ties among the estimated propensities, leading to uneven divisions among the classes. Using 10 classes does seem to distinguish better among different rates than does 5 classes. The estimated rates within classes in the last five columns of Table 16.8 are fairly similar regardless of the method of calculation.

Figure 16.1 shows boxplots of the estimated propensities from the model for the 5 and 10 class breakdowns. The ranges are fairly wide in each of the 5 classes but noticeably less within each of the 10 classes. As a further diagnostic, we can check whether balance was achieved for the covariates in the 10-class breakdown. The following R code creates an indicator for whether a person is in the Army National Guard and checks balance:

```
v1 <- rep(0,nrow(sofr.d1.elig))
v1 <- sofr.d1.elig$xsrrcr == 1        # Army National Guard
chk <- glm(v1 ~as.factor(p.class.10) + as.factor(resp) +
              as.factor(p.class.10)*as.factor(resp),
          family=binomial(link = "logit"),
        data = sofr.d1.elig)
anova(chk, test="Chisq")
```

Part of the output of the anova statement is

```
                                         Df  Deviance  P(>Chi)
as.factor(p.class.10)                     9  12764.8   < 2.2e-16  ***
as.factor(resp)                           1      8.9   0.00288    **
as.factor(p.class.10):as.factor(resp)     9     24.5   0.00364    **
```

Similar checks (not shown here) reveal that the interaction term is significant when predicting whether a person is in the Army Reserve, is in pay group E1-E3, or is a non-Hispanic White. As a result, the model with two-way interactions does not achieve statistical balance. In part, this is probably due to the extremely large sample in which small effects test out as statistically significant, and, in part, to misspecification of the model itself. In particular, there may be higher order interactions. Using a regression tree may be one way of finding these.

Regression Tree

Using the same four variables as above—service, pay group, gender, and race/ethnicity—we fit a CART model with this code:

```
t1  <-  rpart(resp ~ xcpay1r + xreth4r + xsexr + xsrrcr,
                  method = "class",
                  control = rpart.control(minbucket = 250, cp=0),
                  data = sofr.d1.elig)
```

Table 16.8: Ranges of estimated response propensities for 5 and 10 classes along with five estimates of response propensity within each class.

Range of estimated propensities	Number of sample persons	Unweighted mean	Weighted mean	Unweighted response rate	Weighted response rate	Median response propensity
5 classes						
[0.098, 0.213]	13,568	0.160	0.172	0.160	0.201	0.166
(0.213, 0.300]	13,539	0.262	0.251	0.260	0.267	0.264
(0.300, 0.446]	13,077	0.375	0.387	0.380	0.416	0.375
(0.446, 0.569]	13,213	0.520	0.509	0.521	0.517	0.520
(0.569, 0.735]	12,751	0.631	0.635	0.629	0.651	0.628
10 classes						
[0.098, 0.166]	8,218	0.140	0.147	0.141	0.177	0.137
(0.166, 0.213]	5,350	0.191	0.198	0.190	0.226	0.190
(0.213, 0.264]	7,600	0.244	0.242	0.236	0.256	0.238
(0.264, 0.300]	5,939	0.285	0.284	0.290	0.313	0.281
(0.300, 0.360]	6,053	0.336	0.335	0.346	0.350	0.326
(0.360, 0.446]	7,024	0.409	0.403	0.409	0.437	0.398
(0.446, 0.519]	6,182	0.481	0.472	0.478	0.494	0.484
(0.519, 0.569]	7,031	0.555	0.551	0.558	0.542	0.555
(0.569, 0.627]	6,136	0.595	0.599	0.588	0.611	0.591
(0.627, 0.735]	6,615	0.665	0.667	0.667	0.687	0.649
Total	66,148					

Predicted Probabilities by 5 adjustment classes
Dots represent class mean

Predicted Probabilities by 10 adjustment classes
Dots represent class mean
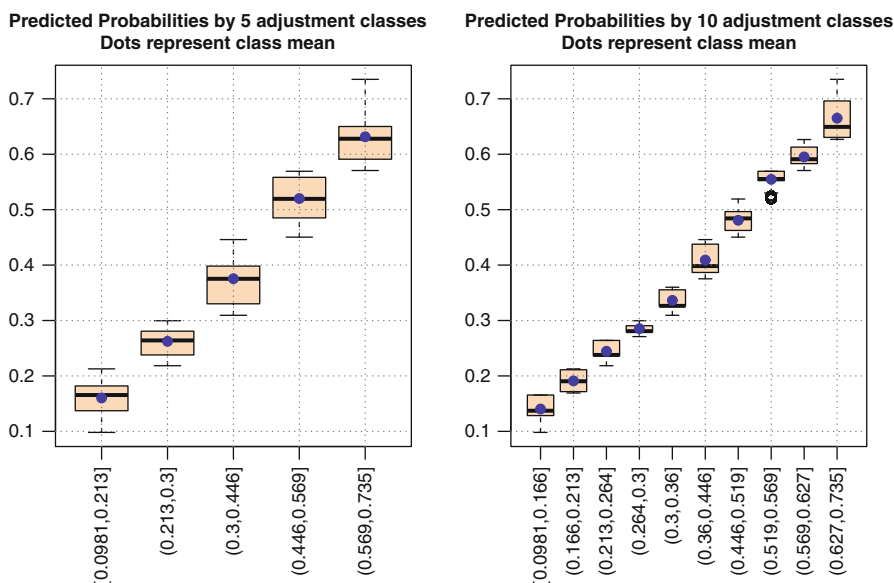
Fig. 16.1: Boxplots of estimated response propensities grouped into 5 and 10 classes. Propensity model estimated based on the four variables available for respondents and nonrespondents using a model with main effects and all two-way interactions. A dot marks the average propensity in each class.

The tree with 13 terminal nodes is shown in Fig. 16.2. As is apparent from the figure, the structure has some complicated combinations. Table 16.9 gives the descriptions of the nodes. The CART classes are numbered differently by the print method than in the object t1$where. The highest ranking officers have the highest response rates; this is reflected in class 25 containing pay groups O4-O6 which has a response rate of 0.631 (unweighted) and 0.672 (weighted). Enlisted personnel did not respond well—CART put all E1-E6's in class 2, which has unweighted and weighted rates of 0.272 and 0.321. Among higher paid personnel, Marines are some of the poorest responders. For example, class 21, containing E7-E9, non-Hispanic Whites in the Marine Corps Reserve had an unweighted rate of 0.385 (0.410 weighted). The numbers of persons in the CART classes range from 296 to 41,380, which are obviously far from the more nearly equal-sized classes in the propensity analysis. Note that 41,380 of the 66,148 eligible (62.6%) are in the same class and are assigned the same response rate.

Since the classes formed by the regression tree seem to capture the complexity of the response process better than the logistic model, the classes in Table 16.9 will be used for nonresponse adjustment. We used weighted response rates to make the weight adjustment in each class. The weighted values are shown in the last column of Table 16.9. As mentioned in Chap. 13, not all practitioners will agree on whether the weighted or unweighted rates
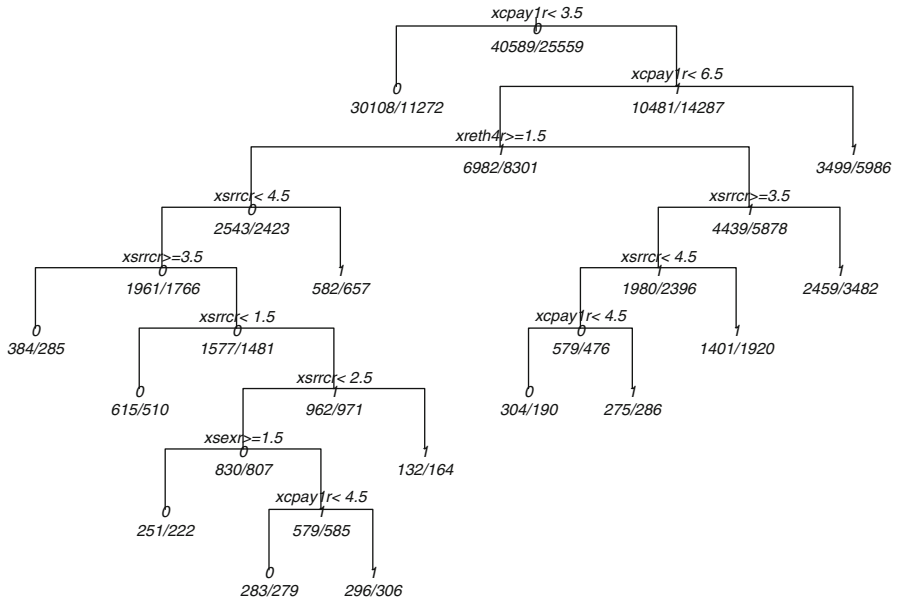
Fig. 16.2: Regression tree to predict response based on the four variables available for respondents and nonrespondents.

should be used. Using the weighted rates is, in a sense, a compromise solution. Conditional on the classes formed, the weighted rates are model unbiased under a model in which every person in a class has a common probability of responding. They are also approximately unbiased estimates of the population response rates in repeated sampling given the particular set of classes used.

16.7 Calibration to Population Counts

The final weighting step in this project will be calibration to some of the available population counts. The statistical function that calibration serves here is mainly to reduce standard errors. Since military administrative records should be accurate, there should be no systematic over- or undercoverage to be corrected. In addition, calibration has some cosmetic appeal here. Having estimated counts exactly equal to ones from administrative personnel records will give the survey results face validity—a feature that may be important to many data users. There are two major operational questions that must be addressed:

Table 16.9: Nonresponse adjustment classes created using a regression tree.

CART class (t1$where)(print)	CART class	Description	No. of persons	Unweighted response rate	Weighted response rate
2	2	E1-E3, E4, E5-E6	41,380	0.272	0.320
7	48	E7-E9, W1-W5, O1-O3, Minority, Marine Corps Reserve	669	0.426	0.406
9	98	E7-E9, W1-W5, O1-O3, Minority, Army National Guard	1,125	0.453	0.481
12	396	E7-E9, W1-W5, O1-O3, Minority, Army Reserve, Female	473	0.469	0.471
14	794	E7-E9, Minority, Army Reserve, Male	562	0.496	0.485
15	795	W1-W5, O1-O3, Minority, Army Reserve, Male	602	0.508	0.517
16	199	E7-E9, W1-W5, O1-O3, Minority, Naval Reserve	296	0.554	0.525
17	25	E7-E9, W1-W5, O1-O3, Minority, Air National Guard, Air Force Reserve	1,239	0.530	0.544
21	104	E7-E9, non-Hispanic White, Marine Corps Reserve	494	0.385	0.410
22	105	W1-W5, O1-O3, non-Hispanic White, Marine Corps Reserve	561	0.510	0.509
23	53	E7-E9, W1-W5, O1-O3, non-Hispanic White, Air National Guard, Air Force Reserve	3,321	0.578	0.612
24	27	E7-E9, W1-W5, O1-O3, non-Hispanic White, Army National Guard, Army Reserve, Naval Reserve	5,941	0.586	0.586
25	7	O4-O6	9,485	0.631	0.672

- Which variables and/or combinations of variables should be used for calibration?
- How should missing values for the calibrating variables be handled in the sample file and the file of population counts?

The code for completing the analyses sketched below is in the file `16.3 Solution calibration adj.R` on the web site.

Other questions that we will not address here, but would be important in a real survey, are:

- For what time period should population counts be made when there is a delay between sample selection and data collection?
- Which persons should be counted to get the controls?

Administrative record databases are typically updated periodically—once a month, once a quarter, etc. There may also be a lag between the time period of the database and the time at which it is available for tabulation. This means that population counts may not be for the time period when data are collected. In addition, data collection may extend across two or more updates of the administrative data. For example, there might be a 2-month lag between sample selection and data collection, the survey period may last 10 weeks, and the administrative records may be updated once a month. When such a lag occurs, the persons who are surveyed will be the "survivors," i.e., the ones who were in the frame when the sample was selected and are still eligible when data are collected. No new entrants to the population would be included in the sample. If the population counts are made close to the time of data collection and include all persons who are eligible based on the survey rules, then the counts would include the new entrants who had no chance of being sampled. If we calibrate to these counts, we are saying that the attitudes of the new entrants can be predicted by those of the sample persons who have been in the population longer. Another option would be to tabulate the control counts using only persons who have been in the military for at least two months, if that is the amount of lag between sampling and data collection. In some surveys, like those of the US household population, such selective tabulations may not be feasible.

In this project, we will use the population counts as given in the `RCCPDS57.sas7bdat` file. As noted in Chap. 12, this file came from matching the sampling frame to the most current personnel file available as of the start of the data collection period. That is, the counts are those of the survivors. Thus, these counts should cover only eligible cases.

Identifying Variables to Use

The file of population counts contains combinations of service, gender, pay group, race/ethnicity, education, marital status, and length of activation. All

of these are categorical and can be used singly or in any number of combinations. We could, for example, use only the marginal counts of service, pay group, and gender. Or, we could use service × pay group and service × gender or service × pay group × gender. Some modeling is a useful approach to guide the decision. The goal will be to determine one set of weights that is reasonably efficient for the important variables measured in the survey. We have six analysis variables listed in Table 16.10 (`RA006a`, `RA006B`, `RA008`, `RA115`, `RA118`, and `RA112RA`) to aid in making the decision.

To do the modeling we created several binary variables. Satisfaction with compensation (`RA006A`) and type of work (`RA006B`) were coded as satisfied/very satisfied = 1 and 0 otherwise. Likelihood of reenlisting (`RA008`) was coded as likely/very likely = 1 and 0 otherwise. Preparation for job (`RA115`) was coded as well prepared/very well prepared = 1 and 0 otherwise. Level of stress (`RA118`)was coded as more than usual/much more than usual = 1 and 0 otherwise. Finally, days in compensated status (`RA112RA`) was used as a continuous variable.

Rather than fitting binary regressions where the form of the predictors is specified in advance, we again used regression trees to allow the algorithm to identify the more important variables and combinations of levels for prediction. Since the intention to reenlist is a key variable in this survey, we present those results here. Figure 16.3 shows the regression tree for predicting whether a person is likely/very likely to reenlist. The code for computing the tree and drawing the figure is

```
t1 <- rpart(ra008R ~ xsrrcr + xsexr + xcpay1r + xreth4r +
                sred + srmarst + xact2r,
            method = "class",
            control = rpart.control(minbucket = 250, cp=0),
            data = datafile)

plot(t1, uniform=TRUE, compress=TRUE, margin = 0.1, branch=0)
text(t1, use.n=TRUE, all=TRUE,
     digits=15,
     cex=1.2,
     pretty=1.2,
     fancy=TRUE,
     fwidth=0.7,
     xpd = TRUE,
     font = 3)
```

The descriptions of the variables, `xsrrcr`, `xsexr`, etc., were given in Table 16.1. The parameter branch=0 in `plot` gives a tree with V-shaped branches, which, in this case, makes the branch labels easier to read. Pay group, branch of service, whether a person had been activated for 30 days or more, and marital status are included in the tree; gender, race/ethnicity, and education are not.

Table 16.10 summarizes which variables were included in the trees for predicting the six analysis variables. Gender was selected only to predict whether people felt prepared to do their jobs. Examination of the individual

Table 16.10: Variables included in regression trees for predicting six analysis variables.

| Name | Analysis variable | Predictors | | | | | | |
		Service	Gender	Pay group	Race/ ethnicity	Education	Marital status	Activation
RA006A	Compensation	✓		✓	✓	✓		✓
RA006B	Type of work	✓		✓		✓	✓	✓
RA008	Reenlist	✓		✓			✓	✓
RA115	Preparation		✓	✓	✓			✓
RA118	Stress	✓		✓	✓			✓
RA112RA	Paid status	✓		✓	✓	✓	✓	✓
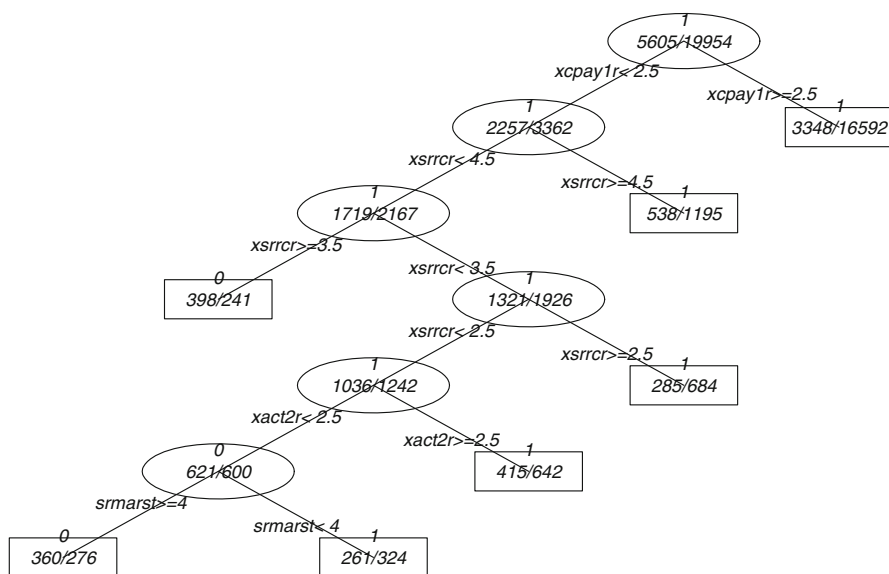
Fig. 16.3: Regression tree for predicting likelihood of reenlisting.

trees shows that service $\times$ pay and service $\times$ activation interactions are always present. Often there are more complicated interactions, as in Fig. 16.3, where there is a combination of service, pay group, activation, and marital status. However, including 3-way and 4-way interactions would lead to samples that are very thin in some combinations of levels even though there are over 25,000 respondents. Based on these results, we decided to use a calibration model with:

- Main effects for service, gender, pay group, race/ethnicity, education, marital status, and activation;
- Interactions for service $\times$ pay and service $\times$ activation.

Although gender only appears once in Table 16.10, we include it for the cosmetic benefit of matching the administrative record count for males and females.

Imputing for Missing Values

Tables 16.6 and 16.7 showed that the file from which population counts were made had missing values for some persons for service, race/ethnicity, gender, education, marital status, and activation. The percentage of persons with missing values ranged from 0.04% for service to 3.3% for education. To impute for the missing values, we need only impute a covariate value whenever it was missing in the `RCCPDS57.sas7bdat` file. For example, there were 159 records of this type in the file that had a missing value for service:

```
service gender pg_group raceth educcat marit activatd
      .        2        2        2        4        5        3
```

To impute the missing service, a random draw is made from the allowable codes in proportion to the population code counts for the non-missing records. The R code for doing the population count imputations is in the function, `impute`, in the file "16.3 Solution calibration adj.R".

The sample file also has some records with missing data on the covariates that will be used for calibration. Table 16.7 shows that 2.1% of the 25,559 respondents are missing education, 0.1% are missing marital status, and 0.5% are missing the activation field. Any missing value for a sample respondent was imputed with a random draw from the allowable codes for a variable. The draws were made in proportion to the distribution among the codes for persons with non-missing data. The R code for doing the sample imputations is in the function, `impute.sam`, which is also in "16.3 Solution calibration adj.R".

These imputation methods are straightforward and could be criticized as not accounting for any multivariate relationships among different variables. Given the small amount of missing data for all variables, we elected to keep the methods simple.

GREG Estimation

Using the files of sample respondents and population counts with all missing values imputed, we calibrated to the population totals using a GREG estimator. When using the `calibrate` function in R survey, some care is needed to be sure that the vector of population totals is in exactly the same order as is being used internally by `calibrate`. The function `model.matrix` will create the model matrix of covariates that `calibrate` uses for a particular formula. In this application, we check the order with

```
        # check how design matrix is formed in calibrate
 mm <- model.matrix(~ as.factor(xsrrcr) * as.factor(xcpay1r)
                    + as.factor(xsrrcr) * as.factor(xact2r)
                    + as.factor(sred)
                    + as.factor(xsexr)
                    + as.factor(xreth4r)
                    + as.factor(srmarst),
                    data = sofr.cal)
 dimnames(mm)[[2]]
```

The last statement lists the column names of the model matrix. The interactions are in "row-major" order. For example, the first five values of the service × pay grade interaction are

```
 "as.factor(xsrrcr)2:as.factor(xcpay1r)2"
```

```
"as.factor(xsrrcr)3:as.factor(xcpay1r)2"
"as.factor(xsrrcr)4:as.factor(xcpay1r)2"
"as.factor(xsrrcr)5:as.factor(xcpay1r)2"
"as.factor(xsrrcr)6:as.factor(xcpay1r)2"
```

That is, service is incremented before pay grade. The code for putting the population controls in the correct order and for computing the GREG weights follows. Prior to this code counts for service × pay grade and service × activation were made and stored in the objects `svc.pay1` and `svc.act1`:

```
          # reorder the pop totals for the interaction terms
          # to match way that calibrate creates model matrix
svc.pay1 <- svc.pay[order(svc.pay[,2]),]
svc.act1 <- svc.act[order(svc.act[,2]),]
del1 <- svc.pay1[,1]==1 | svc.pay1[,2]==1
del2 <- svc.act1[,1]==1 | svc.act1[,2]==1

pop.tots <- c(N,
              svc[-1,2],
              pay[-1,2],
              activated[-1,2],
              educ[-1,2],
              gender[-1,2],
              raceth[-1,2],
              marital[-1,2],
              svc.pay1[!del1,3],
              svc.act1[!del2,3])

 sam.lin.ub  <-  calibrate(design = sofr.cal.dsgn,
                   formula = ~as.factor(xsrrcr)*as.factor(xcpay1r)
                   + as.factor(xsrrcr) * as.factor(xact2r)
                   + as.factor(sred)
                   + as.factor(xsexr)
                   + as.factor(xreth4r)
                   + as.factor(srmarst),
                   population = pop.tots,
                   bounds = c(-Inf,Inf),
                   calfun = c("linear") )
```

Table 16.11 gives some summary statistics on the weights after each step in the process. The mean weight is about the same before and after the GREG step, while the range is larger for the GREG weights than for the nonresponse-adjusted weights. The sum of the weights is smallest after the GREG step (801,809) accounting for the fact that some persons became ineligible between sampling and data collection and that the control totals are for the survivors only.

In this solution, weight trimming was not used, although some practitioners might consider it. Although the range of final weights is fairly large—1.199 to 613.4—the base weights began with a wide range owing to the highly differential sampling rates that were used. The base weights were adjusted to reflect substantially different response rates among some types of personnel.

Table 16.11: Summary of weights and counts of persons after each step.

Weighting step	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Sum	Persons
Base	1	2.201	5.049	12.14	14.27	178.3	870,373	71,701
Adjusted for unknown eligibility	1.023	2.251	5.05	12.3	14.59	182.3	813,342	66,148
Adjusted for nonresponse	1.521	4.746	14.63	31.82	34.31	514.7	813,342	25,559
GREG	1.199	4.672	13.43	31.37	30.91	613.4	801,809	25,559

Consequently, the final weights have a wide range. This is necessary to correct nonresponse bias for some subgroups. However, the 99th quantile of the final weights is about 385 while the maximum final weight is 613.4. Trimming of the largest 1% of weights might reduce SE's for full population estimates without introducing too much bias, but estimates for the subgroups with very low response rates might then be biased. As usual, we are faced with conflicting goals with no unique way of achieving them.

16.8 Writing Output Files

The resulting file with the GREG weights can be written to comma delimited (csv) text files for use in other statistical software. The code below appends the GREG weight to the file, selects fields for output, and writes the text files. The `write.foreign` function in the `foreign` package (R Core Team and contributors worldwide, 2012a) will also write code to be used in importing the csv files into a few other packages. We illustrate the process below for SAS and Stata:

```
        # append GREG weights to data file of 25,559 respondents
 sofr.cal$d3 <- weights(sam.lin.ub)
        # specify fields for the text, SAS, and Stata files
 fields <- c("rec.id", "nr.class", "respstat", "stratum",
            "nsamp", "nstrat", "v.strat",
            "srmarst", "sred", "xsrrcr", "xact2r",
            "xreth4r", "xsexr", "xcpay1r",
            "ra006a", "ra006b", "ra008", "ra115",
            "ra118", "ra112ra",
            "pred.logit", "p.class.10", "unwt.rr", "wt.rr",
            "d0", "d1", "a1",
            "d2", "a2",
            "d3")

 write.foreign(df = sofr.cal[, fields],
        datafile = paste(file_loc2, "sofr.cal.sas.csv", sep=""),
```

```
        codefile = paste(file_loc2, "sofr.sas", sep=""),
        package = "SAS")

 write.foreign(df = sofr.cal[, fields],
        datafile = paste(file_loc2, "sofr.cal.stata.csv", sep=""),
        codefile = paste(file_loc2, "sofr.ado", sep=""),
        package = "Stata")
```

The variable, file_loc2, is a text string specifying the folder where
the output files will be written. The reader can consult the programs,
16.1 Solution bwt-unknown adj.R, 16.2 Solution NR adj.R,
and 16.3 Solution calibration adj.R, to see how the different
variables were created.

   Although the data can be imported into statistical packages other than
R, a worry is that the other packages do not have built-in procedures that
recognize that the weights were computed via the GREG procedure. This,
typically, means that linearization variance estimates will be computed using
the ultimate cluster method discussed in Chap. 15 that does not use the cor-
rect set of residuals. As a result, linearization SEs computed from the other
packages will not generally be correct. This problem can be avoided if repli-
cation is used. In that case, the set of replication weights can appropriately
reflect the different steps in weighting, particularly the type of calibration that
was used. The replicate weights are included with the data set, and a pack-
age like SAS or Stata needs only to be told which method of replication was
used—jackknife, BRR, or the bootstrap—in order to produce legitimate SE's.

16.9 Example Tabulations

Finally, in this section, we present a few simple tabulations using the file
with the final weights. The associated R code is in the file 16.4 Example
tabulations.R. The proportions of persons responding in the categories
of the reenlistment question (ra008) can be estimated with

```
             # proportions for re-enlistment item
 reenlist <- svymean(~ as.factor(ra008), design = sam.lin.ub,
                    na.rm = TRUE)


             # format with row labels
 print(ftable(reenlist,
        rownames = list(c("Very unlikely",
                          "Unlikely",
                          "Neither likely nor unlikely",
                          "Likely",
                          "Very likely")
                  ) ), digits = 3)
```

The function `ftable` allows labels to be used for the printed output. The results, even with the use of `ftable` which improves the appearance, are not beautiful:

```
Very unlikely                    mean   0.06608
                                 SE     0.00357
Unlikely                         mean   0.10924
                                 SE     0.00428
Neither likely nor unlikely mean   0.09280
                                 SE     0.00386
Likely                           mean   0.31757
                                 SE     0.00611
Very likely                      mean   0.41431
                                 SE     0.00619
```

If a table is needed for a report, one option is to import the output into a spreadsheet where more attractive formatting can be applied. Suppose that the result of `print(ftable(reenlist, ...))` is saved to an object called `out`. Code that will convert `out` to a data frame, put the proportion and SE side by side, and write the result to a file called `table.csv` is

```
out <- data.frame(out)
out <- cbind(out[1:5,], out[6:10,])
out <- out[, c(1,3,6)]
dimnames(out)[[2]] <- c("Response", "Proportion", "SE")
write.csv(out, file = "c:\\table.csv")
```

Part IV
Other Topics

Chapter 17
Multiphase Designs

Sample designs are developed and estimators are chosen to efficiently fulfill specified analysis plans. Efficiency is generally defined to encompass three primary areas—accurate estimates (*bias*) with high levels of precision (small standard errors) calculated from data collected with procedures that make economical use of the study funds without exceeding the specified budget (*cost*). Sections 3.1 and 3.2 and Chap. 15 detail the gains achieved in precision if auxiliary information that is highly associated with the analysis variables can be used. This includes, for example, auxiliary variables used (i) in sampling as a stratification variable or to construct the measure of size for a probability proportional to size (*pps*) design or (ii) in estimation with a regression (or ratio) estimator. However, what if the only available sampling frame does not have useful auxiliary information? Without the auxiliary information, how might the statistician address concerns that the inflated sample size required for the specified level of precision will exceed the study budget?

One solution for these issues used by statisticians in various fields is known in general terms as a *multiphase design*. In the following sections, we provide a definition (Sect. 17.1) to differentiate this type of sample design from others discussed in this book, as well as real-life examples of multiphase designs (Sect. 17.2). Having established a working definition of multiphase designs, we examine the components needed to develop both base and analysis weights (Sect. 17.3). The weights are then used in the presentation of a few point estimates and variances (Sect. 17.4), borrowing formulas discussed in other chapters of this text along with a few summarized from published research. Methods to determine overall sample size and allocation to phases are given in Sect. 17.5 along with the methods used to justify the need for a multiphase study when these surveys sometimes require a lengthy data collection period. We conclude this chapter with a brief discussion of software available for sample selection and analysis (Sect. 17.6).

17.1 What is a Multiphase Design?

Most major sampling textbooks contain a discussion of *two-phase designs*. These designs use at least two, sequential sampling frames:

(1) A population frame that covers the target population
(2) A frame containing auxiliary (population) information and survey responses for a random sample selected from the population frame

Thinking of a generic survey may clarify the features that distinguish a multiphase design.

Consider a survey where data are collected through a relatively inexpensive mode on a random sample of units drawn from a sampling frame that covers the target population. Call this the *phase 1 sample* selected from a *phase 1 sampling frame*. Information collected in the first phase along with auxiliary data from the phase 1 sampling frame form the *second-phase sampling frame*. Data are then collected from a random subsample of phase 1 sample units, referred to as a *phase 2 sample*. Data collection in the second phase typically involves a more expensive methodology than used in the first phase.

The standard textbook discussion includes only two design phases and units in both phases selected via single-stage sampling. Extending the design to complex sampling within the first phase or to three or more phases complicates the theoretical derivations and variance estimation (as well as record-keeping procedures in actual surveys) but does serve a purpose as discussed later in this chapter. Regardless of the number of phases, the type of analytic unit is the *same* in all phases (e.g., persons).

The distinctive characteristic of multiphase designs is the selection of at least one random subsample drawn from an initial sample as highlighted by our generic two-phase survey above. The subsampling may occur once or multiple times, much like a multistage design (Chaps. 9 and 10). In fact, multistage designs are a specialized type of multiphase design. Let us revisit our generic two-phase survey from above; suppose that the second-phase units are selected from clusters of units randomly sampled in the first phase. Särndal et al. (1992, Sect. 4.3.1) classify this study as a two-stage design *if and only if* two properties are satisfied—*independence* and *invariance*. The independence property indicates that the phase 2 units are randomly selected from each phase 1 cluster independent of the other sample clusters. The invariance property is slightly more complicated and focuses on the theory of repeated sampling. In (slightly theoretical) words this means that the phase 2 sampling mechanism (e.g., sampling scheme, selection probabilities) for the units within a particular phase 1 cluster is not influenced by the presence or absence of other phase 1 clusters in repeated implementations of the phase 1 sampling mechanism. This "no peeking rule" states that the phase 2 units are selected regardless of the phase 1 results obtained for other clusters. As noted in Särndal et al. (1992), the theoretical expectation and variance of an estimator taken with respect to the sample designs implemented in each

phase does not change with the particular phase 1 sample selected in a single draw.

A simple way to think about the independent and invariance properties is that if the sample design for the second stage is specified in advance and does not change regardless of which set of first-stage units is selected (or the results from the first-stage units), then the survey is conducted through a two-stage design. Otherwise, the design is a two-phase design. Not only does the violation of independence or invariance change the design label from multistage to multiphase, it also can affect the variance formula used for the point estimator of interest. We postpone the variance discussion until later in this chapter. There the distinction between multistage and multiphase designs is made more concrete through an example. The difference can be subtle as we hope to illustrate.

Example 17.1 (US Education Surveys). The *Education Longitudinal Study of 2002* (ELS:2002)[1] and the *High School Longitudinal Study of 2009* (HSLS:09),[2] both conducted under contract with the National Center for Education Statistics (NCES) located in the US Department of Education's Institute of Education Sciences, focus on understanding students' chosen paths from early high school into the postsecondary education years (i.e., university) and their workforce careers. These surveys incorporate student population counts by race/ethnicity group into sampling rates to first select schools (sampling stage 1) and then to randomly select students independently within each sampled school (sampling stage 2). The population information is obtained from publicly available NCES files containing data collected one to two years prior. With the dated frame information the percent distribution by racial group found at a participating school can differ from the NCES sampling frame percentages. However, the initial sampling rates may remain unless certain rules are violated as noted below.

When the design is executed as planned, it is clearly two-stage. However, some changes can be made at the second stage without turning the design into two phases. If the student sample size using the pre-set stage-2 sampling rate exceeds maximums set for the study,[3] then statisticians will typically adjust the sampling rates with a given school using the updated information. Assume that these adjustments are made independently within each school (independence property) and would have been introduced regardless of the distribution of other schools in the sample (invariance property). Hence, the claim of a two-stage design (i.e., specialized two-phase design) still holds. ∎

---

[1] http://nces.ed.gov/surveys/els2002/

[2] http://nces.ed.gov/surveys/hsls09/

[3] Sampling rates are typically set to limit the variation in the base weights and to limit the burden placed on the participating schools as measured by the student sample size.

Example 17.2 (US Education Surveys, revisited). Keeping with the previous school example, suppose that after collecting data in less than half of the sample schools the statistician projects that the study will obtain an insufficient number of participating students in one race/ethnicity group to meet the requirements specified in the analysis plan. If the researcher decides to collectively adjust the pre-set sampling rates for the remaining schools to select a larger sample for the underpowered group, then the independence and invariance properties are violated. Consequently, the two-stage design label is no longer valid. Said another way, the changes are introduced mid-data collection to address problems introduced by the unanticipated, random response pattern exhibited in the sample. A thorny issue is how to estimate a variance in this case. Strictly speaking a variance estimator specialized for two-phase sampling should be used. In practice, however, the *two-stage* (rather than *two-phase*) variance estimator may be used. The two-stage formula may be adequate depending on the degree to which the independence and invariance properties are relaxed. Section 17.4.2 discusses variance estimation issues in more detail. ■

In this section, we provided you with a general example to differentiate multistage and multiphase studies. In the next section, we turn to examples of real-world projects to distinguish three types of multiphase designs.

17.2 Examples of Different Multiphase Designs

Multiphase studies are known by different names depending on the purpose of the design. The three types of the multiphase studies discussed in this section, and in the remainder of the chapter, are *double sampling for stratification*, *nonrespondent subsampling*, and *responsive designs*. An overview of each design is discussed below, along with associated surveys found in the literature. Details on weighting and variance estimation for these studies are provided in the subsequent sections.

17.2.1 Double Sampling for Stratification

Lohr (1999) and others note that two-phase sampling, also known as double sampling, was first introduced by Neyman (1938) as a method for obtaining important auxiliary information from a large sample of units by way of a relatively inexpensive method and then using this data to subsample the units for a more intense and expensive data collection procedure. Double sampling for stratification is a specific type of two-phase design where auxiliary information obtained from the phase 1 data collection is used in combination with the phase 1 frame information to form phase 2 design strata within which independent samples are selected.
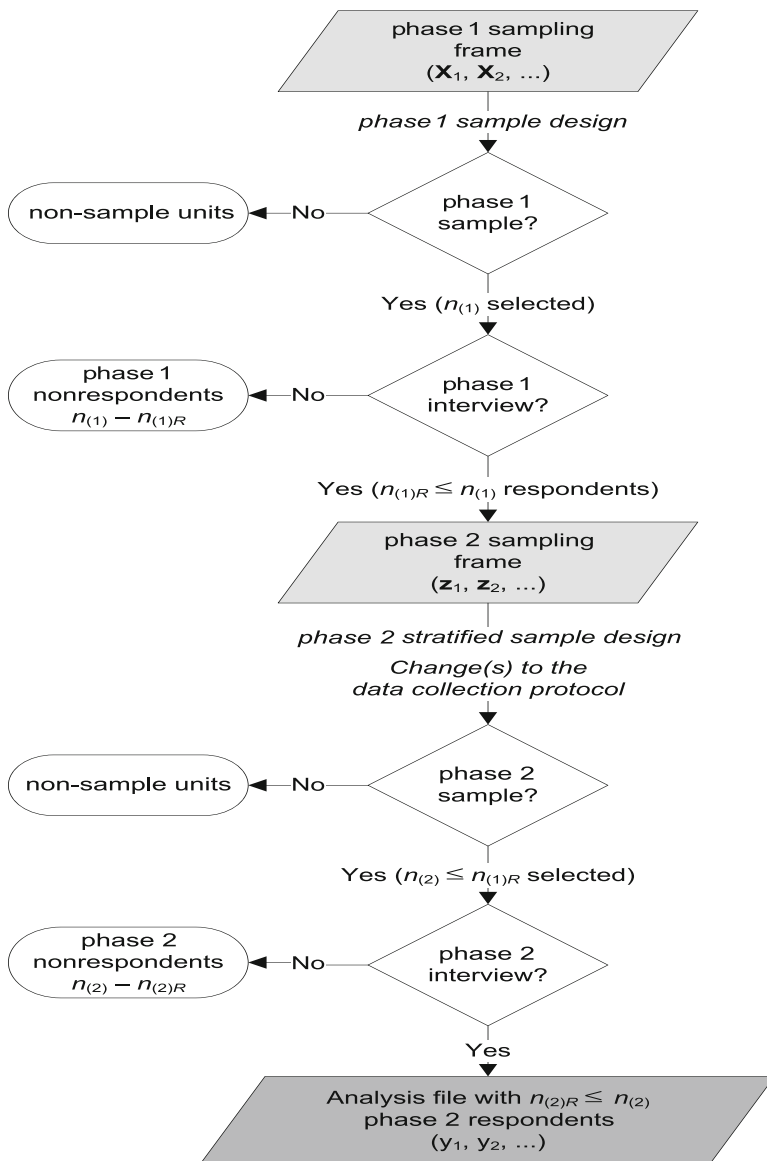
Fig. 17.1: Transition of sample cases through the states of a survey under a double sampling for stratification design.

These words are translated into a picture shown in Fig. 17.1 to demonstrate the transition of cases from state to state within a survey. In words, a phase 1 sample of size $n_{(1)}$ is selected from an available sampling frame of size N.

The particular sample is chosen through a random sampling scheme that uses a set of auxiliary variables $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots)'$, where $\mathbf{x}_g$ denotes a vector of values for the gth variable of length N. Examples of auxiliary variables are type of business used for stratification in an establishment survey or the number of students by race/ethnicity in a school used for *pps* sampling in an education survey. (Note that $\mathbf{X}$ is a vector of N ones when the phase 1 units are drawn via simple random sampling.) Additional auxiliary data $\mathbf{Z} = (\mathbf{z}_1, \mathbf{z}_2, \ldots)'$, including process information known as *paradata*, are obtained from the $n_{(1)R}$ respondents and included on the phase 2 sampling frame. Examples of paradata include call or contact history records, with number of contacts and results from prior contacts, and field observations, such as the presence of toys in the yard to signify an occupied housing unit most likely with children (Kreuter et al. 2010).

The phase 1 auxiliary information ($\mathbf{z}$) and most likely the original frame information ($\mathbf{X}$) are used to develop the phase 2 design. For example, in a household survey, the age and race/ethnicity ($\mathbf{z}$) of each person in the phase 1 households along with household income and renter status might be determined in an initial interview. That information could then be used to stratify the phase 2 sample. A total of $n_{(2)}$ units ($\leq n_{(1)R}$) are randomly selected for a second-phase data collection under a protocol that typically differs from the first phase (e.g., different mode of data collection). The key analysis variables $\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2, \ldots)'$ are then collected from as many phase 2 sample members as possible.

Several two-phase (or double-sampling) designs are cited in the literature. We list five here, but there are many more:

(1) Researchers working to develop a case definition for undiagnosed symptoms in US personnel serving in the 1991 Persian Gulf War surveyed a stratified simple random sample of Gulf War-era veterans (Iannacchione et al. 2011). Based on survey responses to the US military health survey (USMHS) in the first phase, respondents were classified as likely having or not having a certain type of illness. Blood specimens were requested from randomly sampled phase 1 respondents within the illness strata and analyzed using expensive tests. Thus, the critical analytic variable for the USMHS two-phase study was linked to biological data collected only from the phase 2 respondents.

(2) Another example comes from the European Pain in Cancer (EPIC) survey. For this telephone survey, a phase 2 sample was selected from phase 1 respondents screened for significant levels of cancer-linked pain so researchers could better estimate the prevalence and severity of chronic pain and the utility of various treatment regimes to improve quality of life (Breivik et al. 2009).

(3) The quarterly retail commodity survey (QRCS) conducted by Statistics Canada is used to obtain "detailed information on retail commodity sales" on a subsample of companies selected for the monthly survey of retail

trade (MRTS). Updated information collected through the (phase 1) MRTS is used to re-stratify the sample prior to drawing the QRCS phase 2 sample (Hidiroglou 2001).

(4) The Encuesta de Actividades de Niños, Niñas y Adolescentes (EANNA)[4] or Survey of Children and Adolescents is a two-phase sample designed to measure gender equality and child labor in Chile. In phase 1 a national sample of addresses is selected and the ages of persons in the households are determined. In the second phase, children and adolescents are stratified into the age groups 5–8, 9–11, 12–14, and 15–18 and a subsample is selected from each stratum.

(5) A few studies have used a sampling frame formed from the respondents to the National Health Interview Survey (NHIS).[5] For example, Cycle 5 of the National Survey of Family Growth (NSFG-V) subsampled from 1993 NHIS respondents to produce national estimates of fertility practices and sexual health of women in the US 15–44 years of age (Potter, et al 1998). The sample design for the subsequent cycles of NSFG to date is a 4-stage area probability sample with a nonresponse follow-up phase described in the next section. (Lepkowski et al. 2010). Another example is the Medical Expenditure Panel Survey household component (MEPS-HC) where national estimates of health insurance coverage and health care expenditures are produced from a subsample of respondents to the previous years' NHIS (Ezzati-Rice et al. 2008).

(6) One final two-phase example presented in this introductory section is for the birds. Researchers used an inexpensive and somewhat inaccurate method to estimate the density of nesting birds in a large sample of geographic areas in Alaska (Bart and Earnst 2002). Intensive methods were conducted within a phase 2 sample to estimate a measurement error adjustment. This adjustment was then applied to the complete phase 1 sample to estimate the population bird-nesting density. In all of the examples, a phase 2 sampling frame did not exist prior to the first-phase study.

17.2.2 Nonrespondent Subsampling

All students of survey research have been exposed to theory and methodology that assumes 100 % participation from the sample units. However, we know that nonresponse is a very real fact of survey life and must be addressed before, during, and after the conduct of a study. For example, pre-data collection discussions may focus on the use of incentives only after a specified number of attempts to either contact a sample member or convert them to

[4] http://www.lanacion.cl/eanna-primera-radiografia-de-los-ninos-y-adolescentes-de-chile/noticias/2012-02-15/133220.html

[5] http://www.cdc.gov/nchs/nhis.htm

a participant. During data collection, the project team will review records to ensure that hard-to-reach cases are contacted at varying times of the day and week to increase the contact rate. Finally, with edited survey responses in hand, sampling weight adjustments have been shown to reduce nonresponse errors (see, e.g., Chap. 15 of this text; Särndal et al. 1992; Kott 2006).

Procedures to adjust for potential nonresponse bias are unnecessary only if the nonrespondents are no different than the respondents on the set of important analytic variables for the study. This is referred to as ignorable nonresponse or missing completely at random (MCAR discussed in Sect. 13.5) in references such as Little and Rubin (2002). Few researchers are willing to blindly make this assumption since many studies do not have data that can be used to verify similarities between the respondents and nonrespondents (e.g., administrative records) other than frame information. To limit the dependence on weight adjustments to correct any bias due to nonresponse, many researchers make every attempt to maximize the response rate along with the quality of the data.

Achieving response rates at or above those used in the sample size calculations is also important to meet the analytic objectives set for the study (see Chap. 6). If fewer completed questionnaires are obtained than desired, then statistical tests may be underpowered or estimates for certain subgroups may be unstable. Problems with bias and low respondent sample sizes might suggest the need to change the study protocol during data collection to include the use of (larger) incentives, more call-backs, abbreviated questionnaires, differing contact or data collection methods, and the like (see, e.g., Dillman et al. 2009). Most of the changes introduced will add burden to the project budget as well as the length of the data collection. What if the project budget is not large enough to handle these more intensive modes of data collection for all nonrespondents? Nonrespondent subsampling is a proposed method to quantify differences between initial (phase 1) respondents and nonrespondents, to lower nonresponse bias, and to increase the number of study participants.

Example 17.3 (Potential bias due to nonresponse). This simple example illustrates why a subsample of nonrespondents should be selected if it is feared that they may be different from the set of initial respondents. Suppose the population can be divided into two strata—one stratum contains cases that respond to the initial phase of data collection and the other stratum includes cases that do not. Denote the proportions of the population in the two strata by W_1 and $W_2 = 1 - W_1$ and the population means by $\bar{y}_{U1}$ and $\bar{y}_{U2}$, respectively. The population mean is $\bar{y}_U = W_1\bar{y}_{U1} + W_2\bar{y}_{U2}$. A simple random sample is selected and only cases in stratum 1 respond (by definition). If the population mean is estimated by the sample mean, $\bar{y}_1$, then the expected value of $\bar{y}_1$ is $\bar{y}_{U1}$, i.e., $E(\bar{y}_1) = \bar{y}_{U1}$. Now, assume that $\bar{y}_{U2} = k\bar{y}_{U1}$. The relative bias (*relbias*) of $\bar{y}_1$ as an estimator of $\bar{y}_U$ is calculated as

$$relbias(\bar{y}_1) = \frac{E(\bar{y}_1) - \bar{y}_U}{\bar{y}_U} \qquad (17.1)$$
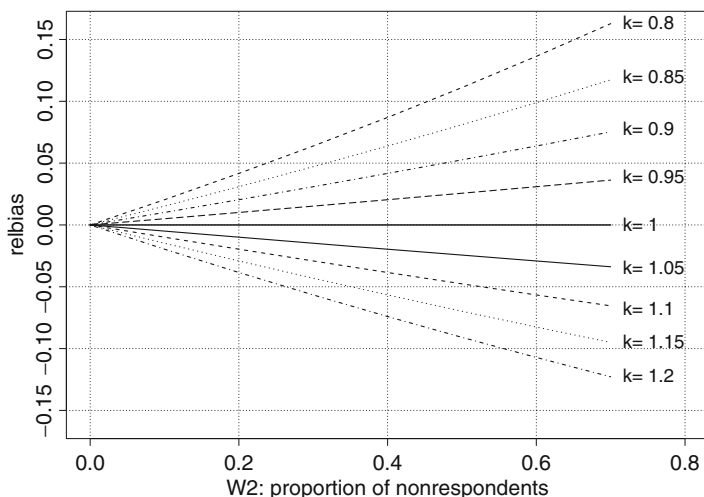
Fig. 17.2: Relationship of the relbias of an estimated population mean to the means of respondents and nonrespondents.

and for this example is easily found to be

$$relbias\left(\bar{y}_1\right) = \frac{W_2\left(1-k\right)}{1-W_2\left(1-k\right)}.$$

Figure 17.2 graphs the *relbias* of the mean of the respondents versus the nonresponse proportion for values of k ranging from 0.8 to 1.2. The *relbias* can be either positive or negative depending on whether the mean of the nonrespondents is less $(k < 1)$ or more $(k > 1)$ than that of the respondents. The absolute value of the *relbias* increases as the proportion of nonrespondents increases and as the value of k becomes farther from 1. Since the mean of the nonrespondents is unknown, the only symptom of potential bias in this example is the proportion of nonrespondents in the sample. ∎

Although there is no bias in Example 17.3 if $k = 1$, as long as the response rate is not zero, it is natural to be apprehensive that estimates are biased any time that there is nonresponse. Selecting a subsample of nonrespondents is one way to try and get representation of that group and to avoid nonresponse bias. The example above is oversimplified because there is probably some randomness to whether a given unit responds, implying that a population cannot be cleanly classified into respondent and nonrespondent strata. However, the example is realistic enough to show that nonresponse bias is something to worry about.

A study that includes a nonrespondent subsampling, also known as a *nonresponse follow-up study* (NRFU) or *double sampling for nonresponse* (DSNR), involves the selection of a random subsample of phase 1 nonrespondents. Often the study team will use different, more expensive data collection methods than used for the first phase with the goal of

obtaining complete cooperation from the phase 2 subsample. Provided that the number of phase 2 respondents is sizeable, in theory researchers are able to test for differences in the phase 1 respondents and nonrespondents, as well as reduce the nonresponse bias. The definition of "sizeable" is based on the associated power calculations to determine this detectable difference (see Chap. 4).

Figure 17.3 contains a pictorial representation of this type of two-phase design. Comparing this figure with Fig. 17.1, we are able to see the differences between double sampling for stratification and double sampling for nonresponse. As before, a phase 1 sample of size $n_{(1)}$ is selected from a sampling frame using a set of auxiliary variables $\mathbf{X}$. Questionnaire responses $(\mathbf{y}_{11}, \mathbf{y}_{12}, \ldots)$ and other auxiliary information $(\mathbf{z})$ are collected from $n_{(1)R}$ respondents during the first phase, leaving $(n_{(1)} - n_{(1)R}) > 0$ sample cases with no interview data. Given that $n_{(1)R} > 0$, the project team needs to justify the inclusion of a second phase that could extend data collection. The reasons may include an insufficient respondent sample size $n_{(1)R}$ for the analytic goals or an indication of sizeable (estimated) nonresponse bias using $\mathbf{X}$. If this analysis suggests that another phase would not be cost effective, then the study analysis file is finalized with only the phase 1 responses. Conversely, if it is determined that a second phase can be conducted with the available funds and is needed to meet the analytic goals, the statistician randomly selects a subsample of $n_{(2)}$ cases using the complete set of phase 1 nonrespondents. This design generally includes auxiliary information used in the phase 1 sample design as well as useful paradata obtained during phase 1 data collection such as number of contacts for each case. The phase 2 sample is then fielded typically with a different data collection protocol (e.g., mode, incentive, abbreviated questionnaire) than implemented in the first phase. The phase 1 and phase 2 respondent data are then combined to produce the analysis data file of $n_R = n_{(1)R} + n_{(2)R}$ records from which population estimates are produced.

Double sampling for stratification and double sampling for nonresponse are similar in that the phase 1 data collection produces stratifying information used for the second-phase sample design. However, the mathematics needed to analyze the two designs is different. Doubling sampling for stratification assumes that, for a given phase 1 sample, the same strata would always be formed. If response is treated as random, then the split between respondents and nonrespondents in a given set of phase 1 units will vary. Examples of the doubling sampling for stratification design include the identification of adults with a rare medical condition or households containing children within a certain age range. Response status is the primary stratifier for the latter design. Put in terms of the double sampling label, the phase 2 sampling rate for the phase 1 respondent stratum is 1 (i.e., sampled with certainty for the study) and the sampling for the phase 1 nonrespondent stratum is generally less than 1. See Sect. 17.5.2 for methods to determine the sampling rates.
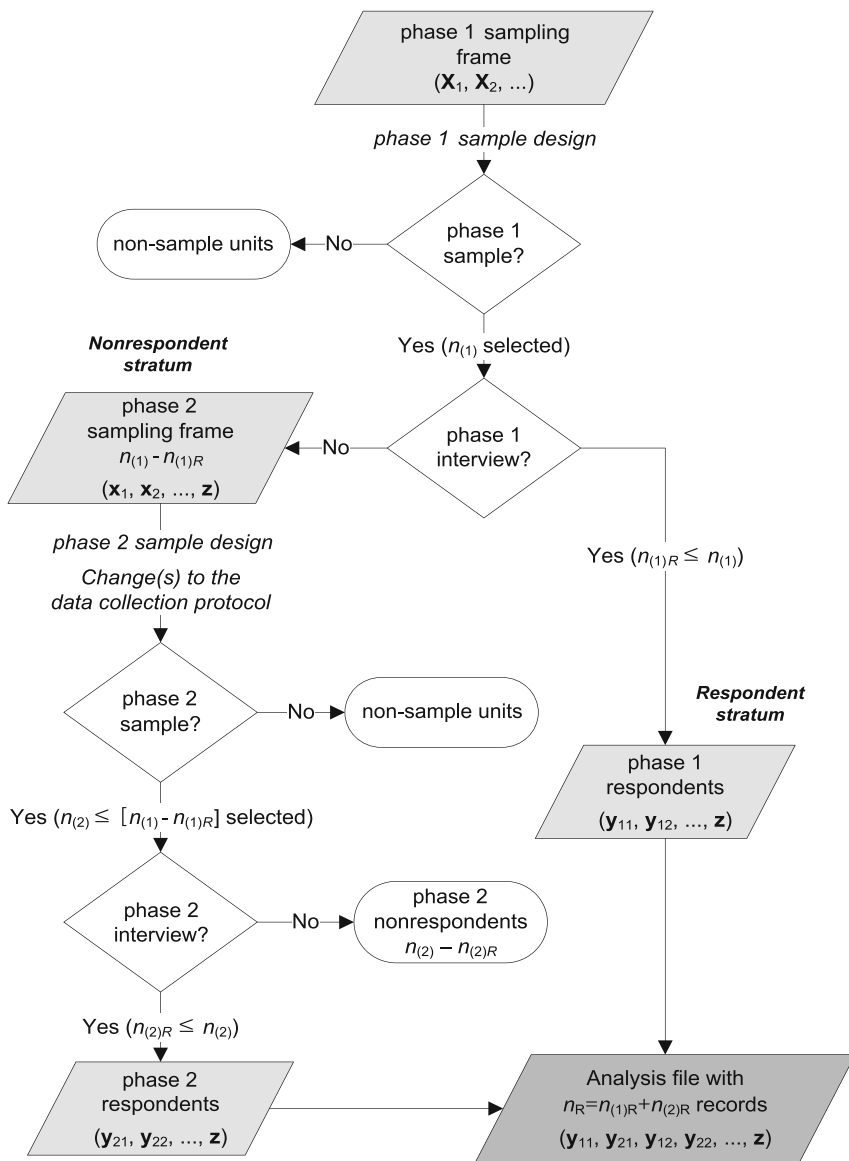
Fig. 17.3: Transition of sample cases through the states of a survey under a double sampling for nonresponse design.

DSNR designs of the kind depicted in Fig. 17.3 are included in many studies. Four examples are provided below:

1. The *Tenth Anniversary Gulf War Veterans Health Survey* was conducted to estimate the prevalence of certain adverse health conditions in US

military personnel who served in the 1991 Persian Gulf War (Singh et al. 2004, 2005). After obtaining a low response rate from a mail survey using a stratified simple random sample, 1,000 nonrespondents were randomly subsampled for a telephone follow-up. The nonrespondent sample size was determined by the available project funds and the sample allocated to the phase 1 strata to minimize a set of variance constraints.

2. The *General Social Survey*[6] tracks "societal change in the United States" and facilitates comparisons with other countries through a shared set of questions. Some question sets are static (referred to as the core modules), while other modules are newly introduced to capture data on timely issues. The instrument is primarily administered in a face-to-face setting to a national sample of adults ages 18 and over randomly selected through a complex, multistage design. Nonrespondent subsampling has been a GSS design component since approximately the mid-1990s to increase the respondent pool and to lower nonresponse bias.[7]

3. The American Community Survey (ACS) is an on-going, national household survey conducted by the US Census Bureau.[8] ACS staff collect interview data first by mail and then by telephone for nonrespondents who have not returned a completed questionnaire. Finally, a subsample of nonrespondents and households not contacted through the other modes (e.g., mail returned undelivered) is chosen for an in-person visit.

4. The European Social Survey (ESS)[9] is implemented in over 30 countries with the goal of evaluating cross-sectional changes in social attitudinal and behavioral patterns within and across these nations. In 2006, four countries implemented a nonresponse survey (ESS-NRS) to estimate nonresponse bias levels and correlates for the ESS (Matsuo et al. 2010). For example, the Belgian ESS-NRS sampled all nonrespondents at the doorstep immediately upon receiving a survey refusal and requested information only for an abbreviated seven-question instrument. Unlike the two-stage Belgian survey, the researchers on the Norwegian ESS-NRS implemented a two-phase design by subsampling ESS nonrespondents at different rates based on the nonresponse severity category. The Norwegian ESS sample was selected from a population registry with a one-stage systematic design.

Nonresponse is the primary motivation for two-phase designs with a nonrespondent follow-up or subsample. In the next section, we briefly review a specialized design that has become popular in our recent history.

---

[6] http://www3.norc.org/GSS+Website

[7] http://www.irss.unc.edu/odum/jsp/content_node.jsp?nodeid=83

[8] http://www.census.gov/acs/www/

[9] http://www.europeansocialsurvey.org/

17.2.3 Responsive Designs

The term *responsive design* was coined by Groves and Heeringa (2006) to describe a particular type of survey that attempts to uncover and implement the best combination of survey conditions (incentives, mode, contact time, etc.) to maximize participation. The best survey conditions could include, for example, a mail survey with a 5-dollar incentive (referred to as a pre-incentive) followed by a reminder postcard, or a telephone interview after 8pm during a weekday. As hinted here, the best conditions likely differ for various groups of people.

Responsive designs employ two or more survey conditions. As shown in Fig. 17.4 for a two-phase responsive design, the $n_{(1)}$ phase 1 sample cases are uniquely assigned to one of the $D_{(1)}(\geq 1)$ survey conditions, i.e., $n_{(1)} = \sum_{d=1}^{D_{(1)}} n_{(1)d}$, where $n_{(1)d}$ is the number of phase 1 cases given survey condition d. The assignment of cases to conditions may be random or may be informed by prior research. If $D_{(1)} = 1$, then all cases receive the same survey condition. The phase 1 interview is conducted, resulting in $n_{(1)R} = \sum_{d=1}^{D_{(1)}} n_{(1)dR}$ respondents with $n_{(1)dR}$ representing the number of respondents from the dth survey condition and a total of $n_{(1)} - n_{(1)R}$ nonrespondents. Note that Groves and Heeringa (2006) and many other researchers call this the first phase of the study whether or not they plan to subsample for the second phase. For example, consider a survey that will use either mail/Web or telephone to collect responses where the literature is not suggestive of the preferred mode. The $D_{(1)} = 4$ survey conditions could be: (i) a mail questionnaire sent with a small incentive; (ii) a mail questionnaire that includes the option of completing the interview on the Web along with a promised incentive upon completion; (iii) a telephone interview where the sample member was initially mailed an incentive along with information about the study; and (iv) a partial interview conducted by phone with the remainder being completed on the Web and a promised incentive.

Information gained in the first phase such as paradata or screener responses along with sampling frame values are used to create the set of $D_{(2)} = 1$ phase 2 survey conditions. If $D_{(1)}$ is large by construction, then the phase 2 set may be a subset of the cost-effective phase 1 conditions. The phase 1 information is also used to assign the phase 2 nonrespondent subsample to the conditions. Using the $D_{(1)} = 4$ example above, results from the first phase might suggest that the "partial telephone interview" survey condition is more cost effective than the other conditions developed for the study. Then, either all or a subsample of the phase 1 respondents would be administered with this condition with a goal of improving participation. This analysis is discussed in more detail after we have an understanding of the larger picture. Respondents from the phase 2 interview are combined with the phase 1 respondents to form the analysis file.
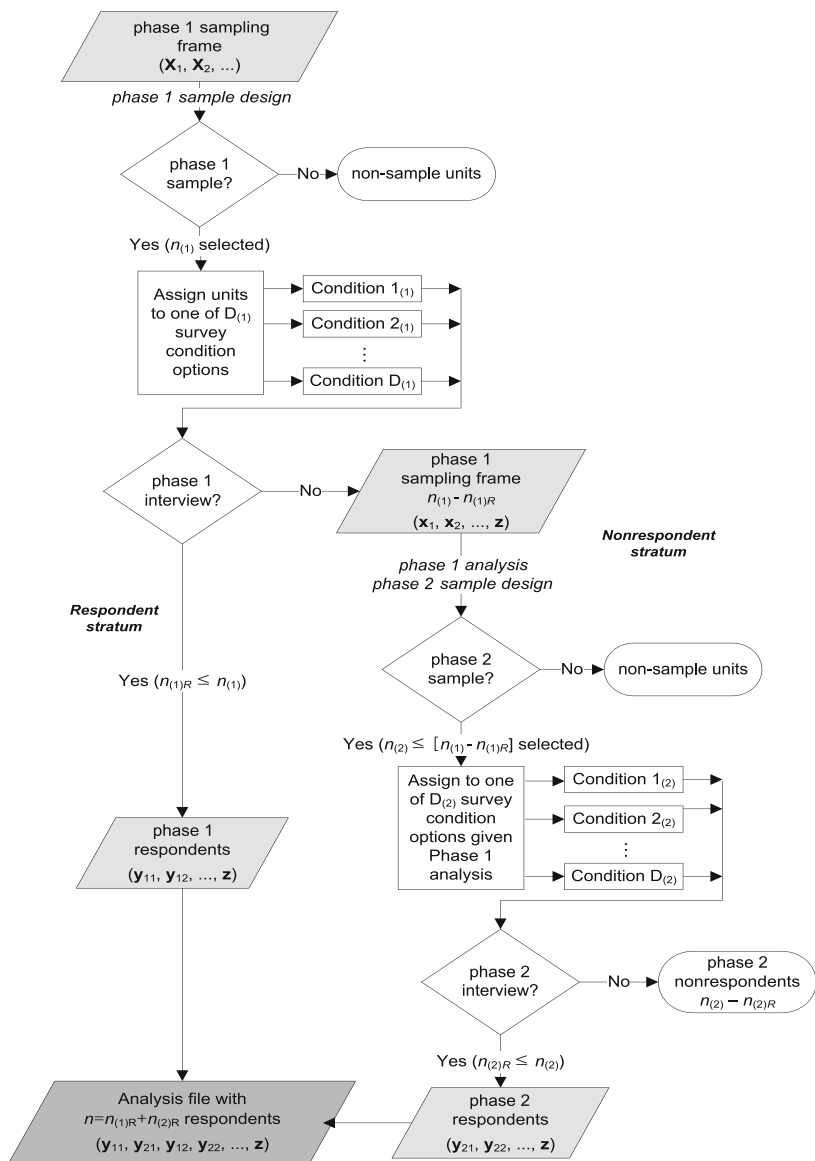
Fig. 17.4: Flow of sample cases through a simulated two-phase responsive design.

Now that we have reviewed the overall picture for this type of design, we can venture back to the phase 1 analysis briefly mentioned earlier. During data collection, various measures of completion, quality, and cost are typically monitored. Measures of completion may include rates of completed cases (e.g., interviews, biospecimens), the projected response rates given the flow
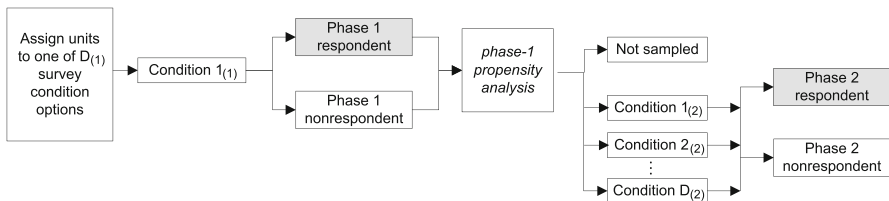
Fig. 17.5: Flow of responsive-design sample cases assigned to survey condition $1_{(1)}$ in phase one.

of completed cases by important reporting subgroups, and the probability of obtaining a response from the remaining nonrespondents (response propensities) given sampling frame data and possibly paradata. Quality measures, discussed in detail in Chap. 18, may include nonresponse bias analysis and estimates of precision for a set of important study variables again by important reporting subgroups. Measuring cost efficiency of the "essential survey conditions" for the current phase is the third prong of the analysis. This may include, along with the response propensities, an analysis that suggests which phase 1 condition is best suited for a certain set of phase 1 nonrespondents. For example, as shown in Fig. 17.5, nonrespondents in phase 1 condition $1_{(1)}$ may be assigned to one of $D_{(2)}$ conditions based on this analysis. The results may also suggest the characteristics included in the phase 2 conditions such as the size of the increased incentive amount. Once the project statistician has compiled the analysis, he/she uses a set of predefined decision rules to determine (i) when the phase 1 conditions are no longer cost effective, (ii) that more data collection *is* required to meet the analytic objectives, and (iii) if sampling or data collection features need to be revisited. At this point, a new phase of the study design is introduced.

Many would agree that most surveys have some responsive design characteristics. For example, interviewers use a variety of techniques to solicit cooperation from sample members, and increasing levels of incentive can be used given the history of refusal for a given sample unit. However, here we reserve the label "responsive design" for NRFU studies with varying survey conditions in at least the second phase (i.e., $D_{(2)} > 1$). One survey that fits this definition and is the first cited as a responsive design is the National Survey of Family Growth (NSFG), Cycles 6 and higher (Groves and Heeringa 2006).[10]

The NSFG is sponsored by the National Center for Health Statistics, United States Department of Health and Human Services, and was initially designed to collect information on fertility and health for the noninstitutionalized population of women aged 15–44 years selected through an area probability sample. Beginning with Cycle 6, a corresponding sample of males aged

---

[10] http://www.cdc.gov/nchs/nsfg.htm

15–44 years was also selected to obtain estimates on fatherhood and involve-
ment with their children. As described in Axinn et al. (2011) and Groves and
Heeringa (2006), the Cycle-6 responsive design was implemented in three
phases that include both subsampling of cases and changes to the data col-
lection protocol. Subsampling strata were developed based on results from
response propensity models. In addition, periodic analyses of paradata and
important study estimates were conducted throughout the data collection
phase with a "dashboard" system (see, e.g., Lepkowski et al. 2010) in an
attempt to predict a cost-effective point in the data collection window to
change study phases.

17.2.4 General Multiphase Designs

We conclude this section with a brief discussion of a general multiphase
design before moving on to the particulars of multiphase survey weights. As
suggested by our discussion at present, subsampling (and changes to the
initial set of essential survey conditions) can occur multiple times within the
data collection window. Figure 17.6 shows the general set-up for a multi-
phase design with an unlimited number of phases. The number of phases is
naturally limited by time and project funds. For some the number of contacts
especially after an initial refusal may be limited by an oversight committee
(e.g., Institutional Review Board or IRB) who, among other responsibil-
ities, protects sample members from coercion and excessive participant
burden.

17.3 Survey Weights

17.3.1 Base Weights

The form of the base (or design) weights for a multiphase design follows
the recipe described in Chap. 13 for multistage designs. The first-phase base
weights $d_{(1)0k}$ are calculated to reflect the sampling design as if the design
contained only a single phase. For example, with a stratified two-stage design,
the unconditional base weight for the kth (second-stage) element is

$$d_{(1)0k} = \pi_{(1)hi}^{-1} \pi_{(1)k|hi}^{-1} , \qquad (17.2)$$

i.e., the inverse selection probability for the ith cluster in stratum h $\left(\pi_{(1)hi}^{-1}\right)$
multiplied by the inverse selection probability for the kth element in cluster
hi, conditional on the cluster being sampled in the first stage $\left(\pi_{(1)k|hi}^{-1}\right)$. The
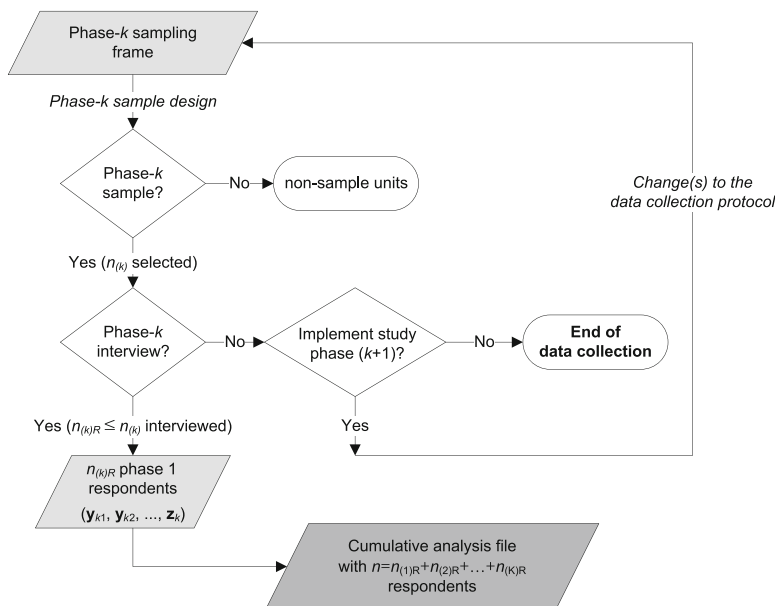
Fig. 17.6: Transition of sample cases through the states of a survey under a general multiphase design.

subscript in parentheses denotes the associated phase of sampling, i.e., "(1)" represents probabilities of selection associated with phase one. The analytic units may be selected from strata within each cluster; we suppress the additional second-stage stratum indicator from the notation only for simplicity. This two-stage notation used here implies that the elements are selected only with respect to the selection of cluster hi and no other clusters in the sample. Therefore, the independence and invariance properties are preserved so that the "two-stage" label is appropriate. This differs from the second-phase design as discussed below.

The second-phase base weights are calculated conditional on the first-phase outcome, the third conditional on the second, and so on. The *unconditional* base weight for the kth second-phase sample unit has the following general form:

$$d_{(2)0k} = d_{(1)0k}\pi_{(2)k|(1)}^{-1} \ , \tag{17.3}$$

where $d_{(1)0k}$ is the first-phase base weight defined in Eq. (17.2) and $\pi_{(2)k|(1)}$ is the phase 2 selection probability for the kth unit conditional on the phase 1 information. In other words, the subscripted "|(1)" says that the phase 2 sample was randomly chosen from the frame generated by the phase 1 sample. Note that expression (17.3) indicates that the kth unit is selected in *both* phases, regardless of the type of multiphase design. The phase 2 sample design may include stratification and random selections within

multiple stages resulting in a complex algorithm for constructing $\pi_{(2)k|(1)}$. The notation in expression (17.3) remains somewhat simplistic for this general discussion. The following examples will provide the design-specific forms for the weight components.

Example 17.4 (Weights for a two-phase stsrs/stsrs design). Consider a study similar to the EPIC survey mentioned in Sect. 17.2.1. The analysis plan developed at the start of the project states that the study will examine factors associated with a high quality of life among cancer patients with moderately high levels of pain experienced during treatment. The sampling plan dictates that all cancer treatment centers within a region of the country are to be selected with certainty.[11]

Having defined the sampling strata, cancer patients are randomly chosen for the study using sampling rates defined from historic patient-enrollment statistics:

$$\pi_{(1)h} = n_{(1)h} / N_{(1)h},$$

where $n_{(1)h}$ is the number of sample members selected from treatment center h ($h=1, \ldots, H$), $N_{(1)h}$ is the number of new patients expected to enter center h for treatment, and $d_{(1)0hk} = \pi_{(1)h}^{-1}$ is the associated phase 1 base weight. The newly recruited $n_{(1)}$ ($= \sum_{h=1}^{H} n_{(1)h}$) sample members are asked to complete a short (screener) questionnaire. Within their third week of treatment, the sample members are administered a 20 m questionnaire to collect health information as well as inputs for scales on pain threshold and quality of life. The following indicator variable was generated from the second phase 1 interview results:

$$\delta_{(1)dk} = \begin{cases} 1 & \text{if sample member } k \text{ has characteristic } d, \\ 0 & \text{otherwise,} \end{cases}$$

where domain 1 ($d = 1$) represents those patients experiencing at least a moderately high level of pain (high score on the pain scale) who also report a high quality of life (high score on the quality of life scale), and domain 2 ($d = 2$) identifies patients with moderately high levels of pain but experience a low quality of life. Sample members who have no or low levels of pain constitute the third subgroup within our phase 1 sample and are not eligible for the phase 2 study. The project statistician determined the sampling rates

---

[11] The cancer treatment centers under this design are treated as the first-stage strata for point and variance estimation because all and not a sample of centers are included in the study. As an aside, mathematical modelers would label this "cancer treatment variable" a fixed effect. If a subset of centers were randomly chosen, then these first-stage clusters (PSUs) usually would be modeled with random effects.

within each of the two second-phase strata to meet the comparative analysis objectives. Therefore, the conditional phase 2 selection probabilities are

$$\pi_{(2)hd|(1)} = n_{(2)hd}/n_{(1)hd},$$

where $n_{(2)hd}$ $(n_{(2)hd} \leq n_{(1)hd})$ is the number of cancer patients randomly selected for the phase 2 sample out of the total number of phase 1 patients identified as members of stratum d, i.e., $n_{(1)hd} = \sum_{k \in s_{(1)h}} \delta_{(1)dk}$, through the phase 1 questionnaire. A second questionnaire is then administered to the $n_{(2)}$ $(= \sum_{h=1}^{H} \sum_{d=1}^{2} n_{(2)hd})$ phase 2 subsample members to gather detailed information on issues related to social support, religiosity, and home life.

Combining the two selection probabilities, the unconditional phase 2 base weight for the kth sample member is

$$d_{(2)0hdk} = d_{(1)0hk} \, \pi_{(2)hd|(1)}^{-1} = \frac{N_{(1)h}}{n_{(1)h}} \frac{n_{(1)hd}}{n_{(2)hd}}.$$

As a quality check, the sum of the base weights for all phase 1 sample members, regardless of their phase 2 eligibility status, equals the total number of cancer patients receiving treatment in cancer treatment centers within the designated region of the country, i.e.,

$$\sum_{k \in s_{(1)}} N_{(1)h}/n_{(1)h} = \sum_{h=1}^{H} \sum_{k \in s_{(1)h}} N_{(1)h}/n_{(1)h} = \sum_{h=1}^{H} N_{(1)h} = N_{(1)}.$$

The sum of the phase 2 base weights estimates the total number of cancer patients with moderate to high levels of pain during treatment. ∎

Example 17.5 (Weights for an stsrs NRFU). Let's recast the study design from Example 17.4 as one that includes two interviews (screener and *combined* pain threshold and social support questionnaire). All $n_{(1)}$ phase 1 sample members responded to screener, but only a portion, $n_{(1)R}/n_{(1)}$, responded to the larger interview. An initial analysis conducted by the project statistician identified a significant difference in the estimates for the $n_{(1)R}$ respondents and $n_{(1)\bar{R}}$ $(= n_{(1)} - n_{(1)R})$ nonrespondents calculated from the screener and administrative record data, i.e., there is the potential for nonresponse bias (see Sect. 13.5). Consequently, $2H$ conditional subsampling rates were developed for the treatment centers with the following form:

$$\pi_{(2)hd|(1)} = \begin{cases} n_{(2)hd}/n_{(1)hd} & \text{if } d = \bar{R}, \text{ a phase1 } NR \text{ stratum in center } h, \\ 1 & \text{if } d = R, \text{ the phase1 } R \text{ stratum,} \end{cases}$$

where NR stands for nonrespondents and R for respondents. An abbreviated version of the phase 1 instrument is developed and administered to the subsample of $n_{(2)}$ patients selected from the $n_{(1)\bar{R}}$ nonrespondents. Using the phase 1 base weights defined in the previous example, the

resulting unconditional base weights included on the preliminary analysis file of $n_{(1)} = n_{(1)R} + n_{(2)}$ records are

$$
\begin{aligned}
d_{(2)0hdk} &= d_{(1)0hk}\, \pi_{(2)hd|(1)}^{-1} \\
&= \begin{cases}
\frac{N_{(1)h}}{n_{(1)h}}\, \frac{n_{(1)hd}}{n_{(2)hd}}, & \text{phase 1 } NRs \text{ selected for phase 2 ,} \\
0, & \text{phase 1 } NRs \text{ } not \text{ selected for phase 2,} \\
\frac{N_{(1)h}}{n_{(1)h}}, & \text{phase 1 } Rs.
\end{cases}
\end{aligned}
$$

∎

17.3.2 Analysis Weights

Adjustments such as those for nonresponse are applied to the base weights to form the final multiphase analysis weights. The adjustment procedures for creating the adjustment factors follow the material presented in Chaps. 13 and 14 for multistage designs. This information is summarized in a series of steps below for the two-phase and the NRFU designs. The disposition response status for cases in these two designs was shown in Figs. 17.2 and 17.3, respectively. The section is concluded by a brief discussion of weights for general multiphase designs.

Two-Phase Designs. The first step in constructing analysis weights for phase 2 respondents within a two-phase design is to develop the phase 1 analysis weights. In keeping with the notation used in Chap. 13, the phase 1 weights $w_{(1)k}$ take the following form:

$$
w_{(1)k} = d_{(1)0k}\, a_{(1)1k}\, a_{(1)2k}\, g_{(1)k} \tag{17.4}
$$

where $d_{(1)0k}$ is the base weight calculated as the inverse probability of selection for the phase 1 sample, $a_{(1)1k}$ is an adjustment for unknown eligibility status, $a_{(1)2k}$ is an adjustment for nonresponse applied to the base weight adjusted for unknown eligibility ($d_{(1)1k} = d_{(1)0k}a_{(1)1k}$), and $g_{(1)k}$ is the calibration adjustment made to the adjusted base weights, $d_{(1)2k} = d_{(1)0k}\, a_{(1)1k}\, a_{(1)2k}$, using controls generated from the population. Any respondents who are classified as ineligible for the phase 2 study are removed from the sampling frame. The weights and the phase 1 questionnaire data are used in the selection of the second-phase sample.

After data have been collected from the responding phase 2 sample members, the final unconditional, phase 2 analysis weight is similarly constructed as follows:

$$
w_{(2)k} = w_{(1)k}\, a_{(2)0k|(1)}\, a_{(2)1k|(1)}\, a_{(2)2k|(1)} \tag{17.5}
$$

where $w_{(1)k}$ is the final phase 1 weight specified in Expression (17.4), $a_{(2)0k|(1)}$ is the adjustment for subsampling conditional on the responses from phase one, and $a_{(2)1k|(1)}$ and $a_{(2)2k|(1)}$ are adjustments for unknown eligibility and nonresponse strictly associated with the phase 2 sample. An adjustment for unknown eligibility might apply if, for example, some sample members are unlocatable during the phase 2 data collection window. Expression (17.5) could also include a second calibration adjustment, $g_{(2)k|(1)}$, that includes controls associated with the population of interest as well as estimated counts tabulated from the phase 1 responses and the final phase 1 weights. Through a general regression estimator (GREG; see Chap. 14), these calibration adjustments could be made simultaneously or sequentially. We discuss this issue in more detail in Sect. 17.3. The set of units that would be used for estimation would generally be the eligible respondents in phase 2, $s_{(2)R}$. Or if the control totals used for the phase 2 calibration include ineligibles, then the set of units used for analysis is $s_{(2)R} \cup s_{(2)IN}$ where $s_{(2)IN}$ represents the set of phase 2 sample units known to be ineligible.

NRFU Designs. As shown in Fig. 17.3, (NRFU) or double sampling for nonresponse studies differ from two-phase designs in that all or a portion of the data collected in phase 1 are also collected in the second phase. The final phase 1 weights

$$w_{(1)k} = d_{(1)0k}a_{(1)1k}g_{(1)k} \qquad (17.6)$$

are calculated by adjusting the base weight ($d_{(1)0k}$) for any unknown eligibility ($a_{(1)1k}$) and then calibrating to population control totals ($g_{(1)k}$) prior to selecting the subsample of $n_{(2)}$ ($< n_{(1)\bar{R}}$) nonrespondents for follow-up in phase 2. After the phase 2 data collection has ended, the input weights for the phase 2 sample members ($w_{(1)k}$) are corrected for subsampling ($a_{(2)0k|(1)}$) and any unknown eligibility ($a_{(2)1k|(1)}$) or nonresponse ($a_{(2)2k|(1)}$), resulting in adjusted weights of the form:

$$d_{(2)2k} = w_{(1)k}a_{(2)0k|(1)}a_{(2)1k|(1)}a_{(2)2k|(1)}, \qquad (17.7)$$

Note that the unconditional base weight for the phase 2 sample cases is defined by the first two components above, i.e.,

$$d_{(2)0k} = w_{(1)k}a_{(2)0k|(1)}. \qquad (17.8)$$

The study analysis file will contain questionnaire values for the $n_{(1)R}$ phase 1 respondents as well as the $n_{(2)R}$ ($\leq n_{(2)}$) phase 2 respondents. A final calibration adjustment ($g_{(2)k|(1)}$), using population control totals, may be applied to all respondent records to generate the unconditional phase 2 analysis weights

$$w_{(2)k} = \begin{cases} w_{(1)k}a_{(2)0k|(1)}a_{(2)1k|(1)}a_{(2)2k|(1)}g_{(2)k|(1)}, \\ \qquad\qquad \text{for the phase 1 nonrespondents selected for phase 2} \\ w_{(1)k} \times g_{(2)k|(1)}, \quad \text{for the phase 1 respondents} \end{cases}$$

$$(17.9)$$

for components defined for expression (17.7). Unlike the other two-phase design, the weights, in general, are calibrated to only population controls and not to data collected in the first phase.

General Multiphase Designs. Weights for survey designs with more than two phases (see, e.g., Figs. 17.4 and 17.3) follow the prescription described above. Appropriate weight adjustments are applied to each subsample and associated outcomes from the phase-specific data collection efforts. A word of caution needs to be voiced at this point—as the number of phases of subsampling increases, so does the variability in the resulting analysis weights. The Sect. 14.4 discussion highlighted the potential damage that widely varying weights can do to survey estimates, making the precision so poor that results from the study may not be published. In addition to weight smoothing techniques, optimal subsampling procedures have also been developed to minimize this problem. We discuss a few of these later in Sect. 17.5.

Example 17.6 (Base weights for an stsrs NRFU). Suppose a project statistician develops the allocation for a single-stage stratified design, assuming that the response rate will be sufficient to obtain at least 1,000 respondents as required by the analysis plan. A proportional allocation is chosen because of the limited information available on the relative sizes of the population variances across four sampling strata. The following table contains the population counts and sample size by stratum, along with the estimated number of respondents given a response rate of at least 61 %. The overall sampling fraction, used in each of the strata, was 0.022 or 1,650 / 75,000, resulting in an equal probability sample with identical base weight of 45.45 (1/0.22).

Data collection proceeded with the randomly selected sample of cases. How-

Stratum	Pop size (N_h)	Sample size (n_h)	Estimated respondents (r_h)
1	12,882	284	173
2	27,332	601	366
3	18,361	404	246
4	16,425	361	220
Overall	75,000	1,650	1,005

ever, the perceived conservative response rate of 61 % proved too optimistic—data were collected from only 972 sample members (57.7 unweighted response rate). A subsequent analysis determined that the existing set of responses produced inadequate precision based on the approved analysis plan, and that continuation of the current study protocol would be ineffective. The team received the appropriate authorization to introduce a more expensive data collection methodology than initially implemented including a monetary

incentive for participation. Because of the increased data collection cost and dwindling project funds, the team determined that this second phase could be implemented on at most 120 phase 1 nonrespondents. The sampling statistician drew an *srs* of equal size within the phase 1 sampling strata.

	Phase 1			Phase 2		
Stratum	Sample size	Respondents	Response rate	Frame size	Sample	Sampling fraction
h	$n_{(1)h}$		(%)	$N_{(2)h}$	$n_{(2)h}$	(%)
1	284	227	79.9	57	30	52.6
2	601	270	44.9	331	30	9.1
3	404	222	55.0	182	30	16.5
4	361	253	70.1	108	30	27.8
Overall	1,650	972	58.9	678	120	17.7

The phase-specific base weights, equal within strata, were constructed as follows:

	Phase 1		Phase 2		
Stratum	Respondents	Base weight	Sample size	Subsample	Base weight
h	$n_{(1)h}$	$d_{(1)0h}$	$n_{(2)h}$	$a_{(2)0h(1)}$	$d_{(2)0h}$
1	227	45.4	30	1.9	86.2
2	270	45.5	30	11.0	501.8
3	222	45.4	30	6.1	275.7
4	253	45.5	30	3.6	163.8
Overall	972		120		

∎

Example 17.7 (Nonresponse-adjusted analysis weights for an stsrs NRFU). Continuing with Example 17.6, 45 of the 120 sample members participated in the phase 2 data collection. Even though only a 37.5 % unweighted, conditional phase 2 response rate was achieved (= 45/120), a total of 1,017 completed cases were processed for the final analysis file.

Weights for the 45 respondents were adjusted for nonresponse specific to the second phase using a standard weighting class adjustment (see Sect. 13.5) with the unconditional base weights defined in expression (17.5) within each of four design strata. The summarized results are provided below.

Stratum	Phase 1 respondents		Phase 2 respondents				
	Sample size	Adjusted weight	Sample size	Base weight	Sub- sample	Non- response	Adjusted weight
h	$n_{(1)h}$	$w_{(2)k}$	$n_{(2)h}$	$d_{(1)0h}$	$a_{(2)0h}$	$a_{(2)2h}$	$w_{(2)k}$
1	227	45.4	4	45.4	1.9	7.50	646.4
2	270	45.5	22	45.5	11.0	1.36	684.2
3	222	45.4	12	45.4	6.1	2.50	689.3
4	253	45.5	7	45.5	3.6	4.29	702.0
Overall	972		45				

The adjusted weight for the phase 1 respondents is approximately the same as the base weight, i.e., $d_{(1)0h} \doteq 45.4$ for each respondent. The adjusted weight for the phase 2 respondents is the phase 1 base weight multiplied by the conditional subsampling and nonresponse weights, $a_{(2)0h}$ and $a_{(2)2h}$, respectively. Notice that, even though the weights for the phase 1 respondents are all equal and the weights for the NRFU cases are similar to each other (ranging from 646.4 to 702), there is quite a bit of weight variation in the full set of responding cases (45.4 to 702). Whether the nonresponse follow-up is statistically efficient or not should be evaluated using the data collected in the survey. ∎

17.4 Estimation

Now that the analysis weights have been defined for multiphase design, we turn to the construction of point and variance estimates produced from the study data.

17.4.1 Descriptive Point Estimation

The form of the descriptive point estimates such as means and totals from multiphase designs follows the same formula specified for other designs. The following examples are discussed in the literature for two-phase design, primarily to demonstrate the efficiency (or inefficiency) of certain variance estimates. (*i*) The double expansion estimator (DEE; Kott and Stukel 1997) for a population total in a two-phase design is calculated as

$$\hat{t}_{(2)y} = \sum_{k \in s_{(2)}} w_{(2)k}\, y_k \qquad (17.10)$$

where $w_{(2)k}$ is the unconditional phase 2 weight defined in expression (17.9); y_k is the characteristic of interest; and $s_{(2)}$ signifies the second-phase sample (and consequently any design characteristics such as stratification and clus-

tering). An associated estimator, found by Kott and Stukel (1997) to have smaller variances than the DEE, is known as the reweighted expansion estimator (REE). Expressed as a mean for a two-phase design, the REE has the following form:

$$\bar{y}_{(2)} = \frac{1}{\hat{N}_{(2)}} \sum_{h=1}^{H} \hat{N}_{(1)h} \bar{y}_{(2)h}, \tag{17.11}$$

where $\hat{N}_{(2)} = \sum_{k \in s_{(1)}} \sum_{k \in s_{(2)}} w_{(2)k}$, the estimated number of units in the target population using weights generated from expression (17.9); $h = 1, \ldots, H$ indexes the mutually exclusive groups, such as strata, associated with the two-phase sample design; $\hat{N}_{(1)h} = \sum_{k \in s_{(1)}} w_{(1)k}$, the estimated number of units in stratum h with $w_{(1)k}$ defined in expression (17.6); and $\bar{y}_{(2)h} = \left(\sum_{k \in s_{(2)h}} w_{(2)k} \right)^{-1} \sum_{k \in s_{(2)h}} w_{(2)k} y_k$ is the estimate of the population mean in stratum h based on the phase 2 sample and the unconditional weights.

Bias of the Estimators

One final note before we move to variance estimators. The unconditional design-based expectation of a multiphase estimator is evaluated as a function of the conditional expectations within each successive phase of the design (see, e.g., Casella and Berger 2002, Theorem 4.4.3). The formula for a two-phase design is

$$E\left(\hat{\theta}\right) = E_{(1)} \left[E_{(2)} \left[\hat{\theta} \middle| (1) \right] \right]. \tag{17.12}$$

Working first with the innermost bracketed term, the expectation of the generic point estimator $\hat{\theta}$ is evaluated with respect to the second-phase sample design conditional on the components of the phase 1 design, e.g., the sample size is fixed. The expectation of the resulting estimator is then evaluated treating the phase 1 sample selection as random. Using this same partitioning, the expression above can be expanded to more than two phases by evaluating the conditional expectation and substituting into the previous equation, e.g.,

$$E_{(2)} \left[\hat{\theta} \middle| (1) \right] = E_{(2)} \left[E_{(3)} \left[\hat{\theta} \middle| (2) \right] \middle| (1) \right].$$

In addition to quantifying the theoretical bias of an estimator, this equality is useful in building variance estimators as shown in the next section.

Example 17.8 (Expectation of a Two-phase Estimator of a Total). Consider the two-phase estimator $\hat{\theta} = \hat{t}_{(2)y}$ of the population total $t_y = \sum_{k \in U} y_k$, where $\hat{t}_{(2)y} = \sum_{k \in s_{(2)}} d_{(2)0k} \; y_k$ and the unadjusted base weight $d_{(2)0k}$ defined as $d_{(2)0k} = \pi_{(1)k}^{-1} \pi_{(2)k|(1)}^{-1}$, a function of the unconditional phase 1 and

conditional phase 2 selection probabilities as defined in expression (17.3). The unconditional expectation of the population estimate is

$$E\left(\hat{t}_y\right) = E_{(1)}\left[E_{(2)}\left[\sum_{k\in s_{(2)}} d_{(2)0k}\, y_k\middle|\,(1)\right]\right]$$

$$= E_{(1)}\left[E_{(2)}\left[\sum_{k\in U} d_{(2)0k}\, I_{(2)}\, y_k\middle|\,(1)\right]\right].$$

where $I_{(2)}$ is a binary variable to identify the population units selected for the phase 2 sample. Note that selection in the phase 2 sample is a function of the phase 1 selection and conditional phase 2 selection, i.e., $I_{(2)} = I_{(1)} \times I_{(2|1)}$. Substituting the formula for $d_{(2)0k}$ and $I_{(2)}$, we have

$$E\left(\hat{t}_y\right) = \sum_{k\in U}\left(\pi_{(1)k}^{-1}\pi_{(2)k|(1)}^{-1}\right) E_{(1)}\left[I_{(1)}\right] E_{(2)}\left[I_{(2)|(1)}\middle|\,(1)\right] y_k$$

$$= \sum_{k\in U}\left(\pi_{(1)k}^{-1}\pi_{(2)k|(1)}^{-1}\right)\pi_{(1)k}\pi_{(2|1)k}y_k$$

$$= t_y.$$

Therefore, $\hat{t}_y$ is an unbiased estimator of t_y. This assumes that the frame used for the phase 1 sample covers the whole population. If there are frame undercoverage problems and nonresponse at either or both phases, unbiasedness depends on assumptions about the nonresponse and coverage mechanisms along with the properties of the steps (like calibration) taken to correct those problems. ∎

17.4.2 Variance Estimation

Variance estimation techniques for general surveys were covered in Chap. 15. Once augmented, the same approach is useful for multiphase designs discussed in this section. As with the previous chapter, this section includes a discussion of (Taylor Series) linearization and replication variances.

Linearization Variance Estimators

The procedure for developing a multiphase variance estimator for a generic point estimate, $\hat{\theta}$, begins with the derivation of the unconditional formula (see, e.g., Casella and Berger 2002, Theorem 4.4.7):

$$V\left(\hat{\theta}\right) = V_{(1)}\left[E_{(2)}\left[\hat{\theta}\middle|\,(1)\right]\right] + E_{(1)}\left[V_{(2)}\left[\hat{\theta}\middle|\,(1)\right]\right] \tag{17.13}$$

where, similar to expression (17.12), $E_{(1)}$ and $V_{(1)}$ are the theoretic expectation and variance with respect to the phase 1 sample design and $E_{(2)}\left[\hat{\theta}\middle|\,(1)\right]$

and $V_{(2)}\left[\hat{\theta}\middle|(1)\right]$ are the corresponding quantities for the phase 2 sample design conditional on the realized phase 1 sample. Evaluating expression (17.13) for the complete design results in a Taylor series linearization variance component that accounts for the random selection within each phase.

Example 17.9 (Variance of an estimated total for a generic two-phase design). Consider the estimator of a total, $\hat{t}_{(2)y} = \sum_{k \in s_{(2)}} d_{(2)0k} \; y_k$, discussed in Example 17.8 desired for a two-phase survey where the sample design for each phase is classified only in generic terms. The first thing to note is that, similar to the decomposition for a two-stage design, the estimator can be expressed as a function of phase 1, phase 2 and population terms. Namely,

$$\hat{t}_{(2)y} - t_y = \left(\hat{t}_{(1)y} - t_y\right) + \left(\hat{t}_{(2)y} - \hat{t}_{(1)y}\right)$$
$$= \hat{D}_{(1)} + \hat{D}_{(2)}$$

where $\hat{t}_{(1)y} = \sum_{k \in s_{(1)}} d_{(1)0k} \; y_k$, the population estimate using the phase 1 data; and $\hat{D}_{(1)} = \left(\hat{t}_{(1)y} - t_y\right)$ and $\hat{D}_{(2)} = \left(\hat{t}_{(2)y} - \hat{t}_{(1)y}\right)$ represent the error associated with the phase 1 and phase 2 random sampling designs, respectively. The variance of $\hat{t}_{(2)y}$ is then evaluated as

$$Var\left(\hat{t}_{(2)y}\right) = Var\left(\hat{t}_{(2)y} - t_y\right)$$
$$= V_{(1)} \left\{ E_{(2)} \left[\hat{t}_{(2)y} - t_y\middle|(1)\right] \right\} + E_{(1)} \left\{ V_{(2)} \left[\hat{t}_{(2)y} - t_y\middle|(1)\right] \right\}.$$

Working with the innermost formulas, we have

$$E_{(2)}\left[\hat{t}_{(2)y} - t_y\middle|(1)\right] = E_{(2)}\left[\hat{D}_{(1)} + \hat{D}_{(2)}\middle|(1)\right] = \hat{D}_{(1)}$$

assuming that the phase 2 estimator, $\hat{t}_{(2)y}$, is an unbiased estimator of the phase 1 estimator, $\hat{t}_{(1)y}$ (conditional on the phase 1 sample). The phase 2 variance, given the phase 1 sample, is $V_{(2)}\left[\hat{t}_{(2)y} - t_y\middle|(1)\right] = V_{(2)}\left[\hat{t}_{(2)y}\middle|(1)\right]$. Thus, the two-phase sampling variance for the estimated total is defined as

$$V\left(\hat{t}_{(2)y}\right) = V_{(1)}\left[\hat{t}_{(1)y}\right] + E_{(1)}\left\{V_{(2)}\left[\hat{t}_{(2)y}\middle|(1)\right]\right\}. \qquad (17.14)$$

Consequently, the variance of the two-phase estimator, $\hat{t}_{(2)y}$, will be larger than the variance of the population total tabulated as if all the data were obtained in the first phase. But, as noted earlier, the point of doing a second phase is to either use methods that would be too costly to apply to all first phase units or to target the sample in a way that would not be feasible using a single-phase sample.

A general formulation of expression (17.14) is given in Result 9.3.1 of Särndal et al. (1992) and recast as follows using the notation specific to this chapter:

$$V\left(\hat{t}_{(2)y}\right) = \sum\sum_{s_{(2)}} \frac{\Delta_{(1)kl}}{\pi_{(2)kl}} \frac{y_k}{\pi_{(1)k}} \frac{y_l}{\pi_{(1)l}}$$

$$+ \sum\sum_{s_{(2)}} \frac{\Delta_{(2)kl|(1)}}{\pi_{(2)kl|(1)}} \frac{y_k}{\pi_{(2)k}} \frac{y_l}{\pi_{(2)l}} \qquad (17.15)$$

where $\pi_{(2)kl}$ and $\pi_{(2)kl|(1)}$ are the unconditional and conditional phase 2 inclusion probabilities, respectively; $\pi_{(2)kl} = \pi_{(1)kl}\ \pi_{(2)kl|(1)}$, the unconditional joint inclusion probability defined as the product of the phase 1 probability and the conditional phase 2 probability; $\Delta_{(1)kl} = \pi_{(1)kl} - \pi_{(1)k}\ \pi_{(1)l}$ and $\Delta_{(2)kl|(1)} = \pi_{(2)kl|(1)} - \pi_{(2)k|(1)}\ \pi_{(2)l|(1)}$, the phase 1 and (conditional) phase 2 joint inclusion probabilities for units k and l, respectively. An explicit formula for the population variance in expression (17.14) is defined once the sample designs in each phase are specified. In general, a design consistent sample estimate of the variance is obtained by substituting the sample estimates for the population values. The next example provides such a specialization for one type of two-phase design. ∎

Example 17.10 (Variance for an srs/stsrs two-phase design, Example 17.9 continued). Consider the double sampling for stratification design where the phase 1 design is an *srs* of size $n_{(1)}$ and a second-phase random sample of size $n_{(2)} = \sum_{h=1}^{H} n_{(2)h}$ is selected from the newly identified strata. First, note that the DEE estimated population total from Example 17.9 can be reexpressed as a function of estimated stratum means:

$$\hat{t}_{(2)y} = \sum_{h=1}^{H} \sum_{k \in s_{(2)h}} d_{(2)0k}\ y_k = \sum_{h=1}^{H} \sum_{k \in s_{(2)h}} \left(\frac{N}{n_{(1)}} \frac{n_{(1)h}}{n_{(2)h}}\right) y_k$$

where h indexes the strata identified from the phase 1 sample, $w_{(1)h} = \left(n_{(1)h}/n_{(1)}\right)$, and $\hat{\bar{y}}_{(2)h} = \sum_{k \in s_{(2)h}} \left(y_k/n_{(2)h}\right)$. Therefore, expression (17.15) is evaluated as

$$V\left(\hat{t}_{(2)y}\right) = N^2 \left[\left(1 - f_{(1)}\right) \frac{S^2}{n_{(1)}} + E_{(1)} \left(\sum_{h=1}^{H} w_{(1)h}^2 \left(1 - f_{(2)}\right) \frac{s_{(1)h}^2}{n_{(2)h}}\right)\right]$$

with the phase-specific sampling fractions, $f_{(1)} = n_{(1)}/N$ and $f_{(2)h} = \left(n_{(2)h}/n_{(1)h}\right)$; the population sampling variance, $S^2 = (N-1)^{-1} \sum_{k \in U} (y_k - \bar{y})^2$, and mean, $\bar{y} = N^{-1} \sum_{k \in U} y_k$; and the phase 1 sampling variance

$$s_{(1)h}^2 = \left(n_{(1)h} - 1\right)^{-1} \sum_{k \in s_{(1)h}} \left(\hat{y}_{(1)k} - \hat{\bar{y}}_{(1)h}\right)^2$$

with mean $\hat{\bar{y}}_{(1)h} = n_{(1)h}^{-1} \left(\sum_{k \in s_{(1)h}} \hat{y}_{(1)k}\right)$ where $\hat{y}_{(1)k} = d_{(1)0k}\ y_k$. The second term in $V\left(\hat{t}_{(2)y}\right)$ is left as an expectation because $w_{(1)h}$ and $n_{(2)h}$ are random variables. Estimates of the variance components due to first, and second-phase sampling are given by Rao (1973) and Särndal et al. (1992, Sect. 9.4) as

$$\hat{V}_1 = \frac{1 - f_{(1)}}{n_{(1)}} \left[\sum_{h=1}^{H} w_{(1)h} \left(1 - \frac{1}{n_{(2)h}} \frac{n_{(1)} - n_{(1)h}}{n_{(1)} - 1} \right) s_{(2)h}^2 \right.$$
$$\left. + \frac{n_{(1)}}{n_{(1)} - 1} \sum_{h=1}^{H} w_{(1)h} \left(\hat{\bar{y}}_{(2)h} - \hat{\bar{y}}_{(2)} \right)^2 \right]$$

and $\hat{V}_2 = \sum_{h=1}^{H} w_{(1)h}^2 \left(1 - f_{(2)h} \right) \frac{s_{(2)h}^2}{n_{(2)h}}$. Adding these and assuming that the first-phase sampling fraction, $f_{(1)}$, is small and that $\left(n_{(1)h} - 1 \right) / \left(n_{(1)} - 1 \right) \doteq w_{(1)h}$, the estimated variance of $\hat{t}_{(2)y}$ is

$$v \left(\hat{t}_{(2)y} \right) \cong N^2 \left[\frac{1}{n_{(1)}} \sum_{h=1}^{H} w_{(1)h} \left(\hat{\bar{y}}_{(2)h} - \hat{\bar{y}}_{(2)} \right)^2 + \sum_{h=1}^{H} w_{(1)h}^2 \left(\frac{s_{(2)h}^2}{n_{(2)h}} \right) \right],$$

where $\hat{\bar{y}}_{(2)} = \sum_{h=1}^{H} w_{(1)h} \hat{\bar{y}}_{(2)h}$, $\hat{\bar{y}}_{(2)h} = \sum_{k \in s_{(2)h}} y_k / n_{(2)h}$, and $s_{(2)h}^2 = \left(n_{(2)h} - 1 \right)^{-1} \sum_{k \in s_{(2)h}} \left(\hat{y}_{(1)k} - \hat{\bar{y}}_{(1)h} \right)^2$. ■

We give a numerical illustration of *srs*/*stsrs* two-phase sampling for stratification in Example 17.12.

It should be apparent from the previous examples that as the sample designs become more complex, so does the variance estimator. This also holds true with an increase in the number of sampling phases. Software for computing two-phase variance estimates to date is limited and currently non-existent for multiphase designs. Because researchers must develop and program the formula, many instead turn to replicate variances that are in general easier to implement.

Replication Variance Estimators

Replicate variance estimators, such as the jackknife, are applicable to a variety of sample designs and estimators. As discussed in Chap. 15, the variance estimate is a function of the deviation of A replicate estimates, $\hat{\theta}_{(2)}^{(r)}$, calculated with the replicate weights, $w_{(2)}^{(r)}$, from an aggregate value, $\hat{\theta}_{(2)}^{(*)}$,

$$v \left(\hat{\theta}_{(2)} \right) = \frac{1}{C} \sum_{r=1}^{A} \left(\hat{\theta}_{(2)}^{(r)} - \hat{\theta}_{(2)}^{(*)} \right)^2,$$

where C is a constant that depends on the method of replication (jackknife, *BRR*, or bootstrap). The aggregate value, $\hat{\theta}_{(2)}^{(*)}$, could be generated as the average of the replicate estimates, $\hat{\theta}_{(2)}^{(*)} = R^{-1} \sum \hat{\theta}_{(2)}^{(r)}$, or using the complete phase 2 data estimate and the original analysis weight (full-sample weight).

Kott and Stukel (1997) and Kim and Yu (2011) discuss the theoretical and empirical properties of the jackknife variance estimator while Fuller (1998) studied balanced repeated replication. The work of Kim et al. (2006) covered replication variances but did not focus on a specific replicate form. To date, no research has been implemented on the bootstrap estimators for multiphase designs.

The generic process for creating the two-phase replicate weights is summarized in three steps:

(1) Identify a sample unit or group of units (for a delete-a-group variance estimator) from the analysis data file and set their analysis weights to zero. The remaining units are classified as the replicate subsample.
(2) Next, adjust the base weights for the subsampling implemented in step 1 to form the replicate base weight.
(3) Finally, reapply any weight adjustments used to produce the full-sample weights to calculate the final replicate analysis weight.

For a jackknife variance the three steps are repeated R times so that each unit is excluded once to form a replicate weight. A random group variance estimate is similar in that units are randomly grouped and all or a random subset of the groups are removed to form the replicate weights.

As implemented with a single-phase sample (e.g., see Valliant 1993, 2004), the weight adjustments such as nonresponse and calibration are newly applied to each replicate so that the variance will capture any additional random properties other than the sampling process. For example, if the phase 2 weights are calibrated to a set of phase 1 estimates, then new estimated controls are calculated for each replicate prior to this adjustment. Additional replicate adjustments have been investigated including one to capture the variation in the phase 1 estimated controls (Fuller 1998) and a non-negligible phase 1 finite population correction (Korn and Graubard 1999[12], Lee and Kim 2002) as well as a correction for bias that is inherent in the jackknife (Kim and Yu 2011).

The three steps above are further specialized for attributes of the phase 1 design. For example, if the phase 1 sample design is clustered and the jackknife is used, then clusters are deleted to form replicates and the weights for all units in a deleted phase 1 cluster are set to zero in step 1. As noted in Kim et al. (2006), if a consistent variance estimator is available for the estimates under the phase 1 design, then this property will hold for a multiphase extension.

Example 17.11 (Variance for two-phase design with cluster sampling). Consider a study that requires estimates generated from psychological tests administered in person. Data from an initial battery of questions (phase one) were used to ensure that the in-person sample (phase two) includes female

---

[12] Also, see the panel discussion on the appropriate uses of an *fpc* at http://web.cos.gmu.edu/~wss/wss070328paper.pdf.

head of households with relatively good health and quality of life (QoL) as well as those with physical maladies or poor QoL. A sample of m_h area segments is selected from region h ($h = 1, 2, 3, H = 4$) for phase one; all female head of households in the sample segments were included in an initial telephone interview.

The second-phase *srs* was selected from three strata ($G=3$) within each segment with strata defined by the categorization of the phase 1 scale score of high, medium, and low QoL. All unknowns, i.e., nonrespondents, were grouped in the "medium" stratum based on prior research. It should be noted that the clustered nature of the phase 1 design not only enabled a cost-effective methodology for conducting the phase 2 in-person interview, but also permitted in-person follow-up with phase 1 nonrespondents.

The project statistician chose to calculate replicate two-phase weights for the analytic data file. The construction of the base weights began before implementation of the phase 2 data collection. Suppose the full-sample base weight defined for cluster i in phase 1 stratum h is $d_{(1)0hi}$. When cluster (st) is deleted, the replicate jackknife base weight was created as

$$d_{(1)0hi}^{(st)} = \begin{cases} d_{(1)0hi} & h \neq s, \ i \neq t, \\ d_{(1)0hi} \times (m_h/m_h - 1) & h = s, \ i \neq t, \\ 0 & h = s, \ i = t. \end{cases}$$

Note that all members of the cluster were selected with certainty in the phase 1 sample, so that $w_{(1)hi}$ was applied to all sample units in cluster hi. The conditional base weight for the second-phase interview was defined as $d_{(2)0hij|(1)} = n_{(1)hig}/n_{(2)hig}$ for sample member j in nested stratum hig ($j \in s_{hig}$), where $n_{(1)hig}$ is the number of eligible, phase 1 sample members in stratum g within cluster hi (i.e., the number of phase 1 female head of households) and $n_{(2)hig}$ is the corresponding phase 2 sample size. Combining the two, the unconditional phase 2 replicate base weight for phase 2 sample member j in phase 2 stratum hig was calculated as $d_{(2)0hij}^{(st)} = d_{(1)0hi}^{(st)} \times d_{(2)0hij|(1)}$. ∎

The jackknife variance estimator for either the DEE or REE has a negative bias which, at least in some cases, is negligible (Kim and Yu 2011). For example, if clusters are selected by *srswor*, the bias is small when the first phase sampling fraction is small. When the first phase sampling fraction is not negligible, Kim and Yu (2011) give methods of construction replication estimates that remove the bias.

A Comment on Complex Multiphase Designs

Literature to date primarily focuses on what can be classified as "single-stage" two-phase designs. These designs include, for example, a single-stage of selection (phase one) followed by a second single stage of selection (phase two) such as the *srs/stsrs* design discussed in Example 17.10. This paradigm

follows the one used to develop the original theory that results in variance components that are familiar to those with at least one sampling course. However, literature that includes linearization variance estimators for more complex designs that include clustering in the first phase, such as with Example 17.11, is limited. More research has been done on replication. Review of a few methodology reports, such as the NSFG-V (Potter et al. 1998) and the American Time Use Survey (Bureau of Labor Statistics 2012), indicates that the study (unconditional phase 2) weights should be used with standard software that accounts for the (phase 1) clustering. This suggests that only the first component in expression (17.14) is accounted for in the variance estimate. If true, the implication of ignoring the "within phase 2" variance component (i.e., potential underestimation of the variance) requires additional research.

17.4.3 Generalized Regression Estimator (GREG)

The use of strong auxiliary data to select a sample is part of the justification for multiphase surveys. If not already available, this important information is collected in earlier phases for subsequent ones. The generalized regression estimator or GREG discussed in Chap. 15 taps auxiliary information to both reduce bias and variance of the estimates. In this section, GREGs produced from multiphase designs are discussed.

GREG Weights and Point Estimation

Kim and Yu (2011), along with Särndal and Lundström (2005) and Särndal et al. (1992), discuss the benefits of regression estimators related to bias reduction and improved efficiency in precision estimates over the expansion estimators. The formula, reproduced from Chap. 14, for calculating a GREG for a population total is

$$
\begin{aligned}
\hat{t}_{yGREG} &= \hat{t}_y + \left(\mathbf{t}_x - \hat{\mathbf{t}}_x\right)^{'} \hat{\mathbf{B}} \\
&= \sum_{k \in s} \left[1 + \left(\mathbf{t}_x - \hat{\mathbf{t}}_x\right)^{'} \left(\mathbf{X}^T \mathbf{D} \mathbf{V}^{-1} \mathbf{X}\right)^{-1} \mathbf{x}_k / v_k\right] d_k y_k,
\end{aligned}
\tag{17.16}
$$

where $\mathbf{t}_x$ is the vector of control totals, $\hat{\mathbf{t}}_x$ is the associated vector of sample estimates, g_k is the term in brackets in the second line, and $w_k = g_k d_k$. The term g_k is sometimes referred to as the g-weight or calibration weight. Focusing on a two-phase design for convenience, we have a total of three weight calibration scenarios. Namely, the phase 2 weights are calibrated (i) to phase 1 estimated control totals only, (ii) to population control totals only, or (iii) to the phase 1 and population controls simultaneously.

Calibration to the phase 1 estimated control totals, scenario 1 above, should be considered for situations where there is no information (or unreliable estimates) for the target population of interest. With this procedure, the

(adjusted) base weights $d_{(2)k}$ for the phase 2 study respondents are calibrated to satisfy the constraints

$$\sum_{k \in s_{(2)R}} w_{(2)k}\, \mathbf{z}_{(2)k} = \hat{\mathbf{t}}_{(1)z}, \qquad (17.17)$$

where $\mathbf{z}_{(2)k}$ is a vector of variables collected in phase 1 and populated with information for the phase 2 respondents $(s_{(2)R})$; $w_{(2)k}$ is the resulting calibrated analysis weight; and $\hat{\mathbf{t}}_{(1)z} = \sum_{k \in s_{(1)R}} w_{(1)k}\, z_{(1)k}$ is the vector of estimated controls calculated from the phase 1 respondents. Thus, within expression (17.16), we have $\mathbf{t}_x = \hat{\mathbf{t}}_{(1)z}$ and $\hat{\mathbf{t}}_x = \hat{\mathbf{t}}_{(2)z} = \sum_{k \in s_{(2)R}} d_{(2)k}\, \mathbf{z}_{(2)k}$.

A survey sample can also be calibrated to estimates from an independent survey. The Health and Retirement Study,[13] for example, calibrates its weights to family composition distributions estimated from the Current Population Survey,[14] which is an independent household survey. For calibration to estimated controls, the ideal situation is to have highly precise survey estimates from a study that is much larger than the survey requiring calibration (Dever and Valliant 2010).

When the group of phase 1 respondents is insufficient in size to produce efficient estimates for calibration, researchers must follow, if possible, scenario 2 and adjust the weights to population controls. Here the (adjusted) base weights $d_{(2)k}$ for the phase 2 study respondents are calibrated to satisfy the constraints

$$\sum_{k \in s_{(2)R}} w_{(2)k}\, \mathbf{x}_{(2)k} = \mathbf{t}_x, \qquad (17.18)$$

where $\mathbf{x}_{(2)k}$ is a vector of variables known for the study sample and containing data for the phase 2 respondents; $w_{(2)k}$ is the resulting final calibrated analysis weight; and $\mathbf{t}_x$ is the vector of population totals as defined in expression (17.16). For studies with a NRFU, this more traditional style of weight calibration might be appropriate especially if the phase 1 and phase 2 respondents appear to differ on characteristics relevant to the study.

Särndal and Lundström (2005, Chap. 8) discuss the use of control totals estimated from the first-phase sample initially or simultaneously with population control totals (calibration scenario 3). We refer to these as *sequential calibration* and *simultaneous calibration*, respectively, and adapt their discussion for a two-phase design. A two-step sequential calibration for the phase 2 respondents is produced as follows:

Step 1: Calibrate the (adjusted) base weights for the phase 2 study respondents to estimated totals from phase 1 as defined for scenario 1 above. The

---

resulting calibrated weight is defined as $w_{(2)k|(1)}$ and labeled as "conditional" because it relies on the sample drawn for phase one.

<u>Step 2</u>: Simultaneously calibrate the adjusted weights from step 1 for the study respondents, $w_{(2)k|(1)}$, to satisfy the constraints $\sum_{k \in s_{(2)R}} w_{(2)k} \mathbf{x}_k^* = \hat{\mathbf{t}}_x^*$, where $\mathbf{x}_k^* = \left(\mathbf{z}_{(2)k}, \ \mathbf{x}_{(2)k} \right)^T$ and $\hat{\mathbf{t}}_x^* = \left(\hat{\mathbf{t}}_{(1)z}, \mathbf{t}_x \right)^T$ with the component vectors defined for expressions (17.17) and (17.18), respectively. While preserving the estimated-control calibration constraint specified in step 1, the step 2 procedure additionally forces the respondents estimates to equal the population controls.

For K-phase studies with $K{=}2$, the control total vector could be expanded to include estimates from the K-1 design phases, $\hat{\mathbf{t}}_x^* = \left(\hat{\mathbf{t}}_{(1)z}, \hat{\mathbf{t}}_{(2)z}, \ldots, \hat{\mathbf{t}}_{(K-1)z}, \mathbf{t}_x \right)'^T$. Under this scenario, the second-phase respondent base weight, adjusted for any sample loss in the first phase, would be calibrated to population totals and estimates calculated from the phase 1 questionnaire responses. To date the benefits of one scenario over another are not well defined and warrant further research.

GREG Linearization Variance Estimators

GREG variance estimation is premised on a (linear) model containing the auxiliary information that effectively represents the population characteristic being estimated. An effective model is one that leads to an estimator with a smaller variance than would be obtained by not using the auxiliaries. Said differently, this model will result in small residuals, the deviation from the value of y and the estimated value of y, i.e., $e_k = y_k - \hat{y}_k$ is small. The residual is the key component to the GREG variance estimator.

The GREG estimator of a total, $\hat{t}_{yGREG}$ given in Eq. (17.16), is written in terms of a single-phase design using only population-based auxiliary information $\mathbf{X}$. As discussed above, auxiliary information, an important component in calibration, is obtained from the various phases of the design along with any population sources. If we consider the simultaneous weight calibration to population controls and to controls estimated from the first phase, then the GREG two-phase variance estimator is a function of two estimated residuals:

Phase 1: $e_{(1)k} = y_k - \mathbf{x}_{(2)k}^T \hat{\mathbf{B}}_{(1)}$

$$\hat{\mathbf{B}}_{(1)} = \left(\sum_{k \in s_{(2)}} \frac{w_{(2)k} \mathbf{x}_{(2)k} \mathbf{x}_{(2)k}^T}{\sigma_{(1)k}^2} \right)^{-1} \sum_{k \in s_{(2)}} \frac{w_{(2)k} \mathbf{x}_{(2)k} y_k}{\sigma_{(1)k}^2}$$

Phase 2: $e_{(2)k} = y_k - \mathbf{x}_k^{*T} \hat{\mathbf{B}}_{(2)}$

$$\hat{\mathbf{B}}_{(2)} = \left(\sum_{k \in s_{(2)}} \frac{w_{(2)k} \mathbf{x}_k^* \mathbf{x}_k^{*T}}{\sigma_k^2} \right)^{-1} \sum_{k \in s_{(2)}} \frac{w_{(2)k} \mathbf{x}_k^* y_k}{\sigma_k^2},$$

where $\mathbf{x}_k^* = \left(\mathbf{z}_{(2)k}, \ \mathbf{x}_{(2)k}\right)^T$, a vector of auxiliary values for the respondent sample taken from the phase 2 and phase 1 data collections, respectively; and the models for each phase are specified with an assumed variance of σ_k^2 and $\sigma_{(1)k}^2$. Note if either model variance is assumed to be a constant value (i.e., $\sigma_k^2 \equiv \sigma^2$ for all $k \in s_{(1)}$), then the quantities in the numerator and denominator cancel, thereby producing a more familiar form of the regression coefficients. As discussed in Sect. 9.7 of Särndal et al. (1992), the associated variance estimator takes the general form

$$
\begin{aligned}
v\left(\hat{t}_{(2)y}\right) = {} & \sum \sum\nolimits_{s_{(2)}} \hat{\Delta}_{(1)kl} \left(g_{(1)k}\hat{e}_{(1)k} \ g_{(1)l}\hat{e}_{(1)l}\right) \\
& + \sum \sum\nolimits_{s_{(2)}} \hat{\Delta}_{(2)kl} \left(g_{(2)k}\hat{e}_{(2)k} \ g_{(2)l}\hat{e}_{(2)l}\right),
\end{aligned}
\tag{17.19}
$$

where $\hat{e}_{(1)k} = e_{(1)k}/w_{(1)k}$ and $\hat{e}_{(2)k} = e_{(2)k}/w_{(2)k}$ are estimated model residuals; the phase-specific g-weights are

$$
g_{(1)k} = 1 + \left(\mathbf{t}_x - \hat{\mathbf{t}}_x\right)' \left(\sum\nolimits_{k \in s_{(2)}} \frac{w_{(1)k}\mathbf{x}_{(2)k}\mathbf{x}_{(2)k}'}{\sigma_{(1)k}^2}\right)^{-1} \frac{\mathbf{x}_{(2)k}}{\sigma_{(1)k}^2} \text{ and}
$$

$$
g_{(2)k} = 1 + \left(\hat{\mathbf{t}}_x^* - \hat{\mathbf{t}}_{(2)x}^*\right)' \left(\sum\nolimits_{k \in s_{(2)}} \frac{w_{(2)k}\mathbf{x}_k^*\mathbf{x}_k^{*T}}{\sigma_{(2)k}^2}\right)^{-1} \frac{\mathbf{x}_k^*}{\sigma_{(2)k}^2}.
$$

The estimated covariances for units selected in the phase 1 and phase 2 sample designs are designated as $\hat{\Delta}_{(1)kl}$ and $\hat{\Delta}_{(2)kl}$. As in Sect. 17.4.2, the first component in expression (17.19) is the phase 1 variance contribution and the second component is associated with the second phase.

GREG Replicate Variance Estimators

We make brief mention of the replication variance estimators for the GREG since they follow the same steps that are discussed previously. Mainly, any calibration introduced in the weights must be independently implemented within each replicate.

17.5 Design Choices

A basic design choice is whether to use more than one phase or stick to a single phase. The situations where multiphase sampling can profitably be used are discussed in Sect. 17.5.1. If two phases are used, how to allocate the sample to the phases must be determined. Section 17.5.2 covers sample size calculation when double sampling for stratification and when doing a nonresponse follow-up study.

17.5.1 Multiphase Versus Single Phase

Multiphase sampling involves additional complications and costs compared to a single-phase design. An administrative system must be developed to track the cases and distinguish ones assigned to different phases and their dispositions (i.e., response status) after each phase. Special programs must be written to translate the final dispositions into the analysis weights. As noted earlier and discussed in more detail in Sect. 17.6, few options exist with variance estimation software for multiphase designs at the date of this writing, and the current options have limited capabilities. These issues raise the question of whether using a multiphase design is worth the trouble. There are at least three important uses of multiphase designs that were mentioned in previous sections to distinguish them from single-phase designs:

(1) To improve precision of estimates
(2) To obtain target sample sizes in some analytic subgroups
(3) To attempt to reduce nonresponse biases through a NRFU study

Each of these is discussed below. Hansen et al. (1953a, chap. 11, Sect. 3) give another discussion of these uses and their efficiencies compared to other options.

Neyman (1938) introduced the idea that precision could be increased by collecting data in phase 1 to use as stratification variables for the phase 2 sample, or as covariates for phase 2 regression estimates. Hansen et al. (1953a) illustrate that if stratifying information can be collected in phase 1 that is highly effective in separating units into homogeneous groups, then double sampling for stratification with an optimal allocation to the phase 2 strata can produce variances of estimated means that are much less than would be obtained with no strata. For example, think of sampling businesses. If there is no information on business size, employee counts might be collected in the first phase and used to create strata for phase 2 sampling. Similarly, if an auxiliary variable that is highly correlated with the analysis variables can be collected in phase 1, then it can be used in constructing efficient regression estimators of a total using the phase 2 responses. Hansen and Tepping (1990) give an example of this in quality control of governmental welfare programs. Hansen et al. (1953a) give conditions under which there will be gains when using double sampling for stratification or regression estimation. The applications where these gains accrue are fairly specialized and are more likely to occur in surveys of businesses and institutions.

Kalton and Anderson (1986) discuss several techniques for sampling rare populations. Without good information on the prevalence within the population, they highlight several examples of mail screening questionnaires sent to households to identify adults with a particular characteristic (e.g., disability). Individuals may then be sampled at revised rates based on the screening information to obtain target sample sizes in the subgroups. When the goal is to obtain target sample sizes of certain groups and estimates with which to set

the sampling rates are unavailable, then there is little choice in the USA but to screen and subsample. Commercial lists used in address-based sampling may help target some domains of persons (e.g., race/ethnicity given the surname[15]) but the lists are currently incomplete and the estimated accuracy is not documented. In other countries, e.g., ones in Scandinavia, population registries allow very accurate identification of some types of people. But, they too are limited by what items are on the registry.

Thus, for uses (1) and (2) above, multiphase sampling is clearly useful and may be the only way to achieve the goals of some surveys. When following up nonrespondents, the decision is less clear-cut. In some surveys a NRFU study may be the only way to achieve a desired weighted response rate (see, e.g., AAPOR 2011; Singh et al. 2004) or a target sample size of respondents. This is especially true if the survey is being conducted under contract and the contract specifies the minimum response rate or sample size that is to be achieved. In some surveys the only hope of adding more respondents after the first phase is to change the mode of collection, start offering incentives to participate, or both for a subsample. In such cases the survey perforce becomes multiphase.

17.5.2 Sample Size Calculations

Sample size calculations for multiphase designs follow many of the techniques already discussed in this book. We discuss a few approaches below to orient the thought process starting with double sampling for stratification, then surveys with a nonresponse follow-up, and finally responsive designs.

Double Sampling for Stratification Designs

Sample size calculations with double sampling for stratification designs (and multiphase designs in general) are conducted using various approaches. The methods depend on whether population estimates for a key analytic variable are known by strata (i) during the design of the phase 1 study or (ii) only after phase 1 data are collected. Cochran (1977, Sect. 12.3) discusses an optimal allocation to the phase 2 strata by minimizing the variance subject to a linear cost model. We demonstrate the technique through an example.

Example 17.12 (Sample size calculation for srs/stsrs design with population estimates). Consider a mental health pilot study that will be conducted through a computer-assisted telephone interview (CATI). The instrument contains a set of psychological questions (VDF-14) to identify serious mental illness that have been validated within a clinical setting but not with CATI.

Though validated, the cost of conducting the complete study in a clinic setting is cost prohibitive and therefore this lower-cost option is being investigated as a viable alternative. A double sampling for stratification design was proposed where a subsample of CATI respondents (phase one) will be asked to participate in a second interview by a trained psychologist (phase two).

Phase 1 respondents will be grouped into one of four strata based on the mental health score produced as a linear combination of responses to the VDF-14. An equal number of persons will be assigned to each stratum ($W_h = 0.25$). Additionally, estimated proportions of serious mental illness by stratum (P_h) were calculated from a series of small clinical studies to assess the sensitivity of the CATI questions. The associated population variances were tabulated using the standard $P_h (1 - P_h)$ formula and included in the table below with the other information. The results (Neyman allocation) are provided in the final column and justified below.

Stratum	W_h	P_h	S_h^2	Neyman allocation
1	0.25	0.02	0.0196	31
2	0.25	0.12	0.1056	72
3	0.25	0.37	0.2331	107
4	0.25	0.54	0.2484	110
Overall		0.26	0.1936	320

Cochran (1977) and Neyman (1938) give the two-phase variance when phase 1 is a simple random sample, phase 2 is *stsrs*, and an optimal allocation to strata is used in the second phase. The sampling fractions at both stages are assumed to be negligible. The optimal proportion of the phase 2 sample to assign to stratum h for estimating the population mean is $n_{(2)h}/n_{(2)} = W_h S_h / \sum_h W_h S_h$. The formula for the variance of an estimated mean with this allocation is

$$V_{opt} = \frac{\sum_h W_h (P_h - P)^2}{n_{(1)}} + \frac{\left(\sum_h W_h S_h\right)^2}{n_{(2)}} = \frac{V_{(1)}}{n_{(1)}} + \frac{V_{(2)}}{n_{(2)}},$$

where $S_h = \sqrt{S_h^2}$. A phase 1 CATI interview is estimated to be $1/5^{th}$ of the cost associated with a phase 2 clinical interview. In particular, suppose that the linear cost model used in the cost-variance optimization takes the form $C = c_{(1)} n_{(1)} + c_{(2)} n_{(2)}$ where $c_{(1)} = \$10$, and $c_{(2)} = \$50$. If all interviews were conducted in person by a clinician, suppose that the study could only afford 400 interviews, i.e., $\$20,000/c_{(2)}$. Cochran (1977) then gives the following expression for the subsampling rate that minimizes the variance expression above:

$$\frac{n_{(2)}}{n_{(1)}} = \sqrt{\frac{V_{(2)}}{V_{(1)}} \bigg/ \frac{c_{(2)}}{c_{(1)}}}.$$

The formulas for the phase 1 and phase 2 sample sizes that minimize V_{opt} subject to a fixed total cost C are

$$n_{(1)} = \frac{C}{c_{(1)} + c_{(2)}\sqrt{K}},$$

$$n_{(2)} = n_{(1)}\sqrt{K},$$

where

$$K = \left(V_{(2)}/V_{(1)}\right) / \left(c_{(2)}/c_{(1)}\right).$$

Using the population estimates above, the variance components are calculated as $V_{(1)} = 0.0419$ and $V_{(2)} = 0.1307$ so that $n_{(2)}/n_{(1)} = 0.79$. Using the formulas above, the optimal phase 1 and phase 2 sample sizes are $n_{(1)} = 404$ and $n_{(2)} = 319$. The phase 2 sample size would then be distributed across the four strata with the Neyman allocation as shown in the table above. The size of the simple random sample with total cost C with each unit costing $c_{(2)}$ is $n_{srs} = C/c_{(2)}$. The variance of an *srs* of that size (neglecting an *fpc*) is $V_{srs} = S^2/n_{srs}$ where S^2 is the population unit variance. The gain, if any, from double sampling is V_{opt}/V_{srs}. In this example, $V_{opt}/V_{srs} = 1.06$. Although there is actually a small loss in the precision of the estimated population mean by using double sampling, the real goal is often to get certain sample sizes in the strata. If so, double sampling can accomplish that, and the Neyman allocation is probably not what is needed.

The R function, dub, in `PracTools` will compute the results for this example. Its inputs are:

c1	Cost per unit in phase 1
c2	Cost per unit in phase 2
Ctot	Total variable cost
Nh	Vector of stratum population counts or proportions
Sh	Vector of stratum population standard deviations
Yh.bar	Vector of stratum population means

The inputs and call to the function for this example are:

```
Wh <- rep(0.25,4)
Ph <- c(0.02,0.12,0.37,0.54)
Sh <- sqrt(Ph*(1-Ph))
c1 <- 10
c2 <- 50
Ctot <- 20000
dub(c1, c2, Ctot, Nh=Wh, Sh, Yh.bar=Ph)
```

∎

If no information about the characteristic of interest is available during the planning stage, then the statistician may use a variable that is highly

correlated (or believed to be) with the analysis variable and continue with the technique discussed above. Otherwise, the phase 2 sample size and allocation to strata are created only after data are analyzed from the first phase using procedures discussed in Part I of the text. Liu and Aragon (2000), for example, note that the design effect of the weights (i.e., unequal weighting effect) is minimized if a probability proportional to the phase 1 weight is used to draw the phase 2 sample.

Nonresponse Follow-Up Designs

We return to the NRFU application introduced in Sect. 17.2.2 and present an example based on Särndal et al. (1992, examples 15.4.4 and 15.4.5). Suppose that an initial *srswor*, $s_{(1)}$ is selected followed by an *srswor* subsample of the nonrespondents. In this uncomplicated situation, we can determine the sample sizes for the phases to either (i) minimize a relvariance for a fixed cost or (ii) minimize the cost for a fixed relvariance. Suppose that an initial sample of size $n_{(1)}$ is selected. There are $n_{(1)R}$ respondents and $n_{(1)NR}$ nonrespondents. The proportions of respondents and nonrespondents in the phase 1 sample are

$$p_{(1)R} = n_{(1)R}/n_{(1)} \text{ and } p_{(1)NR} = n_{(1)NR}/n_{(1)}.$$

A NRFU sample $s_{(2)}$ of $n_{(2)}$ units is selected by simple random sampling from the $n_{(1)NR}$ phase 1 nonrespondents. Data on the survey variables are collected on the initial respondents and on the participating units in the NRFU sample. Notice that this is different from two-phase applications where only respondent data collected in the second phase are used in estimation. The base weights for the sample units are

$$d_{(2)k} = \begin{cases} \frac{N}{n_{(1)}} & k \in s_{(1)R}, \\ \frac{N}{n_{(1)}} \frac{n_{(1)NR}}{n_{(2)}} & k \in s_{(2)}. \end{cases}$$

Some units in the phase 2 sample will also be nonrespondents, so that only $n_{(2)R}$ will respond. A nonresponse-adjusted weight, using an overall correction, is then

$$w_{(2)k} = \begin{cases} \frac{N}{n_{(1)}} & k \in s_{(1)R}, \\ \frac{N}{n_{(1)}} \frac{n_{(1)NR}}{n_{(2)}} \frac{n_{(2)}}{n_{(2)R}} & k \in s_{(2)R}. \end{cases} \qquad (17.20)$$

Using the weights in expression (17.20), the estimator of the population total of a variable y is

$$\hat{t}_{(2)y} = \sum_{s_{(1)R}} \frac{N}{n_{(1)}} y_k + \sum_{s_{(2)R}} \frac{N}{n_{(1)}} \frac{n_{(1)NR}}{n_{(2)R}} y_k \qquad (17.21)$$
$$= N \left[p_{(1)R} \, \bar{y}_{(1)R} + p_{(1)NR} \, \bar{y}_{(2)R} \right]$$

where $\bar{y}_{(1)R} = \sum_{s_{(1)R}} y_k/n_{(1)R}$, the unweighted mean of the phase 1 respondents, and $\bar{y}_{(2)R} = \sum_{s_{(2)R}} y_k/n_{(2)R}$, the unweighted mean for the $n_{(2)R}$ respondents in the subsample. The population mean is estimated by $\hat{\bar{y}} = \hat{t}_{(2)y}/N$. Note that this estimator does leave room for the possibility that the first- and second-phase respondents do represent groups whose population means are different, as in Example 17.3.

Assuming that response is a random process and that each sample unit independently has a probability θ of responding to phase 1, the numbers of respondents and nonrespondents, $n_{(1)R}$ and $n_{(1)NR}$, are random. Modifying the argument in Särndal et al. (1992, example 15.4.5) slightly, the variance of $\hat{\bar{y}}$, which is a special case of Eq. (17.15), can be found as

$$V\left(\hat{\bar{y}}\right) = \frac{1 - f_{(1)}}{n_{(1)}} S_{yU}^2 + E_{(1)} E_{RD}\left(p_{(1)NR}^2 \frac{1 - f_{(2)R}}{n_{(2)}} S_{y(1)NR}^2 \middle| s_{(1)}\right),$$

where E_{RD} is the expectation with respect to the phase 1 and 2 response distributions, $S_{y(1)NR}^2$ is the unit variance among phase 1 nonrespondents, and $f_{(2)} = n_{(2)R}/n_{(1)NR}$ is the responding fraction of the phase 1 nonrespondents. Because $n_{(1)NR}$ is random, we set the achieved second phase sampling fraction to a constant, $\nu = f_{(2)R}$, which will allow optimal values of $n_{(1)}$ and ν to be found. Note that ν includes both the initial subsampling rate of phase 1 nonrespondents and the proportion of the phase 2 subsample that responds. If the unit variance among nonrespondents is the same as the unit variance of all units, $S_{y(1)NR}^2 = S_{yU}^2$, the relvariance of the mean can be shown to be

$$CV^2\left(\hat{t}_{(2)y}\right) = \frac{CV_{yU}^2}{n_{(1)}}\left[1 - f_{(1)} + \frac{1 - \nu}{\nu}\left(1 - \theta\right)\right], \qquad (17.22)$$

where CV_{yU}^2 is the unit relvariance in the population (exercise 17.1 asks you to derive this result and the ones below).

Now, suppose that c_0 is the total of fixed costs that do not depend on sample size, c_1 is the cost per unit in phase 1 averaged over respondents and nonrespondents and that c_2 is the unit cost of collecting and processing data from a unit in phase 2. The linear cost function is expressed as

$$C = c_0 + c_1 n_{(1)} + c_2 n_{(2)}.$$

Because $n_{(2)}$ is not a constant due to the randomness of response in the first phase, we compute the expected cost to use in the optimization:

$$E_{RD}\left(C - c_0\right) = c_1 n_{(1)} + c_2 \nu\left(1 - \theta\right) n_{(1)}. \qquad (17.23)$$

The optimum value of ν that either minimizes the relvariance (17.22) for a fixed cost or minimizes cost for a fixed relvariance is

$$\nu_{opt} = \sqrt{\frac{c_1}{c_2 \theta}}.$$

For this to be a feasible value, we need $c_1/c_2 \leq \theta$. Thus, the phase 1 unit cost may have to be substantially less than that of phase 2 if the phase 1 response probability is low. The optimal value of the phase 1 sample for a fixed cost is found by substituting ν_{opt} in the cost function:

$$n_{(1)opt} = \frac{C - c_0}{c_1 + c_2 \nu_{opt} (1 - \theta)}. \tag{17.24}$$

When the relvariance is fixed at a value of CV_0^2, the optimum value is

$$n_{(1)opt} = \frac{1}{\nu_{opt}} \frac{1 - \theta (1 - \nu_{opt})}{\frac{CV_0^2}{CV_{yU}^2} + \frac{1}{N}}.$$

Selecting a nonresponse follow-up sample can be disturbingly inefficient compared to just selecting a larger *srs* in the first place. The relvariance of an estimated mean from an *srs*, neglecting the *fpc*, is $CV_{srs}^2 (\bar{y}_{srs}) = CV_{yU}^2/n_{srs}$. Setting this equal to Eq. (17.22) and solving for n_{srs} gives $n_{srs} = n_{(1)} \left[\theta + (1 - \theta)/\nu - f_{(1)} \right]^{-1}$. If only θ of these units respond, the required initial *srs* size is

$$n_{srs} = \frac{n_{(1)}}{\theta} \left[\theta + \frac{1 - \theta}{\nu} - f_{(1)} \right]^{-1}. \tag{17.25}$$

Assuming that the unit cost for the *srs* is c_1 and that θ respond, the total cost of the *srs* of n_{srs} units will be $C_{srs} = c_1 n_{srs}$. The ratio of the two-phase cost to the *srs* cost is then

$$\frac{C_{tot}}{C_{srs}} = \frac{n_{(1)}}{n_{srs}} \left[1 + \frac{c_2}{c_1} \nu (1 - \theta) \right]. \tag{17.26}$$

This calculation does assume that, within phases, all units are equally likely to respond, which may be unrealistic. The chance of responding may depend on demographic characteristics, and the demographic composition of the phase 2 subsample may be different from that of the phase 1 sample. In Chap. 13 we looked at some techniques that will account for such demographic differences when making nonresponse adjustments. For getting an idea of the sample sizes needed for the first and second phases of a NRFU design, the simpler calculations above are still useful.

The R function `NRFUopt` in Appendix C will calculate the values of ν_{opt} and $n_{(1)opt}$ for either a fixed cost or a target coefficient of variation. The function accepts the following parameters:

Ctot	Total variable cost
c1	Cost per unit in phase 1
c2	Cost per unit in phase 2
theta	Probability of response for each unit
CV0	Target coefficient of variation for the estimated total or mean
CVpop	Unit coefficient of variation
N	Population size; default is Inf
type.sw	Type of allocation
	``cost'' = target total variable cost
	``cv'' = target coefficient of variation

In addition to v_{opt} and $n_{(1)opt}$, the outputs from the function include the expected size of the second-phase sample, the *srs* size from Eq. (17.25), and the cost ratio in Eq. (17.26).

Example 17.13 (Optimal sample sizes for a fixed budget). Suppose that the budget for total variable costs is \$100,000, the unit costs for phase 1 and 2 are \$50 and \$200, the probability of response is 0.5, and the unit coefficient of variation is 1. The target coefficient of variation for the mean is 0.05. The function call with these parameter values is

```
NRFUopt(Ctot=100000, c1=50, c2=200, theta=0.5, CV0=NULL,
        CVpop=1, type.sw="cost")
```

The output is

```
$allocation
[1] "fixed cost"
$'Total variable cost'
[1] 1e+05
$'Response rate'
[1] 0.5
$CV
[1] 0.0382
$v.opt
[1] 0.7071
$n1.opt
[1] 828
$'Expected n2'
[1] 293
$'Expected total cases (2-phase)'
[1] 1121
$'srs sample for same cv'
[1] 1373
$'Cost Ratio: Two phase to srs'
[1] 1.457
```

The anticipated CV is 0.0382 with sample sizes of 828 for phase 1 and 293 in phase 2 for a total of 1,121. The subsampling fraction of the phase 1

nonrespondents is 0.7071. To obtain a *CV* of 0.0382 by selecting a larger initial *srs*, we would need to select 1,373 of whom 0.5*1,373 = 687 would be expected to respond. Note that the second sampling fraction, 0.7071, is fairly high. If phase 2 nonresponse is more than 30 %, the solution above will not be feasible. Also, note that the two-phase sample would be more expensive than an initial *srs* of 1,373 by a factor of 1.457. ∎

The preceding example seems to imply that we would be better off to select a larger initial sample that anticipates how much nonresponse there will be. We used this method in Chap. 6 to adjust sample sizes. However, a larger initial sample is not always a solution. For example, an unexpectedly low response rate may be obtained in phase 1. Also, the initial mode of data collection may reach a limit of its effectiveness. For instance, in a mail-out of paper questionnaires, the response rate may be 30 %, but a final response rate of 50 % is required. If more mailings will result in few if any additional responses, then a nonresponse follow-up sample with a different mode will be needed if there is any hope of obtaining 50 % response.

Example 17.14 (Optimal sample sizes for a target CV). In a two-phase NRFU study, suppose that a *CV* of 0.10 is desired for the estimated mean. The unit costs for the two phases are $c_1 = \$75$ and $c_2 = \$150$. The unit *CV* in the population is 3 and a response rate to the first phase is anticipated to be 70 %. Determine the allocation of the sample to both phases and the estimated variable cost of the survey. The function call and its results are:

```
NRFUopt(Ctot=NULL, c1=75, c2=150, theta=0.7, CV0=0.10,
          CVpop=3, type.sw="cv")
$allocation
[1] "fixed CV"
$'Total variable cost'
[1] 107320.2
$'Response rate'
[1] 0.7
$CV
[1] 0.1
$v.opt
[1] 0.8452
$n1.opt
[1] 949
$'Expected n2'
[1] 241
$'Expected total cases (2-phase)'
[1] 1190
$'srs sample for same cv'
[1] 1286
$'Cost Ratio: Two phase to srs'
[1] 1.113
```

The expected cost is about $107,320 with 1,190 units split between 949 phase 1 units and 241 second-phase units. A single phase *srs* of 1,286 would be needed to yield the same *CV* of 0.10. ∎

Responsive Designs

Key to the responsive design is the inability to plan during the design stage of the project for the points in time for which a change needs to be made to the essential survey conditions. For example, two months into the data collection for Study X, the team decides based on analyzing the current state of the project to send an additional incentive to hopefully increase participation. However, little information has been published to date on specific decision rules for invoking the next phase in a responsive design. We sketch the general procedures below based on our personal experience starting with the study viewpoint from at least three different angles:

(1) *Response propensity.* The project team monitors the response rates and response propensities throughout the data collection period. The indicators (and possibly response model covariates) may include a combination of frame information, paradata, "on the ground" information from the interviewers, past experience, and time/funding in the remaining data collection period. Through a best and worst case scenario, the team identifies a point at which the required sample size (overall and within subgroups) either analytically or contractually is unlikely to be met given the current sample.

(2) *Nonresponse bias analysis.* Some project teams may conduct periodic nonresponse bias analyses with variables known for respondents and nonrespondents (see, e.g., Ingels et al. 2011). The results may suggest certain subgroups are underperforming and areas are in need of additional attention from the field staff.

(3) *Precision of key estimates.* In addition to response propensity and nonresponse bias analysis, a set of key estimates may be analyzed using the current data. Especially with subgroup analysis, low levels of precision in the estimates may suggest the release of additional sample or the need to change the methods for soliciting participation.

Common results among these and other analyses may signal that funds used for "business as usual" will be wasted. At this point, the project team can decide to (i) end data collection, (ii) release members of the reserve sample, or (iii) implement a procedural change on a subsample of the nonresponding cases. Any decision must also include the remaining funds available for data collection. As noted for double sampling, special care must be taken to ensure that any subsampling does not introduce bias by purposively selecting those that, relatively speaking, are more likely to response.

17.6 R Software

We conclude this chapter with a discussion of software. No software exists for explicitly drawing multiphase samples because the sample for phase $r + 1$ depends on information gathered from the rth phase. Consequently, the

sample selection must be uniquely implemented within each phase using software developed for single-phase designs. The same is said for multistage designs where samples are drawn sequentially within each stage. Some of these procedures were discussed in other chapters of this text and are not repeated here.

Only one software package was available for analyzing two-phase designs during the time this text was developed. The R programming language includes functions for analyzing data from a two-phase design under the assumption that the first-phase units were drawn either by *srs* or through a clustered design. As with other survey designs, a two-phase R survey object must be constructed prior to conducting the analysis using the `twophase`[16] function.

Example 17.15 (Analyzing a srs/srs Two-Phase Survey Design Object in R). Borrowing the `pbc` data from the R library, the following code is used to develop a design appropriate R object. These data are from a Mayo Clinic trial in primary biliary cirrhosis (PBC) of the liver conducted between 1974 and 1984. The R `survival` package describes the data set in more detail. The `subset` option in the code below identifies the phase 2 sample units, which are defined to be persons with missing values of the `trt` variable:

```
    # two-phase simple random sampling.
data(pbc, package="survival")
pbc$id <- 1:nrow(pbc)
d2pbc <- twophase(id=list(~id, ~id), data=pbc,
          subset = ~I(is.na(trt)))
```

∎

Example 17.16 (Two-phase sampling for stratification using the NHIS population). Suppose that an initial *srswor* sample of $n = 2,000$ persons is selected from the `nhis.large` population. The age of each person is ascertained and the first phase is stratified into five groups: < 18 years, 18–24 years, 25–44 years, 45–64 years, and 65+. A stratified phase 2 sample was selected with $n_{(2)h}$=100 in each stratum with the idea being that equal precision is desired for analyses of persons in the different age groups. Note that this is different from the examples in Chap. 10 where fixed sampling rates were set in advance for subgroups in a two-stage area sample. Although the rates in those examples were designed to produce certain target sample sizes, the within cluster rates could be determined in advance. In this example, the second-phase rates depend on how many persons were found in the age groups in the first phase. R code for selecting the two-phase sample and estimating the proportion of persons is shown below. The function in the `survey` package that handles

---

[16] http://rss.acs.unt.edu/Rdoc/library/survey/html/twophase.html

two-phase samples is twophase. A data frame (`p1.dat`) in this example must be constructed that has a record for each phase 1 element with indicators for whether an element was in the second-phase sample or not (`p1.dat$p2`). The parameter

```
subset = ~p2
```

in the call to `twophase` specifies the field that identifies the second-phase elements.

In this example, 5.89 % of people delayed medical care in the previous 12 months because of cost. The youngest (< 18) and oldest ($65+$) were less likely (5 % and 2 %) to delay care than persons in the other groups:

```
attach("nhis.large.RData", pos=2)
require(sampling)
require(survey)
set.seed(1716768836)

nhis <- as.data.frame(nhis.large)
n1 <- 2000
N <- nrow(nhis.large)

    # recode delay.med to be 0,1
nhis$delay.med <- abs(nhis$delay.med-2)
    # select a phase-1 sample of n1
sam <- sort(sample(1:N, n1))
p1.dat <- nhis[sort(sam), ]
    # Phase-1 weights
p1.dat$p1wts <- rep(N/n1, n1)
n2 <- rep(100,5)
p2.str.sam <- strata(data.frame(p1.dat),
                     stratanames = c("age.grp"),
                     size = n2,
                     method = "srswor")
   # set a T/F variable for whether person is in phase-2 sample
p1.dat$p2 <- FALSE
p1.dat$p2[p2.str.sam$ID_unit] <- TRUE
   # Phase-2 conditional weights
p1.dat$p2wts <- 0
p1.dat$p2wts[p2.str.sam$ID_unit] <- 1/p2.str.sam$Prob
   # 2-phase design object
d2.nhis <- twophase(id = list(~ID, ~ID),
        data = p1.dat,
        strata=list(NULL, ~age.grp),
        weights = list(~p1wts, ~p2wts),
        subset = ~p2,
        method = "approx")

mns <-svymean(~factor(delay.med), design = d2.nhis, na.rm = TRUE)
ftab <- ftable(mns, rownames=list(delay.med = c("No","Yes")))
round(ftab,4)
```

```
delay.med
No        mean   0.9411
          SE     0.0114
Yes       mean   0.0589
          SE     0.0114


age.mns <- svyby(formula= ~delay.med, by=~age.grp,
           FUN=svymean, design = d2.nhis, na.rm=TRUE)
round(age.mns,4)

 age.grp delay.med se.delay.med
       1    0.0500       0.0219
       2    0.0800       0.0272
       3    0.0612       0.0243
       4    0.0800       0.0272
       5    0.0200       0.0140
```

Estimates of the components of variance due to phases ($\hat{V}_1$ and $\hat{V}_2$ in Example 17.10) can be extracted with the following code, which applies to the overall estimate of the proportion who delayed medical care:

```
V <- vcov(svymean(~factor(delay.med), design = d2.nhis, na.rm = TRUE))
V1 <- attr(V, "phases")$phase1
V2 <- attr(V,"phases")$phase2
```

In this case, `V1= 2.79e-05` and `V2= 1.028e-04` so that the second-phase accounts for about 79 % of the variance of $\hat{t}_{(2)y}$.

How does the stratified, double sample compare to an *srs* of $n = 500$ for the overall estimate? If we had selected 500 persons by *srs* and obtained an estimate of 0.0589, the standard error would have been

$$\sqrt{(0.058943\,(1 - 0.058943)\,/500)} = 0.0105$$

compared to 0.0114 above. Thus, the double sample is slightly less precise, but the expected numbers of persons in the five age groups in an *srs* of 500 are 139, 47, 142, 116, and 57. Ages 18–24 and 65+ have fewer than the target of 100. Two-phase sampling gave overall precision about the same as an *srs* of the same size but allowed the numbers of sample persons in each age group to be controlled. Of course, screening to determine ages costs money that would not be spent in an *srs*. ∎

Note, in addition to allowing the use of standard survey functions, like `svymean`, the `calibrate` function will produce GREG weights for a two-phase design. However, the calibration is currently reserved only for the phase 2 units.

Exercises

17.1. Consider a nonresponse follow-up study in which the first phase is selected by *srswor* from the population and the second phase is selected by *srswor* from the phase 1 nonrespondents. Notation is defined in Sect. 17.5.2.

(a) Show that the double expansion estimator is
$\hat{t}_{(2)y} = N \left[p_{(1)R} \bar{y}_{(1)R} + p_{(1)NR} \bar{y}_{(2)R} \right]$.
(b) Beginning with the expression

$$V\left(\hat{t}_{(2)y} \right) = \frac{1 - f_{(1)}}{n_{(1)}} S_{yU}^2 + E_{(1)} E_{RD} \left(p_{(1)NR}^2 \frac{1 - f_{(2)}}{n_{(2)}} S_{y(1)NR}^2 \middle| s_{(1)} \right)$$

show that the variance equals

$$V\left(\hat{t}_{(2)y} \right) = \frac{S_{yU}^2}{n_{(1)}} \left[1 - f_{(1)} + \frac{1 - \nu}{\nu} \left(1 - \theta \right) \right].$$

where $\nu = f_{(2)R}$ is the fixed second phase achieved sampling fraction (i.e., the number of phase 2 respondents divided by the number of phase 1 nonrespondents) and θ is the probability that any unit responds. Assume that whether a unit responds is independent of any other unit.
(c) Show that if the cost function is $C = c_0 + c_1 n_{(1)} + c_2 n_{(2)}$ where $n_{(2)}$ is treated as random, then $E_{RD} \left(C - c_0 \right) = c_1 n_{(1)} + c_2 \nu \left(1 - \theta \right) n_{(1)}$.
(d) Show that the optimal value of the phase2 subsampling fraction is $\nu_{opt} = \sqrt{c_1 / c_2 \theta}$.
(e) If the variance is minimized subject to a fixed total (expected) cost, then show that $n_{(1)opt} = \frac{C - c_0}{c_1 + c_2 \nu_{opt} (1 - \theta)}$.
(f) If the cost is minimized for a fixed value, CV_0, of the coefficient of variation of $\hat{t}_{(2)y}$, then

$$n_{(1)opt} = \frac{1}{\nu_{opt}} \frac{1 - \theta \left(1 - \nu_{opt} \right)}{\frac{CV_0^2}{CV_{yU}^2} + \frac{1}{N}}.$$

17.2. Suppose that the budget for total variable costs in a NRFU study is $500,000, the unit costs for phase 1 and 2 are $25 and $200, the probability of response is 0.3, and the unit coefficient variation is 1. Find the optimal allocation for a two-phase sample to minimize the coefficient of variation of the estimated mean. Discuss the results.

17.3. In a two-phase NRFU study, suppose that a CV of 0.10 is desired for the estimated mean. The unit costs for the two phases are $c_1 = \$75$ and $c_2 = \$350$. The unit CV in the population is 2 and a response rate to the first phase is anticipated to be 40 %. Determine the allocation of the sample to both phases and the estimated variable cost of the survey.

17.4. In a two-phase NRFU study, suppose that a CV of 0.10 is desired for the estimated mean. The unit costs for the two phases are $c_1 = \$75$ and $c_2 = \$150$. The unit CV in the population is 2 and a response rate to the first phase is anticipated to be 40%. That is, the assumptions are the same as in Exercise 17.3, except that the phase 2 cost is much less. Determine the allocation of the sample to both phases and the estimated variable cost of the survey. Discuss your results.

17.5. Use the `nhis.large` population to study double sampling for stratification. Select an initial *srswor* of $n = 2,000$ persons. In R use initialize the random number generator with set.seed(1716768836). The age of each person is ascertained and the first phase is stratified into five groups: < 18 years, 18–24 years, 25–44 years, 45–64 years, and 65+. A stratified phase two sample was selected with $n_{(2)h}=100$ in each stratum.

(a) Estimate the proportion of persons and the SEs of the proportion of persons that had an overnight hospital stay in the previous 12 months.
(b) What proportion of the variance in (a) was due to phase 1 and phase 2?
(c) Estimate the proportions and SEs for the five age groups.
(d) How do the SEs in (a) and (c) compare to an *srs* of $n = 500$ selected in a single phase?

17.6. Consider a situation where an initial wave of data collection is attempted. Some units respond and others do not. Suppose the population can be divided into two strata—one of cases that respond to the initial phase and another of the cases that do not. Denote the proportions of the population in the two strata by W_1 and $W_2 = 1 - W_1$ and the population means by $\bar{y}_{U1}$ and $\bar{y}_{U2}$. A simple random sample is selected and only cases in stratum 1 respond. Now, assume that $\bar{y}_{U2} = k\,\bar{y}_{U1}$. Show that the *relbias* of $\bar{y}_1$ as an estimator of $\bar{y}_U$ is

$$relbias\,(\bar{y}_1) = \frac{W_2\,(1-k)}{1 - W_2\,(1-k)}.$$

17.7. In this problem we revisit Example 17.7.

(a) Calculate the unequal weighting effect for the final weights $w_{(2)k}$. Why might this be important to examine.
(b) What suggested changes would you implement if you had the current results as your historical information?

17.8. A double sampling for stratification design is proposed for a study with a telephone screener questionnaire in phase one. A subsample of respondents will be administered a longer, in-depth questionnaire in the second phase. Phase 1 is *srs* and phase 2 is *stsrs*. The following population estimates are provided by the two strata:

Stratum	N_h	W_h	P_h
1	1,580	0.79	0.19
2	430	0.21	0.52
Total	2,010		

(a) Determine the overall sample sizes for the first and second phases of the design using the method described in Example 17.12 with an overall cost value of $C = \$10,000$, $c_1 = 10$, and $c_{(2)} = \$100$. Comment on your findings. Is there a gain from using double sampling with an optimal allocation to strata compared to selecting an *srs* with the same total cost? Why or why not? Assume that each unit in the *srs* costs $c_{(2)}$. If there is no gain, why might double sampling still be used?

(b) How do your results change if C=\$10,000 but the cost of phase 2 data collection is double (i.e., $c_{(2)} = \$200$). Comment on your findings.

Chapter 18
Process Control and Quality Measures

So far we have described a wide variety of tools and tasks necessary for sampling and weighting. Key to a successful project, however, is not only the mastery of the tools, and knowing which tool to use when, but also the monitoring of the actual process, as well as the careful documentation of the steps taken, and the possibility to replicate each of those steps. For any project, certain quality control measures should be taken prior to data collection during sample frame construction and sample selection and after data collection during editing, weight calculation, and database construction. Well-planned projects are designed so that quality control is possible during the data collection process and that steps to improve quality can be taken before the end of the data collection period. Obviously the specific quality control measures will vary by the type of project conducted. For example, repeated longitudinal data collection efforts allow comparisons to prior years, whereas one-time cross-sectional surveys often suffer from uncertainty with respect to procedures and outcomes. However, we have found a core set of tools to be useful for almost all survey designs and will introduce those in this chapter. We do want to emphasize that while it is tempting to think that assurance of reproducibility and good documentation is only worth the effort for complex surveys that will be repeated, in our experience, even the smallest survey "runs" better when the tools introduced here are used.

The material in this chapter is only scratching the surface of what can be done and focuses in particular on elements of key relevance to researchers. This chapter is organized into three distinct time periods of a survey: pre-data collection (study design, frame construction, and sample selection), mid-data collection (monitoring techniques and performance rates), and post-data collection (editing, weighting, specification writing, and documentation). We highly recommend reading the Quality Guidelines provided by various statistical agencies and other organizations such as Eurostat (Aitken et al. 2004), the US Office of Management and Budget (2006),[1] Canada (2009), the

---

[1] http://www.whitehouse.gov/sites/default/files/omb/inforeg/statpolicy/standards_stat_surveys.pdf

United Kingdom's Office for National Statistics,[2] and the American Association for Public Opinion Research[3]; reports from large survey projects such as CAHPS Hospital Survey[4] and the Programme for International Student Assessment (PISA)[5]; and textbooks and other sources such as Biemer and Lyberg (2003), Blasius and Thiessen (2012), and the Cross-Cultural Survey Guidelines hosted at the University of Michigan.[6]

18.1 Design and Planning

Project design and planning is a critical first step to ensure the timely administration of the survey and the gathering of high-quality data. The order and interdependencies of the study tasks should be understood and specified at the beginning of the project. Several charting tools come in handy during the overall project planning stage. Ones that are often used are Gantt charts, Critical Path Method, and flowcharts.

Gantt charts and charts known as Critical Path Methods are designed both to visualize the time dependency of various project tasks, and to reflect how the delay in one project step will impact the final outcome. Gantt charts are a mixture of tables and graphs and list one task in each row of the chart. Next to each task the estimated begin and end dates are entered as well as the duration of the project. A graphical representation of the time this task takes is the signature part of a Gantt chart (see Fig. 18.1). The horizontal axis in the graphical representation is time, either in absolute time or in time since beginning of the project. The time resolution depends on the project and can be days, weeks, or months. The individual rows of a Gantt chart can be linked with each other. Thus, if one of the task takes longer (or shorter) than expected, the remaining rows can change accordingly. The Gantt chart should be updated regularly throughout the duration of the project.

Figure 18.1 displays a portion of a Gantt chart we used for a project in 2011 at the Institute for Employment Research (IAB) in Germany. The second column in this chart represents a list of all tasks necessary for the project, followed by an indication of start and end date, from which duration days are computed (hint: important to not forget holidays and vacation times). The visual display is on the right-hand side of the graph, where the two shades indicate the level of completion of these individual tasks. While easy to create and understand, Gantt charts have been criticized for their heavy grid layout,

---

[2]http://www.ons.gov.uk/ons/guide-method/best-practice/gss-best-practice/gss-quality-good-practice/index.html

[3] http://www.aapor.org/Best_Practices1.htm

[4] http://www.hcahpsonline.org/home.aspx

[5] http://nces.ed.gov/surveys/pisa/pdf/2011025.pdf

[6] http://ccsg.isr.umich.edu/quality.cfm

WBS	Tasks	Task Lead	Start	End	Duration (Days)	% Complete	Working Days	Days Complete	Days Remaining
1	Quex Development		31.05.11	13.07.11	44	62%	32	27	17
1.1	Deliver draft to LINK		31.05.11	31.05.11	1	100%	1	1	0
1.2	Deliver final to LINK		14.06.11	14.06.11	1	100%	1	1	0
1.3	Programming		15.06.11	07.07.11	23	75%	17	17	6
1.4	Testing		08.07.11	13.07.11	6	0%	4	0	6
1.5	Develop Interviewer Questionnaire		24.06.11	30.06.11	7	0%	5	0	7
2	Advance Letters		14.06.11	22.07.11	39	88%	29	34	5
2.1	Develop & Review with IAB Frau Ross		14.06.11	16.06.11	3	100%	3	3	0
2.2	Deliver final pretest letter to LINK		17.06.11	17.06.11	1	100%	1	1	0
2.3	LINK prints pretest letters		20.06.11	24.06.11	5	100%	5	5	0
2.4	Develop main study adv letter		20.06.11	06.07.11	17	100%	13	17	0
2.5	Deliver final main study letter to LINK		08.07.11	08.07.11	1	100%	1	1	0
2.6	LINK prints main study letters		11.07.11	15.07.11	5	0%	5	0	5
2.7	LINK personalizes main study letters		18.07.11	22.07.11	5	0%	5	0	5
3	Pretest Sample		17.05.11	09.06.11	24	100%	18	24	0
3.1	Delivery from ITM		17.05.11	17.05.11	1	100%	1	1	0
3.2	QC & Cleanup		18.05.11	06.06.11	20	100%	14	20	0
3.3	Delivery to LINK		09.06.11	09.06.11	1	100%	1	1	0
4	Sample		10.06.11	06.07.11	27	73%	19	19	8
4.1	Delivery from ITM		10.06.11	10.06.11	1	100%	1	1	0
4.2	QC & Cleanup		11.06.11	21.06.11	11	100%	7	11	0
4.3	Delivery to LINK		22.06.11	28.06.11	7	100%	5	7	0
4.4	Phone # Research (n=5,000)		30.06.11	06.07.11	7	0%	5	0	7
5	Pretest		07.07.11	05.08.11	30	6%	22	1	29
5.1	Mail advance letters (n=600)		07.07.11	07.07.11	1	100%	1	1	0

Fig. 18.1: Example Gantt chart (using MS Project software)—filter question project at IAB.

the sparseness of the data display, and their inability to show clearly the relative importance of individual tasks (Tufte 1990; DeMeyer et al. 2002).

Relative importance and dependency of tasks is visualized more clearly in charts based on the Critical Path Method. Critical Path Methods use network diagrams to show the sequence and dependencies of each of the tasks.[7] They clearly show which tasks can occur simultaneously and which need to be finished before other tasks can start. The longest path determines the overall length of the project. If activities outside the critical path speed up or slow down, the total project time does not change. The amount of time that a noncritical path activity can be delayed without delaying the entire project is referred to as slack time. The Critical Path Method was developed for fairly complex but routine activities. For less routine projects, estimates of completion times are unstable, which limits the usefulness of the Critical Path Method.

Flowcharts are often used in project planning to visualize steps within a task. Flowcharts are semantic representations of an algorithm or a process. Flowcharts can be used for technical aspects of the project, such as weighting, but are fairly useful in other parts of the project as well (e.g., visualizing the flow of questionnaires, or detailing recruitment steps and nonresponse follow-up procedures). Figure 18.2 shows the beginning of a study design flowchart

---

[7] A free online lecture on using CPM with a survey example can be accessed here: http://gunston.gmu.edu/healthscience/ProjectManagementInIT/ CriticalPathMethod.asp

as it is used in the Best Practices Manual at the Survey Research Operations center at the University of Michigan in the USA.

Although not always used for survey research, standardized flowchart symbols have been developed in the context of computer programming (International Organization for Standardization 1985). For example, boxes are used to represent tasks (or processes) and diamonds are used for decision points. Each branch leaving a diamond shows the actions following each outcome at the decision point. Figure 18.2 uses task boxes and decisions diamonds. The *Handbook on Improving Quality by Analysis of Process Variables*,[8] published by Eurostat, shows a series of flowcharts for each step in the survey process. We saw a flowchart for weighting in Chap. 13. For programming, this flowchart would need to be specified in much more detail. Flowcharts are very useful in providing a high-level overview of the process and its interconnections. However unlike the Critical Path Method they do not give an indication how a delay in one task will affect other tasks.

18.2 Quality Control in Frame Creation and Sample Selection

After constructing or acquiring a sample frame, survey statisticians are well advised to perform a series of quality control checks on the files. Those quality control checks typically involve identifying and excluding duplicates as well as erroneous records, verifying that the count in the frame matches what is known to be present in the overall population and possibly within certain subgroups, and comparing the distribution of variables on the frame with other sources for the population. In some situations, it may be possible to check frame data for consistency with other frames or administrative data. For example, surveys that use the US Postal Service Delivery Sequence File, mentioned in Chaps. 1 and 10, should check if there are areas in the file that are undercovered compared to census housing counts (Iannacchione 2011).

Variables on the frame that will be used for sampling should also be checked for missing or unallowable values. In a frame of schools, for example, variables that may be used for the sample design are the number of students enrolled in each school, which might be a measure of size for *pps* sampling, and the grade range of each school, which may be used for stratification or to exclude ineligible sample units. Frames of hospitals, households, and businesses will have different types of checks that should be made on design variables. When data are missing, imputations may be needed before the frame can be used for sample selection. We will briefly return to the topic of editing frame data in Sect. 18.5.

---

[8] http://www.processdox.com/pix/ImprovingQuality.pdf
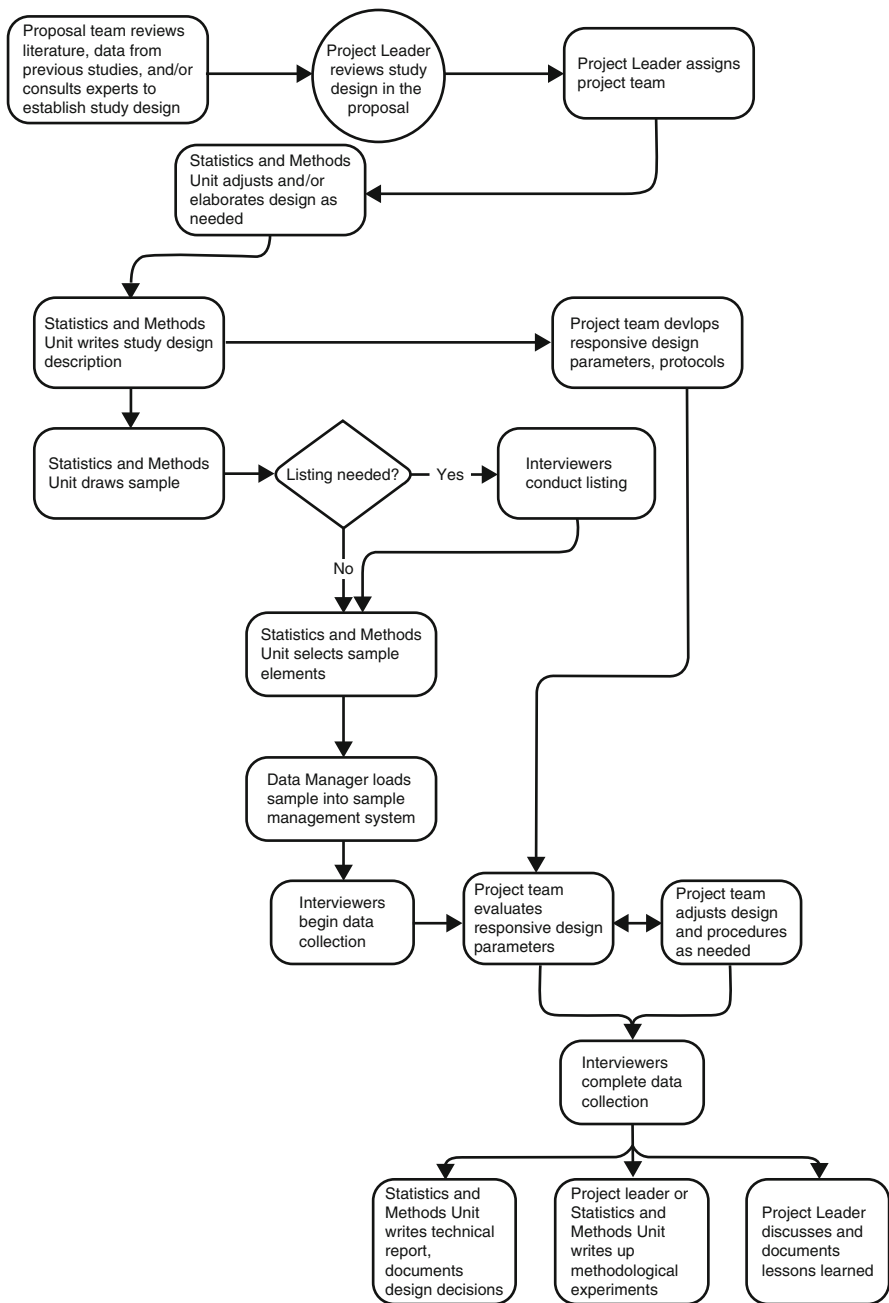
Fig. 18.2: Example flowchart—study design and sampling from SRO best practice manual.

Many software packages allow straightforward checks for duplicates.[9] However, if frames include names and likely typos, then record linkage software should be used for de-duplication (Herzog et al. 2007). Two free software packages specifically geared toward use in official statistics and survey research are Matcher-2 (Porter and Winkler 1997) and the Merge ToolBox (Schnell et al. 2004). Some governmental agencies have developed their own software for matching like Statistics Canada's Generalized Record Linkage System (GRLS) (Thomas 1999, Willenborg and Heerschap 2012).

During the field period some assessments on the quality of the frame can be made based on what is found for the sample, for example, decisions on out-of-scope units can be verified, and missed units added (Eckman and O'Muircheartaigh 2011). If addresses are released in replicates (see Chap. 6), the composition and number of the household members at the beginning of each replicate should be similar; if not, interviewer learning effects might affect how the screening process is done. In some surveys, interviewers are instructed to select one respondent at random in each sample household. Quality checks for several European surveys showed clearly that the selected household members were disproportionably female, suggesting that interviewer "randomly selected" the contact person as respondents (Kohler 2007).

Sample selection can be an involved process that requires its own quality control checks. Basic checks are whether the complete frame has been processed for sampling, whether the desired number of sample units has been selected, and whether the selection probabilities of units can be computed and, if so, whether have they been recorded. As described in Sect. 18.7, specifications should be written that clearly explain all steps in frame construction, cleaning, and sample selection.

18.3 Monitoring Data Collection

Successful data collection efforts require close monitoring of the sample during production. Such monitoring can help identify potential shortfalls in achieving the desired outcome such as, for example, a specified response rate or other goals associated with sampling and data quality. To do the monitoring, key process variables need to be identified. These are usually variables that can vary with each repetition of the process and have a strong effect on the quality of the survey. Examples of process variables are disposition codes for the contact attempts (see Chap. 6), measures of resources used, or coding errors. In Sect. 18.4, we will list a series of such indicators. While indicators could be monitored in tables as part of reports, graphical displays are often more efficient for monitoring.

---

[9] In R this would be: `duplicated(x, incomparables = FALSE, ...)`.
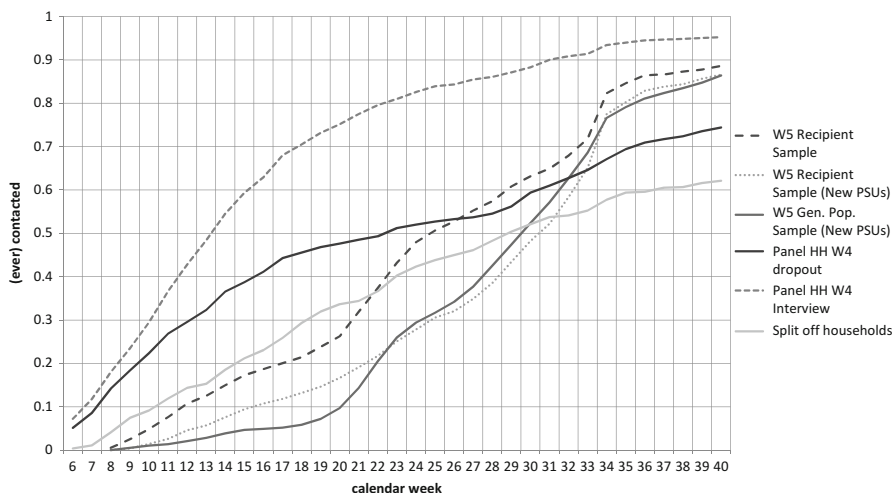
Fig. 18.3: Contact rates for each subsample by calendar week in the PASS survey at the Institute of Employment Research, Germany (Müller, 2011).

Common in industry applications are process control charts (Deming 1982). Their use in surveys is less common, despite the fact that one of its main proponents, Deming himself, worked at the US Census Bureau between 1939 and 1945. However, the steady increase in computer-aided data collection procedures has also increased the data flow during data collection. Consequently, we see a revived interest in statistical process control and related charts to monitor and manage fieldwork procedures (Jans et al. 2013). In their simplest form, charts to monitor ongoing fieldwork display key process variables in the development over days or weeks of the fieldwork period. More informative are displays by relevant subgroups, such as the chart in Fig. 18.3. Here we see contact rates per calendar week divided by subsamples; it can be seen from the chart that panel households were contacted at a much higher rate than households sampled as refreshment cases in this panel survey. Such differential contact rates can have strong effects on the overall survey quality.

Ideally those charts inform interventions. The result of a successful intervention can be seen in Fig. 18.4 which displays response rates by important subgroups in the National Survey of Family Growth. Starting with the third week of data collection Hispanic males between 20 and 44 years of age were found to be lagging behind in response rates. An intervention was launched where interviewers were asked to increase their effort on those cases, and traveling interviewers with bilingual capabilities were sent to segments containing sampled cases in those subgroups. As a result of this intervention the coefficient of variation in response rates among the relevant subgroups decreased (Kirgis and Lepkowski 2010).
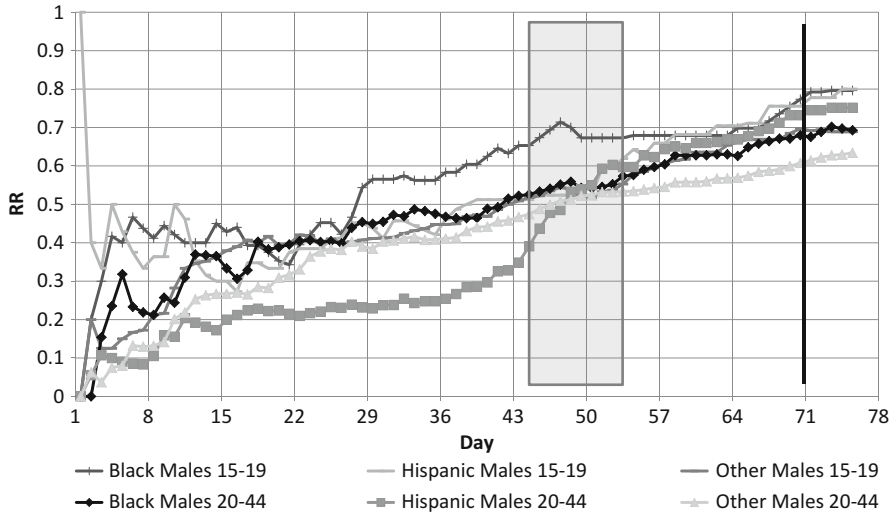
Fig. 18.4: Cumulative response rates by subgroups in the national survey of family growth, intervention was launched during the grey area (Lepkowski et al., 2010).

For a fieldwork manager and those monitoring the sample during data collection, it is important to not react to "normal" variation in key process indicators. It would be a waste of resources to intervene if the process is still in control. Thus, a typical feature of control charts, as they are propagated in the statistical process control literature, is their ability to separate common and special causes that influence a given process. This separation is important because the action step required to address special causes is very different from those that address common causes. A good example for surveys is interview response times over the course of the field period. In many surveys interview time goes down as the interviewers get more and more used to administering a given survey (Olson and Peytchev 2007). If such a reduction in interviewing time is threatening interviewing quality, management has to intervene and change the system. An example for a special cause would be an individual interviewer or local area in which a reduction (or an increase) in interviewing time is visible. Here a one-time, local intervention by the operating staff might be sufficient.

A chart that displays both common cause variation and special cause variation is the Shewhart (1931) chart (see example in Fig. 18.5). Here control limits (usually three times the standard deviation of the key process variable denoted in the figure as a dashed line) are displayed alongside an x-axis that groups the data in a meaningful way. Such grouping is often done for certain time intervals (here days in field), but geographic areas, sample portions, or interviewers could form the x-axis as well. If the variation does not display a typical pattern and falls within the control limits, then the variation is said
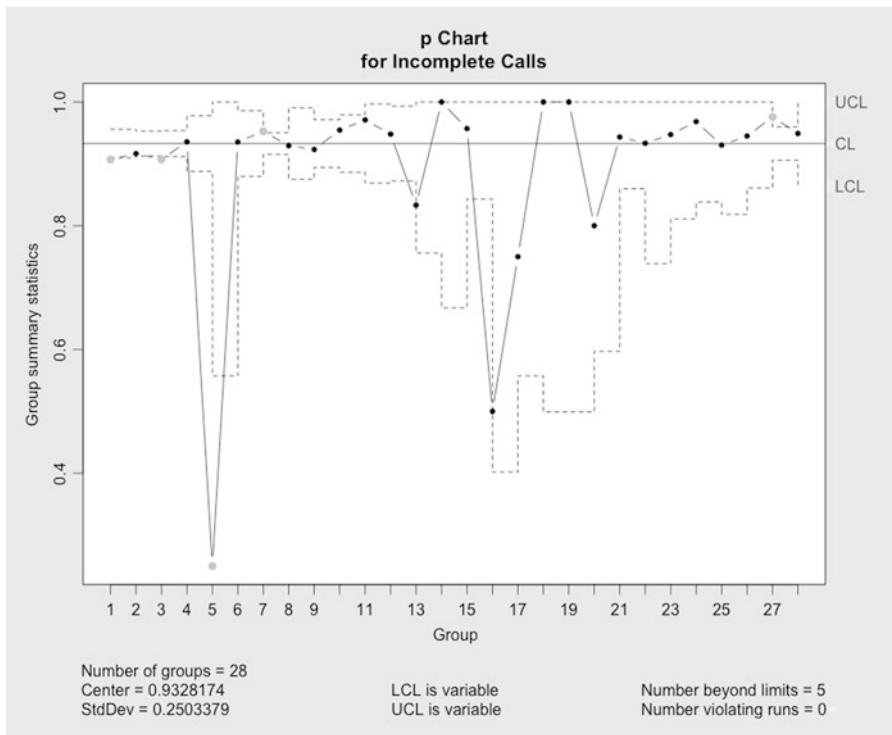
Fig. 18.5: Proportion of incomplete calls by days in field. Data from Joint Program in Survey Methodology (JPSM) practicum survey 2011.

to be due to a common cause. However if there are deviations outside the control limits or if there is variation in a typical pattern, those are said to be due to a special cause.

The Shewhart graph in Fig. 18.5 displays the proportions of calls for each day in the field. Cases can receive multiple calls per day (for instance if the first call was busy). On most days the proportion of incomplete calls is well above 80 %. On day five the number of incomplete calls is unusually low. The chart does not judge good or bad; it only indicates what is common and what is unusual. In this case, the variability is very high because very few calls were made on day five. Many of them could have been prearranged appointments causing the number of incomplete calls to be less than the expected limit for the given sample size. But it could also be the case that an unusually low proportion of incomplete calls is the result of a programming error or technological problem (say, cases mistakenly being coded as complete), interviewer error, or even falsification by interviewers. So when numbers are unusually good, it is possible that they are too good to be true and in fact are indicative of some sort of underlying problem.

For optimal use, key process input variables should be specified prior to data collection, together with set thresholds. Fieldwork should be stopped if those key process variables exceed the thresholds or in the language of process control "go out of control." Which key process variables are monitored in any given survey will be a function of the survey itself and its design. However, care should be taken to select indicators that are meaningful with respect to the outcome quality, and not just those that are easy to measure or readily available (Morganstein and Marker 1997).

18.4 Performance Rates and Indicators

Earlier we showed contact and response rates in Figs. 18.3 and 18.4. These rates are two important performance rates that most surveys track. Whether they are computed during data collection or at the end of the data collection efforts, performance indicators are important quality control tools. The American Association of Public Opinion Research (AAPOR) has provided standard definitions for the computation or estimation of such performance rates, many of them related to the proper specification of response rates. Much broader are the terms and rates specified by the Data Documentation Initiative (DDI), which is designed to document and manage data across the entire life cycle from specification of survey design features to survey outcomes and archiving (www.ddialliance.org). It is important to note that not all researchers follow the definitions provided by AAPOR or DDI. Consequently, it is advisable to communicate a common understanding within the project team and essential to use standard terms for comparing outcomes across surveys. Many journals require performance rates to be explicitly described and the DDI or AAPOR document can easily be referenced in study reports and journal articles. Chapter 6 has definitions and explanations for the four most common rates: location rate, eligibility rate, cooperation rate, and response rate.

The rates discussed in Chap. 6, in particular response rates, are very popular outcome goals set by clients. However, there is not necessarily a link between response rates and nonresponse bias, which is the actual point of concern for most clients. Groves and Peytcheva (2008) review 59 methodological studies which were designed to estimate the magnitude of nonresponse bias on a variety of statistics. They found very little relationship between response rate and bias. Thus, while the response rates will be asked for, they only carry limited amount of information about survey quality. In response, attempts have been made to develop alternative measures that capture additional information about the composition of the responding sample. Those rates can also be tracked during data collection, given that auxiliary information is available about respondents and nonrespondents.

R-Indicators

One set of indicators that describe the respondent composition relative to the sample composition are called Representativity Indicators or R-indicators (http://www.risq-project.eu/). They are designed to capture imbalances in response propensities between subgroups of sampled units. In its simple form the estimated R-indicator for a survey with sample size n is proportional to the standard deviation of the response propensities for individuals estimated using a set of covariates. Assuming equal sampling probabilities it is expressed through

$$\hat{R}(\rho) = 1 - 2\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(\hat{\rho}_i - \hat{\bar{\rho}})^2}, \tag{18.1}$$

where $\hat{\rho}_i$ are the individual response propensities and $\hat{\bar{\rho}}$ is the average response propensity over all sample cases (Schouten and Cobben 2007; Bethlehem et al. 2011; Schouten et al. 2009).[10] The R-indicator uses available information on both respondents and nonrespondents to estimate response propensities, either through logistic regression models or classification trees. If all the response propensities were equal, then the nonrespondents would be missing completely at random (MCAR) as described in Sect. 13.5. Thus, the larger $\hat{R}(\rho)$ is, the more the data depart from MCAR.

Balance Indicators

Similar in spirit to the R-indicator is the Q^2 indicator developed by Särndal and Lundström (2008), which is defined as the variance of the predicted inverse response probabilities. Smaller values of the Q^2 indicator imply that there may be more work needed on the weight adjustments to correct for potential nonresponse bias. In both cases the potential for nonresponse bias can be assessed *only* for those variables that are available for both respondents and nonrespondents. This is a strong limitation of both approaches. Often variables that are available for both respondents and nonrespondents are not strongly related to the survey outcome variables (and those are the ones where bias is feared). Nevertheless, R-indicators and other balance indicators are used to monitor the incoming respondent pool. Similar to tracking response rates for subgroups (as shown in Sect. 18.3) those indicators can help reallocate recruitment efforts. Note that the nonresponse adjustments could be made using the same covariates used to estimate response propensities in $\hat{R}(\rho)$. If these covariates are good predictors of response and of survey analysis variables, then nonresponse bias can be reduced by weighting. But, removing

---

[10] In the case of unequal sampling variance this equation changes to reflect the design weights. Sampling weights are not included in the estimation of the propensity models but are used when the variance is constructed.

imbalance between the respondents and nonrespondents during data collection can reduce the variation in nonresponse adjustment weights and lessen the burden on weighting to correct nonresponse bias.

The National Survey of Family Growth uses the Fraction of Missing Information (FMI) indicator to track fieldwork progress. The FMI seeks to measure uncertainty about values imputed for missing elements (Rubin 1987; Little and Rubin 2002). We have not covered imputation procedures in this book and will therefore not go into detail on how FMI is estimated. However, a detailed explanation for the survey setting can be found in Wagner (2010).

Interviewer-Specific Indicators

Often overlooked, but very important, is the role of interviewers in face-to-face and telephone surveys, in particular with respect to the indicators discussed in the previous section. In addition to conducting the survey itself, interviewers do play an important role in recruitment and within household respondent selection. Many of the performance rates discussed here vary significantly among interviewers. To monitor interviewers, it is therefore useful to compute missing data rates by interviewer, several statistics by interviewer such as average interview length, cost per interview, refusal conversion rates, level of effort by interviewer, etc. In face-to-face surveys interviewers work often only in one geographical area. Therefore, variation in response rates could be due to variation in respondent characteristics or to the geographic clustering as well as the interviewers. In telephone surveys, cases are usually randomly assigned to interviewers and several interviewers might have "touched" a case before the respondent agrees to participate. Nevertheless, the few studies that did allow a separation of interviewer effects and effects from other sources show strongly the role interviewers play both with respect to measurement error (O'Muircheartaigh and Campanelli 1998, Schnell and Kreuter 2005) as well as nonresponse (O'Muircheartaigh and Campanelli 1999, Durrant and Steele 2009, West and Olson 2010).

Interviewer-specific indicators can take various shapes and forms. West and Groves (2013) recently developed propensity-adjusted interviewer scoring indicators that take respondents' covariate information into account. For effects of interviewers on survey responses, interviewer-specific design effects can be useful indicators (Kreuter et al. 2010). Figure 18.6 shows for 18 interviewers in a CATI survey the relative contribution of each interviewer to the overall design effect. More specifically the intraclass correlation coefficient ρ is estimated 18 times, each time leaving out all interviews conducted by one of the interviewers. The horizontal line in Fig. 18.6 shows the average ρ with all interviewers. There are two outliers in this graph. When removing interviewer #3, ρ dropped from an average of 0.0130 to 0.009 (Interviewer IDs mark the plot symbols). With an average workload of 75 the effective sample size in a survey of 1600 cases would be 816 without interviewer #3 and only
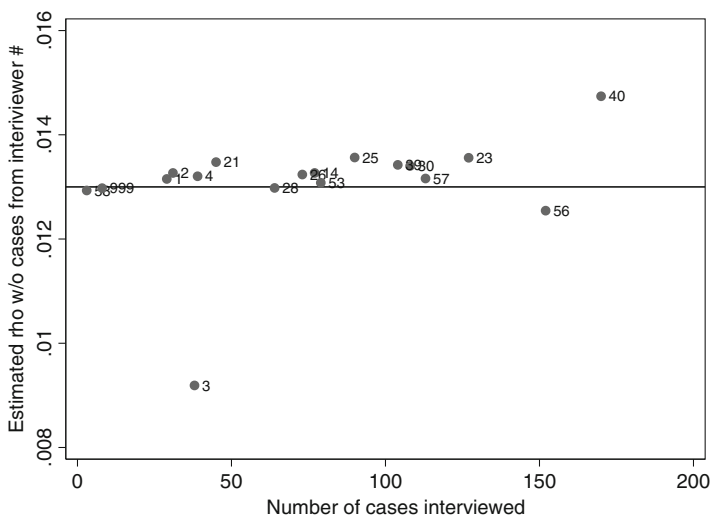
Fig. 18.6: Interviewer contribution to rho in the DEFECT telephone survey, based on Kreuter (2002); survey data are described in Schnell and Kreuter (2005).

209 with this interviewer.[11] Upon further examination of the interviewing staff, this interviewer was found to be the only male interviewer among a staff of female telephone interviewers in a survey on fear of crime.

18.5 Data Editing

Data editing is a common quality control step. While some surveys, and in particular surveys conducted through statistical agencies, suffer from over-editing (Lyberg et al. 1997), there are various editing steps that need to be done in (almost) every survey. Clean data facilitate sample selection, the creation of analysis weights, analysis tables, and the final project data set. Some edits also need to be done on a flow basis, for example, to quickly identify problems with interviewers, to check skip patterns in electronic questionnaires in the first days of the field period, and to check if all variables forming an index are adequately captured. Some editing might also be needed to feed into the monitoring charts discussed above. Thus, ideally, the edit specifications are developed during the planning phase of a project and will be updated as the project progresses. In general it is fair to say that the data-editing phase is only as good as the specifications. Consistent with the scope of this book we will not talk in detail about specifications to edit questionnaire variables but focus on those that are relevant for creating a sampling frame, eligibil-

---

[11] $1600/(1 + 0.09*(74)) = 208.88$; and $1600/(1 + 0.0130*(74)) = 815.49$

ity variables, disposition codes, and weighting variables. Our suggestions for writing the actual specifications are discussed in Sect. 18.7.

18.5.1 Editing Disposition Codes

In Chap. 6 we introduced disposition codes (see Table 6.2) used to compute or estimate performance rates. While these codes seem straightforward, in practice they often are not. Two points should be discussed with a client: first, the mapping of detailed disposition codes into one of the seven categories and, second, the hierarchy of outcome codes to determine a current or final status.

Mapping

When mapping survey-specific disposition codes to those used in the standardized rate definitions, assignments may differ as a function of target populations. For example, some studies exclude institutionalized persons. Thus, a person who is (temporarily) institutionalized would in one survey be classified as "other non-interview" while in the other the same person would be "not eligible." Second, researchers may express different preferences on how assignments should be executed; this is particularly true for the use of partial interviews. Addressing these issues ahead of time is important. For example, the sample disposition codes recorded for the May 2004 Status of Forces Survey of Reserve Component Members (SOFReserves), a survey conducted by Defense Manpower Data Center (Defense Manpower Data Center 2004) of Military Reservists, are provided in Table 18.1. If these disposition codes are also used to tailor fieldwork recruitment during the data collection, it would be advisable to differentiate between refusals and deployed personnel. Both of these codes are currently lumped into one disposition category: 8. Depending on the survey and the mode of data collection, the number of disposition codes can be rather large.

It is useful to specify ahead of time how disposition codes can be grouped to later compute study performance rates. In the mapping task it is important to capture all outcomes seen in the survey. Thus, in some instances, assignments made prior to data collections will need to be revised once field data are available. Survey statisticians should review the disposition code mapping to make sure that all assignments needed for weighting can be made. Even prior to data collection it is important that supervisors and data collectors understand what is later needed for weighting purposes. Once data are collected a case designation to a specific disposition code can change given the amount of data provided by the respondent (i.e., the classification into partial completes vs. nonrespondents classification) and the quality of the data provided (data reset to missing after failing edit/consistency checks).

Table 18.1: Sample dispositions for the May 2004 SOF reserves study

Disposition code	Description
1	Ineligible–based on check of updated personnel records
2	Ineligible–self/proxy report, deceased, ill, incarcerated, separated
3	Ineligible–survey self report
4	Complete eligible response
5	Incomplete eligible response
8	Refused–refusal, deployed, other refusal
9	Blank (returned questionnaire)
10	Postal nondelivery (PND)
11	Other nonrespondent

Hierarchy

More difficult and often more important than the decisions about mapping are the decisions regarding the hierarchy of outcome codes. Many sample cases will be contacted repeatedly throughout the survey and data collectors differ in their translations of preliminary response status codes to final case outcome codes. If the most recent status code is used to determine the final outcome, then the assignment is straightforward, though it might not properly reflect the case. For example, if a sample unit was successfully contacted early in the field period but subsequent contact attempts failed to lead to an interview and the final contact attempt is a noncontact, some researchers would count such a case as a refusal whereas others would count this as a noncontact. Variations in how the coding decisions are made can make comparisons among performance rates from different surveys difficult. If decisions are made based on the entire history of outcome codes, a priority coding can be very useful. Here one should agree with the client upfront about the hierarchy of codes. For example, if a sample unit that had one refusal in their history and no successful refusal conversion is recorded, then this case would be classified as a refusal even if the last outcome code was a noncontact. A detailed discussion on effects of various outcome codes in particular when comparing surveys across countries is given by Blom (2008).

18.5.2 Editing the Weighting Variables

In the editing process survey statisticians also need to ensure that relevant weighting variables are available either by matching to the sampling frame

or because they have been collected in the interview. Matching to the frame should be straightforward, though in practice files coming out of the field may not have the proper identification variables that allow a match to the sampling frame. If matching to frame information is planned, the need for these variables should be clearly communicated to field managers. If variables used for weighting are based on respondents' answers during the interview, then prior to data collection, care should be taken that the questions asked in the survey match those of benchmarking surveys. Even for demographic variables this seems like a straightforward task, but is often not that simple. For example, the questions on race/ethnicity in Fig. 18.7 were included in the 2010 US decennial census.

If a survey does not ask for race/ethnicity in exactly the same way, the survey estimate of, say, the number of Hispanics in the population will not be comparable to the census count. In that case, calibrating survey weights to census counts may actually introduce bias rather than reduce it.

18.6 Quality Control of Weighting Steps

Check Weighting Variables

Before starting the weighting process you need to check that a clean file of weighting variables is available. This means weighting variables should have no illegal codes and no missing values (or have those imputed if need be); all stratum codes and PSUs used in sampling should be clearly indicated; domain identifiers need to be present if different sampling rates were used for domains (e.g., different sampling rates for varying racial and ethnic groups, different age groups); and variables not used in sampling must be present if they are planned for poststratification or other types of calibration. When planning the weighting steps, care should be taken that weight variables are edited prior to any edits of the substantive questionnaire variables, so that weighting can proceed.

Check Selection Probabilities

Selection probabilities, also known as inclusion probabilities, are the building block for weights in most surveys. In general, selection probabilities must be between (0,1). This range check applies to most surveys. On the other hand, some designs allow units to be selected more than once (see Sect. 13.3). This can happen in samples where units are selected with probability proportional to size and some units are very large or in samples selected with replacement. In such cases, the selection probability is replaced with the expected number of hits, which can be greater than one. You need to repeat

> → NOTE: Please answer BOTH Question 8 about Hispanic origin and Question 9 about race. For this census, Hispanic origins are not races.
>
> **8. Is Person 1 of Hispanic, Latino, or Spanish origin?**
>   - ☐ **No**, not of Hispanic, Latino, or Spanish origin
>   - ☐ **Yes, Mexican**, Mexican Am., Chicano
>   - ☐ **Yes, Puerto Rican**
>   - ☐ **Yes, Cuban**
>   - ☐ **Yes, another Hispanic, Latino, or Spanish origin** — *Print origin, for example, Argentinean, Colombian, Dominican, Nicaraguan, Salvadoran, Spaniard, and so on.* ↗
>
> **9. What is Person 1's race?** *Mark ☒ one or more boxes.*
>   - ☐ **White**
>   - ☐ **Black, African Am., or Negro**
>   - ☐ **American Indian or Alaska Native** — *Print name of enrolled or principal tribe.* ↗
>
>   - ☐ **Asian Indian**   ☐ **Japanese**   ☐ **Native Hawaiian**
>   - ☐ **Chinese**   ☐ **Korean**   ☐ **Guamanian or Chamorro**
>   - ☐ **Filipino**   ☐ **Vietnamese**   ☐ **Samoan**
>   - ☐ **Other Asian** — *Print race, for example, Hmong, Laotian, Thai, Pakistani, Cambodian, and so on.* ↗   ☐ **Other Pacific Islander** — *Print race, for example, Fijian, Tongan, and so on.* ↗
>
>   - ☐ **Some other race** — *Print race.* ↗

Fig. 18.7: Ethnicity and race questions used in the 2010 decennial census.

this check for each stage of the sampling design—PSU, SSUs, etc.—and store the selection probabilities in separate fields, while also creating an additional field with the product of them all. If the design is a self-weighting design for some subgroups, you need to check the equality of the selection probabilities within each group. The National Health and Nutrition Examination Survey (NHANES), for example, is self-weighting with respect to age, gender, and race/ethnicity domains; the Commercial Buildings Energy Consumption Survey (CBECS) is self-weighting within building size and building usage categories. In practice, it will be often necessary to allow for exceptions. Good documentation of selection probabilities is required in international surveys that have to assure comparability across countries. The Programme for the International Assessment of Adult Competencies (PIAAC) study issued by

the Office of Economic Cooperation and Development (OECD)[12] provides
the following instructions for information to be collected for each stage of
selection:

- A list of variables used for stratification and their categories
- Procedures used to construct the sampling frame and to stratify and select
  sampling units
- The definition of sampling unit
- Data sources used for forming sampling units
- Average, minimum and maximum cluster size
- List of certainty units, such as large primary sampling units
- Measure of size for the sampling units, as well as minimum measure of size
- A description of units collapsed to obtain the minimum measure of size
- A formula describing the probability of selection for each sampling unit
- A sample selection worksheet that provides the following details:

  - Target population totals for each level of stratification
  - Number of sampling units on the frame for each level of stratification
  - Total measure of size on the frame for each level of stratification
  - Target sample size, or rate, prior to sampling for each stratum
  - Actual sample size for each stratum, for certainty and noncertainty
    units
  - Weighted sample estimates for each level of stratification, where the
    weight is equal to the inverse of the overall selection probability for the
    sampling unit of the current selection stage

Exact Checks on Record Counts and Weight Sums

When checking weights, it is helpful to remember that the numbers of records
within an input file entering each step must equal the number of records on
the output file exiting the step plus any records discarded. Likewise, the sums
of the incoming weights and outgoing weights in each step must balance. For
example, the sum of the weights in the input file that has respondents and
nonrespondents in it must equal the sum of the weights of the respondents
in the outgoing file after certain types of nonresponse adjustments have been
made.

Statistical Checks

Finally the sum of the weights should be an estimate of the number of units
in the population. You therefore should compare the sum of the weights to
the external population count. We suggest you make this check after each

---

[12] PIAAC Technical Standards and Guidelines, Second Draft presented at the Meet-
ing of the National Project Managers, 23–27 March 2009, Barcelona, Spain.

weighting step (see Chaps. 13 and 14): base weight, adjustments for unknown eligibility, and adjustments for nonresponse. If calibration is used, the estimate of the total for each calibration variable should exactly equal its control total and should have a standard error of 0, as we saw in Chap. 14. For example, if age group poststrata are used in a household survey, the estimated total number of persons in each age group should equal the population control total. If a method of estimating variances is used that properly accounts for the poststratification, the SEs of the estimated total number of persons in each poststratum should be 0.

Another excerpt from the PIAAC instruction regarding general quality control procedures in the weighting steps is the following:

- The quality checks will be performed after each step in the weighting process. The checks will include:
  - Reviewing the distribution of weights at each stage to identify any missing or extreme values.
  - Computing the weighted frequencies of important survey characteristics after each weighting adjustment to show how each adjustment affects the estimates for key survey variables. In addition, weighted frequencies will be compared to reliable external totals.
  - Reviewing a random listing of records for abnormalities.
  - Producing the mean, median, minimum, and maximum and checking for each jackknife replicate weight after each weight adjustment.
  - After the final weights are produced, producing preliminary standard errors and design effects on survey variables as a check on the replicate weights.

Although these are intended for PIAAC and include some items specifically for the methods used in that program (e.g., checks on replicate weights), the general steps apply to many surveys.

18.7 Specification Writing and Programming

Writing good specifications will save a lot of work later on, avoid ambiguities in the steps that must be completed, and, in the end, lead to better documented projects. In smaller projects it is tempting to not even write specifications at all because the same person would be writing the specification as well as the programming code. Nevertheless, taking the extra step of writing specifications allows for effective communication with the team, for good documentation of the work, and for making changes afterwards. When specification memos are written, it is useful to just write one memo for each task. Having a standard format for the names of files that contain the memos is a good practice. The actual file name of a specification memo should include indicators for the task within a sequence of tasks, include the version

Project Memo Log

Memo	Date	Author	Reviewer	Subject	Programs
W1.1	12/18/06	R. Valliant	S. Miller	Weighting Plan	*None*
W2.1	01/17/07	J. Dever	F. Kreuter	Weighting Final Report	*None*
W3.2	02/26/07	S. Miller	R. Valliant	Specifications for 1st-stage weights	1 PSU Weights.sas
W4.1	02/19/07	R. Valliant	J. Dever	Specifications for 2nd-stage weights	2 SSU Weights.sas
W5.1	04/03/07	D. Gilbert	R. Valliant	Preliminary weights file	3 Sample Weights File.sas

Memo file names = <number>.<version>task

Memo/Program fields contain hyperlinks to specified files in shared project folder

Fig. 18.8: Project memo log.

M E M O R A N D U M

TO: ____   **DATE:** *<date of each version>*
FROM: ____
CC: *<project director>*, *<project manager>*, *<quality assurance task leader>*
SUBJECT: *<project name>* – *<process description>*: *<task title>*
KEY WORDS: ____

Overview
This memorandum provides specifications for ...... The task to create ..... involves the following steps:

Input Files
      <filenames and location>

Output Files
      <filenames and location>

SubTask 1 – *<title>*
      <description, variable names, specs to create/check variable, formats>

SubTask 2 – *<title>*
      <etc.>

Formats

Variable Name	Variable Type	Variable Label	Format	Value & Label
GENDER	NUM	Respondent's Gender	gen_.	1 = Male 2 = Female

Fig. 18.9: Example memo.

number of the memo, and indicate the purpose of the task. For example, a file name *S1.2.doc* may contain the second version of a specification memo related to the first task in the sampling section. A master file should provide good mapping between each individual task, the program, and preferably the programmer as well. An example for such mapping is given in Fig. 18.8. A template for the memo is given in Fig. 18.9.

We advise writing separate programs that perform individual tasks. Modular programming allows tracking changes to parts of the process. Figure 18.10 shows an example of header comment statements that might be used in a SAS program written for a specific task. Of course, how a task is defined is a matter of taste. Tasks can be large or small depending on the organizational style of the programmer. As an example, consider a school survey. "Construct sample frame" could be a task, but it may be more manageable to break it into several steps:

- Download latest school universe from the Department of Education web site.
- Eliminate ineligible schools based on survey eligibility criteria.
- Check file for missing data.
- Create stratum codes.
- Write output file.

One or more of these steps might deserve a separate task number and program, depending on the details required for a step. An example of part of a flowchart for weight calculation in the National Assessment of Education Progress (NAEP) is shown later in Fig. 18.11. (Weighting includes a number of other steps that follow the continuation block at the bottom of the chart but are not shown here.) Particular tasks in the flowchart, like W0, W1, and WP0, have their own specification memos. The specifications for these tasks can be saved in files whose names include the task numbers as shown in Fig. 18.8.

Within each program, comments should be included to highlight subtasks and their purpose. Likewise, programs need comments to link different operations to the steps in the specification such as those suggested in the project memo log (Fig. 18.8). Numbers in program file names can be used to keep steps in sequence (see the last column of Fig. 18.8). We recommend creating and keeping output log files from programs including program headers, as illustrated in Fig. 18.10, that indicate the task of the program (and the corresponding log file). General rules about documentation of programming codes and effective coding can be found in Long (2009) and Kohler and Kreuter (2012).

18.8 Project Documentation and Archiving

A quality project requires proper documentation. Such documentation needs to be well organized so that decisions are recorded and can be retraced later on. If issues arise, well-documented projects can address those easily. The project documentation should at any point in time be ready for audit. When thinking of how to structure your documentation it might help to develop one system of documentation for external and one for internal use.

```
/**************************************************************************/
/* FILE:     1 Process Frame.sas                                        */
/* PROJECT: 2006 Medicinal Laughter Study                               */
/* DATE:     09/20/07                                                   */
/* AUTHORS: J. Dever, Sampling Task Leader                              */
/* SPECS:    S5.1.doc                                                   */
/* PURPOSE: Process sampling frame file, tabulate frame counts, and create */
/*          stratification variables.                                   */
/* INPUT:    PFML0907.sas7bdat (old = PFML0707.sas7bdat)                */
/* OUTPUT:   Frame0907.sas7bdat                                         */
/* REVISED: 10/31/07 Reran program with updated frame file (S5.2.doc).   */
/**************************************************************************/
```

Fig. 18.10: Program header (SAS file).

External Documentation

The external documentation needs to include a sample design report with sufficient details to make the work reproducible and defensible. The sample design report should also allow a comparison of the contractual promises to the final outcome (and reasons for any change). That means a sample design report would include information on the target population, the sampling frame, the sample size, the sample design, the sample selection, the response rates (and other prespecified performance indicators), sample monitoring, and sample quality control steps. Likewise the weighting reports would include all details on designing the weights and adjustments and importantly an evaluation of the final weights (e.g., weight variation). Given that the client will receive a data file, the external report should also include a layout of the analysis file and a codebook. The overall layout of the file needs to be determined with the client (we recommend beforehand) but usually includes all edited questionnaire items, imputed values and imputation flags, sample weights and disposition codes, final weights, and individual adjustment weights. Depending on the confidentiality agreements, variables used alone or in combination to identify a participant need to be masked. The codebook itself describes the file layout, variable names and labels, value labels, etc. Sometimes codebooks are created in the form of an annotated questionnaire. Good example codebooks can be found for the General Social Survey (GSS, http://www3.norc.org/GSS+Website) and the National Health Interview Survey (http://www.cdc.gov/nchs/nhis/nhis_questionnaires.htm).
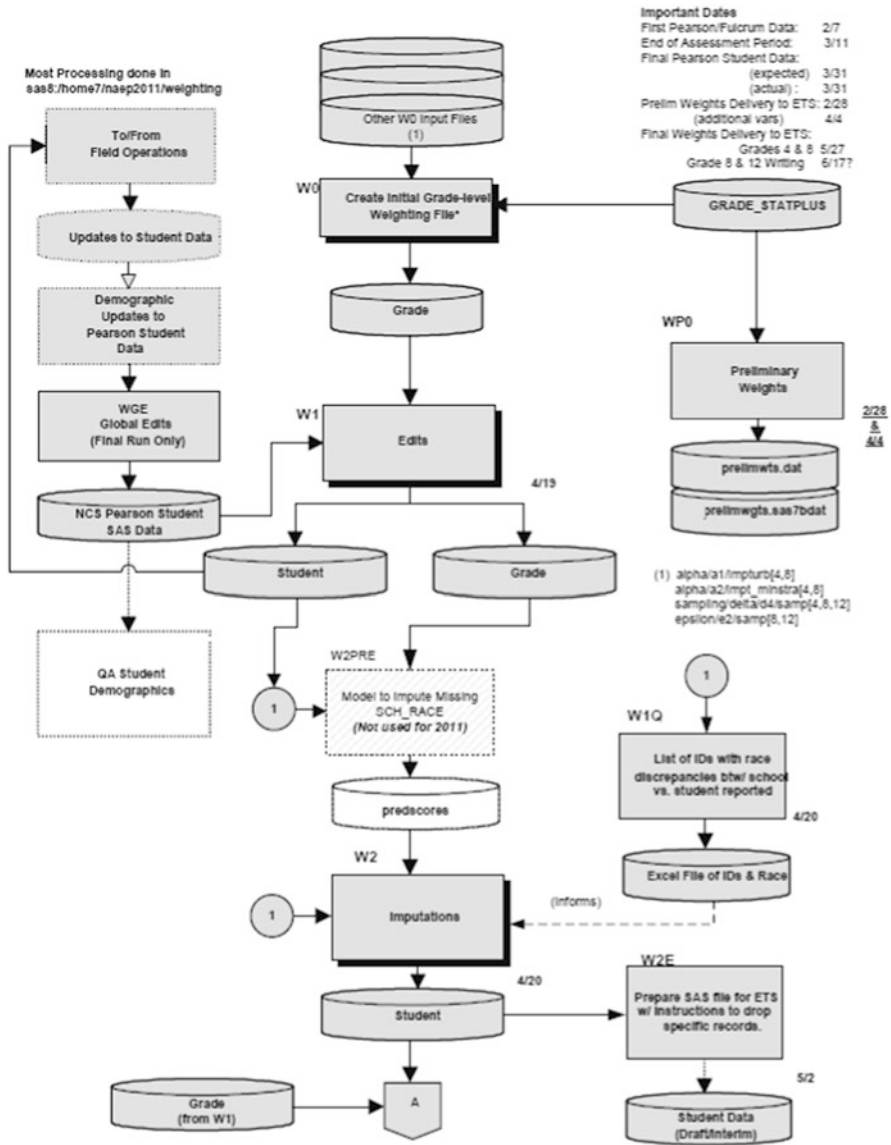
Fig. 18.11: Flowchart for weighting in the NAEP survey.

Internal Documentation

Internal documentation is typically much more detailed than external documentation. Specification memos are included along with a dictionary of the files that contain memos and programs. Intermediate files that are created during frame creation, sample selection, field data collection, and weight computation will all be part of internal documentation.

Appendix A
Notation Glossary

This appendix collects much of the notation used in chapters of this book. More detailed descriptions can be found in the chapters that are referenced below.

Sample Design and Sample Size for Single-Stage Surveys (Chap. 3)

General Notation

N = number of elements in the population

$\bar{y}_U = \sum_{i=1}^{N} y_i \big/ N$ = finite population mean of an analysis variable y

$t_U = \sum_{i=1}^{N} y_i$ = population total of an analysis variable y

$\bar{y}_s = \sum_{i=1}^{n} y_i / n$ = sample mean of a variable y

$S^2 = \sum_{i=1}^{N} (y_i - \bar{y}_U)^2 \big/ (N - 1)$ = population variance or *unit variance* of y

$CV_U = S / \bar{y}_U$ = population (or unit) coefficient of variation of y.

$CV_U^2 = S^2 / \bar{y}_U^2$ = population (or unit) relvariance

$\hat{\theta}$ is an estimator of some population parameter, e.g., a total or mean,

$E\left(\hat{\theta}\right)$ = expected value of $\hat{\theta}$ in repeated sampling under a particular sampling design

$V\left(\hat{\theta}\right)$ = variance of an estimator $\hat{\theta}$

$SE\left(\hat{\theta}\right) = \sqrt{V\left(\hat{\theta}\right)}$ = standard error of $\hat{\theta}$

$v\left(\hat{\theta}\right)$ = estimator of $V\left(\hat{\theta}\right)$

$se\left(\hat{\theta}\right) = \sqrt{v\left(\hat{\theta}\right)}$ = estimated standard error of $\hat{\theta}$

$CV\left(\hat{\theta}\right) = \sqrt{V\left(\hat{\theta}\right)}\Big/\theta = $ coefficient of variation (CV) of $\hat{\theta}$

$cv\left(\hat{\theta}\right) = \sqrt{v\left(\hat{\theta}\right)}\Big/\hat{\theta} = $ sample estimator of $CV\left(\hat{\theta}\right)$

$relvar\left(\hat{\theta}\right) = \left[cv\left(\hat{\theta}\right)\right]^2 = $ estimated relvariance of $\hat{\theta}$

$V\left(\bar{y}_s\right) = \left(\frac{1}{n} - \frac{1}{N}\right)S^2 = $ variance of $\bar{y}_s$ in a simple random sample selected without replacement ($srswor$)

$\hat{t} = N\bar{y}_s$ is an estimator of the population total of y in an $srswor$

$p_U = $ population proportion of a 0–1 characteristic; $q_U = 1 - p_U$

Single-Stage Sampling

$srswor = $ simple random sample selected without replacement

$srswr = $ simple random sample selected with replacement

$ppswr = $ probability proportional to some measure of size selected with replacement

$n = $ number of sample elements

$pwr = $ "probability with replacement," used to refer to any design in which the first-stage units are selected with replacement

$\bar{y}_s = \sum_{i=1}^{n} y_i/n = $ mean of y in a simple random sample of n elements

$p_s = \sum_{i \in s} y_i/n$ is an estimator of p_U from a simple random sample of n elements

$p_i = $ one-draw selection probability for unit i in a sample selected with varying probabilities

$\pi_i = $ selection probability of unit i in a sample of size n selected with varying probabilities

$\pi_{ij} = $ joint selection probability of units i and j in a sample of size n selected with varying probabilities

$\hat{\bar{y}}_\pi = N^{-1}\sum_{i \in s} y_i/\pi_i = \pi$-estimator of the population total of y; also called the Horvitz–Thompson estimator

$\hat{\bar{y}}_{pwr} = \frac{1}{Nn}\sum_s \frac{y_i}{p_i} = $ "probability with-replacement" estimator of a total when the sample is selected with varying probabilities and with replacement

$\hat{\bar{y}}_r = \hat{\bar{y}}_\pi + \sum_{j=1}^{p} b_j\left(\bar{x}_{Uj} - \hat{\bar{x}}_{\pi j}\right) = $ general regression estimator of a total; $\bar{x}_{Uj}$ is the population mean of an auxiliary variable x_j ($j=1, \ldots, p$); $\hat{\bar{x}}_{\pi j}$ is the π-estimator of the total of x_j; b_j is an estimator of the slope on x_j in a regression of y on all p x's

Stratified Single-Stage Sampling

$stsrswor = $ stratified simple random sample selected without replacement in each stratum

N_h = number of population elements in stratum h

$W_h = N_h/N$ = population proportion of units in stratum h

y_{hi} = value of an analysis variable for unit i in stratum h

$S_h^2 = \sum_{i=1}^{N_h} (y_{hi} - \bar{y}_{Uh})^2 \big/ (N_h - 1)$ = population or unit variance in stratum h

U_h= set of all units in the population from stratum h

$\bar{y}_{Uh} = \sum_{i=1}^{N_h} y_{hi} \big/ N_h$ = population mean in stratum h

$\bar{y}_U = \sum_{h=1}^{H} W_h \bar{y}_{Uh}$ = population mean expressed as a weighted sum of stratum means

s_h = sample of elements from stratum h

n_h = number of sample elements from stratum h in a stratified simple random sample

$\bar{y}_{hs} = \sum_{i \in s_h} y_{hi}/n_h$ = sample mean of elements in stratum h

$\bar{y}_{st} = \sum_{h=1}^{H} W_h \bar{y}_{sh}$ = estimated mean when an *stsrs* is selected

$p_{st} = \sum_{h=1}^{H} W_h p_{sh}$ = estimated proportion of units with a characteristic when an *stsrs* is selected; p_{sh} = proportion of units in the sample in stratum that have the characteristic

$V(\bar{y}_{st}) = \sum_{h=1}^{H} W_h^2 \frac{1-f_h}{n_h} S_h^2$ = variance in *stsrswor* of $\bar{y}_{st}$ where $f_h = n_h/N_h$

c_h = cost per element of all costs that vary with the number of sample elements

Designing Multistage Samples (Chap. 9)

Two-Stage Sampling

U = universe of PSUs

M = number of PSUs in universe

U_i = universe of elements in PSU i

N_i = number of elements in the population for PSU i

$N = \sum_{i \in U} N_i$ = total number of elements in the population

π_i = selection probability of PSU i

π_{ij} = joint selection probability of PSUs i and j

m = number of sample PSUs

n_i = number of sample elements in PSU i

s = set of sample PSUs

s_i = set of sample elements in PSU i

y_{ik} = analysis variable for element k in PSU i

$\bar{y}_U$ = mean per element in the population

$\bar{y}_{Ui}$ = mean per element in the population in PSU i

$t_U = \sum_{i \in U} \sum_{k \in U_i} y_k$ = population total of an analysis variable y

$\hat{t}_\pi = \sum_{i \in s} \frac{\hat{t}_i}{\pi_i}$ = π-estimator of the population total of y where $\hat{t}_i = (N_i/n_i) \sum_{k \in s_i} y_k$. The sample design is two-stage with PSUs selected with varying probabilities and elements selected with equal probability within each PSU

$S_{U1}^2 = \frac{\sum_{i \in U}(t_i - \bar{t}_U)^2}{M-1}$ = variance among PSU totals with t_i being the population total of y in PSU i

$\bar{t}_U = \sum_{i \in U} t_i / M$ is the mean total per PSU

$S_{U2i}^2 = \frac{\sum_{k \in U_i}(y_k - \bar{y}_{Ui})^2}{N_i - 1}$ = unit variance of y among the elements in PSU i

$V\left(\hat{t}_\pi\right) = \frac{M^2}{m} \frac{M-m}{M} S_{U1}^2 + \frac{M}{m} \sum_{i \in U} \frac{N_i^2}{n_i} \frac{N_i - n_i}{N_i} S_{U2i}^2$
= variance of the π-estimator in a two-stage sample in which PSUs are selected by *srswor* and units within sample PSUs are selected by *srswor*

$B^2 = S_{U1}^2 / \bar{t}_U^2$ = unit relvariance among PSU totals

$W^2 = \frac{1}{M\bar{y}_U^2} \sum_{i \in U} S_{U2i}^2$ = within-PSU relvariance among elements

$\hat{t}_{pwr} = \frac{1}{m} \sum_{i \in s} \frac{\hat{t}_i}{p_i}$ = *pwr*-estimator of a total when the PSUs are selected with replacement; $\hat{t}_i = \frac{N_i}{n_i} \sum_{k \in s_i} y_{ik}$ is the estimated total for PSU i from a simple random sample and p_i is the 1-draw selection probability of PSU i.

When m PSUs are selected with replacement and an *srswor* of size n_i is selected in sample PSU i, the variance of $\hat{t}_{pwr}$ is

$$V\left(\hat{t}_{pwr}\right) = \frac{1}{m} \sum_{i \in U} p_i \left(\frac{t_i}{p_i} - t_U\right)^2 + \sum_{i \in U} \frac{N_i^2}{m p_i n_i} \left(1 - \frac{n_i}{N_i}\right) S_{U2i}^2 \qquad \text{(A.1)}$$
$$\equiv V_{PSU} + V_{SSU}.$$

Special case of the variance of $\hat{t}_{pwr}$ when m PSUs are selected with replacement and $\bar{n}$ elements are selected by *srswor* in each PSU:

$$V\left(\hat{t}_{pwr}\right) = \frac{S_{U1(pwr)}^2}{m} + \frac{1}{m\bar{n}} \sum_{i \in U} \left(1 - \frac{\bar{n}}{N_i}\right) \frac{N_i^2 S_{U2i}^2}{p_i}.$$

Special case of the relvariance of $\hat{t}_{pwr}$ when PSUs are selected with replacement, $\bar{n}$ elements are selected by *srswor* in each PSU, and the within-PSU sampling fraction, $\bar{n}/N_i$, is negligible:

$$\frac{V\left(\hat{t}_{pwr}\right)}{t_U^2} \doteq \frac{B^2}{m} + \frac{W^2}{m\bar{n}} = \frac{B^2 + W^2}{m\bar{n}} \left[1 + \delta\left(\bar{n} - 1\right)\right]$$
$$B^2 = S_{U1(pwr)}^2 \Big/ t_U^2$$
$$= \text{between relvariance component}$$

$$W^2 = \frac{1}{t_U^2} \sum_{i \in U} N_i^2 \frac{S_{U2i}^2}{p_i}$$

$$= \text{within relvariance component} \qquad (A.2)$$

$$\delta = B^2 / \left(B^2 + W^2\right)$$

$$= \text{measure of the homogeneity of elements within PSUs}$$

Variance Component Estimation in Two-Stage Sampling

In a design in which PSUs are selected by *ppswr* and elements within PSUs are selected by simple random sampling, the variance of the *pwr*-estimator is $V\left(\hat{t}_{pwr}\right) = V_{PSU} + V_{SSU}$ as defined in (A.1). Estimators of the variance components are

$$v_{SSU} = \sum_{i \in s} \frac{\hat{V}_i}{\left(\pi_i^*\right)^2}$$

$$= \text{an estimator of } V_{SSU} \text{ in (A.1)}$$

$$v_{PSU} = \frac{1}{m\left(m-1\right)} \sum_{i \in s} \left(\frac{\hat{t}_{i\pi}}{p_i} - \hat{t}_{pwr}\right)^2 - \sum_{i \in s} \frac{1 - \pi_i^*}{\left(\pi_i^*\right)^2} \hat{V}_i$$

$$= \text{an estimator of } V_{PSU} \text{ in (A.1)}$$

where $\hat{V}_i = \frac{N_i^2}{n_i}\left(1 - f_i\right)\hat{S}_{2i}^2$ where $\pi_i^* = m p_i$, $\hat{S}_{2i}^2 = \left(n_i - 1\right)\sum_{k \in s_i}\left(y_k - \bar{y}_{si}\right)^2$, and $\bar{y}_{si} = \sum_{k \in s_i} y_k / n_i$.

Estimators of the between- and within-relvariance components in (A.2) for a *ppswr/srs* sample are

$$\hat{B}^2 = \frac{1}{\hat{t}_{pwr}^2}\left\{\frac{1}{(m-1)}\sum_{i \in s}\left(\frac{\hat{t}_{i\pi}}{p_i} - \hat{t}_{pwr}\right)^2 - \sum_{i \in s}\frac{1 - \pi_i^*}{m p_i^2}\hat{V}_i\right\} \text{ and}$$

$$W^2 = \frac{1}{\hat{t}_{pwr}^2}\sum_{i \in s}\frac{N_i^2 \hat{S}_{2i}}{m p_i^2}$$

Three-Stage Sampling

$U_i = $ population of SSUs in PSU i

$U_{ij} = $ population of elements in PSU/SSU ij

$N_i = $ population number of SSUs in PSU i

$N = \sum_{i \in U} N_i = $ total number of SSUs in the population

$m = $ number of sample PSUs

$n_i = $ number of sample SSUs; Chap. 9 gives results applicable to *srswor* sampling of SSUs

$Q_{ij} = $ population number of elements in PSU/SSU ij

$Q_i = \sum_{j \in U_i} Q_{ij}$ = total number of elements in PSU i in the population is Q

q_{ij} = number of elements selected by *srswor* from PSU/SSU ij

$\bar{y}_{sij} = \sum_{k \in s_{ij}} y_k \big/ q_{ij}$, the sample mean of elements in SSU ij

$t_{ij} = \sum_{k \in U_{ij}} y_k$ being the population total for PSU/SSU ij

$\hat{t}_{ij} = Q_{ij}\bar{y}_{sij}$, the estimated total for SSU ij assuming that an equal probability sample is selected within the SSU

$\hat{t}_{i\pi} = \frac{N_i}{n_i} \sum_{j \in s_i} \hat{t}_{ij}$, the estimated total for PSU i assuming that SSUs are selected by *srs*

$\hat{t}_{\pi} = \frac{M}{m} \sum_{i \in s} \frac{N_i}{n_i} \sum_{j \in s_i} \frac{Q_{ij}}{q_{ij}} \sum_{k \in s_{ij}} y_k$ = π-estimator if an *srswor* (or another type of equal probability sample) is selected at each stage.

$S_{U1}^2 = \frac{\sum_{i \in U}(t_i - \bar{t}_U)^2}{M-1}$ = variance among PSU totals

$S_{U2i}^2 = \frac{1}{N_i-1} \sum_{j \in U_i} (t_{ij} - \bar{t}_{Ui})^2$ = unit variance of SSU totals in PSU i where

$\quad \bar{t}_{Ui} = \sum_{j \in U_i} t_{ij} \big/ N_i$ is the average total per SSU in PSU i

$S_{U3ij}^2 = \frac{1}{Q_{ij}-1} \sum_{k \in U_{ij}} (y_k - \bar{y}_{Uij})^2$ = unit variance among elements in PSU/SSU ij where $\bar{y}_{Uij} = \sum_{k \in U_{ij}} y_k \big/ Q_{ij}$

Relvariance of the π-estimator in three-stage sampling when each stage of sampling is *srswor*:

$$\frac{V(\hat{t}_{\pi})}{t_U^2} = \frac{1}{t_U^2} \left\{ \frac{M^2}{m} \frac{M-m}{M} S_{U1}^2 + \frac{M}{m} \sum_{i \in U} \frac{N_i^2}{n_i} \frac{N_i - n_i}{N_i} S_{U2i}^2 + \frac{M}{m} \sum_{i \in U} \frac{N_i}{n_i} \sum_{j \in U_i} \frac{Q_{ij}^2}{q_{ij}} \frac{Q_{ij} - q_{ij}}{Q_{ij}} S_{U3ij}^2 \right\}.$$

A special case of the relvariance of the π-estimator in three-stage sampling follows. The requirements are that all stages are *srswor* the same number of SSUs, $\bar{n}$, is selected from each PSU, the same number of elements, $\bar{\bar{q}}$, is selected from each SSU, the number of SSUs in the population for each PSU is the same, $N_i = \bar{N}$, and the number of elements in the population in each SSU is the same, $Q_{ij} = \bar{\bar{Q}}$:

$$\frac{V(\hat{t}_{\pi})}{t_U^2} = \frac{1}{\bar{\bar{y}}_U^2} \left\{ \frac{1}{m} \frac{M-m}{M} S_1^2 + \frac{1}{m\bar{n}} \frac{\bar{N}-\bar{n}}{\bar{N}} S_2^2 + \frac{1}{m\bar{n}\bar{\bar{q}}} \frac{\bar{\bar{Q}}-\bar{\bar{q}}}{\bar{\bar{Q}}} S_3^2 \right\}$$

where $\bar{\bar{y}}_U = \sum_{i \in U} \sum_{j \in U_i} \sum_{k \in U_{ij}} y_k \big/ M\bar{N}\bar{\bar{Q}}$

$\quad S_1^2 = (M-1)^{-1} \sum_U (\bar{\bar{y}}_{Ui} - \bar{\bar{y}}_U)^2$; $\bar{\bar{y}}_{Ui} = t_i \big/ \bar{N}\bar{\bar{Q}}$ is the mean per element in PSU i

$\quad S_2^2 = \sum_{i \in U} \sum_{j \in U_i} (\bar{y}_{Uij} - \bar{\bar{y}}_{Ui})^2 \big/ M(\bar{N}-1)$

$\quad \bar{y}_{Uij} = \sum_{k \in U_{ij}} y_k \big/ \bar{\bar{Q}}$ is the mean per element in SSU ij; and

$\quad S_3^2 = \sum_{k \in U_{ij}} (y_k - \bar{y}_{Uij})^2 \big/ M\bar{N}(\bar{\bar{Q}}-1)$

Another special case of the relvariance when all stages are selected by *srswor* is

$$\frac{V\left(\hat{t}_\pi\right)}{t_U^2} = \left\{\frac{1}{m}\frac{M-m}{M}B^2 + \frac{1}{m\bar{n}}\frac{\bar{N}-\bar{n}}{\bar{N}}W_2^2 + \frac{1}{m\bar{n}\bar{\bar{q}}}\frac{\bar{\bar{Q}}-\bar{\bar{q}}}{\bar{\bar{Q}}}W_3^2\right\}$$

with $B^2 = S_1^2/\bar{\bar{y}}_U^2$, $W_2^2 = S_2^2/\bar{\bar{y}}_U^2$, and $W_3^2 = S_3^2/\bar{\bar{y}}_U^2$

When PSUs are selected with replacement, the *pwr*-estimator of a total is:
$\hat{t}_{pwr} = \frac{1}{m}\sum_{i\in s}\frac{\hat{t}_i}{p_i}$
where $\hat{t}_i$ is a design-unbiased estimator of the total for PSU i and p_i is the 1-draw probability of selection of PSU i.

Relvariance of the *pwr*-estimator in a three-stage sample in which the first stage is selected with varying probabilities and with replacement and the second and third stages are selected by *srswor*:

$$\frac{V\left(\hat{t}_{pwr}\right)}{t_U^2} = \frac{1}{t_U^2}\left\{\frac{S_{U1(pwr)}^2}{m} + \frac{1}{m}\sum_{i\in U}\frac{N_i^2}{p_i n_i}\frac{N_i-n_i}{N_i}S_{U2i}^2+\right.$$
$$\left.\frac{1}{m}\sum_{i\in U}\frac{1}{p_i}\frac{N_i}{n_i}\sum_{j\in U_i}\frac{Q_{ij}^2}{q_{ij}}\frac{Q_{ij}-q_{ij}}{Q_{ij}}S_{U3ij}^2\right\} \qquad (A.3)$$
$$\equiv \frac{1}{t_U^2}\left\{V_{PSU} + V_{SSU} + V_{TSU}\right\}$$

$S_{U1(pwr)}^2 = \sum_{i\in U}p_i\left(\frac{t_i}{p_i}-t_U\right)^2$ is the unit variance appropriate for the *pwr*-estimator when the first stage is selected with replacement
$\frac{V\left(\hat{t}_{pwr}\right)}{t_U^2} \doteq \frac{B^2}{m} + \frac{W_2^2}{m\bar{n}} + \frac{\bar{\bar{Q}}-\bar{\bar{q}}}{\bar{\bar{Q}}}\frac{W_3^2}{m\bar{n}\bar{\bar{q}}}$ is a special case of the three-stage sampling relvariance in a *ppswr/srswor/srswor* design that assumes that $\bar{n}$ SSUs are sampled in each sample PSU, the sampling fractions of SSUs in each PSU, $\bar{n}/N_i$, are small, and $\bar{\bar{q}}$ elements in each sample SSU. The components are $B^2 = S_{U1(pwr)}^2/t_U^2$, $W_2^2 = \frac{1}{t_U^2}\sum_{i\in U}N_i^2 S_{U2i}^2/p_i$, and $W_3^2 = \frac{1}{t_U^2}\sum_{i\in U}\frac{N_i}{p_i}\sum_{j\in U_i}Q_{ij}^2 S_{U3ij}^2$

A rewritten version of the relvariance in a *ppswr/srswor/srswor* design is

$$\frac{V\left(\hat{t}_{pwr}\right)}{t_U^2} \doteq \frac{\tilde{V}}{m\bar{n}\bar{\bar{q}}}\left\{k_1\delta_1\bar{n}\bar{\bar{q}} + k_2\left[1 + \delta_2\left(\bar{\bar{q}}-1\right)\right]\right\}, \qquad (A.4)$$

where
$k_1 = \tilde{V}_1/\tilde{V}$ with $\tilde{V} = \frac{1}{Q-1}\sum_{i\in U}\sum_{j\in U_i}\sum_{k\in U_{ij}}\left(y_k-\bar{y}_U\right)^2/t_U^2$ being the unit variance of y in the population
$k_2 = \tilde{V}_2/\tilde{V}$ with $\tilde{V}_2 = W_2^2 + \left(\bar{\bar{Q}}-1\right)W_3^2/\bar{\bar{Q}}$
$\delta_1 = \frac{B^2-W^2/\bar{Q}}{\tilde{V}_1}$ is a measure of homogeneity of elements within PSUs (i.e., ignoring SSU membership); $\tilde{V}_1 = B^2 + \left(\bar{Q}-1\right)W^2/\bar{Q}$, $\bar{Q} = \sum_{i\in U}Q_i/M$ is the average number of elements per PSU

$W^2 = \frac{1}{t_U^2} \sum_{i \in U} Q_i^2 S_{U3i}^2 / p_i$ with $S_{U3i}^2 = \frac{1}{Q_i - 1} \sum_{j \in U_i} \sum_{k \in U_{ij}} (y_k - \bar{y}_{Ui})^2$ and

$\bar{y}_{Ui} = \sum_{j \in U_i} \sum_{k \in U_{ij}} y_k / Q_i$, i.e., S_{U3i}^2 is the element-level variance among all elements in PSU i

$\delta_2 = \frac{W_2^2 - W_3^2 / \bar{Q}}{\tilde{V}_2}$ is a measure of homogeneity of elements within the SSUs

Variance Component Estimation in Three-Stage Sampling

Direct estimates of the components of the variance of the *pwr*-estimator of the total in (A.3) can be made from a sample. The estimates presented below are for the case of *ppswr* sampling of m PSUs and simple random sampling of n_i SSUs in PSU i and q_{ij} elements in SSU ij. First, define

$\hat{S}_{2ai}^2 = (n_i - 1)^{-1} \sum_{j \in s_i} \left(\hat{t}_{ij} - \bar{\hat{t}}_i \right)^2$, the sample variance among estimated SSU totals, where $\bar{\hat{t}}_i = \sum_{j \in s_i} \hat{t}_{ij} / n_i$ and $\hat{t}_{ij} = Q_{ij} \bar{y}_{sij}$, the estimated total for SSU ij

$\hat{S}_{3ij}^2 = (q_{ij} - 1)^{-1} \sum_{k \in s_{ij}} (y_k - \bar{y}_{sij})^2$, the sample variance among elements in SSU ij

$\hat{V}_{3ij} = \frac{Q_{ij}^2}{q_{ij}} \frac{Q_{ij} - q_{ij}}{Q_{ij}} \hat{S}_{3ij}^2$, the estimated variance of the estimated total $\hat{t}_{ij}$ for SSU ij

$\hat{S}_{3ij}^2 = (q_{ij} - 1)^{-1} \sum_{k \in s_{ij}} (y_k - \bar{y}_{sij})^2$, the sample variance among elements in SSU ij, an estimator of S_{U3ij}^2

$\hat{V}_{3ij} = \frac{Q_{ij}^2}{q_{ij}} \frac{Q_{ij} - q_{ij}}{Q_{ij}} \hat{S}_{3ij}^2$, the estimated variance of the estimated total $\hat{t}_{ij}$ for SSU ij;

$\hat{S}_{2bi}^2 = \frac{1}{n_i} \sum_{j \in s_i} \hat{V}_{3ij}$

$\hat{S}_{2i}^2 = \hat{S}_{2ai}^2 - \hat{S}_{2bi}^2$, an estimator of S_{2Ui}^2 the variance of the SSU totals;

$\hat{S}_{1a}^2 = \frac{1}{m-1} \sum_{i \in s} \left(\frac{\hat{t}_{i\pi}}{p_i} - \hat{t}_\pi \right)^2$;

$\hat{S}_{1b}^2 = \frac{1}{m} \sum_{i \in s} \frac{N_i^2}{p_i n_i} \left[(1 - f_{2i}) \hat{S}_{2ai}^2 + f_{2i} S_{2bi}^2 \right]$ where $f_{2i} = n_i / N_i$

$\hat{S}_1^2 = \hat{S}_{1a}^2 - \hat{S}_{1b}^2$, an estimator of $S_{U1(pwr)}^2$

The estimator of the third-stage component,

$$V_{TSU} = \frac{1}{m} \sum_{i \in U} \frac{1}{p_i} \frac{N_i}{n_i} \sum_{j \in U_i} \frac{Q_{ij}^2}{q_{ij}} \frac{Q_{ij} - q_{ij}}{Q_{ij}} S_{U3ij}^2,$$

in (A.3) is

$$v_{TSU} = \sum_{i \in s} \frac{1}{(mp_i)^2} \frac{N_i^2}{n_i^2} \sum_{j \in s_i} \hat{V}_{3ij}.$$

The estimator of the second-stage component,

$V_{SSU} = \frac{1}{m} \sum_{i \in U} \frac{N_i^2}{p_i n_i} \frac{N_i - n_i}{N_i} S_{U2i}^2$, in (A.3) is

$$v_{SSU} = \sum_{i \in s} \frac{1}{(mp_i)^2} \frac{N_i^2}{n_i} \frac{N_i - n_i}{N_i} \hat{S}_{2i}^2.$$

The estimator of the first-stage component, $V_{PSU} = S_{U1(pwr)}^2 \big/ m$, in (A.3) is

$$v_{PSU} = \hat{S}_1^2 \big/ m.$$

Special case of the relvariance of the *pwr*-estimator in three-stage sampling for the case of the same number of sample SSUs, $\bar{n}$, and the same number of sample elements, $\bar{\bar{q}}$, and population elements, $\bar{\bar{Q}}$, in each SSU; an *ad hoc fpc*, $(\bar{N} - \bar{n}) \big/ \bar{N}$, is used on the second-stage component where $\bar{N}$ is the average number of SSUs per PSU:

$$\frac{v\left(\hat{t}_{pwr}\right)}{\hat{t}_{pwr}^2} = \frac{\hat{B}^2}{m} + \frac{\bar{N} - \bar{n}}{\bar{N}} \frac{\hat{W}_2^2}{m\bar{n}} + \frac{\bar{\bar{Q}} - \bar{\bar{q}}}{\bar{\bar{Q}}} \frac{\hat{W}_3^2}{m\bar{n}\bar{\bar{q}}},$$

where $\hat{B}^2 = \frac{\hat{S}_1^2}{\hat{t}_{pwr}^2}$,

$\hat{W}_2^2 = \frac{1}{\hat{t}_{pwr}^2} \sum_{i \in s} \frac{N_i^2}{mp_i^2} \hat{S}_{2i}$, and

$\hat{W}_3^2 = \frac{1}{\hat{t}_{pwr}^2} \left\{ \sum_{i \in s} \frac{1}{mp_i^2} \frac{N_i^2}{n_i} \sum_{j \in s_i} Q_{ij}^2 \hat{S}_{3ij}^2 \right\}.$

Plug-in estimators of the measures of homogeneity in (A.4) are:

$\delta_1 = \frac{\hat{B}^2 - \hat{W}^2 / \bar{Q}}{\hat{V}_1}$ with

$\hat{V}_1 = \hat{B}^2 + \left(\bar{Q} - 1\right) \hat{W}^2 \big/ \bar{Q}$ and

$\hat{W}^2 = \frac{1}{\hat{t}_{pwr}^2} \sum_{i \in s} \frac{Q_{ij}^2 \bar{S}_i^2}{mp_i^2}$, where

$\tilde{S}_i^2 = \left(\sum_{j \in s_i} \sum_{k \in s_{ij}} w_k \right)^{-1} \sum_{j \in s_i} \sum_{k \in s_{ij}} w_k \left(y_k - \hat{\bar{y}}_i \right)^2$

$\hat{\bar{y}}_i = \sum_{j \in s_i} \sum_{k \in s_{ij}} w_k y_k \big/ \sum_{j \in s_i} \sum_{k \in s_{ij}} w_k$

$\hat{\delta}_2 = \frac{\hat{W}_2^2 - \hat{W}_3^2 / \bar{\bar{Q}}}{\hat{\bar{V}}_2}$ with $\hat{\bar{V}}_2 = \hat{W}_2^2 + \left(\bar{\bar{Q}} - 1\right) \hat{W}_3^2 \big/ \bar{\bar{Q}}$

Cost Functions in Two-Stage and Three-Stage Sampling

$C = C_0 + C_1 m + C_2 m\bar{n}$ is a cost function for two-stage sampling with $C_0 =$ costs that do not depend on the number of sample PSUs and elements; $C_1 =$ cost per sample PSU; $C_2 =$ cost per element

$C = C_0 + C_1 m + C_2 m\bar{n} + C_3 m\bar{n}\bar{\bar{q}}$ is a cost function for three-stage sampling with C_0, C_1 defined as in two-stage sampling; C_2 is the cost per SSU; and C_3 is the cost per element

Basic Steps in Weighting (Chap. 13)

$d_{0i} = \pi_i^{-1} =$ base weight for unit i in a single-stage sample, computed as the inverse of the selection probability, π_i. If stratification is used, a subscript h is added to give d_{0hi}

$d_{0ij} = \pi_{ij}^{-1} =$ base weight for element j in cluster i in a two-stage sample where $\pi_{ij} = \pi_i \pi_{j|i}$ with

$\pi_i =$ selection probability of cluster i

$\pi_{j|i} =$ selection probability of element j within cluster i

$s =$ initial set of all sample units

$s_{IN} =$ set of units in s that are known to be ineligible

$s_{ER} =$ set of units that are eligible respondents

$s_{ENR} =$ set of units that are eligible nonrespondents

$s_{KN} =$ set of units whose eligibility is known ($s_{IN} \cup s_{ER} \cup s_{ENR}$, where $\cup$ denotes the union of one or more sets)

$s_{UNK} =$ set whose eligibility is unknown

$a_{1b} = \dfrac{\sum_{i \in s_b} d_{0i}}{\sum_{i \in s_{b,KN}} d_{0i}} =$ weight adjustment for unknown eligibility, assuming that elements are placed into $b = 1, \ldots, B$ adjustment classes; s_b is the set of all sample elements in cell b; $s_{b,KN}$ is the set of elements whose eligibility status is known

$d_{1i} = a_{1b} d_{0i} =$ adjusted weight for unit i in $s_{b,KN}$

$a_{2c} = \dfrac{\sum_{i \in s_{c,E}} d_{1i}}{\sum_{i \in s_{c,ER}} d_{1i}} =$ weight class adjustment for nonresponse, assuming that elements are placed into $c = 1, \ldots, C$ adjustment classes; $s_{c,E}$ is the set of cases known to be eligible in class c; $s_{c,ER}$ is the set of eligible respondents in class c

The weight for unit i in the initial sample, after the adjustments for unknown eligibility and nonresponse, depends on whether the unit is an eligible respondent, a known ineligible, or a nonresponding or unknown-eligibility case:

$$d_{2i} = \begin{cases} d_{1i} a_{2c} & i \in s_{c,ER}, \\ d_{1i} & i \in s_{IN}, \\ 0 & i \in s_{UNK} \cup s_{ENR}, \end{cases}$$
$$= \begin{cases} d_{0i} a_{1b} a_{2c} & i \in s_{b,KN} \cap s_{c,ER}, \\ d_{0i} a_{1b} & i \in s_{b,KN} \cap s_{IN}, \\ 0 & i \in s_{UNK} \cup s_{ENR}, \end{cases}$$

Calibration (Chap. 14)

The GREG estimator of the population total of y is

$$\hat{T}_{yGREG} = \hat{t}_y + \left(\mathbf{t}_x - \hat{\mathbf{t}}_x \right)^T \hat{\mathbf{B}}$$
$$= \sum_{i \in s} \left[1 + \left(\mathbf{t}_x - \hat{\mathbf{t}}_x \right)^T \left(\mathbf{X}^T \mathbf{D} \mathbf{V}^{-1} \mathbf{X} \right)^{-1} \mathbf{x}_i / v_i \right] d_i y_i,$$

where $\hat{t}_y = \sum_s d_i y_i$ is the estimator of the total based on the input weights

$t_x = (t_{x1}, \ldots, t_{xp})^T$ is the $p \times 1$ vector of population (or control) totals of p auxiliaries using the number of rows by the number of column matrix notation (the superscript T denotes the transpose of a vector)

$\hat{t}_x = \sum_s d_i x_i$ is the estimate of totals of the x's based on the d_i weights (these can be base weights or weights adjusted for unknown eligibility and nonresponse)

$\mathbf{x}_i$ is the $p \times 1$ vector of auxiliary values for the ith sample unit

$\mathbf{D} = diag\,(d_i)$ is the $n \times n$ diagonal matrix of input weights

$$X = \begin{pmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{pmatrix} \text{ is the } n \times p \text{ matrix of auxiliaries for the } n \text{ sample units}$$

$\hat{\mathbf{B}} = \left(\mathbf{X}^T \mathbf{D} \mathbf{V}^{-1} \mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{D} \mathbf{V}^{-1} y$ with $y = (y_1, \ldots, y_n)^T$ being the vector of y's for the sample units

$\mathbf{V} = diag\,(v_i)$ is an $n \times n$ diagonal matrix of values associated with the variance parameters in an underlying linear model

The GREG weight for element i is

$$w_i = d_i g_i$$
$$\equiv d_i \left[1 + \left(\mathbf{t}_x - \hat{\mathbf{t}}_x\right)^T \left(\mathbf{X}^T \mathbf{D} \mathbf{V}^{-1} \mathbf{X}\right)^{-1} \mathbf{x}_i / v_i\right].$$

The term in brackets is called the g-weight.

Variance Estimation (Chap. 15)

To estimate design-based variances, the design of the sample must be considered more explicitly than when computing weights. Consequently, the notation must include the stages of sampling that were used.

Sample design: sample is selected with varying probabilities and with replacement (*ppswr*). The *pwr*-estimator of the mean is $\hat{\bar{y}}_{pwr} = \frac{1}{Nn} \sum_s \frac{y_i}{p_i}$ where p_i is the 1-draw selection probability. Its variance is estimated with

$$v\left(\hat{\bar{y}}_{pwr}\right) = \frac{1}{N^2} \frac{1}{n(n-1)} \sum_{i \in s} \left(\frac{y_i}{p_i} - \hat{t}_{pwr}\right)^2.$$

Sample design: PSUs are stratified and a two-stage sample is selected. The PSUs are selected with replacement. The *pwr*-estimator of the mean is

$\hat{\bar{y}}_{pwr} = N^{-1} \sum_h n_h^{-1} \sum_{i \in s_h} y'_{hi}/p_{hi}$ where

p_{hi} = the 1-draw probability of selection of PSU i in stratum h

$y'_{hi} = \sum_{k \in s_{hi}} d_{k|hi} y_{hik}$ = estimated total just for units in PSU hi.

s_{hi} = set of sample units in PSU hi

$d_{k|hi}$ = weight for unit k in PSU hi that expands the PSU sample to only the population of that PSU.

$d_k = d_{k|hi}/p_{hi}$ = full weight for unit k in s_{hi} where $d_{k|hi}$ is sometimes referred to as the conditional within-PSU weight for unit k. The *pwr* variance formula for $\hat{\bar{y}}_{pwr}$ is

$$v\left(\hat{\bar{y}}_{pwr}\right) = \frac{1}{N^2} \sum_h \frac{1}{n_h(n_h-1)} \sum_{i \in s_h} \left(\frac{y'_{hi}}{p_{hi}} - \hat{t}_{pwr,h}\right)^2,$$

where $\hat{t}_{pwr,h} = n_h^{-1} \sum_{s_h} y'_{hi}/p_{hi}$,

$\hat{Y}_{hi} = \sum_{k \in s_{hi}} d_k y_k$, and

$\hat{\bar{Y}}_h = n_h^{-1} \sum_{i \in s_h} \hat{Y}_{hi}$.

Jackknife Variance Estimator

Sample design: A single-stage of size n is selected. n replicate estimates are formed in the basic jackknife method by dropping one unit at a time and reweighting the remaining units. The jackknife variance estimator of the estimated total, $\hat{t} = \sum_{k \in s} d_k y_k$, used in this book is

$$v_J\left(\hat{t}\right) = \frac{n-1}{n} \sum_{i=1}^{n} \left(\hat{t}_{(i)} - \hat{t}\right)^2, \tag{A.5}$$

where

$\hat{t}_{(i)} = \sum_{k \in s(i)} d_{k(i)} y_k$ = the estimated total for a variable y based on replicate i

$d_{(k)i} = \frac{n}{n-1} d_k$ = weight for unit k that is retained for replicate i

$s(i)$ denotes the set of sample units excluding unit i

Formula (A.5) also applies to a nonlinear estimator $\hat{\theta} = f\left(\hat{t}_1, \ldots, \hat{t}_p\right)$ that is a differentiable function of the vector of estimated totals, $\hat{t} = \left(\hat{t}_1, \ldots, \hat{t}_p\right)^T$.

Sample design: A stratified multistage with n PSUs is selected; any number of stages can be used within the PSUs. The jackknife variance estimator of a differentiable function of estimated totals, $\hat{\theta} = f\left(\hat{t}_1, \ldots, \hat{t}_p\right)$, is

$$v_J\left(\hat{\theta}\right) = \sum_h \frac{n_h - 1}{n_h} \sum_{i \in s_h} \left(\hat{\theta}_{(hi)} - \hat{\theta}\right)^2,$$

where

$s_h = $ the set of sample PSUs in stratum h

$\hat{\theta}_{(hi)}$ is the estimate from replicate hi found by dropping all sample units in PSU hi and reweighting the remaining sample units

The adjusted base weight for unit k when PSU hi is deleted is

$$d_{k(hi)} = \begin{cases} 0 & \text{if unit } k \text{ is in PSU } i \text{ in stratum } h, \\ \frac{n_h}{n_h-1}d_k & \text{if unit } k \text{ is in stratum } h \text{ but not in PSU } i, and \\ d_k & \text{if unit } k \text{ is not in stratum } h. \end{cases}$$

Balanced Repeated Replication (BRR) Variance Estimator

BRR is mainly used in PSU samples but applies generally when the sample is stratified and two first-stage units are selected in each stratum. Suppose that the full sample estimator is $\hat{\theta} = f\left(\hat{t}_1, \ldots, \hat{t}_p\right)$, a differentiable function of a vector of estimated totals. Replicate subsamples are formed by identifying half-samples using the method prescribed in Sect. 15.4.2. The standard BRR variance estimator is

$$v_{BRR}\left(\hat{\theta}\right) = A^{-1} \sum_{\alpha=1}^{A} \left(\hat{\theta}_\alpha - \hat{\theta}\right)^2,$$

where

$\hat{\theta}_\alpha = f\left(\hat{t}_{1\alpha}, \ldots, \hat{t}_{p\alpha}\right)$ where $\hat{t}_{j\alpha}$ is the estimated total for the j-th variable based on the units in half-sample α

$A = $ number of replicates

The replicate weights for the standard BRR are

$$d_{k(\alpha)} = \begin{cases} 0 & \text{if unit } k \text{ is in a PSU that is not in the half-sample,} \\ 2d_k & \text{if unit } k \text{ is in a PSU that is in the half-sample.} \end{cases}$$

The Fay BRR uses all units in the sample to calculate replicate estimates. Weights for units in replicates are

$$d_{k(\alpha)} = \begin{cases} \rho d_k & \text{if unit } k \text{ is in a PSU that is not in the half-sample,} \\ (2-\rho)d_k & \text{if unit } k \text{ is in a PSU that is in the half-sample,} \end{cases}$$

where $0 \leq \rho < 1$.

Bootstrap Variance Estimator

The bootstrap is implemented by selecting an *srswr* of m_h PSUs from the n_h initial sample PSUs in stratum h:

$m_{hi}^* =$ number of times that PSU i is selected from stratum h
$\sum_{i=1}^{n_h} m_{hi}^* = m_h$; $m_{hi}^* = 0$ for PSUs not selected for the bootstrap sample.

The replicate weight for each sample unit k within the initial sample PSUs $(k \in s_{hi})$ is

$$d_k^* = d_k \left(\left\{ 1 - \sqrt{\frac{m_h}{(n_h-1)}} \right\} + \sqrt{\frac{m_h}{(n_h-1)}} \frac{n_h}{m_h} m_{hi}^* \right) \tag{A.6}$$
$$= d_k B_{hi},$$

where B_{hi} is defined by the last equality. This is computed for units in all sample PSUs, not just those in the bootstrap sample.
The Rao–Wu bootstrap variance estimator is

$$v_{boot} \left(\hat{\theta} \right) = \frac{1}{B} \sum_{b=1}^{B} \left(\hat{\theta}_{(b)} - \hat{\theta} \right)^2,$$

where $\hat{\theta}_{(b)}$ is the estimate from bootstrap sample b computed using the weights in (A.6).

Multiphase Designs (Chap. 17)

Multiphase designs refer to sample designs in which two or more phases are used to select the sample. Generally, information is collected on an initial set of units (the first phase) and used to select a subsample of units for the next phase (the second phase). This pattern can be continued to subsequent phases. Subscripts in parentheses are used to denote phases:

$n_{(1)}$, $n_{(1)R} =$ numbers of initial sample units selected in phase 1 and the number that respond
$n_{(2)}$, $n_{(2)R} =$ numbers of sample units selected in phase 2 and the number that respond
$n_{(p)d} =$ number of phase p units given survey condition d. A survey condition might be defined by whether an incentive was offered and, if so, the amount of the incentive
$D_{(p)} =$ number of survey conditions used in phase p

The notation below refers to a sample design in which a stratified sample of clusters is selected followed by a sample of elements within each sample cluster

$\pi_{(1)hi}$ = selection probability for the ith cluster in stratum h in phase 1

$\pi_{(1)k|hi}$ = selection probability for the kth element in cluster hi conditional on the cluster being sampled in the first stage

$d_{(1)0k} = \pi_{(1)hi}^{-1} \times \pi_{(1)k|hi}^{-1}$ = base weight of an element in the phase 1 sample

$\pi_{(2)k|(1)}$ = phase 2 selection probability for the kth unit conditional on being selected in phase 1

$d_{(2)0k} = d_{(1)0k}\,\pi_{(2)k|(1)}^{-1}$ = base weight of unit k in phase 2

Analysis weights for phase 1 elements may be computed if data are collected from them that can be analyzed separately. In that case, an analysis weight, $w_{(1)k}$, can be computed as

$w_{(1)k} = d_{(1)0k}a_{(1)1k}a_{(1)2k}g_{(1)k}$ = analysis weight for an element in the phase 1 responding sample where

$a_{(1)1k}$ = adjustment for unknown-eligibility status of element k

$a_{(1)2k}$ = adjustment for nonresponse applied to the base weight adjusted for unknown eligibility $d_{(1)1k} = d_{(1)0k} \times a_{(1)1k}$

$g_{(1)k}$ = calibration adjustment made to the adjusted base weights, $d_{(1)2k} = d_{(1)0k}\,a_{(1)1k}\,a_{(1)2k}$, using controls generated from the population

$s_{(2)R}$ = set of eligible sample respondents in phase 2

After data have been collected from the responding phase 2 sample members, the final unconditional, phase 2 analysis weight can be constructed for the elements in $s_{(2)R}$ as follows:

$$w_{(2)k} = w_{(1)k}\,a_{(2)0k|(1)}\,a_{(2)1k|(1)}\,a_{(2)2k|(1)}\,g_{(2)k|(1)},$$

where

$w_{(1)k}$ = final phase 1 weight

$a_{(2)0k|(1)}$ = adjustment for subsampling conditional on the responses from phase one, $a_{(2)1k|(1)}$ = adjustment for unknown eligibility strictly associated with the phase 2 sample

$a_{(2)2k|(1)}$ = nonresponse strictly associated with the phase 2 sample

$g_{(2)k|(1)}$ = calibration adjustment applied to the adjusted weights $d_{(2)2k} = w_{(1)k}\,a_{(2)0k|(1)}\,a_{(2)1k|(1)}\,a_{(2)2k|(1)}$

$\hat{t}_{(2)y} = \sum_{k \in s_{(2)R}} w_{(2)k}\,y_k$ = double expansion estimator of the population total of y

Variance Estimation in a Two-Phase Design

Consider the double sampling for stratification design where the phase-one design is an *srs* of size $n_{(1)}$ and a second-phase random sample of size $n_{(2)} = \sum_{h=1}^{H} n_{(2)h}$ is selected from the newly identified strata. The variance of the double expansion estimator is

$$
\begin{aligned}
V\left(\hat{t}_{(2)y}\right) &= N^2 \left[\left(1 - f_{(1)}\right) \frac{S^2}{n_{(1)}} + E_{(1)} \left(\sum_{h=1}^{H} w_{(1)h}^2 \left(1 - f_{(2)}\right) \frac{s_{(1)h}^2}{n_{(2)h}} \right) \right], \\
&= V_1 + V_2
\end{aligned}
$$

(A.7)

where $E_{(1)}$ is the expectation over the phase 1 sample design

$f_{(1)} = n_{(1)}/N$ = phase 1 sampling fraction

$f_{(2)h} = \left(n_{(2)h}/n_{(1)h}\right)$ = fraction of the phase 1 sample in stratum h that is sampled for phase 2

$S^2 = (N-1)^{-1} \sum_{k \in U} (y_k - \bar{y}_U)^2$ = the population (unit) variance

$s_{(1)h}^2 = \left(n_{(1)h} - 1\right)^{-1} \sum_{k \in s_{(1)h}} \left(\hat{y}_{(1)h} - \hat{\bar{y}}_{(1)h}\right)^2$
   = phase 1 unit variance among phase 1 sample units in stratum h

$\hat{\bar{y}}_{(1)h} = n_{(1)h}^{-1} \left(\sum_{k \in s_{(1)h}} \hat{y}_{(1)k} \right)$
   = mean of $\hat{y}_{(1)k} = d_{(1)0k}\, y_k$ among phase 1 elements

Estimates of the components of variances associated with phases 1 and 2 in (A.7) are

$$
\hat{V}_1 = \frac{1 - f_{(1)}}{n_{(1)}} \left[\sum_{h=1}^{H} w_{(1)h} \left(1 - \frac{1}{n_{(2)h}} \frac{n_{(1)} - n_{(1)h}}{n_{(1)} - 1} \right) s_{(2)h}^2 + \right.
$$
$$
\left. \frac{n_{(1)}}{n_{(1)} - 1} \sum_{h=1}^{H} w_{(1)h} \left(\hat{\bar{y}}_{(2)h} - \hat{\bar{y}}_{(2)} \right)^2 \right]
$$

$$
\hat{V}_2 = \sum_{h=1}^{H} w_{(1)h}^2 \left(1 - f_{(2)h}\right) \frac{s_{(2)h}^2}{n_{(2)h}},
$$

where

$\hat{\bar{y}}_{(2)} = \sum_{h=1}^{H} w_{(1)h}$
$\hat{\bar{y}}_{(2)h}$,

$\hat{\bar{y}}_{(2)h} = \sum_{k \in s_{(2)h}} y_k \Big/ n_{(2)h}$

$s_{(2)h}^2 = \left(n_{(2)h} - 1\right)^{-1} \sum_{k \in s_{(2)h}} \left(\hat{y}_{(1)k} - \hat{\bar{y}}_{(1)h}\right)^2$

Adding $\hat{V}_1$ and $\hat{V}_2$ and assuming that the first phase sampling fraction, $f_{(1)}$, is small and that $\left(n_{(1)h} - 1\right) / \left(n_{(1)} - 1\right) \doteq w_{(1)h}$, the estimated variance of $\hat{t}_{(2)y}$ is

$$
v\left(\hat{t}_{(2)y}\right) \cong N^2 \left[\frac{1}{n_{(1)}} \sum_{h=1}^{H} w_{(1)h} \left(\hat{\bar{y}}_{(2)h} - \hat{\bar{y}}_{(2)} \right)^2 + \sum_{h=1}^{H} w_{(1)h}^2 \left(\frac{s_{(2)h}^2}{n_{(2)h}} \right) \right].
$$

Appendix B
Data Sets

Several data sets are used in this book for examples. This appendix gives a brief description of each. These data files are also included in the companion R package, `PracTools`.

Domainy1y2

A small data set with 30 observations and two variables, `y1` and `y2`, is used in an exercise.

hospital

The `hospital` data are from the National Hospital Discharge Survey conducted by the US National Center for Health Statistics. The survey collects characteristics of inpatients discharged from non-federal short-stay hospitals in the United States. This population is from the January 1968 survey and contains observations on 393 hospitals.

Variable	Description
y	Number of patients discharged by the hospital in January 1968
x	Number of inpatient beds in the hospital

labor

This population is a clustered population of 478 persons extracted from the September 1976 Current Population Survey (CPS) in the United States. The clusters are compact geographic areas used as one of the stages of sampling in the CPS and are typically composed of about four nearby households. The units within clusters for this illustrative population are individual persons.

Variable	Description
h	Stratum of clusters
hsub	Substratum (each stratum contains two substrata)
cluster	Cluster (or segment) number. Each segment is a small group of persons living near each other
person	Person number
age	Age
agecat	Age category
	1 = 19 years and under
	2 = 20–24
	3 = 25–34
	4 = 35–64
	5 = 65 years and over
race	Race (1 = non-Black; 2 = Black)
sex	Gender (1 = male; 2 = female)
HourPerWk	Usual number of hours worked per week
WklyWage	Usual amount of weekly wages (in 1976 US dollars)
y	An artificial variable generated to follow a model with a common mean. Persons in the same cluster are correlated. Persons in different clusters are uncorrelated under the model

MDarea.pop

A data set of 403,997 persons based on the 2000 decennial US Census for Anne Arundel County in the state of Maryland. Person records were generated based on counts from the 2000 census. Individual values for each person were generated using models. Groupings to form the variables PSU and SSU were done after sorting the census file by tract and block group within tract.

Variable	Description
PSU	Primary sampling unit; a grouping of block groups (BLKGROUP) which has about 5,000 persons
SSU	Secondary sampling units; a grouping of block groups which has about 1,000 persons
TRACT	A geographic area defined by the Census Bureau. Tracts generally have between 1,500 and 8,000 people but have a much wider range in Anne Arundel county
BLKGROUP	Block group. A geographic area defined by the Census Bureau. Block groups generally have between 600 and 3,000 people
Hispanic	Hispanic ethnicity (1=Hispanic; 2=non-Hispanic)
Gender	Gender (1 = male; 2 = female)
Age	23 level age category: 1 = under 5 years 2 = 5–9 years 3 = 10–14 years 4 = 15–17 years 5 = 18–19 years 6 = 20 years 7 = 21 years 8 = 22–24 years 9 = 25–29 years 10 = 30–34 years 11 = 35–39 years 12 = 40–44 years 13 = 45–49 years 14 = 50–54 years 15 = 55–59 years 16 = 60–61 years 17 = 62–64 years 18 = 65–66 years 19 = 67–69 years 20 = 70–74 years 21 = 75–79 years 22 = 80–84 years 23 = 85 years and over
person	Counter for person within tract/block group/Hispanic/gender/age combination
y1	Artificial continuous variable
y2	Artificial continuous variable
y3	Artificial continuous variable
ins.cov	Medical coverage: 0 = person does not have medical insurance coverage 1 = person has medical insurance coverage
hosp.stay	Overnight hospital stay: 0 = person did not have an overnight hospital stay in last 12 months 1 = person did have an overnight hospital stay in last 12 months

nhis

The National Health Interview Survey (NHIS) is used to monitor health conditions in the US Data are collected through personal household interviews. Only demographic variables are included in this subset which was collected in 2003. The nhis data set contains observations for 3,911 persons. The file contains only persons 18 years and older.

Variable	Description
ID	Identification variable
stratum	Sample design stratum (1–100)
psu	Primary sampling unit, numbered within each stratum (1,2)
svywt	Survey weight
sex	Gender (1 = male; 2 = female)
age	Age, continuous
age_r	Recoded age:
	3 = 18–24 years
	4 = 25–44 years
	5 = 45–64 years
	6 = 65–69 years
	7 = 70–74 years
	8 = 75 years and older
hisp	Hispanic ethnicity:
	1 = Hispanic
	2 = non-Hispanic
marital	Marital status:
	1 = separated
	2 = divorced
	3 = married
	4 = single/never married
	5 = widowed
	9 = unknown marital status
parents	Parent(s) of sample person present in the family:
	1 = mother, no father
	2 = father, no mother
	3 = mother and father
	4 = neither mother nor father
parents_r	Parent(s) of sample person present in the family recode (1 = yes; 2 = no)

Variable	Description
educ	Education:
	1 = 8th grade or less
	2 = 9-12th grade, no high school diploma
	3 = high school graduate
	4 = general education development (GED) degree recipient
	5 = some college, no degree
	6 = associate's degree, technical or vocational
	7 = associate's degree, academic program
	8 = bachelor's degree (BA, BS, AB, BBA)
	9 = master's, professional, or doctoral degree
educ_r	Education recode:
	1 = high school, general education development degree (GED), or less
	2 = some college
	3 = bachelor's or associate's degree
	4 = master's degree and higher
race	Race (1 = White; 2 = Black; 3 = other)
resp	Respondent (0 = nonrespondent; 1 = respondent)

nhis.large

The National Health Interview Survey (NHIS) is used to monitor health conditions in the US Data are collected through personal household interviews. Demographic variables and a few health-related variables are included in this subset. The nhis.large data set contains observations on 21,588 persons. nhis.large is a set of 21,588 persons extracted from the 2003 US survey. The file contains only persons 18 years and older.

Variable	Description
ID	Identification variable
stratum	Sample design stratum (1–100)
psu	Primary sampling unit, numbered within each stratum (1,2)
svywt	Survey weight
sex	Gender (1 = male; 2 = female)
age.grp	Age group:
	1 = < 18 years
	2 = 18–24 years
	3 = 25–44 years
	4 = 45–64 years
	5 = 65+
hisp	Hispanic ethnicity:
	1 = Hispanic
	2 = non-Hispanic White
	3 = non-Hispanic Black
	4 = non-Hispanic all other race groups
parents	Parents present in the household:
	1 = mother, father, or both present
	2 = neither present
educ	Highest level of education attained:
	1 = high school graduate, graduate equivalence degree less
	2 = some college
	3 = bachelor's or associate's degree
	4 = master's degree or higher
	NA = missing
race	Race:
	1 = White
	2 = Black
	3 = all other race groups
inc.grp	Family income group:
	1 = < $20K
	2 = $20,000-$24,999
	3 = $25,000-$34,999
	4 = $35,000-$44,999
	5 = $45,000-$54,999
	6 = $55,000-$64,999
	7 = $65,000-$74,999
	8 = $75K+
	NA = missing
delay.med	Delayed medical care in last 12 months because of cost:
	1 = yes;
	2 = no;
	NA = missing
hosp.stay	Had an overnight hospital stay in last 12 months:
	1 = yes;
	2 = no;
	NA = missing

Variable	Description
doc.visit	During 2 Weeks before interview, did person see a doctor or other health care professional at a doctor's office, a clinic, an emergency room, or some other place? (excluding overnight hospital stay)? 1 = yes 2 = no
medicaid	Covered by medicaid, a governmental subsidy program for the poor: 1 = yes 2 = no NA = missing
notcov	Not covered by any type of health insurance 1 = Yes; 2 = No; NA = missing
doing.lw	What was person doing last week? 1 = working for pay at a job or business 2 = with a job or business but not at work 3 = looking for work 4 = working, but not for pay, at a job or business 5 = not working and not looking for work NA = missing
limited	Is the person limited in any way in any activities because of physical, mental, or emotional problems? 1 = limited in some way 2 = not limited in any way NA = missing

smho.N874

The 1998 Survey of Mental Health Organizations (SMHO) was conducted by the US Substance Abuse and Mental Health Services Administration. It collected data on mental health care organizations and general hospitals that provide mental health care services, with an objective to develop national- and state-level estimates for total expenditure, full-time equivalent staff, bed count, and total caseload by type of organization. The population omits one extreme observation in the smho98 population and contains observations on 874 facilities.

Variable	Description
EXPTOTAL	Total expenditures in 1998
BEDS	Total inpatient beds
SEENCNT	Unduplicated client/patient count seen during year
EOYCNT	End of year count of patients on the role
FINDIRCT	Hospital receives money from the state mental health agency (1=yes; 2=no)
hosp.type	Hospital type: 1 = psychiatric 2 = residential or veterans 3 = general 4 = outpatient, partial care 5 = multi-service, substance abuse

smho98

The 1998 SMHO was conducted by the US Substance Abuse and Mental Health Services Administration. It collected data on mental health care organizations and general hospitals that provide mental health care services, with an objective to develop national- and state-level estimates for total expenditure, full-time equivalent staff, bed count, and total caseload by type of organization.

Variable	Description
STRATUM	Sample design stratum 1 = psychiatric hospital 2 = residential 3 = general hospital 4 = military veterans 5 = partial care or outpatient 6 = multiservice or substance abuse
BEDS	Total inpatient beds
EXPTOTAL	Total expenditures in 1998
SEENCNT	Unduplicated client/patient count seen during year
EOYCNT	End of year count of patients on the role
Y_IP	Number of inpatient visits during year
OPCSFRST	Number of outpatients on the roles on the first day of the reporting year
OPCSADDS	Number of outpatients admitted, readmitted, or transferred to the organization during the reporting year for less than a 24 h period and not overnight
OPCSVIST	Number of outpatient visits during the reporting year for less than a 24 h period and not overnight
EMGWALK	Number of emergency walk-ins during the reporting year
PSYREHAB	Number of visits for psychiatric rehabilitation services
IPCSADDS	Number of residential patients added during the reporting year or patients admitted for more than a 24 h period

Appendix C
R Functions Used in this Book

Many examples in the book were developed using the R programming language (R Core Team, 2012). Below we provide a brief overview of R including steps to download a new or updated version (Sect. C.1). The functions in the R package `PracTools` are listed in Sect. C.2.

C.1 R Overview

R encompasses a statistical language and a full graphical user interface (RGui) with data manipulation, analysis, and graphic capabilities. The software is available free to all from the R website, http://www.r-project.org/.

Documentation and Resources

Free, downloadable user's manuals are located on the R website under the Documentation—Manuals link.[1] Additionally, all R functions have an associated help screen that is seen by (i) typing a question mark and the name of the function on the R command line or (ii) By using the `help` function (e.g., `?mean` or `help(mean)`) in the RGui or (iii) By running an "R Site Search" on the website (see the R Project—Search link). The website also contains an abbreviated list of books on various topics for users including guides for translating SAS or Stata concepts and code into R. An example of a comprehensive text used by the authors is Crawley (2007).

---

[1] http://cran.r-project.org/manuals.html.

Download a New Version of R

To obtain a new version of R, access the website and select "Download, Packages–CRAN" from the list on the left side of the screen. Select a mirror site near you and the version of R most appropriate for your computer system (e.g., Windows). Select the "base" and then the "download" links.

R Packages/Libraries

Many functions have been created by R users, vetted by the R Core Team, and made available to all in the R community. These user-written functions are organized into packages also referred to as libraries. A few key packages used in this book are listed below.

R package	Purpose of associated functions
alabama	Nonlinear optimization (Varadhan, 2010)
doBy	Summary statistics by specified subgroups (Højsgaard and Halekoh, 2012)
foreign	Import/export data created from/to other software such as SAS or Stata (R Core Team and contributors worldwide, 2012a)
graphics	Graphics (R Core Team and contributors worldwide, 2012b)
lme4	Linear mixed effects for estimating variance components (Bates et al., 2012)
nlme	Linear and nonlinear mixed-effect models (Pinheiro and Bates, 2000)
pps	Selection of samples from finite populations (Gambino, 2005)
quadprog	Quadratic programming (Turlach and Weingessel, 2011)
reshape	Data frame manipulation (Wickham, 2011)
rpart	Classification and regression tree (CART) analysis (Therneau et al., 2012)
sampling	Selection of samples from finite populations (Tillé and Matei, 2012)
samplingbook	Sample size and estimation currently for single-stage designs (Manitz, 2012)
stats	Statistical functions including classical hypothesis test and regression (R Core Team and contributors worldwide, 2012c)
survey	Analysis of complex survey data (Lumley, 2012)
survival	Data files and analytic functions for survival analysis (Therneau, 2012)

Functions within most packages are available for use only after the library has been installed from a selected CRAN mirror and accessed during an R session. To install an external package not included with the base installation, choose "Install Package(s)" from the "Packages" menu within RGui, select a local CRAN mirror, and then choose one or more packages from the resulting list. A few libraries, such as MASS, are automatically loaded when an R session begins. Other installed packages are accessed using either the require or the library functions, e.g., require(survey) or library(survey).

Updating R

The R base package is occasionally updated with no set schedule. Users should regularly check the R website for a new release of the base package along with updates to the function packages. The most recent version of the software available for download is listed in the "News" section on the main R Web page. To upgrade to the latest version, first uninstall the current version of R from your system and then install the latest version of R from the website. Note that even though the software has been uninstalled, the R folder containing the previously downloaded function packages still remains.

As with the base package, the function packages are periodically updated to include new functions or enhancements to old functions. Previously downloaded function packages are updated using either the `update.packages()` function or selecting the appropriate filename from the "Packages/Update Packages..." RGui list. With an updated version of the base package, simply copy the function packages from the old R folder to the new folder prior to running the updates.

Creating and Executing R Code

R code is executed interactively (through the RGui) in one of three ways:

(1) By entering the code line-by-line, pressing the enter key after each line entry
(2) By copying and pasting a complete set of code developed in a text editor
(3) By including a complete R program using the `source(''file name'')` function

Additionally, R programs can be executed in batch mode. There are several text editors that are designed to work closely with R—RWinEdt,[2] Tinn-R,[3] and RStudio[4] are three. R also comes with a built-in editor that has fewer capabilities. RWinEdt is an R package that uses the WinEdt[5] editor, which also is a popular choice for editing with the LaTex[6] typesetting language. These specialized editors have several nice features, including highlighting of matching parentheses, brackets, and braces; ability to highlight, copy, and paste R code directly to the R Console; and accenting R reserved words, like function names and operators.

---

[2] http://cran.r-project.org/web/packages/RWinEdt/.

[3] http://www.sciviews.org/Tinn-R/.

[4] http://rstudio.org/.

[5] http://www.winedt.com/.

[6] http://miktex.org/ or http://www.latex-project.org/.

C.2 Author-Defined R Functions

Functions developed by the authors for use with this textbook are detailed in alphabetical order below. These functions are available in the `PracTools` library available for download at the book website and the main R website. Following the lead of the R help files, each description below contains:

- The function name and summary of its purpose
- Syntax along with a description of the arguments
- The value(s) returned by the function
- Example code

More details for each function can be found in the help files for `PracTools`. Other useful functions can be found in, for example, Valliant et al. (2000).

`BW2stagePPS`—Relvariance components for 2-stage sample

Description
Compute components of relvariance for a sample design where primary sampling units (PSUs) are selected with probabilities proportional to size (*pps*) and elements are selected via simple random sampling (*srs*). The input is an entire sampling frame.

Usage
```
BW2stagePPS(X, pp, psuID)
```

Arguments

X	data vector; length is the number of elements in the population.
pp	vector of 1-draw probabilities for the PSUs. This vector must be as long as X. Each element in a given PSU should have the same value in pp. PSUs must be in the same order as in X.
psuID	vector of PSU identification numbers. This vector must be as long as X. Each element in a given PSU should have the same value in psuID. PSUs must be in the same order as in X.

Value
List object with elements:

B2	between-PSU unit relvariance
W2	within-PSU unit relvariance
B2+W2	sum of between- and within-relvariance estimates
delta	intraclass correlation estimated as $B^2/(B^2 + W^2)$

Examples

```
data(MDarea.pop)
    # Use PSU and SSU variables to define psu's
pp.PSU <- table(MDarea.pop$PSU) / nrow(MDarea.pop)
pp.SSU <- table(MDarea.pop$SSU) / nrow(MDarea.pop)

    # components with psu's defined by the PSU variable
BW2stagePPS(MDarea.pop$y1, pp=pp.PSU, psuID=MDarea.pop$PSU)
    # components with psu's defined by the SSU variable
BW2stagePPS(MDarea.pop$y1, pp=pp.SSU, psuID=MDarea.pop$SSU)

    # Use census tracts and block groups to define psu's
trtBG <- 10*MDarea.pop$TRACT + MDarea.pop$BLKGROUP
pp.trt <- table(MDarea.pop$TRACT) / nrow(MDarea.pop)
pp.BG <- table(trtBG) / nrow(MDarea.pop)
    # components with psu's defined by tracts
BW2stagePPS(MDarea.pop$ins.cov, pp=pp.trt,
        psuID=MDarea.pop$TRACT)
    # components with psu's defined by block groups
BW2stagePPS(MDarea.pop$ins.cov, pp=pp.BG, psuID=trtBG)
```

BW2stagePPSe—Estimated relvariance components for 2-stage sample

Description
Estimate components of relvariance for a sample design where primary sampling units (PSUs) are selected with probabilities proportional to size (*pps*) and elements are selected via simple random sampling (*srs*). The input is a sample selected in this way.

Usage
```
BW2stagePPSe(Ni, ni, X, psuID, w, m, pp)
```

Arguments

Ni	vector of number of elements in the population of each sample PSU; length is the number of PSUs in the sample.
ni	vector of number of sample elements in each sample PSU; length is the number of PSUs in the sample. PSUs must be in the same order as in X.
X	data vector for sample elements; length is the number of elements in the sample. These must be in PSU order. PSUs must be in the same order as in X.
psuID	vector of PSU identification numbers. This vector must be as long as X. Each element in a given PSU should have the same value in psuID.

w   vector of full sample weights. This vector must be as long as X.
                Vector must be in the same order as X.
m   number of sample PSUs
pp   vector of 1-draw probabilities for the PSUs. This vector must
                be as long as X. Each element in a given PSU should have the
                same value in pp. Vector must be in the same order as X.

Value

List object with elements:

   Vpsu   estimated between-PSU unit variance
   Vssu   estimated within-PSU unit variance
   B2   estimated between-PSU unit relvariance
   W2   estimated within-PSU unit relvariance
   delta   intraclass correlation estimated as $B^2/(B^2 + W^2)$

Examples

```
require(sampling)
require(reshape)      # has function that allows renaming vars
data(MDarea.pop)
Ni <- table(MDarea.pop$TRACT)
m <- 20
probi <- m*Ni / sum(Ni)
    # select sample of clusters
sam <- cluster(data=MDarea.pop, clustername="TRACT", size=m,
            method="systematic",
            pik=probi, description=TRUE)
    # extract data for the sample clusters
samclus <- getdata(MDarea.pop, sam)
samclus <- rename(samclus, c(Prob = "pi1"))
    # treat sample clusters as strata and select srswor from
      each
s <- strata(data = as.data.frame(samclus),
            stratanames = "TRACT",
            size = rep(50,m), method="srswor")
    # extracts the observed data
samdat <- getdata(samclus,s)
samdat <- rename(samdat, c(Prob = "pi2"))
    # extract pop counts for PSUs in sample
pick <- names(Ni) %in% sort(unique(samdat$TRACT))
Ni.sam <- Ni[pick]
pp <- Ni.sam / sum(Ni)
wt <- 1/samdat$pi1/samdat$pi2

BW2stagePPSe(Ni = Ni.sam, ni = rep(50,20), X = samdat$y1,
            psuID = samdat$TRACT, w = wt,
            m = 20, pp = pp)
```

BW2stageSRS—Relvariance components for 2-stage sample

Description
Compute components of relvariance for a sample design where primary sampling units (PSUs) and elements are selected via simple random sampling (*srs*). The input is an entire sampling frame.

Usage
BW2stageSRS(X, psuID)

Arguments

X	data vector; length is the number of elements in the population.
psuID	vector of PSU identification numbers. This vector must be as long as X. Each element in a given PSU should have the same value in psuID. PSUs must be in the same order as in X.

Value
List object with elements:

B2	between-PSU unit relvariance
W2	within-PSU unit relvariance
unit relvar	unit relvariance for population
delta full	intraclass correlation

Examples

```
data(MDarea.pop)
    # psu's are defined by PSU variable
BW2stageSRS(abs(MDarea.pop$Hispanic-2),
      psuID=MDarea.pop$PSU)
    # psu's are defined by SSU variable
BW2stageSRS(abs(MDarea.pop$Hispanic-2),
      psuID=MDarea.pop$SSU)
```

BW3stagePPS—Relvariance components for 3-stage sample

Description
Compute components of relvariance for a sample design where primary sampling units (PSUs) are selected with probabilities proportional to size and with replacement (*ppswr*) and secondary sampling units (SSUs) and elements within SSUs are selected via simple random sampling (*srs*). The input is an entire sampling frame.

Usage
```
BW3stagePPS(X, pp, psuID, ssuID)
```

Arguments

X	data vector; length is the number of elements in the population.
pp	vector of 1-draw probabilities for the PSUs. This vector must be as long as X and will contain the same value for the units that are in the same PSU. PSUs must be in the same order as in X.
psuID	vector of PSU identification numbers. This vector must be as long as X. Each element in a given PSU should have the same value in psuID. PSUs must be in the same order as in X.
ssuID	vector of SSU identification numbers. This vector must be as long as X. Each element in a given SSU should have the same value in ssuID. PSUs and SSUs must be in the same order as in X. ssuID should have the form psuID||(ssuID within PSU).

Value

List object with elements:

B	between-PSU unit relvariance
W	within-PSU unit relvariance computed as if the sample were two stage
W2	unit relvariance among SSU totals
W3	unit relvariance among elements within PSU/SSUs
delta1	homogeneity measure among elements within PSUs
delta2	homogeneity measure among elements within SSUs

Examples
```
data(MDarea.pop)
M <- length(unique(MDarea.pop$PSU))
    # srs/srs/srs design
pp.PSU <- rep(1/M,M)
BW3stagePPS(X=MDarea.pop$y1, pp=pp.PSU,
   psuID=MDarea.pop$PSU, ssuID=MDarea.pop$SSU)
    # ppswr/srs/srs design
pp.PSU <- table(MDarea.pop$PSU) / nrow(MDarea.pop)
BW3stagePPS(X=MDarea.pop$y1, pp=pp.PSU,
   psuID=MDarea.pop$PSU, ssuID=MDarea.pop$SSU)
```

BW3stagePPSe—Estimated relvariance components for 3-stage sample

Description

Estimate components of relvariance for a sample design where primary sampling units (PSUs) are selected with probabilities proportional to size and with replacement (*ppswr*) and secondary sampling units (SSUs) and elements within SSUs are selected via simple random sampling (*srs*). The input is a sample.

Usage

```
BW3stagePPSe(dat, v, Ni, Qi, Qij, m)
```

Arguments

dat	data frame for sample elements with PSU and SSU identifiers, weights, and analysis variable(s). The data frame should be sorted in hierarchical order: by PSU and SSU within PSU. Required names for columns: psuID = PSU identifier ssuID = SSU identifier. These must be unique, i.e., numbering should not restart within each PSU. Setting ssuID = psuID||(ssuID within PSU) is a method of doing this. w1i = vector of weights for PSUs w2ij = vector of weights for SSUs (PSU weight*SSU weight within PSU) w = full sample weight
v	name or number of column in dat with variable to be analyzed.
Ni	m-vector of number of SSUs in the population in the sample PSUs; m is number of sample PSUs.
Qi	m-vector of number of elements in the population in the sample PSUs
Qij	vector of numbers of elements in the population in the sample SSUs
m	number of sample PSUs

Value

List object with elements:

Vpsu	estimated between-PSU unit variance
Vssu	estimated second-stage unit variance among SSU totals
Vtsu	estimated third-stage unit variance

B	estimated between-PSU unit relvariance
W	estimated within-PSU unit relvariance computed as if the sample were two-stage
W2	estimated unit relvariance among SSU totals
W3	estimated third-stage unit relvariance among elements within PSU/SSUs
delta1	estimated homogeneity measure among elements within PSUs
delta2	estimated homogeneity measure among elements within SSUs

Examples

See the help file for `BW3stagePPSe` in the `PracTools` package for an extended example.

---

clusOpt2—Compute optimal sample sizes for a two-stage sample

Description

Compute the sample sizes that minimize the variance of the *pwr*-estimator, the "p-expanded with-replacement" estimator developed by Hansen and Hurwitz (1943), of a total in a two-stage sample.

Usage

```
clusOpt2(C1, C2, delta, unit.rv, CV0 = NULL, tot.cost =
NULL, cal.sw)
```

Arguments

C1	unit cost per primary sampling unit (PSU)
C2	unit cost per element
delta	homogeneity measure
unit.rv	unit relvariance or $B^2 + W^2$
CV0	target CV
tot.cost	total budget for variable costs
cal.sw	specify type of optimum 1 = find optimal m.opt for fixed total budget 2 = find optimal m.opt for target CV0

Value

List object with elements:

C1	unit cost per PSU
C2	unit cost per element
delta	homogeneity measure

unit relvar	unit relvariance or $B^2 + W^2$
cost	total budget for variable costs, C–C0
m.opt	optimum number of sample PSUs
n.opt	optimum number of sample elements per PSU
CV	target CV

Examples

```
    # optimum for a fixed total budget
 clusOpt2(C1=750, C2=100, delta=0.05, unit.rv=1,
        tot.cost=100000, cal.sw=1)
    # optimum for a target CV
 clusOpt2(C1=750, C2=100, delta=0.01, unit.rv=1,
        CV0=0.05, cal.sw=2)
```

---

clusOpt2fixedPSU—Optimal number of sample elements per primary sampling unit (PSU) in a two-stage sample

Description
Compute the optimum number of sample elements per PSU for a fixed set of PSUs.

Usage

```
clusOpt2fixedPSU(C1, C2, m, delta, unit.rv, CV0=NULL,
    tot.cost, cal.sw)
```

Arguments

C1	unit cost per PSU
C2	unit cost per element
m	number of sample PSU's (fixed)
delta	homogeneity measure
unit.rv	unit relvariance or $B^2 + W^2$
CV0	target CV
tot.cost	total budget for variable costs
cal.sw	specify type of optimum
	$1 = $ find optimal $\bar{n}$ for fixed total budget
	$2 = $ find optimal $\bar{n}$ for target CV0

Value

List object with elements:

C1	unit cost per PSU
C2	unit cost per element
m	number of (fixed) sample PSUs
delta	homogeneity measure
unit relvar	unit relvariance or $B^2 + W^2$
budget	total budget for variable costs, $C - C_0$
n	optimum number of sample elements per PSU
CV	target CV

Examples

```
    # optima for a vector of budgets
 clusOpt2fixedPSU(C1=500, C2=100, m=100, delta=0.05,
       unit.rv=2, CV0=NULL,
       tot.cost=c(100000, 500000, 10^6),
       cal.sw=1)
    # optima for a target CV and vector of PSU costs
 clusOpt2fixedPSU(C1=c(500,1000,5000), C2=100, m=100,
       delta=0.05, unit.rv=2, CV0=0.05, tot.cost=NULL,
       cal.sw=2)
```

---

clusOpt3—Compute optimal sample sizes for a three-stage sample

Description

Compute the sample sizes that minimize the variance of the *pwr*-estimator of a total in a three-stage sample. The "p-expanded with-replacement" (*pwr*) estimator is discussed in Hansen and Hurwitz (1943).

Usage

```
clusOpt3(unit.cost, delta1, delta2, unit.rv, CV0=NULL,
    tot.cost=NULL, cal.sw)
```

Arguments

unit.cost	vector with three components for unit costs:
	C1 = unit cost per primary sampling unit (PSU)
	C2 = unit cost per secondary sampling units (SSUs)
	C3 = unit cost per element
delta1	homogeneity measure among elements within PSUs
delta2	homogeneity measure among elements within SSUs

unit.rv	unit relvariance or $B^2 + W^2$
CV0	target CV
tot.cost	total budget for variable costs
cal.sw	specify type of optimum
	1 = find optimal m.opt for fixed total budget
	2 = find optimal m.opt for target CV0

Value
List object with elements:

C1	unit cost per PSU
C2	unit cost per SSU
C3	unit cost per element
delta1	homogeneity measure among elements within PSUs
delta2	homogeneity measure among elements within SSUs
unit relvar	unit relvariance
budget	total budget for variable costs
m.opt	optimum number of sample PSUs
n.opt	optimum number of sample SSUs per PSU
q.opt	optimum number of sample elements per SSU
CV	target CV if cal.sw=2 or achieved CV if cal.sw=1

Examples

```
    # optima for a fixed total budget
clusOpt3(unit.cost=c(500, 100, 120), delta1=0.01,
      delta2=0.10, unit.rv=1,
      tot.cost=100000,cal.sw=1)
    # optima for a target CV
clusOpt3(unit.cost=c(500, 100, 120), delta1=0.01,
      delta2=0.10, unit.rv=1,
      CV0=0.01,cal.sw=2)
```

clusOpt3fixedPSU—Compute optimal number of sample secondary sampling units (SSUs) and elements per primary sampling unit (PSU) in a three-stage sample

Description
Compute the sample sizes that minimize the variance of the *pwr*-estimator of a total in a three-stage sample when the PSU sample is fixed. The "p-expanded with-replacement" (*pwr*) estimator is discussed in Hansen and Hurwitz (1943).

Usage

```
clusOpt3fixedPSU(unit.cost, m, delta1, delta2, unit.rv,
        CV0=NULL, tot.cost=NULL, cal.sw)
```

Arguments

unit.cost	3-vector of unit costs:
	C1 = unit cost per PSU
	C2 = unit cost per SSU
	C3 = unit cost per element
m	number of sample PSUs (fixed)
delta1	homogeneity measure among elements within PSUs
delta2	homogeneity measure among elements within SSUs
unit.rv	unit relvariance or $B^2 + W^2$
CV0	target CV
tot.cost	total budget for variable costs, including PSU costs
cal.sw	specify type of optimum. 1 = find optimal m.opt for fixed total budget; 2 = find optimal m.opt for target CV0

Value

List object with elements:

C1	unit cost per PSU
C2	unit cost per SSU
C3	unit cost per element
m	number of sample PSUs (fixed)
delta1	homogeneity measure among elements within PSUs
delta2	homogeneity measure among elements within SSUs
unit relvar	unit relvariance
cost check	budget constraint (tot.cost); used if `cal.sw=1`
cost	computed cost; used if `cal.sw=2`
n	optimum number of sample SSUs per PSU
q	optimum number of sample elements per SSU
CV	achieved CV, used if `cal.sw=1`, or target CV, used if `cal.sw=2`
CV check	computed CV based on optimal sample sizes; used only if `cal.sw=2`

Examples

```
    # optima for a fixed total budget
 clusOpt3fixedPSU(unit.cost=c(500, 100, 120), m=100,
       delta1=0.01, delta2=0.05,
       unit.rv=1, tot.cost=500000,cal.sw=1)
    # optima for a target CV
 clusOpt3fixedPSU(unit.cost=c(500, 100, 120), m=100,
```

```
 delta1=0.01, delta2=0.05,
        unit.rv=1, CV0=0.05,cal.sw=2)
```

---

gamEst—Estimate variance model parameter

Description
Regresses a Y on a set of covariates X where $V_M(y_i) = \sigma^2 \gamma$ and then regresses the squared residuals on log(x) to estimate γ.

Usage
```
gamEst(X1, x1, y1, v1)
```

Arguments

X1	matrix of predictors in the linear model for y1
x1	vector of x's for individual units in the assumed specification of $var(y)$
y1	vector of dependent variables for individual units
v1	vector proportional to $var(y)$

Value
The estimate of γ.

Examples
```
 data(hospital)
 x <- hospital$x
 y <- hospital$y
 X <- cbind(sqrt(x), x)
 gamEst(X1 = X, x1 = x, y1 = y, v1 = x)
```

---

gammaFit—Estimate of variance model parameter γ

Description
Iteratively computes estimate of γ in a model with $E_M(y_i) = x_i^T \beta$ and $V_M(y_i) = \sigma^2 \gamma$.

Usage
```
gammaFit(X, x, y, maxiter = 100, show.iter = FALSE, tol
= 0.001)
```

Arguments

X	matrix of predictors in the linear model for y
x	vector of x's for individual units in the assumed specification of var(y)
y	vector of dependent variables for individual units
maxiter	maximum number of iterations allowed
show.iter	should values of γ be printed of each iteration? TRUE or FALSE
tol	size of relative difference in $\hat{\gamma}$'s between consecutive iterations used to determine convergence. Algorithm terminates when relative difference is less than tol

Value
List object with elements:

g.hat	estimate of γ when iterative procedure stopped
converged	TRUE or FALSE depending on whether convergence was obtained
steps	number of steps used by the algorithm

Examples
```
data(hospital)
x <- hospital$x
y <- hospital$y
X <- cbind(sqrt(x), x)
gammaFit(X = X, x = x, y = y, maxiter=100, tol=0.001)
```

---

nCont—Compute a simple random sample size for an estimated mean

Description
Compute a simple random sample size using either a target coefficient of variation, CV0, or target variance, V0, for an estimated mean.

Usage
```
nCont(CV0=NULL, V0=NULL, S2=NULL, ybarU=NULL, N=Inf,
        CVpop=NULL)
```

Arguments

CV0	target value of coefficient of variation of $\bar{y}_s$
V0	target value of variance of $\bar{y}_s$
S2	unit (population) variance
ybarU	population mean of target variable
N	number of units in finite population
CVpop	unit (population) coefficient of variation

Value
numeric sample size

Examples

```
nCont(CV0=0.05, CVpop=2)
nCont(CV0=0.05, CVpop=2, N=500)
nCont(CV0=0.10/1.645, CVpop=1)

    # sample size for ratio estimator in smho98 pop
data(smho98)
    # extract certainties
cert <- smho98[,"BEDS"] > 2000
tmp <- smho98[!cert, ]
tmp <- tmp[tmp[, "BEDS"] > 0, ]
x <- tmp[,"BEDS"]
y <- tmp[, "EXPTOTAL"]
m <- lm(y ~ 0 + x, weights = 1/x)
ybarU <- mean(y)
S2R <- sum(m$residuals^2/(length(x)-1))
nCont(CV0=0.15, S2=S2R, ybarU=ybarU)
```

---

nDep2sam—Simple random sample size for difference in means

Description
Compute a simple random sample size for difference in means when samples overlap.

Usage

```
nDep2sam(S2x, S2y, g, r, rho, alt, del, sig.level=0.05,
          pow=0.80)
```

Arguments

S2x	unit variance of analysis variable x in sample 1
S2y	unit variance of analysis variable y in sample 2
g	proportion of sample 1 that is in the overlap with sample 2
r	ratio of the size of sample 1 to that of sample 2
rho	unit-level correlation between x and y
alt	should the test be 1-sided or 2-sided; allowable values are `alt="one.sided"` or `alt="two.sided"`.
del	size of the difference between the means to be detected
sig.level	significance level of the hypothesis test
pow	desired power of the test

Value

List object with elements:

n1	sample size in group 1
n2	sample size in group 2
S2x.S2y	unit variances in groups 1 and 2
delta	difference in group means to be detected
gamma	proportion of sample 1 that is in the overlap with sample 2
r	ratio of the size of sample 1 to that of sample 2
rho	unit-level correlation between analysis variables in groups 1 and 2
alt	type of test: one-sided or two-sided
sig.level	significance level of test
power	power of the test

Examples

```
nDep2sam(S2x=200, S2y=200,
         g=0.75, r=1, rho=0.9,
         alt="one.sided", del=5,
         sig.level=0.05, pow=0.80)
```

---

nLogOdds—Calculate simple random sample size for estimating a proportion

Description

Calculate the simple random sample size for estimating a proportion using the log-odds transformation.

Usage

```
nLogOdds(moe.sw, e, alpha=0.05, pU, N=Inf)
```

Arguments

moe.sw	switch for setting desired margin of error (1 = CI half-width on the proportion; 2 = CI half-width on a proportion divided by pU)
e	desired margin of error
alpha	1 - (confidence level)
pU	population proportion
N	number of units in finite population

Value
numeric sample size

Examples

```
nLogOdds(moe.sw=1, e=0.05, alpha=0.05, pU=0.2, N=Inf)
nLogOdds(moe.sw=2, e=0.05, alpha=0.05, pU=0.2, N=Inf)
```

---

nProp—Compute simple random sample (*srs*) size for estimating a proportion

Description
Compute the simple random sample size for estimating a proportion based on different precision requirements.

Usage
```
nProp(CV0 = NULL, V0 = NULL, pU = NULL, N = Inf)
```

Arguments

CV0	target value of coefficient of variation of the estimated proportion
V0	target value of variance of the estimated proportion
pU	population proportion
N	number of units in finite population

Value
numeric sample size

Examples

```
# srs sample size so that CV of estimated proportion
# is 0.05 assuming the population is large and pU=0.01
# Both examples below are equivalent.
```

```
nProp(V0=0.0005^2, N=Inf, pU=0.01) #or
nProp(CV0=0.05, N=Inf, pU=0.01)
    # srswor sample size so that half-width of 2-sided
    # 95% CI is 0.005
nProp(V0=(0.005/1.96)^2, N=Inf, pU=0.01)
```

nProp2sam—Simple random sample size for difference in proportions

Description
Compute a simple random sample size for difference in proportions when
samples overlap.

Usage
```
nProp2sam(px, py, pxy, g, r, alt, sig.level=0.05, pow=
0.80)
```

Arguments
px	proportion in group 1
py	proportion in group 2
pxy	proportion in the overlap has the characteristic in both samples
g	proportion of sample 1 that is in the overlap with sample 2
r	ratio of the size of sample 1 to that of sample 2
alt	should the test be 1-sided or 2-sided; allowable values are alt="one.sided" or alt="two.sided".
sig.level	significance level of the hypothesis test
pow	desired power of the test

Value
List object with elements:
n1	sample size in group 1
n2	sample size in group 2
px.py.pxy	input values of the px, py, pxy parameters
gamma	proportion of sample 1 that is in the overlap with sample 2
r	ratio of the size of sample 1 to that of sample 2
alt	type of test: one-sided or two-sided
sig.level	significance level of test
power	power of the test

Examples
```
nProp2sam(px=0.5, py=0.55, pxy=0.45, g=0.5, r=1,
    alt="two.sided")
```

nPropMoe—Simple random sample (*srs*) size for a proportion based on margin of error

Description
Calculates a simple random sample size based on a specified margin of error.

Usage
```
nPropMoe(moe.sw, e, alpha = 0.05, pU, N = Inf)
```

Arguments

moe.sw	switch for setting desired margin of error (1 = CI half-width on the proportion; 2 = CI half-width on a proportion divided by pU
e	desired margin of error; either $e = z_{1-\alpha/2}\sqrt{V(\bar{y}_s)}$ or $e = z_{1-\alpha/2}\sqrt{CV(\bar{y}_s)}$
alpha	1 - (confidence level)
pU	population proportion
N	number of units in finite population

Value
numeric sample size

Examples
```
    # srs sample size so that half-width of a 95% CI is 0.01
    # population is large and population proportion is 0.04
nPropMoe(moe.sw=1, e=0.01, alpha=0.05, pU=0.04, N=Inf)
    # srswor sample size for a range of margins of error
    # defined as half-width of a 95\% CI
nPropMoe(moe.sw=1, e=seq(0.01,0.08,0.01), alpha=0.05, pU=0.5)
    # srswor sample size for a range of margins of error
    # defined as the proportion that the half-width of a
    # 95% CI is of pU
nPropMoe(moe.sw=2, e=seq(0.05,0.1,0.2), alpha=0.05, pU=0.5)
```

NRFUopt—Sample sizes for a nonresponse follow-up study

Description
Compute optimal values of the first-phase sample size and the second-phase sampling fraction in a two-phase sample.

Usage

```
NRFUopt(Ctot=NULL, c1, c2, theta, CV0=NULL, CVpop=NULL,
    N=Inf, type.sw)}
```

Arguments

Ctot	total variable cost
c1	cost per unit in phase 1
c2	cost per unit in phase 2
theta	probability of response for each unit
CV0	target coefficient of variation for the estimated total or mean
CVpop	unit coefficient of variation
N	population size; default is Inf
type.sw	type of allocation; "cost" = target total variable cost, "cv" = target coefficient of variation

Value

List object with elements:

allocation	type of allocation: either "fixed cost" or "fixed CV"
"Total variable cost"	expected total cost: fixed cost if type.sw="cost" or computed cost if type.sw="cv"
"Response rate"	first-phase response rate
CV	anticipated coefficient of variation (CV) if type.sw="cost" or target CV if type.sw="cv"
v.opt	optimal fraction of first-phase nonrespondents to select for second-phase follow-up
n1.opt	optimal number of units to sample at first phase
"Expected n2"	expected number of respondents obtained at second phase
"Expected total cases (2-phase)"	expected number of respondents across both phases
"srs sample for same cv"	size of single-phase simple random sample (srs) needed to obtain same CV as the two-phase sample
"Cost Ratio: Two phase to srs"	ratio of expected cost for two-phase sample to cost of single-phase srs

Examples

```
    # optima for fixed target CV
NRFUopt(Ctot=NULL, c1=50, c2=200, theta=0.5, CV0=0.05,
```

```
    CVpop=1, type.sw = "cv")
    # optima for fixed total cost
NRFUopt(Ctot=100000, c1=50, c2=200, theta=0.5, CV0=NULL,
    CVpop=1, type.sw = "cost")
```

---

nWilson—Calculate a simple random sample (*srs*) size for estimating a proportion

Description
Calculate a simple random sample size for estimating a proportion using the Wilson method.

Usage
```
nWilson(moe.sw, alpha = 0.05, pU, e)
```

Arguments
moe.sw	switch for setting desired margin of error ($1 =$ CI half-width on the proportion; $2 =$ CI half-width on a proportion divided by pU
alpha	1 - (confidence level)
pU	population proportion
e	desired margin of error; either the value of CI half-width or the value of the half-width divided by pU

Value
numeric sample size

Examples

```
    # srs sample size using Wilson method so that half-width
    # of a 95% CI is 0.01. Population proportion is 0.04
nWilson(moe.sw = 1, pU = 0.04, e = 0.01)
```

---

strAlloc—Allocate a sample to strata

Description
Compute the proportional, Neyman, cost-constrained, and variance-constrained allocations in a stratified simple random sample.

Usage

```
strAlloc(n.tot = NULL, Nh = NULL, Sh = NULL, cost = NULL,
      ch = NULL, V0 = NULL, CV0 = NULL, ybarU = NULL, alloc)
```

Arguments

n.tot	fixed total sample size
Nh	vector of population stratum sizes (N_h) or pop stratum proportions (W_h)
Sh	stratum unit standard deviations (S_h), required unless `alloc = "prop"`
cost	total variable cost
ch	vector of costs per unit in stratum c_h
V0	fixed variance target for estimated mean
CV0	fixed CV target for estimated mean
ybarU	population mean of y $(\bar{y}_U$
alloc	type of allocation; must be one of `"prop"`, `"neyman"`, `"totcost"`, `"totvar"`

Value

numeric vector of stratum sample sizes

Examples

```
    # Neyman allocation Nh <- c(215, 65, 252, 50, 149, 144)
Sh <- c(26787207, 10645109, 6909676, 11085034,
    9817762, 44553355)
strAlloc(n.tot = 100, Nh = Nh, Sh = Sh,
    alloc = "neyman")
    # cost constrained allocation
ch <- c(1400, 200, 300, 600, 450, 1000)
strAlloc(Nh = Nh, Sh = Sh, cost = 100000, ch = ch,
    alloc = "totcost")
    # allocation with CV target of 0.05
strAlloc(Nh = Nh, Sh = Sh, CV0 = 0.05, ch = ch,
    ybarU = 11664181, alloc = "totvar")
```

---

wtdvar—Compute weighted variance

Description

Compute an estimate of a population unit variance from a complex sample with survey weights.

Usage

```
wtdvar(x,w)
```

Arguments
  x   data vector
  w   vector of survey weights; must be same length as X

Value
numeric estimate of population unit variance

Examples

```
x <- c(1:3)
wts <- c(4, 6, 8)
wtdvar(x=x, w=wts)
```

References

AAPOR (2011). Standard definitions: Final dispositions of case codes and outcome rates for surveys, 7th edn. Tech. rep., The American Association for Public Opinion Research, Deerfield, IL, URL http://www.aapor.org/pdfs/standarddefs_4.pdf

Abraham K.G., Maitland A., Bianchi S.M. (2006). Nonresponse in the American Time Use Survey: Who is missing from the data and how much does it matter? *Public Opinion Quarterly* 70(5):676–703

Aitken A., Hörngren J., Jones N., Lewis D., Zilhão M.J. (2004). Handbook on improving quality by analysis of process variables. Tech. rep., Luxembourg, URL http://epp.eurostat.ec.europa.eu/portal/page/portal/quality/documents/HANDBOOK%20ON%20IMPROVING%20QUALITY.pdf

Armitage P., Berry G. (1987). *Statistical Methods in Medical Research*, 2nd edn. Blackwell, Oxford

Axinn W.G., Link C.F., Groves R.M. (2011). Responsive survey design, demographic data collection, and models of demographic behavior. *Demography* 48(3):1127–1149

Bart J., Earnst S. (2002). Double sampling to estimate density and population trends in birds. *The AuK* 119(1):36–45

Bates D., Maechler M., Bolker B. (2012). lme4: Linear mixed-effects models using S4 classes. URL http://CRAN.R-project.org/package=lme4

Bell B., Mohadjer L., Montaquila J., Rizzo L. (1999). Creating a frame of newly constructed units for household surveys. In: Proceedings of the Survey Research Methods Section, American Statistical Association, pp 306–310

Bethlehem J., Cobben F., Schouten B. (2011). *Handbook in Nonresponse in Household Surveys*. John Wiley & Sons, Inc., New Jersey

Biemer P., Lyberg L. (2003). *Introduction to Survey Quality*. John Wiley & Sons, Inc., New Jersey

Blasius J., Thiessen V. (2012). *Assessing the Quality of Survey Data*. SAGE Publications Ltd., London

Blom A. (2008). Measuring nonresponse cross-nationally. ISER Working Paper Series URL http://ideas.repec.org/p/ese/iserwp/2008-41.html, no. 2008–41

Breiman L., Friedman J., Stone C., Olshen R. (1993). *Classification and Regression Trees*. Chapman & Hall, London

Breivik H., Cherny N., Collett B., de Conno F., Filbet M., Foubert A.J., Cohen R., Dow L. (2009). Cancer-related pain: a pan-European survey of prevalence, treatment, and patient attitudes. *Annals of Oncology* 20(8):1420–1433

Brick J.M., Waksberg J., Kulp D., Starer A. (1995). Bias in list-assisted telephone samples. *Public Opinion Quarterly* 59(2):218–235

Brown L., Cai T., Das Gupta A. (2001). Interval estimation for a binomial proportion. *Statistical Science* 16:101–133

Bureau of Labor Statistics (2006). Household Data (A tables, monthly: D tables, quarterly). Employment and Earnings. URL http://www.bls.gov/cps/eetech_methods.pdf

Bureau of Labor Statistics (2009). Economic News Release: CPI Consumer Price Index, June 2009. URL http://www.bls.gov/news.release/cpi.nr0.htm

Bureau of Labor Statistics (2012). American Time Use Survey User's Guide. URL http://stats.bls.gov/tus/atususersguide.pdf

Canada S. (2009). Statistics Canada quality guidelines. Tech. rep., Ottawa CA

Casella G., Berger R. (2002). *Statistical Inference*. Duxbury Press, Pacific Grove CA

Center for Disease Control and Prevention (2005). National Hospital Discharge Survey: 2005 Annual summary with detailed diagnosis and procedure data. Vital and health Statistics (165), URL http://www.cdc.gov/nchs/data/series/sr_13/sr13_165.pdf.

Center for Disease Control and Prevention (2009). National Health and Nutrition Examination Survey: 1999–2010 survey content URL www.cdc.gov/nchs/data/nhanes/survey_content_99_10.pdf

Chromy J.R. (1979). Sequential sample selection methods. In: Proceedings of the Survey Research Methods Section, American Statistical Association, pp 401–406

Chromy J.R., Myers L.E. (2001). Variance models applicable to the NHSDA. In: Proceedings of the Survey Research Methods Section, American Statistical Association

Cochran W. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics* 24:295–313

Cochran W. (1977). *Sampling Techniques*. John Wiley & Sons, Inc., New York

Cohen J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, New Jersey

Cook R.D., Weisberg S. (1982). *Residuals and Influence in Regression*. Chapman & Hall Ltd, London

Council of the European Union (1998). Council regulation (ec) no. 577/98 on the organization of a labour force survey in the community. Official Journal of the European Communities

Council of the European Union (2003). Council regulation (ec) no. 1177/2003 concerning community statistics on income and living conditions. Official Journal of the European Communities

Crawley M. (2007). *The R Book*. John Wiley & Sons, Chichester UK

Czajka J., Hirabayashi S., Little R.J.A., Rubin D.B. (1992). Projecting from advance data using propensity modeling: An application to income and tax statistics. *Journal of Business and Economic Statistics* 10:117–131

D'Agostino R.B. (1998). Propensity score methods for bias reduction for the comparison of a treatment to a non-randomized control group. *Statistics in Medicine* 17:2265–2281

Dantzig G.B. (1963). *Linear Programming and Extensions*. Princeton University Press, Princeton NJ

Deak M.A., Helba C., Lee K., Rockwell D., Perry S., Simmons R.O., D'Amato-Neff A.L., Ferro G., Lappin B.M. (2002). Tabulations of Responses from the 2000 Survey of Reserve Component Personnel: Defense Manpower Data Center Volume 2 Military Plans, Military Training, and Military Unit. URL http://www.dtic.mil/cgi-bin/GetTRDoc?AD=ADA415264&Location=U2&doc=GetTRDoc.pdf

Defense Manpower Data Center (2004). May 2004 Status of Forces Survey of Reserve component members: Administration, datasets, and codebook. Tech. Rep. No. 2004–013, Defense Manpower Data Center, Arlington, VA

DeMeyer A., Loch C.H., Pick M.T. (2002). Managing project uncertainty: From variation to chaos. *MIT Sloan Management Review* 30:60–67

Deming W.E. (1982). *Out of the Crisis*. Cambridge University Press, Cambridge

Dever J.A., Valliant R. (2010). A comparison of variance estimators for poststratification to estimated control totals. *Survey Methodology* 36:45–56

Deville J.C., Särndal C. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association* 87(418):376–382

Dillman D.A., Smyth J.D., Christian L.M. (2009). *Internet, Mail, and Mixed-Mode Surveys: The Tailored Design Method*. John Wiley & Sons, Inc., Hoboken, NJ

Dippo C.S., Fay R.E., Morganstein D.R. (1984). Computing variances from complex samples with replicate weights. In: Proceedings of the Survey Research Methods Section, American Statistical Association, pp 489–494

Durrant G.B., Steele F. (2009). Multilevel modelling of refusal and non-contact in household surveys: Evidence from six UK government surveys. *Journal Of The Royal Statistical Society, Series A* 172(2):361–381

Eckman S. (2010). Errors in Housing Unit Listing and their Effects on Survey Estimates. University of Maryland, College Park MD, URL http://drum.lib.umd.edu//handle/1903/10302

Eckman S., Kreuter F. (2011). Confirmation bias in housing unit listing. *Public Opinion Quarterly* 75(1):139–150

Eckman S., O'Muircheartaigh C. (2011). Performance of the half–open interval missed housing unit procedure. *Survey Research Methods* 5(3):125–131

Efron B. (1982). The Jackknife, the Bootstrap and Other Resampling Plans. SIAM [Society for Industrial and Applied Mathematics], Philadelphia

Efron B., Tibshirani R. (1998). *An Introduction to the Bootstrap*. CRC Press LLC, Boca Raton FL

Ezzati-Rice T., Rohde F., Greenblatt J. (2008). Sample design of the medical expenditure panel survey household component, 1998–2007. Tech. Rep. Methodology Report No. 22, Agency for Healthcare Research and Quality

Fay R.E. (1984). Some properties of estimates of variance based on replication methods. In: Proceedings of the Survey Research Methods Section, American Statistical Association, pp 495–500

Folsom R.E., Singh A.C. (2000). The generalized exponential model for sampling weight calibration for extreme values, nonresponse, and poststratification. In: Proceedings of the Survey Research Methods Section, American Statistical Association, pp 598–603

Folsom R.E., Potter F.J., Williams S.R. (1987). Notes on a composite size measure for self-weighting samples in multiple domains. In: Proceedings of the Survey Research Methods Section, American Statistical Association, pp 792–796

Francisco C., Fuller W.A. (1991). Quantile estimation with a complex survey design. *Annals of Statistics* 19:454–469

Freund R. (1994). Professor George Dantzig: Linear programming founder turns 80. SIAM News

Fuller W.A. (1998). Replication variance estimation for two-phase samples. *Statistica Sinica* 8:1153–1164

Gabler S., Haeder S., Lahiri P. (1999). A model based justification of Kish's formula for design effects for weighting and clustering. *Survey Methodology* 25(1):105–106

Gambino J.G. (2005). pps: Functions for PPS sampling. URL http://CRAN.R-project.org/package=pps

Gelman A., Carlin J., Stern H., Rubin D.B. (1995). *Data Analysis*. Chapman & Hall/CRC., Boca Raton FL

Godambe V.P., Joshi V.M. (1965). Admissibility and Bayes estimation in sampling finite populations - I. *Annals of Mathematical Statistics* 36:1707–1723

Groves R.M. (1989). *Survey Errors and Survey Costs*. John Wiley & Sons, Inc., New York

Groves R.M. (2006). Nonresponse rates and nonresponse bias in household surveys. *Public Opinion Quarterly* 70(5):646–675

Groves R.M., Heeringa S.G. (2006). Responsive design for household surveys: Tools for actively controlling survey errors and costs. *Journal of the Royal Statistical Society, Series A: Statistics in Society* 169(3):439–457

Groves R.M., Peytcheva E. (2008). The impact of nonresponse rates on nonresponse bias. *Public Opinion Quarterly* 72:167–189

Groves R.M., Fowler F., Couper M., Lepkowski J., Singer E., Tourangeau R. (2004). *Survey Methodology*. John Wiley & Sons, Inc., New York

Hansen M.H., Hurwitz W.H. (1943). On the theory of sampling from finite populations. *Annals of Mathematical Statistics* 41:517–529

Hansen M.H., Tepping B.J. (1990). Regression estimates in federal welfare quality control programs (C/R: P864-873). *Journal of the American Statistical Association* 85:856–864

Hansen M.H., Hurwitz W.H., Madow W.G. (1953a). *Sample Survey Methods and Theory*, Volume I. John Wiley & Sons, Inc., New York

Hansen M.H., Hurwitz W.H., Madow W.G. (1953b). *Sample Survey Methods and Theory*, Volume II. John Wiley & Sons, Inc., New York

Hansen M.H., Madow W.G., Tepping B.J. (1983). An evaluation of model-dependent and probability sampling inferences in sample surveys. *Journal of the American Statistical Association* 78:776–793

Harder V., Stuart E., Anthony J. (2010). Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. *Psychological Methods* 15(3):234–249

Hartley H.O., Rao J.N.K. (1962). Sampling with unequal probabilities and with replacement. *Annals of Mathematical Statistics* 33(2):350–374

Hedges L.V., Olkin I. (1985). *Statistical Methods for Meta-analysis*. Academic Press, Orlando

Heiberger R.M., Neuwirth E. (2009) *R Through Excel: A Spreadsheet Interface for Statistics, Data Analysis, and Graphics*. Springer, New York

Henry K.A. (2011). Weight adjustment methods and their impact on sample-based inference. PhD thesis, College Park MD, URL http://drum.lib.umd.edu/handle/1903/12278

Henry K.A., Valliant R. (2009). Comparing sampling and estimation strategies in establishment populations. *Survey Research Methods* 3:27–44

Henry K.A., Testa V.L., Valliant R. (2008). Variance estimation for an estimator of between-year change in totals from two stratified Bernoulli samples. In: Proceedings of the Survey Research Methods Section, American Statistical Association, pp 1108–1115

Herzog T.N., Scheuren F.J., Winkler W.E. (2007). *Data Quality and Record Linkage*. Springer, New York

Hidiroglou, M. A. (2001). Double sampling. *Survey Methodology*, 27, 143–154

Højsgaard S., Halekoh U. (2012). doBy: doBy - Groupwise summary statistics, general linear contrasts, population means (least-squares-means), and other utilities. URL http://CRAN.R-project.org/package=doBy, (contributions from J. Robison-Cox, K. Wright, A. A. Leidi)

Hunter S.R., Bowman K.R., Chromy J.R. (2005). Results of the variance component analysis of sample allocation by age in the National Survey on Drug Use and Health. In: Proceedings of the Survey Research Methods Section, American Statistical Association, pp 3132–3136

Iannacchione V.G. (2011). Research synthesis: The changing role of address-based sampling in surveys. *Public Opinion Quarterly* 75(3):556–576

Iannacchione V.G., Staab J.M., Redden D.T. (2003). Evaluating the use of residential mailing lists in a metropolitan household survey. *Public Opinion Quarterly* 67(2):202–210

Iannacchione V.G., Dever J.A., Bann C.M., Considine K.A., Creel D., Carson C.P., Best H.L., Haley R.W. (2011). Validation of a research case definition of Gulf War illness in the 1991 U.S. military population. *Neuroepidemiology* 37(2):129–140

Ingels S.J., Pratt D.J., Herget D., Dever J.A., Ottem R., Rogers J., Jin Y., Leinwand S. (2011). High School Longitudinal Study of 2009 (HSLS:09) base-year data file documentation (NCES 2011-328). Tech. rep., National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education, Washington DC

Internal Revenue Service (2004). Internal Revenue Bulletin: 2004–20, Meals and Entertainment Expenses. URL http://www.irs.gov/irb/2007-23_IRB/ar10.html

Internal Revenue Service (2007). Cost Segregation Audit Techniques Guide. URL http://www.irs.gov/Businesses/Cost-Segregation-Audit-Techniques-Guide---Table-of-Contents

International Organization for Standardization (1985). Information processing – documentation symbols and conventions for data, program and system flowcharts, program network charts and system resources charts. Tech. rep., Geneva, Switzerland, URL http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=11955

Isaki C.T., Fuller W.A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association* 77(377):89–96

Isaki C.T., Tsay J.H., Fuller W.A. (2004). Weighting sample data subject to independent controls. *Survey Methodology* 30(1):35–44

Jans M., Sirkis R., Morgan D. (2013). Managing data quality indicators with paradata-based statistical quality control tools. In: Kreuter F. (ed) *Improving Surveys with Paradata: Making use of Process Information*, John Wiley & Sons, Inc., New York

Jovanovic B.D., Levy P.S. (1997). A look at the rule of three. *The American Statistician* 51:137–139

Judkins D. (1990). Fay's method of variance estimation. *Journal of Official Statistics* 6:223–239

Judkins D., Van de Kerckhove W. (2003). RECS 2005 optimization. prepared for U.S. Department of Energy, no. 16.3, Task 98-010, contract no.: De-ac01-96e123968. Tech. rep., Westat, Rockville MD

Judkins D., Hao H., Barrett B., Adhikari P. (2005). Modeling and polishing of nonresponse propensity. In: Proceedings of the Survey Research Methods Section, American Statistical Association, pp 3159–3166

Kalton G. (1993). Sampling rare and elusive populations. Tech. Rep. INT-92-P80-16E, Department for Economic and Social Information and Policy Analysis, United Nations

Kalton G., Anderson D. (1986). Sampling rare populations. *Journal of the Royal Statistical Society A* 149:65–82

Kalton G., Maligalig D.S. (1991). A comparison of methods of weighting adjustment for nonresponse. Proceedings of the US Bureau of the Census Annual Research Conference pp 409–428

Kang J.D.Y., Schafer J.L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science* 22(4):523–539

Kass G.V. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistics* 29(2):119–127

Kim J.J., Li J., Valliant R. (2007). Cell collapsing in poststratification. *Survey Methodology* 33(2):139–150

Kim J.K., Yu C.L. (2011). Replication variance estimation under two-phase sampling. *Survey Methodology* 37(1):67–74

Kim J.K., Navarro A., Fuller W.A. (2006). Replication variance estimation for two-phase stratified sampling. *Journal of the American Statistical Association* 101(473):312–320

Kirgis N., Lepkowski J. (2010). A management model for continuous data collection: Reflections from the National Survey of Family Growth, 2006–2010. NSFG Paper No 10–011 URL http://www.psc.isr.umich.edu/pubs/pdf/ng10-011.pdf

Kish L. (1965). *Survey Sampling*. John Wiley & Sons, Inc., New York

Kish L. (1987a). *Statistical Design for Research*. John Wiley & Sons, Inc., New York

Kish L. (1987b). Weighting in deft. The Survey Statistician

Kish L. (1992). Weighting for unequal pi. *Journal of Official Statistics* 8(2):183–200

Kohler U. (2007). Surveys from inside: An assessment of unit nonresponse bias with internal criteria. *Survey Research Methods* 1(2):55–67

Kohler U., Kreuter F. (2012). *Data Analysis using Stata*, 3rd edn. StataPress, College Station TX

Korn E.L. (1986). Sample size tables for bounding small proportions. *Biometrics* 42:213–216

Korn E.L.., Graubard B.I. (1998). Confidence intervals for proportions with small expected number of positive counts estimated from survey data. *Survey Methodology* 24:193–201

Korn E.L., Graubard B.I. (1999). *Analysis of Health Surveys*. John Wiley & Sons, New York

Korn E.L., Graubard B.I. (2003). Estimating variance components by using survey data. *Journal of the Royal Statistical Society, Series B: Statistical Methodology* 65(1):175–190

Kostanich D., Dippo C.S. (2002). Current Population Survey: Design and methodology (technical paper 63RV). Tech. rep., Census Bureau and Bureau of Labor Statistics, Washington, DC

Kott P.S. (1988). Model-based finite population correction for the Horvitz-Thompson estimator. *Biometrika* 75:797–799

Kott P.S. (1999). Some problems and solutions with a delete-a-group jackknife. In: Federal Committee on Statistical Methodology Research Conference,Vol.4, U.S. Bureau of the Census, pp 129–135

Kott P.S. (2001). The delete-a-group jackknife. *Journal of Official Statistics* 17(4):521–526

Kott P.S. (2006). Using calibration weighting to adjust for nonresponse and coverage errors. *Survey Methodology* 32(2):133–142

Kott P.S. (2009). Calibration weighting: Combining probability samples and linear prediction models. In: *Handbook of Statistics*, Volume 29B, Sample Surveys: Inference and Analysis, Elsevier, Amsterdam

Kott P.S., Liu Y. (2009). One-sided coverage intervals for a proportion estimated from a stratified simple random sample. *International Statistical Review* 77:251–265

Kott P.S., Stukel D.M. (1997). Can the jackknife be used with a two-phase sample? *Survey Methodology* 23:81–89

Kreuter F. (2002) *Kriminalitätsfurcht: Messung und methodische Probleme*. Leske and Budrich, Berlin

Kreuter F., Olson K. (2011). Multiple auxiliary variables in nonresponse adjustment. *Sociological Methods and Research* 40:311–332

Kreuter F., Couper M., Lyberg L. (2010). The use of paradata to monitor and manage survey data collection. In: Proceedings of the Survey Research Methods Section, American Statistical Association, pp 282–296

Krewski D., Rao J.N.K (1981). Inference from stratified samples: Properties of the linearization, jackknife, and balanced repeated replication methods. *Annals of Statistics* 9:1010–1019

Lange K. (2004). *Optimization*. Springer, New York

Lee H., Kim J.K. (2002). Jackknife variance estimation for two-phase samples with high sampling fractions. In: Proceedings of the Survey Research Methods Section, American Statistical Association, pp 2024–2028

Lehtonen R., Veijanen A. (1998). Logistic generalized regression estimators. *Survey Methodology* 24:51–55

Lemeshow S., Hosmer D., Klar J., Lwanga S. (1990). *Adequacy of Sample Size in Health Studies*. John Wiley & Sons, Inc., Chichester

Lepkowski J., Axinn W.G., Kirgis N., West B.T., Ndiaye S.K., Mosher W., Groves R.M. (2010). Use of paradata in a responsive design framework to manage a field data collection. NSFG Survey Methodology Working Papers (10–012), URL http://www.psc.isr.umich.edu/pubs/pdf/ng10-012.pdf

Li J., Valliant R. (2009). Survey weighted hat matrix and leverages. *Survey Methodology* 35:15–24

Li J., Valliant R. (2011). Linear regression diagnostics for unclustered survey data. *Journal of Official Statistics* 27:99–119

Liao D., Valliant R. (2012a). Condition indexes and variance decompositions for diagnosing collinearity in linear model analysis of survey data. *Survey Methodology* 38:53–62

Liao D., Valliant R. (2012b). Variance inflation factors in the analysis of complex survey data. *Survey Methodology* 38:189–202

Link M.W., Battaglia M.P., Frankel M.R., Osborn L., Mokdad A.H. (2008). A comparison of address-based sampling (ABS) versus random-digit dialing (RDD) for general population surveys. *Public Opinion Quarterly* 72(1):6–27

Little R.J.A. (1986). Survey nonresponse adjustments for estimates of means. *International Statistical Review* 54(2):139–157

Little R.J.A., Rubin D.B. (2002). *Statistical Analysis with Missing Data*. John Wiley & Sons, Inc., New Jersey

Little R.J.A., Vartivarian S. (2003). On weighting the rates in non-response weights. *Statistics in Medicine* 22:1589–1599

Little R.J.A., Vartivarian S. (2005). Does weighting for nonresponse increase the variance of survey means? *Survey Methodology* 31:161–168

Liu J., Aragon E. (2000). Subsampling strategies in longitudinal surveys. In: Proceedings of the Survey Research Methods Section, American Statistical Association, pp 307–312

Liu J., Iannacchione V.G., Byron M. (2002). Decomposing design effects for stratified sampling. In: Proceedings of the Survey Research Methods Section, American Statistical Association, pp 2124–2126

Lohr S.L. (1999). *Sampling: Design and Analysis*. Duxbury Press, Pacific Grove CA

Long J.S. (2009). *The Workflow of Data Analysis Using Stata*. StataPress, College Station TX

Lu W., Brick J.M., Sitter R. (2006). Algorithms for constructing combined strata grouped jackknife and balanced repeated replications with domains. *Journal of the American Statistical Association* 101:1680–1692

Lumley T. (2010). *Complex Surveys*. John Wiley & Sons, Inc., New York

Lumley T. (2012). survey: analysis of complex survey samples. URL http://CRAN.R-project.org/package=survey

Lyberg L., Biemer P., Collins M., de Leeuw E., Dippo C.S., Schwarz N., Trewin D. (1997). *Survey Measurement and Process Quality*. John Wiley & Sons, Inc., New York

Madsen K., Nielsen H.B., Tingleff O. (2004). Optimization with constraints. Tech. rep., Technical University of Denmark, URL http://www2.imm.dtu.dk/pubdb/views/edoc_download.php/4213/pdf/imm4213.pdf, 2nd edn

Manitz J. (2012). samplingbook: Survey Sampling Procedures. URL http://CRAN.R-project.org/package=samplingbook, (contributions by M. Hempelmann, G. Kauermann, H. Kuechenhoff, S. Shao, C. Oberhauser, N. Westerheide, M. Wiesenfarth)

Matsuo H., Billiet J., Loosveldt G., Berglund F., Kleven Ø. (2010). Measurement and adjustment of non-response bias based on non-response surveys: The case of Belgium and Norway in the European Social Survey round 3. *Survey Research Methods* 4:165–178

McCarthy P.J. (1969). Pseudo-replication: Half-samples. *Review of the International Statistical Institute* 37:239–264

Michie D. (1989). *Problems of computer-aided concept formation. In Applications of Expert Systems 2*. Turing Institute Press/Addison-Wesley

Montaquila J., Bell B., Mohadjer L., Rizzo L. (1999). A methodology for sampling households late in a decade. In: Proceedings of the Survey Research Methods Section, American Statistical Association, pp 311–315

Morgan J.N., Sonquist J.A. (1963). Problems in the analysis of survey data and a proposal. *Journal of the American Statistical Association* 58:415–434

Morganstein D.R., Marker D.A. (1997). Continuous quality improvement in statistical agencies. In: Lyberg L., Biemer P., Collins M., De Leeuw E.D., Dippo C.S., Schwarz N., Trewin D. (eds) *Survey Measurement and Process Quality*, John Wiley & Sons, Inc., New York

Morton K., Chromy J.R., Hunter S.R., Martin P. (2006). 2005 National Survey on Drug Use and Health sample design report. Tech. Rep. Prepared under contract no. 283-2004-00022 by Research Triangle Institute, Substance Abuse and Mental Health Services Administration.

Müller G. (2011). Fieldwork monitoring in pass. Tech. rep., Institut für Arbeitsmarkt und Berufsforschung, URL http://www.iab.de/de/veranstaltungen/konferenzen-und-workshops-2011/paradata.aspx

National Center For Education Statistics (2008). NAEP weighting procedures: 2003 weighting procedures and variance estimation. Tech. rep., National Center for Education Statistics, URL http://nces.ed.gov/nationsreportcard/tdw/weighting/2002_2003/weighting_2003_studtrim.asp

Newcombe R.G. (1998). Two-sided confidence intervals for the single proportion: Comparison of seven methods. *Statistics in Medicine* 17(8):857–872

Neyman J. (1938). Contribution to the theory of sampling human populations. *Journal of the American Statistical Association* 33(201):101–116

Olson K., Peytchev A. (2007). Effect of interviewer experience on interview pace and interviewer attitudes. *Public Opinion Quarterly* 71:273–286

O'Muircheartaigh C., Campanelli P. (1998). The relative impact of interviewer effects and sample design effects on survey precision. *Journal of the Royal Statistical Society, Series A* 161(1):63–77

O'Muircheartaigh C., Campanelli P. (1999). A multilevel exploration of the role of interviewers in survey non-response. *Journal of the Royal Statistical Society, Series A* 162(3):437–446

Pfeffermann D., Skinner C.J., Holmes D.J., Goldstein H., Rasbash J. (1998). Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society Series B-Statistical Methodology* 60(Part 1):23–40

Pinheiro J.C., Bates D.M. (2000). *Mixed-effects Models in S and S-PLUS*. Springer, New York

Porter E.H., Winkler W.E. (1997). Approximate string comparison and its effect in an advanced record linkage system. In: Alvey W., Jamerson B. (eds) Record Linkage – 1997: Proceedings of an International Workshop and Exposition, U.S. Office of Management and Budget, pp 190–199

Potter F.J. (1990). A study of procedures to identify and trim extreme sampling weights. In: Proceedings of the Survey Research Methods Section, American Statistical Association, pp 225–230

Potter F.J. (1993). The effect of weight trimming on nonlinear survey estimates. In: Proceedings of the Survey Research Methods Section, American Statistical Association, pp 758–763

Potter F.J., Iannacchione V.G., Mosher W., Mason R., Kavee J.A. (1998). Sample design, sampling weights, imputation, and variance estimation in the 1995 National Survey of Family Growth. Vital and Health Statistics, National Center for Health Statistics 124(2)

Powell S.G., Baker K.R. (2003). *The Art of Modeling with Spreadsheets: Management Science, Spreadsheet Engineering, and Modeling Craft*. John Wiley & Sons, Inc., New York

R Core Team (2012). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, URL http://www.R-project.org/

R Core Team and contributors worldwide (2012a). foreign: Read Data Stored by Minitab, S, SAS, SPSS, Stata, Systat, dBase,. URL http://CRAN.R-project.org/package=foreign

R Core Team and contributors worldwide (2012b). graphics: R functions for base graphics. URL http://finzi.psych.upenn.edu/R/library/graphics/html/00Index.html

R Core Team and contributors worldwide (2012c). stats: R statistical functions. URL http://finzi.psych.upenn.edu/R/library/stats/html/00Index.html

Rao J.N.K. (1973). On double sampling for stratification and analytical surveys (Corr: V60 p669). *Biometrika* 60:125–133

Rao J.N.K., Shao J. (1999). Modified balanced repeated replication for complex survey data. *Biometrika* 86(2):403–415

Rao J.N.K., Wu C.F.J. (1985). Inference from stratified samples: Second-order analysis of three methods for nonlinear statistics. *Journal of the American Statistical Association* 80:620–630

Rao J.N.K., Wu C.F.J. (1988). Resampling inference with complex survey data. *Journal of the American Statistical Association* 83:231–241

Rizzo L., Kalton G., Brick J.M. (1996). A comparison of some weighting adjustments for panel nonresponse. *Survey Methodology* 22:43–53

Rosenbaum P., Rubin D.B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* 70:41–55

Royall R.M. (1976). Current advances in sampling theory: Implications for human observational studies. *American Journal of Epidemiology* 104:463–473

Royall R.M. (1986). The effect of sample size on the meaning of significance tests. *The American Statistician* 40:313–315

RTI International (2012). SUDAAN Language Manual, Release 11.0. Research Triangle Park NC

Rubin D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, New York

Rust K.F. (1984). Techniques for estimating variances for sample surveys. PhD thesis, University of Michigan, Ann Arbor MI, unpublished

Rust K.F. (1985). Variance estimation for complex estimators in sample surveys. *Journal of Official Statistics* 1:381–397

Saigo H., Shao J., Sitter R. (2001). A repeated half-sample bootstrap and balanced repeated replications for randomly imputed data. *Survey Methodology* 27(2):189–196

Särndal C. (2007). The calibration approach in survey theory and practice. *Survey Methodology* 33(2):99–119

Särndal C., Lundström S. (2005). *Estimation in Surveys with Nonresponse*. John Wiley & Sons, Inc., England

Särndal C., Lundström S. (2008). Assessing auxiliary vectors for control of nonresponse bias in the calibration estimator. *Journal of Official Statistics* 24:167–191

Särndal C., Swensson B., Wretman J. (1992). *Model Assisted Survey Sampling*. Springer, New York

Schlesselman J. (1982). *Case-Control Studies: Design, Conduct, and Analysis*. Oxford University Press, New York

Schnell R., Kreuter F. (2005). Separating interviewer and sampling-point effects. *Journal of Official Statistics* 21(3):389–410

Schnell R., Bachteler T., Bender S. (2004). A toolbox for record linkage. *Austrian Journal of Statistics* 33(1–2):125–133

Schouten B., Cobben F. (2007). R-indexes for the comparison of different fieldwork strategies and data collection modes. Tech. Rep. Discussion Paper 07002, Voorburg, The Netherlands, URL http://www.risq-project.eu/papers/schouten-cobben-2007-a.pdf

Schouten B., Cobben F., Bethlehem J. (2009). Indicators for the representativeness of survey response. *Survey Methodology* 35(1):101–113

Searle S., Casella G., McCulloch C. (1992). *Variance Components*. John Wiley & Sons, New York

Shao J., Sitter R. (1996). Bootstrap for imputed survey data. *Journal of the American Statistical Association* 91:1278–1288

Shewhart W.A. (1931). Economic Control of Quality of Manufactured Product. Van Nostrand Reinhold Co., Princeton, NJ, republished in 1981 by the American Society for Quality Control, Milwaukee, WI

Singh A.C., Dever J.A., Iannacchione V.G. (2004). Composite response rates for surveys with nonresponse follow-up. In: Proceedings of the Survey Research Methods Section, American Statistical Association, pp 4343–4350

Singh A.C., Dever J.A., Iannacchione V.G., Chen S. (2005). Efficient estimation of response rates when a small subsample of nonrespondents is selected for follow-up conversion. In: Federal Committee on Statistical Methodology (FCSM) Conference, Arlington, VA, URL http://www.fcsm.gov/05papers/Singh_Iannacchione_etal_VIIB.pdf

Sitter R. (1992). Comparing three bootstrap methods for survey data. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique* 20:135–154

Smith P.J., Hoaglin D.C., Battaglia M.P., Khare M., Barker L.E. (2005). Statistical methodology of the National Immunization Survey, 1994–2002. Vital and Health Statistics, National Center for Health Statistics 2(138)

Smith T.M.F. (1976). The foundations of survey sampling: A review. *Journal of the Royal Statistical Society A* 139:183–204

Smith T.M.F. (1984). Present position and potential developments: Some personal views, sample surveys. *Journal of the Royal Statistical Society A* 147:208–221

Smith T.M.F. (1994). Sample surveys 1975–1990; an age of reconciliation? *International Statistical Review* 62:5–34

Spencer B.D. (2000). An approximate design effect for unequal weighting when measurements may correlate with selection probabilities. *Survey Methodology* 26(2):137–138

Stuart E. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science* 25(1):1–21

Stukel D.M., Särndal C., Hidiroglou M.A. (1996). Variance estimation for calibration estimators: A comparison of jackknifing versus Taylor linearization. *Survey Methodology* 22(2):117–125

Thayer W.C., Diamond G.L. (2002). Blood Lead Concentrations of U.S. Adult Females: Summary Statistics from Phases 1 and 2 of the National Health and Nutrition Evaluation Survey (NHANES III). URL http://www.epa.gov/superfund/lead/products/nhanes.pdf

Therneau T. (2012). survival: Survival analysis, including penalised likelihood. URL http://CRAN.R-project.org/package=survival

Therneau T., Atkinson B., Ripley B. (2012). rpart: Recursive Partitioning. URL http://CRAN.R-project.org/package=rpart

Thomas B. (1999). Probabilistic record linkage software: A Statistics Canada evaluation of GRLS and Automatch. In: Proceedings of the Survey Research Methods Section, American Statistical Association, pp 187–192

Tillé Y., Matei A. (2012). sampling: Survey Sampling. URL http://CRAN.R-project.org/package=sampling

Tourangeau R., Kreuter F., Eckman S. (2012). Motivated underreporting in screening interviews. *Public Opinion Quarterly* 76(3):453–469

Traugott M.W., Goldstein K. (1993). Evaluating dual frame samples and advance letters as a means of increasing response rates. In: Proceedings of the Survey Research Methods Section, American Statistical Association, pp 1284–1286

Tufte E. (1990). Envisioning Information. Graphics Press, Cheshire CT

Turlach B.A., Weingessel A. (2011). quadprog: Functions to solve Quadratic Programming Problems. URL http://CRAN.R-project.org/package=quadprog

US Census Bureau (1991). The 1990 Census of Population and Housing. Population and Housing Counts: 1790–1990

US Census Bureau (2001a). Housing Characteristics: 2000. Census 2000 Brief. URL http://www.census.gov/prod/www/abs/briefs.html

US Census Bureau (2001b). Population Change and Distribution 1990–2000. Census 2000 Brief. URL http://www.census.gov/prod/www/abs/briefs.html

US Census Bureau (2002). Source and Accuracy of Estimates for Poverty in the United States: 2001. URL http://www.census.gov/prod/2002pubs/p60-219sa.pdf

US Census Bureau (2006) Current Population Survey: Design and Methodology. URL http://www.census.gov/prod/2006pubs/tp-66.pdf

US Census Bureau (2011). 2010 Census Redistricting Data (Public Law 94–171) Summary File. URL http://www.census.gov/prod/cen2010/doc/pl94-171.pdf

Valliant R. (1985). Nonlinear prediction theory and the estimation of proportions in a finite population. *Journal of the American Statistical Association* 80:631–641

Valliant R. (1993). Poststratification and conditional variance estimation. *Journal of the American Statistical Association* 88:89–96

Valliant R. (2004). The effect of multiple weight adjustments on variance estimation. *Journal of Official Statistics* 20:1–18

Valliant R., Rust K.F. (2010). Degrees of freedom approximations and rules-of-thumb. *Journal of Official Statistics* 26:585–602

Valliant R., Dorfman A.H., Royall R.M. (2000). *Finite Population Sampling and Inference: A Prediction Approach*. John Wiley & Sons, Inc., New York

Valliant R., Göksel H., Barrett B. (2003). 2003 Commercial Buildings Energy Consumption survey sample design report, prepared for U.S. Department of Energy under contract no. DE-AC01-96E123968. Tech. rep., Westat, Rockville MD

Valliant R., Brick J.M., Dever J.A. (2008). Weight adjustments for the grouped jackknife variance estimator. *Journal of Official Statistics* 24(3):469–488

Vapnik V.N. (1995) *The Nature of Statistical Learning Theory*. Springer, New York

Varadhan R. (2010). alabama: Constrained nonlinear optimization. URL http://CRAN.R-project.org/package=alabama, (with contributions from Gabor Grothendieck). R package version 2010.10-1

Venables W.N., Ripley B.D. (2002). *Modern Applied Statistics with S*, 4th edn. Springer, New York

Victor R.G., Haley R.W., Willett D.L., Peshock R.M., Vaeth P.C., Leonard D., Basit M., Cooper R.S., Iannacchione V.G., Visscher W.A., Staab J.M., Hobbs H.H., Dallas Heart Study Investigators (2004). The Dallas Heart Study: A population-based probability sample for the multidisciplinary study of ethnic differences in cardiovascular health. *American Journal of Cardiology* 93(12):1473–1480, URL http://www.ncbi.nlm.nih.gov/pubmed/15194016

Wagner J. (2010). The fraction of missing information as a tool for monitoring the quality of survey data. *Public Opinion Quarterly* 74(2):223–243

Waksberg J., Sperry S., Judkins D., Smith V. (1993). National Survey of Family Growth, evaluation of linked design. *Vital Health Statistics* 2 (117), (PHS) 93–1391

Waksberg J., Judkins D., Massey J.T. (1997). Geographic-based oversampling in demographic surveys of the united states. *Survey Methodology* 23:61–71

Weisberg S. (2005). *Applied Linear Regression*, 3rd edn. John Wiley & Sons, New York

Weisstein E.W. (2010). Extreme Value Distribution. URL http://mathworld.wolfram.com/ExtremeValueDistribution.html, from MathWorld–A Wolfram Web Resource

West B.T., Groves R.M. (2013). A propensity-adjusted interviewer performance indicator. *Public Opinion Quarterly* 77:to be published

West B.T., Olson K. (2010). How much of interviewer variance is really nonresponse error variance? *Public Opinion Quarterly* 74(5):1027–1045

Westat (2007). WesVar 4.3 User?s Guide. Westat, Rockville MD, URL www.westat.com

Wickham H. (2011). reshape: Flexibly reshape data. URL http://CRAN.R-project.org/package=reshape

Willenborg L., Heerschap H. (2012). Matching. Tech. rep., The Hague, URL http://www.cbs.nl/NR/rdonlyres/0EDC70A4-C776-43F6-94AD-A173EFE58915/0/2012Matchingart.pdf, method Series no. 12

Wilson E.B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association* 22:209–212

Winston W., Venkataramanan M. (2003). *Introduction to Mathematical Programming*, 4th edn. Duxbury Press, Pacific Grove CA

Winter N. (2002). svr: Stata SurVey Replication package. URL http://faculty.virginia.edu/nwinter/progs/

Wolter K.M. (2007). *Introduction to Variance Estimation*, 2nd edn. Springer, New York

Woodruff R.S. (1952). Confidence intervals for medians and other position measures. *Journal of the American Statistical Association* 47:635–646

Woodward M. (1992). Formulas for sample size, power, and minimum detectable relative risk in medical studies. *The Statistician* 41:185–196

Wright J., Marsden P. (2010). *Handbook of Survey Research*, 2nd edn. Emerald Group Publishing Limited, United Kingdom

Solutions to Selected Exercises

Chapter 3

3.2

(a) Calculate $CV(p_s)$ and $\sqrt{V(p_s)}$ for a sample size of $n = 100$.

```
n <- 100
p <- c(0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7,
       0.8, 0.9, 0.95, 0.99)
SE <- sqrt(p*(1-p)/n)
CV <- SE/p
cbind(p, SE = round(SE,4), CV = round(CV,4))
          p     SE     CV
 [1,]  0.01 0.0099 0.9950
 [2,]  0.05 0.0218 0.4359
 [3,]  0.10 0.0300 0.3000
 [4,]  0.20 0.0400 0.2000
 [5,]  0.30 0.0458 0.1528
 [6,]  0.40 0.0490 0.1225
 [7,]  0.50 0.0500 0.1000
 [8,]  0.60 0.0490 0.0816
 [9,]  0.70 0.0458 0.0655
[10,]  0.80 0.0400 0.0500
[11,]  0.90 0.0300 0.0333
[12,]  0.95 0.0218 0.0229
[13,]  0.99 0.0099 0.0101
```

(b) Graph SEs versus p.
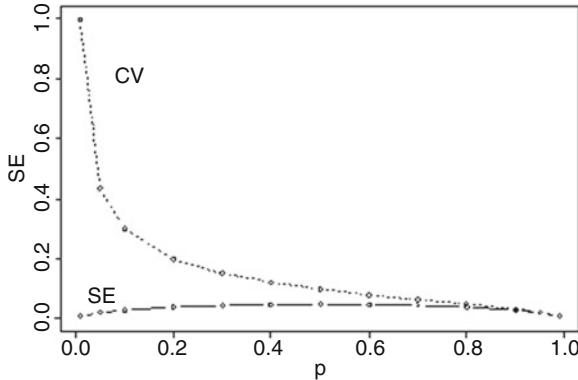
(c) Discuss the differences:

CVs for small p are extremely large, implying that this criterion would be difficult to use for rare characteristics. The relative differences in the SE are smaller over the range of p than for the CV. For setting a precision target for p_s the SE may be a more easily understood criterion than the CV.

3.8

(a) Relvariances of the variables beds and discharges in the hospital population

```
(a) beds, discharges

0.6024728    # unit relvariance of beds
0.5239741    # unit relvariance of discharges
```

(b) Relvariances of the variables total expenditures (EXPTOTAL), number of inpatient beds (BEDS), number of patients seen during 1998 (SEENCNT), the number of clients on the roles at the end of 1998 (EOYCNT), and number of in-patient visits (Y_IP) in the smho98 population.

```
                        var          mean       relvar
beds          4.546172e+04 2.746972e+02  0.6024728
discharges    3.477412e+05 8.146539e+02  0.5239741
smho exp      5.893495e+14 1.166418e+07  4.3317602
smho beds     2.559340e+04 8.389371e+01  3.6363792
smho seen     3.612683e+07 2.259911e+03  7.0737089
smho eoy      1.145212e+07 9.327166e+02 13.1639586
smho yip      2.834705e+08 7.574629e+03  4.9406626
```

3.10

(a) Determine n for $CV(\text{IPV}) = 0.10$.

```
CV0 <- 0.10
attach("C:\\Data\\smho98.RData")
pop <- smho98[smho98\$Y\_IP 0, ]
N <- nrow(pop)
N
```

```
  [1] 484
pk <- pop$BEDS / sum(pop$BEDS)
y <- pop$Y_IP
T <- sum(y)
T
  [1] 6627800
ybarU <- mean(y)
V1 <- sum(pk*(y/pk - T)^2)
V1
 [1] 3.19933e+13
n <- V1 / (N*ybarU*CV0)^2
n <- ceiling(n)
n
 [1] 73
pk1 <- n*pk
summary(pk1)
     Min.  1st Qu.   Median     Mean  3rd Qu.      Max.
 0.001301 0.039040 0.095660 0.150800 0.163300 1.766000

certs <- (1:N)[pk1 $>$= 0.80]
certs # Unit numbers of certainties
  [1] 154 155 156 157 161 179 189 191 192
length(certs)
  [1] 9
n - length(certs)
   64
         # Re-calculate excluding certainties
pk <- pop$BEDS[-certs] / sum(pop$BEDS[-certs])
y <- pop$Y_IP[-certs]
T <- sum(y)
T
  [1] 5706952
V1 <- sum(pk*(y/pk - T)^2)
V1
  [1] 2.552992e+13
nNc <- V1 / (N * ybarU * CV0)^2
nNc <- ceiling(nNc)
nNc
  [1] 59
         # Recheck to see whether there are new certainties.
         # There are none.
summary(nNc * pk)
     Min.  1st Qu.   Median     Mean  3rd Qu.      Max.
 0.001233 0.037000 0.088790 0.124200 0.149200 0.731300
         # total sample size
length(certs) + nNc
  [1] 68
         # check that CV0 met
CV <- sqrt(V1)/(N * sqrt(nNc) * ybarU)
CV
  [1] 0.09924974
```

(b) Repeat part (a) with a CV target of 0.15. There are no certainties. n=33.

(c) Now, suppose that you decide to use a regression estimator of the mean number of inpatient visits. Use a model with no intercept and with the square root of beds and beds itself as predictors. If this model is correct, what is the optimum measure of size to use in a *pps* sample? What sample would be required to obtain an anticipated *CV* of 0.10 with this regression estimator and a sample selected with the optimal MOS? *The optimal MOS for the model $E_M(y) = \beta_1\sqrt{x} + \beta_2 x$, $V_M(y) = \sigma^2 x$ is $\sqrt{x}$.*

```
CV0 <- 0.10
          #Create model variables
x <- pop[, "BEDS"]
y <- pop$Y_IP
rtvBar <- mean(sqrt(x))
vBar <- mean(x)
rtvBar; vBar
 [1] 9.397201
 [1] 115.8926


          #Object containing results of functions of x
          # modeled on y
m <- glm(y ~ 0 + sqrt(x) + x, weights = 1/x)
          #Model results
ybarU <- mean(y)
S2R <- sum(m\$residuals^2/(length(x)-1))
nCont(CV0=0.15, S2=S2R, ybarU=ybarU, N=670)
 [1] 41.88943
          # check for any certainties with n=42 and
          # pp(sqrt(x)) sampling
newPk <- 42 * sqrt(x)/sum(sqrt(x))
summary(newPk)
      Min.   1st Qu.    Median      Mean  3rd Qu.      Max.
  0.009234  0.050580  0.079170  0.086780  0.103400  0.340200
```

There are no certainties with this plan.

(d) Explain any differences in the results for parts (a), (b), and (c). *The sample sizes in (a), (b), and (c) are:*

59 noncertainties plus 9 certainties with pp(x) and CV0 = 0.10
33 with 0 certainties with pp(x) and CV0 = 0.15
42 with 0 certainties with pp(sqrt(x)), a regression estimator, and CV0 = 0.10

With pp(x) and the pi-estimator a smaller sample is naturally required for a target CV of 0.15 than 0.10. If a more efficient regression estimator is used, 42 units rather than 70 are required for CV0=0.10. Thus, sampling with pps does not gain all the efficiency possible from a sample when a strong y-x relationship is present.

3.12

(a) Calculate the design weights for the 50 sample hospitals. How might you verify that the weights were calculated correctly? Show the verification.

```
hosp50 <- read.csv("C:\\Data\\hospital_50.txt", header=TRUE)
wts <- sum(hospital[, "x"]) / 50 / hosp50[, "x"]
N <- nrow(hospital)
n <- nrow(hosp50)
wts <- 1/pik[sam == 1]
sum(wts)
  [1] 442.3302
        # sum of wtd beds should equal pop total of beds
sum(wts*hosp50\$x)
  [1] 107956
sum(hospital\$x)
  [1] 107956
```

(b) Estimate the average number of discharges based on the sample using the π-estimator of the mean.

```
tHat <- sum(wts*hosp50\$y)
ybarHat <- tHat / N
tHat; sum(hospital\$y)
  [1] 317339.5
  [1] 320159
ybarHat; mean(hospital\$y)
  [1] 807.4796
  [1] 814.6539
```

(c) Estimate the sample variance for your estimate in (b) using the formula for with-replacement sampling. If you used more than one estimator in (b), compute the estimated variance of each.

```
y <- hosp50$y
pk <- 1/(n*wts)
V1Hat <- sum( (y/pk -- mean(y/pk))^2 ) / (n-1)
vHat <- V1Hat/N^2/n
vHat
  [1] 918.535
sqrt(vHat)
  [1] 30.30734
sqrt(vHat) / ybarHat
  [1] 0.03753326
```

(d) Estimate the 95 % confidence interval for your estimate in (b).

```
        # 95% CI
LB <- ybarHat - 1.96*sqrt(vHat)
UB <- ybarHat + 1.96*sqrt(vHat)
c(LB, UB)
  [1] 748.0772 866.8820
```

(e) Suppose you want to select a new sample with probabilities proportional to the square root of beds. Estimate the appropriate V_1 for this design. How many sample hospitals would be needed to meet the target $CV\ (\bar{y}_{st})$ = 0.15 with this design?

```
qk <- sqrt(hospital\$x) / sum(sqrt(hospital\$x))
qk <- qk[sam==1]
V1 <- sum(y^2/pk/qk)/n - (mean(y/pk))^2 + vHat
V1
```

```
[1] 13791105407
CV0 = 0.04
ybarU <- mean(hospital\$y)
newN <- V1 / (N * ybarU * CV0)^2
newN
[1] 84.09064
```

3.14

(a) Determine the sample size needed to meet a target $CV{=}0.05$ for the estimated mean of the two analysis variables, y_1 and y_2. Are the estimated sample sizes different? Is so, why?

```
domy1y2 <- read.table("C:\\Data\\Domainy1y2.txt",
                header=TRUE)
ybar1 <- mean(domy1y2\$y1)
ybar2 <- mean(domy1y2\$y2)
s2y1 <- var(domy1y2\$y1)
s2y2 <- var(domy1y2\$y2)
nCont(CV0=0.05, S2=s2y1, ybarU=ybar1, N=100)
 [1] 41.28193
nCont(CV0=0.05, S2=s2y2, ybarU=ybar2, N=100)
 [1] 25.941
s2y1
 [1] 552.5725
s2y2
[1] 706.7866
s2y1/ybar1^2
 [1] 0.1757633
s2y2/ybar2^2
 [1] 0.08756869
```

Sample sizes are different because the unit relvariance of y2 is smaller. Note that the variance of y2 is larger than that of y1, however.

(b) If the target precision level is increased to a $CV{=}0.03$, how do your calculations in (a) change?

```
nCont(CV0=0.03, S2=s2y1, ybarU=ybar1, N=100)
 [1] 66.13528
nCont(CV0=0.03, S2=s2y2, ybarU=ybar2, N=100)
 [1] 49.31539
```

(c) Repeat your calculations in parts (a) and (b) for the proportion of y_1 responses that are less than or equal to 50 ($y_1 \leq 50$).

```
less50 <- rep(0, length(domy1y2$y1))
less50[domy1y2\$y1 $<$= 50] <- 1
ybar1 <- mean(less50)
nProp(CV0=c(0.05,0.03), pU=ybar1, N=100)
 [1] 80.16032 91.81893
```

(d) Repeat your calculations in parts (a) and (b) for the proportion of y_1 responses that are less than or equal to 22 ($y_1 \leq 22$). Compare your results from parts (c) and (d).

```
less22 <- rep(0, length(domy1y2\$y1))
less22[domy1y2\$y1 $<$= 22] <- 1
```

```
ybar1 <- mean(less22)
nProp(CV0=c(0.05,0.03), pU=ybar1, N=100)
  [1] 99.15377 99.69370
```

3.16

(a) Compare the selection probabilities for these two sample designs. For example, compute the mean *pps* selection probability within each stratum and compare it to the *stsrs* selection probabilities.

```
attach("C:\\Data\\smho.N874.RData",pos=2)
hospPop <- smho.N874[smho.N874\$BEDS > 0, ]
x <- sqrt(hospPop$BEDS)
hospPop <- hospPop[order(x), ]
x <- sort(x)
N <- nrow(hospPop)
n <- 50
cumx <- cumsum(x)
H <- 25
size <- cumx[N]/H
brks <- (0:H)*size
strata <- cut(cumx, breaks = brks, labels = 1:H)
Nh <- table(strata)
strSelprobs <- rep(2,H) / Nh
allStrProbs <- NULL
for (h in 1:H){
     allStrProbs <- c(allStrProbs, rep(strSelprobs[h], Nh[h]))
}
        # selection probabilities for pp(sqrt(x))
ppsSelprobs <- n*x / sum(x)
both <- NULL
both <- cbind(stratum = strata, pps = ppsSelprobs,
           stsrs = allStrProbs)
plot(both[, c(2,3)])
abline(0,1)
round(cbind(stsrs = strSelprobs,
        ppsMeans = by(both[,2],strata,mean)),4)
     stsrs ppsMeans
1  0.0263    0.0259
2  0.0357    0.0357
3  0.0417    0.0415
4  0.0465    0.0467
5  0.0526    0.0524
6  0.0571    0.0577
7  0.0625    0.0622
8  0.0667    0.0674
9  0.0714    0.0711
10 0.0769    0.0749
11 0.0769    0.0785
12 0.0833    0.0814
13 0.0833    0.0853
14 0.0909    0.0902
15 0.0952    0.0947
16 0.1000    0.1010
```

```
17 0.1111      0.1076
18 0.1111      0.1134
19 0.1250      0.1232
20 0.1333      0.1377
21 0.1538      0.1477
22 0.1538      0.1583
23 0.1818      0.1835
24 0.2222      0.2070
25 0.2500      0.2704
```

(b) Graph the *stsrs* probabilities versus the *pps* selection probabilities.
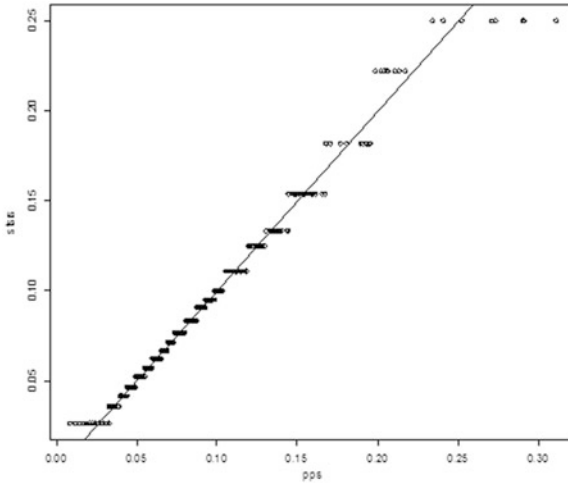
Chapter 4

4.2 Consider Example 4.6 where one-sided tests were used to determine sample sizes with 80 and 90 % power to detect differences in estimates for males and females.

(a) How does the sample size change if $\sigma_d^2 = 200$?

$$\sqrt{\sigma_d^2/2} = \sqrt{200/2} = 10$$

```
power.t.test(power = 0.8,
    delta = 5,
    sd = 10,
    type = "two.sample",
    alt = "one.sided",
    sig.level = 0.05
    )
    # Two-sample t test power calculation
            n = 50.1508
```

```
              delta = 5
                 sd = 10
          sig.level = 0.05
              power = 0.8
        alternative = one.sided
   NOTE: n is number in *each* group
```

```
  power.t.test(power = 0.9,
       delta = 5,
       sd = 10,
       type = "two.sample",
       alt = "one.sided",
       sig.level = 0.05
       )
```

```
      # Two-sample t test power calculation
                n = 69.19782
            delta = 5
               sd = 10
        sig.level = 0.05
            power = 0.9
      alternative = one.sided
```

(b) How does a $\sigma_d^2 = 800$ affect your previous calculation?

$$\sqrt{\sigma_d^2/2} = \sqrt{800/2} = \sqrt{400}$$

```
power.t.test(power = 0.8,
      delta = 5,
      sd = sqrt(400),
      type = "two.sample",
      alt = "one.sided",
      sig.level = 0.05
      )
  # Two-sample t test power calculation
               n = 198.5217
           delta = 5
              sd = 20
       sig.level = 0.05
           power = 0.8
     alternative = one.sided
```

```
 power.t.test(power = 0.9,
      delta = 5,
      sd = sqrt(400),
      type = "two.sample",
      alt = "one.sided",
      sig.level = 0.05
      )
```

```
   # Two-sample t test power calculation
               n = 274.7222
           delta = 5
```

```
            sd = 20
     sig.level = 0.05
         power = 0.9
   alternative = one.sided
```

4.4

(a) The client is interested in determining if the average BMI for children in the first grade (ages 6–7) has increased by 1.5 % from a previously estimated average of 17.5. What is the sample size needed to detect this difference given that the population standard deviation is 0.70?

```
d <- 17.5 * 1.015 - 17.5
power.t.test(power = 0.8,
     delta = d,
     sd = 0.7,
     type = "one.sample",
     alt = "one.sided",
     sig.level = 0.05
     )

#     One-sample t test power calculation
             n = 45.34875
         delta = 0.2625
            sd = 0.7
     sig.level = 0.05
         power = 0.8
   alternative = one.sided
```

(b) How does the sample size change if the client is willing to accept a 3.0 % increase?

```
d <- 17.5 * 1.03 - 17.5
power.t.test(power = 0.8,
     delta = d,
     sd = 0.7,
     type = "one.sample",
     alt = "one.sided",
     sig.level = 0.05
     )

#   One-sample t test power calculation
             n = 12.46081
         delta = 0.525
            sd = 0.7
     sig.level = 0.05
         power = 0.8
   alternative = one.sided
```

(c) How does the sample size change if the client wants to detect a 0.5% increase?

```
d <- 17.5 * 1.005 - 17.5
power.t.test(power = 0.8,
       delta = d,
       sd = 0.7,
       type = "one.sample",
```

```
          alt = "one.sided",
          sig.level = 0.05
          )

    #   One-sample t test power calculation
                n = 397.0399
            delta = 0.0875
               sd = 0.7
        sig.level = 0.05
            power = 0.8
      alternative = one.sided
```

4.6 What simple random sample size would be needed to detect a 10 % decline with a power of 0.90? How would your answer change if the unit relvariance were 6?

```
capGain <- 44000
d <- capGain - 0.9*capGain
unitRv <- 3
sd1 <- sqrt(unitRv * capGain^2)
power.t.test(power = 0.9,
    delta = d,
    sd = sd1,
    type = "one.sample",
    alt = "one.sided",
    sig.level = 0.05
)

#     One-sample t test power calculation
              n = 2570.508
          delta = 4400
             sd = 76210.24
      sig.level = 0.05
          power = 0.9
      alternative = one.sided

d <- capGain - 0.9*capGain
unitRv <- 6
sd1 <- sqrt(unitRv * capGain^2)
power.t.test(power = 0.9,
    delta = d,
    sd = sd1,
    type = "one.sample",
    alt = "one.sided",
    sig.level = 0.05
    )

#     One-sample t test power calculation
              n = 5139.661
          delta = 4400
             sd = 107777.5
      sig.level = 0.05
          power = 0.9
      alternative = one.sided
```

4.8

(a) If the time 1 unemployment rate is anticipated to be 8% and you want to be able to detect a decline of 1.5% points with power 0.8 in a 1-sided, 0.05 level test, how large should the sample be at each time period? Assume that 0.08-0.015=0.065 will be unemployed at both times.

```
p1 <- 0.08
p2 <- p12 <- 0.065
nProp2sam(px=p1,
           py=p2, pxy=p12,
           g=0.75, r=1,
           sig.level = 0.05,
           alt="one.sided")
 #      Two-sample comparison of proportions
 Sample size calculation for overlapping samples
             n1 = 1228
             n2 = 1228
      px.py.pxy = 0.080, 0.065, 0.065
          gamma = 0.75
              r = 1
            alt = one.sided
      sig.level = 0.05
          power = 0.8
```

(b) If you can only afford to sample 500 persons, what will be the power to detect a 1.5% point change?

```
p1 <- 0.08
p2 <- p12 <- 0.065
Sxy <- p12 - p1*p2
Vd <- (p1*(1-p1) + p2*(1-p2) - 2*0.75*1*Sxy) / 500
Z <- 1.645 - (p1-p2)/sqrt(Vd)
1 - pnorm(Z)
 # [1] 0.4768264
```

4.10 The Council of Governments (COG) is an organization in the Washington DC area that is funded by local governments from the District of Columbia and surrounding counties. The COG would like to fund a survey to compare crime rates in the central city to that of one of the suburban counties.

```
cRate <- 1105/100000
 c1 <- 0.75 * cRate
 c2 <- 2*c1
 pow <- seq(0.5, 0.9, 0.05)
 samsize <- vector("numeric", length(pow))
 for (k in 1:length(pow))\{
     samsize[k] <- powerPropTest(n=NULL,
         p1 = c1,
         p2 = c2,
         alt = "one.sided",
         sig.level = 0.05,
```

```
        power = pow[k])\$n
}
out <- cbind(samsize = ceiling(samsize), power = pow)
out
          samsize power
  [1,]      968   0.50
  [2,]     1121   0.55
  [3,]     1288   0.60
  [4,]     1474   0.65
  [5,]     1682   0.70
  [6,]     1923   0.75
  [7,]     2210   0.80
  [8,]     2569   0.85
  [9,]     3060   0.90
```

Chapter 5

5.2 Using the data in Example 5.2 calculate (a) the proportional allocation, (b) the Neyman allocation for estimating total revenue, and (c) the cost-constrained allocation for revenue, assuming a budget of $300,000.

h	Cost constrained n_h	Prop. allocation	Neyman allocation
1	413	350	845
2	318	661	83
3	124	244	80
4	1,397	1,284	465
5	596	308	1,376
Total	2,848	2,848	2,848
Cost	$ 300,000	$ 276,211	$ 362,556
CV			
Revenue	0.047	0.061	0.033
Employees	0.024	0.031	0.016
Research credit	0.021	0.019	0.035
Offshore	0.030	0.031	0.050

Proportional allocation meets the budget constraint. Neyman does not. Proportional CV on revenue is worse than cost-constrained, but Neyman is better (0.033 vs. 0.047). Proportional CV on employees, research credit meets constraints. CV on offshore is close. Neyman meets CV constraint on employees

but not on research credit and offshore. Constraint of $n_h \geq 100$ is met by allocation proportional but violated by Neyman.

5.4 Resolve Example 5.2 with the same CV constraints as in Exercise 5.3 (0.05 on employees, 0.03 on total establishments claiming the research credit, 0.05 on total establishments with offshore affiliates) but revise the objective to be minimizing the total cost.

h	n_h		CV of t.hat
1	129	Revenue	0.09716
2	245	Employees	0.0500
3	108	Research	0.0300
4	511	Offshore	0.0500
5	129		
Total 1,122			

Chapter 9

9.2

(a) Compute the coefficient of variation that you would anticipate from a sample of 20 PSUs, 2 SSUs per PSU, and ten persons per sample SSU.
(b) Repeat the calculation of the coefficient of variation for a sample of 20 PSUs, 5 SSUs per PSU, and four persons per sample SSU.

```
#(a)
p <- 0.32; q <- 1-p
delta1 <- 0.003; delta2 <- 0.174
m <- 20; nbar <- 2; qDbar <- 10
V <- q/p
a <- V/(m*nbar*qDbar)
b <- delta1*nbar*qDbar
c <- 1 + delta2*(qDbar-1)
CV <- sqrt(a*(b+c))
CV
# [1] 0.1181128
#(b)
p <- 0.32; q <- 1-p
delta1 <- 0.003; delta2 <- 0.174
m <- 20; nbar <- 5; qDbar <- 4
V <- q/p
a <- V/(m*nbar*qDbar)
b <- delta1*nbar*qDbar
c <- 1 + delta2*(qDbar-1)
CV <- sqrt(a*(b+c))
```

```
CV
# [1] 0.09167538
```

9.4 Suppose that a two-stage sample is selected and the π-estimator of the total is used for a series of analysis variables. The average number of sample elements per cluster is 23. What are approximate estimates of the measure of homogeneity for design effects equal to 1.1, 1.2, 1.3, ..., 2.7, 2.8, 2.9, and 3.0? How do your answers change if $\bar{n} = 13$?

```
deltaCalc <- function(from,to,by,m){
   deff <- seq(from=1.1,to=3.0,by=0.1)
   nbar <- m
   delta <- (deff-1)/(nbar-1)
   cbind(deff,delta)
}
    #For nbar = 23

deltaCalc(1.1,3.0,0.1,23)
      deff      delta
 [1,]  1.1 0.004545455
 [2,]  1.2 0.009090909
 [3,]  1.3 0.013636364
 [4,]  1.4 0.018181818
 [5,]  1.5 0.022727273
 [6,]  1.6 0.027272727
 [7,]  1.7 0.031818182
 [8,]  1.8 0.036363636
 [9,]  1.9 0.040909091
[10,]  2.0 0.045454545
[11,]  2.1 0.050000000
[12,]  2.2 0.054545455
[13,]  2.3 0.059090909
[14,]  2.4 0.063636364
[15,]  2.5 0.068181818
[16,]  2.6 0.072727273
[17,]  2.7 0.077272727
[18,]  2.8 0.081818182
[19,]  2.9 0.086363636
[20,]  3.0 0.090909091

#For nbar = 13
 deltaCalc(1.1,3.0,0.1,13)
      deff      delta
 [1,]  1.1 0.008333333
 [2,]  1.2 0.016666667
 [3,]  1.3 0.025000000
 [4,]  1.4 0.033333333
 [5,]  1.5 0.041666667
 [6,]  1.6 0.050000000
 [7,]  1.7 0.058333333
 [8,]  1.8 0.066666667
 [9,]  1.9 0.075000000
[10,]  2.0 0.083333333
```

```
[11,]  2.1 0.091666667
[12,]  2.2 0.100000000
[13,]  2.3 0.108333333
[14,]  2.4 0.116666667
[15,]  2.5 0.125000000
[16,]  2.6 0.133333333
[17,]  2.7 0.141666667
[18,]  2.8 0.150000000
[19,]  2.9 0.158333333
[20,]  3.0 0.166666667
```

9.6 Repeat the calculations in Example 9.11 for two-stage sampling using block groups as PSUs in the Maryland population. Use set.seed(-780087528) in R. Select 20 BGs with probabilities proportional to number of persons per tract and 50 persons per BG using *srswor*. Compare your results to those in Example 9.9 where tracts were used as PSUs.

```
attach("C:\\Data\\MDarea.pop.RData", pos=2)
 trtBG <- 10*MDarea.pop$TRACT + MDarea.pop$BLKGROUP
 MDpop <- cbind(MDarea.pop, trtBG)
 require(sampling)
 require(reshape)        # has function that allows renaming
                         # variables
 Ni <- table(MDpop$trtBG)
 m <- 20
 probi <- m*Ni / sum(Ni)
    # select sample of clusters
set.seed(-780087528)
sam <- cluster(data=MDpop, clustername="trtBG", size=m,
               method="systematic", pik=probi, description=TRUE)
    # extract data for the sample clusters
samclus <- getdata(MDarea.pop, sam)
samclus <- rename(samclus, c(Prob = "pi1"))
table(samclus$trtBG)
    # treat sample clusters as strata and select srswor from each
s <- strata(data = as.data.frame(samclus), stratanames = "TRACT",
            size = rep(50,m), method="srswor")
    # extracts the observed data
samdat <- getdata(samclus,s)
samdat <- rename(samdat, c(Prob = "pi2"))
table(samdat$trtBG)
    # extract pop counts for PSUs in sample
pick <- names(Ni) %in% sort(unique(samdat$trtBG))
Ni.sam <- Ni[pick]
pp <- Ni.sam / sum(Ni)
wt <- 1/samdat$pi1/samdat$pi2

BW <- rbind(BW2stagePPSe(Ni = Ni.sam, ni = rep(50,20),
                         X = samdat$y1,
                         psuID = samdat$TRACT, w = wt,
                         m = 20, pp = pp),
            BW2stagePPSe(Ni = Ni.sam, ni = rep(50,20),
                         X = samdat$y2,
                         psuID = samdat$TRACT, w = wt,
```

```
                                    m = 20, pp = pp),
              BW2stagePPSe(Ni = Ni.sam, ni = rep(50,20),
                           X = samdat$y3,
                           psuID = samdat$TRACT, w = wt,
                           m = 20, pp = pp),
              BW2stagePPSe(Ni = Ni.sam, ni = rep(50,20),
                           X = samdat$ins.cov,
                           psuID = samdat$TRACT, w = wt,
                           m = 20, pp = pp),
              BW2stagePPSe(Ni = Ni.sam, ni = rep(50,20),
                           X = samdat$hosp.stay,
                           psuID = samdat$TRACT, w = wt,
                           m = 20, pp = pp)
             )
 round(BW,4)
#              Vpsu           Vssu       B        W   delta
#[1,] 1.369864e+12 1.051959e+12 0.0347   1.3761  0.0246
#[2,] 1.068294e+10 9.071762e+09 0.0226   0.9935  0.0222
#[3,] 7.540980e+11 1.089802e+11 0.0118   0.0884  0.1177
#[4,] 4.256465e+07 2.575943e+07 0.0084   0.2651  0.0309
#[5,] 6.128648e+06 1.045993e+07 0.1449  12.7945  0.0112
```

The results from Example 9.11 are below. When BGs are used as clusters,
the measures of homogeneity are larger.

	Tracts as clusters			BGs as clusters
	B^2	W^2	δ	δ
y1	0.0418	1.3934	0.0291	0.0246
y2	0.0208	1.0416	0.0196	0.0222
y3	0.0101	0.1028	0.0894	0.1177
ins.cov	0.0007	0.3051	0.0023	0.0309
hosp.stay	0.1056	13.9161	0.0075	0.0112

9.8 Use the Maryland population and the function BW3stagePPSe to
compute variance components from a sample of 30 PSUs (tracts), 2 SSUs
(block groups) per tract, and 50 persons per sample SSU. Assume that
tracts are selected with probabilities proportional to the number of per-
sons in the tract and that SSUs and persons are selected via *srs.* Use
set.seed(1696803792) in R. (a) Do the computation for the variables
y2, y3, ins.cov, and hosp.stay. (b) How do your answers compare to the full
population results in Example 9.12? (c) Use the estimated values of δ_1 and
δ_2 to compute the optimum values of m, $\bar{n}$, and $\bar{\bar{q}}$ in a three-stage where
$C_1 = 500$, $C_2 = 100$, $C_3 = 120$, and the total budget for variable costs is
$100,000. How can you estimate the unit relvariance for each variable? (d)
Discuss your results in (c).

```
#(a) Do the computation for the variables y2, y3, ins.cov, and
```

```
    hosp.stay.
    # select 3-stage sample from Maryland population

attach("C:\\Data\\MDarea.pop.RData", pos=2)
trtBG <- 10*MDarea.pop$TRACT + MDarea.pop$BLKGROUP
MDpop <- cbind(MDarea.pop, trtBG)
require(sampling)
require(reshape)       # has function that allows renaming variables
    # make counts of SSUs and elements per PSU
xx <- do.call("rbind",list(by(1:nrow(MDpop),MDpop$trtBG,head,1)))
pop.tmp <- MDpop[xx,]
Ni <- table(pop.tmp$TRACT)
Qi <- table(MDarea.pop$TRACT)
Qij <- table(MDpop$trtBG)
m <- 30          # no. of PSUs to select
probi <- m*Qi / sum(Qi)


#---------------------------------------------------------------------------------------
    # select sample of clusters
 set.seed(1696803792)
sam <- cluster(data=MDpop, clustername="TRACT", size=m,
      method="systematic",
                pik=probi, description=TRUE)
    # extract data for the sample clusters
samclus <- getdata(MDpop, sam)
samclus <- rename(samclus, c(Prob = "p1i"))
samclus <- samclus[order(samclus$TRACT, samclus$BLKGROUP),]


#---------------------------------------------------------------------------------------
    # treat sample clusters as strata and select srswor of block
      groups from each
    # identify psu IDs for 1st instance of each ssuID
xx <- do.call("rbind",list(by(1:nrow(samclus),samclus$trtBG,head,1)))
SSUs <- cbind(TRACT=samclus$TRACT[xx], trtBG=samclus$trtBG[xx],
      BG=samclus$BLKGROUP[xx])
    # select 2 BGs per tract
n <- 2
s <- strata(data = as.data.frame(SSUs), stratanames = "TRACT",
            size = rep(n,m), method="srswor")
s <- rename(s, c(Prob = "p2i"))
    # extract the BG data
    # s contains selection probs of SSUs, need to get those onto
      data file
SSUsam <- SSUs[s$ID\_unit, ]
SSUsam <- cbind(s, SSUsam[, 2:3])
    # identify rows in PSU sample that correspond to sample SSUs
tmp <- samclus$trtBG %in% SSUsam$trtBG
SSUdat <- samclus[tmp,]
SSUdat <- merge(SSUdat, SSUsam[, c("p2i","trtBG")], by="trtBG")
rm(tmp)


#---------------------------------------------------------------------------------------
    # select srswor from each sample BG
n.BG <- m*n
s <- strata(data = as.data.frame(SSUdat), stratanames = "trtBG",
            size = rep(50,n.BG), method="srswor")
s <- rename(s, c(Prob = "p3i"))
samclus <- getdata(SSUdat, s)
del <- (1:ncol(samclus))[dimnames(samclus)[[2]] \%in\%
        c("ID\_unit","Stratum")]
```

```
samclus <- samclus[, -del]

#---------------------------------------------------------------------------------------
    # extract pop counts for PSUs in sample
pick <- names(Qi) \%in\% sort(unique(samclus$TRACT))
Qi.sam <- Qi[pick]
    # extract pop counts of SSUs for PSUs in sample
pick <- names(Ni) \%in\% sort(unique(samclus$TRACT))
Ni.sam <- Ni[pick]
    # extract pop counts for SSUs in sample
pick <- names(Qij) \%in\% sort(unique(samclus$trtBG))
Qij.sam <- Qij[pick]

    # compute full sample weight and wts for PSUs and SSUs
wt <- 1 / samclus$p1i / samclus$p2i / samclus$p3i
w1i <- 1 / samclus$p1i
w2ij <- 1 / samclus$p1i / samclus$p2i
samdat <- data.frame(psuID = samclus$TRACT, ssuID = samclus$trtBG,
                     w1i = w1i, w2ij = w2ij, w = wt,
                     samclus[, c("y1","y2","y3","ins.cov",
                     "hosp.stay")])

#--------------------------------------------------------------
    # call fcn to compute variance component estimates
wtdvar <- function(x, w){
            xbarw <- sum(w*x) / sum(w)
            varw <- sum(w * (x-xbarw)^2) / sum(w)
            varw
}

BW3 <-
 rbind(BW3stagePPSe(dat=samdat, v="y1", Ni=Ni.sam, Qi=Qi.sam,
   Qij=Qij.sam, m),
   BW3stagePPSe(dat=samdat, v="y2", Ni=Ni.sam, Qi=Qi.sam,
   Qij=Qij.sam, m),
   BW3stagePPSe(dat=samdat, v="y3", Ni=Ni.sam, Qi=Qi.sam,
   Qij=Qij.sam, m),
   BW3stagePPSe(dat=samdat, v="ins.cov", Ni=Ni.sam, Qi=Qi.sam,
   Qij=Qij.sam, m),
   BW3stagePPSe(dat=samdat, v="hosp.stay", Ni=Ni.sam, Qi=Qi.sam,
   Qij=Qij.sam, m)
   )
 round(BW3,4)
           Vpsu         Vssu        Vtsu      B      W     W2      W3 delta1 delta2
[1,] 5.634050e+11 1.877326e+12 474697662057 0.0178 1.2553 0.3554 1.5477 0.0138 0.1863
[2,] 7.820266e+09 1.512943e+10   3875989824 0.0221 0.9307 0.2822 1.1299 0.0230 0.1994
[3,] 1.809921e+12 1.752007e+12  44878287984 0.0402 0.0909 0.2458 0.1027 0.3065 0.7052
[4,] 8.769247e+07 1.420352e+08     10750509 0.0250 0.2612 0.2318 0.3162 0.0873 0.4227
[5,] 5.905489e+05 8.930217e+05      4235957 0.0203 12.4614 0.2102 14.9988 0.0014 0.0132
```

(c) Use the estimated values of delta1 and delta2 to compute the
   optimum values of mbar, nbar, and qbarbar in a three-stage where C1=500,
   C2=100, C3=120, and the total budget for variable costs is \$100,000.

```
    # estimate the unit relvariance for each variable
wtdrelvar <- function(x, w){
                xbarw <- sum(w*x) / sum(w)
                varw <- sum(w * (x-xbarw)^2) / sum(w)
                c(mean = xbarw, relvar = varw/xbarw^2)
}
```

```
 rv.y1 <- wtdrelvar(samdat[,"y1"], wt)
 rv.y2 <- wtdrelvar(samdat[,"y2"], wt)
 rv.y3 <- wtdrelvar(samdat[,"y3"], wt)
 rv.inscov <- wtdrelvar(samdat[,"ins.cov"], wt)
 rv.hosp <- wtdrelvar(samdat[,"hosp.stay"], wt)

 round(
     rbind(y1        = rv.y1,
           y2        = rv.y2,
           y3        = rv.y3,
           inscov    = rv.inscov,
           hosp.stay = rv.hosp), 4)

clusOpt3(unit.cost=c(500, 100, 120),
         delta1=0.0138, delta2=0.1863,
         unit.rv=rv.y1[2],
         tot.cost=100000,
         cal.sw=1)

 C1 = 500
 C2 = 100
 C3 = 120
 delta1 = 0.0138
 delta2 = 0.1863
 unit relvar = 1.344438
 budget = 1e+05
 cost check = 1e+05
 m.opt = 31.2
 n.opt = 8.2
 q.opt = 1.9
 CV = 0.0617

clusOpt3(unit.cost=c(500, 100, 120),
         delta1=0.0230, delta2=0.1994,
         unit.rv=rv.y2[2],
         tot.cost=100000,
         cal.sw=1)
 C1 = 500
 C2 = 100
 C3 = 120
 delta1 = 0.023
 delta2 = 0.1994
 unit relvar = 1.002332
 budget = 1e+05
 cost check = 1e+05
 m.opt = 38.4
 n.opt = 6.6
 q.opt = 1.8
 CV = 0.0559

clusOpt3(unit.cost=c(500, 100, 120),
         delta1=0.3065, delta2=0.7052,
         unit.rv=rv.y3[2],
         tot.cost=100000,
         cal.sw=1)
 C1 = 500
 C2 = 100
 C3 = 120
 delta1 = 0.3065
 delta2 = 0.7052
```

```
 unit relvar = 0.1266218
 budget = 1e+05
 cost check = 1e+05
 m.opt = 92.6
 n.opt = 3.4
 q.opt = 0.6
 CV = 0.0301

clusOpt3(unit.cost=c(500, 100, 120),
         delta1=0.0873, delta2=0.4227,
         unit.rv=rv.inscov[2],
         tot.cost=100000,
         cal.sw=1)
 C1 = 500
 C2 = 100
 C3 = 120
 delta1 = 0.0873
 delta2 = 0.4227
 unit relvar = 0.2722481
 budget = 1e+05
 cost check = 1e+05
 m.opt = 61.7
 n.opt = 4.9
 q.opt = 1.1
 CV = 0.0354

clusOpt3(unit.cost=c(500, 100, 120),
         delta1=0.0014, delta2=0.0132,
         unit.rv=rv.hosp[2],
         tot.cost=100000,
         cal.sw=1)
 C1 = 500
 C2 = 100
 C3 = 120
 delta1 = 0.0014
 delta2 = 0.0132
 unit relvar = 12.95215
 budget = 1e+05
 cost check = 1e+05
 m.opt = 13
 n.opt = 6.9
 q.opt = 7.9
 CV = 0.1464
```

9.8 *Discussion. The unit relvariance can be estimated as*

$$RV = \left(\sum_s w_k \right)^{-1} \sum_s w_k \left(y_k - \bar{y}_w \right)^2 \left/ \left[\sum_s w_k y_k \left/ \sum_s w_k \right. \right]^2 \right. .$$

The allocations are different for the five variables because the delta1 and delta2 are quite different. Delta2 is relatively high for y3 and ins.cov. This leads to the optimal $\bar{\bar{q}}$ about 1 for both those variables. For hosp.stay, delta2=0.0132 and $\bar{\bar{q}}=8$. For y1 and y2, $\bar{\bar{q}}=2$. The m_{opt}'s are also quite different. Some sort of compromise is needed since the same allocation will not be optimal for

every variable. If all the variables were equally important, we could average the allocations and use m=45, n̄=6, q̄̄=2, giving a total cost of about $114,000. However, the relative importance to the survey of the five variables would have to be considered, along with (as always) the budget.

Chapter 10

10.2

(a) Total expected sample sizes for the two domains. *Domain 1: 16; domain 2: 28.*
(b) Composite measure of size for each PSU and the total across PSUs. Verify that the grand total equals the total expected sample size.
(c) Selection probability for each PSU.
(d) Domain sampling rate and expected domain sample size within each PSU. Are the expected sample sizes integers? If not, what method can be used for sampling within a PSU that will achieve the desired rate?

	(b)	(c)	(d) Dom. rate	Sampling	Dom. Sample size	
PSU	PSU MOS	PSU prob	1	2	1	2
1	7.5	0.34091	0.14667	0.29333	7.3	14.7
2	11	0.50000	0.10000	0.20000	2.0	20.0
3	10.5	0.47727	0.10476	0.20952	9.4	12.6
4	15	0.68182	0.07333	0.14667	11.7	10.3

(e) Verify that the expected sample sizes for any two of the PSU's sum to the total expected sample size you computed in (a).

Sums of sample sizes for any 2 PSUs	
(1,2)	44.0
(1,3)	44.0
1,4)	44.0
(2,3)	44.0
(2,4)	44.0
(3,4)	44.0

10.4 The two PSUs below are an existing PSU sample selected some years ago. A new survey is to be done in these PSUs.

(a) Compute the expected sample sizes in each domain in each SSU and the total sample size in each SSU across the domains. Assume that rates of 0.03 and 0.01 are used for domains 1 and 2. Note that the population totals for the domains are 5,000 and 2,200 as shown in the table above. *Domain 1: 75; domain 2: 11.*

(b) Compute the composite MOS for each SSU using the method in Sect. 10.5

(c) Compute the SSU selection probabilities assuming that the SSU sample will be selected with probabilities proportional to the composite MOS.

(d) Calculate the within-SSU probabilities required for the sample in each domain to be self-weighting.

(e) Compute the expected workload in each SSU if it were to be sampled. Are these equal? If not, explain why.

(f) Verify that the SSU and within-SSU probabilities computed in (c) and (d) do yield a self-weighting sampling in each domain.

(1) Obtain self-weighting sample but workload is not constant

PSU	PSU prob Pi(i)	SSU	Nij(d) Domain	d=1	d=2 Total count	N(ij,+)	(b) Composite MOS=(sum[f(d)* Nij(d)])/P(i) (D*B6 + E*B7)/B C10*G/G(+) S(ij)	(c) Seg prob P(i)	(d) Within seg probs Pk|ij(d) = f(d)/P(i)/P(j|i) Domain B6/B/H d=1	B7/B/H d=2	(e) E(n) within PSU d=1	d=2	Total	(f) Check self-weighing within domains Overall sel probs B*H*I	B*H*J
1	0.126389	1	1	40	80	120	15.8	0.462	0.513	0.171	20.5	13.7	34.2	0.030	0.010
1	0.126389	2	2	25	45	70	9.5	0.277	0.855	0.285	21.4	12.8	34.2	0.030	0.010
1	0.126389	3	3	35	90	125	15.4	0.451	0.526	0.175	18.4	15.8	34.2	0.030	0.010
1	0.126389	4	4	105	35	140	27.7	0.809	0.293	0.098	30.8	3.4	34.2	0.030	0.010
PSU total						455	68.4								
2	0.280556	1	1	80	180	260	15.0	0.500	0.214	0.071	17.1	12.8	29.9	0.030	0.010
2	0.280556	2	2	40	200	240	11.4	0.381	0.281	0.094	11.2	18.7	29.9	0.030	0.010
2	0.280556	3	3	20	85	105	5.2	0.173	0.619	0.206	12.4	17.6	29.9	0.030	0.010
2	0.280556	4	4	85	150	235	14.4	0.482	0.222	0.074	18.9	11.1	29.9	0.030	0.010
2	0.280556	5	5	110	60	170	13.9	0.464	0.230	0.077	25.3	4.6	29.9	0.030	0.010
PSU total						1010	59.9								
Pop Totals (includes all PSUs)				5000	2200	7200									

(g) Determine a sampling scheme for SSUs and units within SSUs that will give an equal workload in each SSU. Carry out the calculations for SSU and within-SSU selection probabilities and verify that the total expected sample size across the two domains is the same in every SSU.

(h) Does the scheme you designed in (g) lead to a self-weighting sample? Why or why not? Support your answer with calculations.

(2) Obtain a constant workload in each PSU but sample is not self-weighting

(g) PSU	PSU prob Pi(i)	Seg	Nij(d) Domain	d=1	d=2 Total count	N(ij,+)	Composite MOS=(sum[f(d)* Nij(d)])/P(i) (D*B6 + E*B7)/B C10*G/G(+) S(ij)	Seg prob P(i)	Within seg probs Pk|ij(d) n.bar*f(d)/P(i)/S'(ij) Domain F9*B6/B/G d=1	F9*B7/B/G d=2	Equal workloads E(n) within PSU d=1	d=2	Total	(h) Check self-weighing within domains Overall sel probs B*H*I	B*H*J
1	0.126389	1	1	40	80	120	15.8	0.462	0.323	0.108	12.9	8.6	21.5	0.019	0.006
1	0.126389	2	2	25	45	70	9.5	0.277	0.538	0.179	13.4	8.1	21.5	0.019	0.006
1	0.126389	3	3	35	90	125	15.4	0.451	0.331	0.110	11.6	9.9	21.5	0.019	0.006
1	0.126389	4	4	105	35	140	27.7	0.809	0.184	0.061	19.4	2.2	21.5	0.019	0.006
PSU total						455	68.4								
2	0.280556	1	1	80	180	260	15.0	0.500	0.154	0.051	12.3	9.2	21.5	0.022	0.007
2	0.280556	2	2	40	200	240	11.4	0.381	0.202	0.067	8.1	13.4	21.5	0.022	0.007
2	0.280556	3	3	20	85	105	5.2	0.173	0.445	0.148	8.9	12.6	21.5	0.022	0.007
2	0.280556	4	4	85	150	235	14.4	0.482	0.159	0.053	13.5	8.0	21.5	0.022	0.007
2	0.280556	5	5	110	60	170	13.9	0.464	0.165	0.055	18.2	3.3	21.5	0.022	0.007
PSU total						1010	59.9								
Pop Totals (includes all PSUs)				5000	2200	7200									

Chapter 13

13.2 Find the following:

(a) Selection probabilities for the three sample PSUs
(b) Within-PSU sampling rates needed to achieve the desired overall sampling rates
(c) Base weights for each unit
(d) Expected number of sample persons in each PSU by race/ethnic group and in total

				(a)	(b) Within-PSU rates		(c) Unit weights		(d) Expected no. in sample		
		Non-His-panic		PSU	Non-His-panic		Non-His-panic		Non-His-panic		Total PSU
PSU	M_i	White	Other	prob π_i	White	Other	White	Other	White	Other	sample
1	1,000	800	200	0.3000	0.0333	0.1333	100	25	26.7	26.7	53.3
2	850	400	450	0.2550	0.0392	0.1569	100	25	15.7	70.6	86.3
3	150	110	40	0.0450	0.2222	0.8889	100	25	24.4	35.6	60.0
Population total M_+	10,000										

13.4 The following table gives sums of weights for samples of establishments in three cities that were classified as being in retail trade based on yellow page listings

1. Adjust the weights separately in each city first for unknown eligibility and then for nonresponse. Show your calculations in each step.
2. What is the estimated total number of eligible units in each city and across all cities?
3. What is the estimated number of ineligible establishments on the sampling frame?
4. In what circumstance would it be reasonable to combine all three cities together to make the adjustments for unknown eligibility and nonresponse? Do those circumstances hold here?

	a(1)											a(2)			
	Sum of weights						Weights adjusted for unknown eligibilty					Weights adjusted for nonresponse			
	Eligible		Known ineligible	Unknown eligibility	Total	$\frac{\sum_{KN} w_k}{\sum w_k}$	Eligible		Known ineligible	Total	$\frac{\sum_{ER} w_k}{\sum_E w_k}$	Eligible		Known ineligible	Total
	R	NR					R	NR				R	NR		
City															
1	50	46	11	17	124	0.8629	57.9	53.3	12.7	124	0.5208	111.3	0	12.7	124
2	77	89	19	12	197	0.9391	82.0	94.8	20.2	197	0.4639	176.8	0	20.2	197
3	44	31	8	23	106	0.7830	56.2	39.6	10.2	106	0.5867	95.8	0	10.2	106
Total	171	166	38	52	427		196.1	187.7	43.2	427		383.8		43.2	427

(d) In what circumstance would it be reasonable to combine all three cities together to make the adjustments for unknown eligibility and nonresponse? Do those circumstances hold here? *If the cities all had the same rates of known eligibility and response, they could be combined. This is not true here because the known rates are 0.86, 0.94, and 0.78; the RRs are 0.52, 0.46, and 0.59.*

13.6 (a) <u>Logistic</u>

```
glm(formula = resp ~ age + as.factor(sex) + as.factor(hisp) +
     as.factor(race), family=binomial(link="logit"), data=nhis)
 Coefficients:
                  Estimate Std. Error z value Pr(>z)
 (Intercept)      1.014972   0.114182   8.889 $<$ 2e-16 ***
 age             -0.009686   0.002039  -4.749 2.04e-06 ***
 as.factor(sex)2 -0.077060   0.069995  -1.101   0.2709
 as.factor(hisp)2 0.404795   0.088047   4.598 4.28e-06 ***
 as.factor(race)2 -0.212043  0.098400  -2.155   0.0312 *
 as.factor(race)3 -0.352277  0.160388  -2.196   0.0281 *
 ---
 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
 AIC: 4813.5

 L.hat <- glm.logit$linear.predictors
    # transform link values to probability scale
 pred.logit <- exp(L.hat) / (1 + exp(L.hat) )
```

<u>Probit</u>

```
 glm(formula = resp ~ age + as.factor(sex) + as.factor(hisp) +
     as.factor(race), family=binomial(link="probit"), data=nhis)
 Coefficients:
                  Estimate Std. Error z value Pr(>z)
 (Intercept)      0.622283   0.069330   8.976 $<$ 2e-16 ***
 age             -0.005824   0.001237  -4.710 2.48e-06 ***
 as.factor(sex)2 -0.046346   0.042339  -1.095   0.2737
 as.factor(hisp)2 0.245814   0.053728   4.575 4.76e-06 ***
 as.factor(race)2 -0.128363  0.059762  -2.148   0.0317 *
 as.factor(race)3 -0.216234  0.098442  -2.197   0.0281 *
 AIC: 4813.6

 L.hat <- glm.probit$linear.predictors
 pred.probit <- pnorm(L.hat)
```

<u>cloglog</u>

```
 glm(formula = resp ~ age + as.factor(sex) + as.factor(hisp) +
     as.factor(race), family=binomial(link="cloglog"), data=nhis)
 Coefficients:
                  Estimate Std. Error z value Pr(>z)
 (Intercept)      0.271632   0.068058   3.991 6.57e-05 ***
 age             -0.005551   0.001211  -4.583 4.59e-06 ***
 as.factor(sex)2 -0.044086   0.041044  -1.074   0.2828
 as.factor(hisp)2 0.240590   0.053616   4.487 7.21e-06 ***
 as.factor(race)2 -0.124046  0.058554  -2.118   0.0341 *
 as.factor(race)3 -0.219619  0.099917  -2.198   0.0279 *
 AIC: 4814

 L.hat <- glm.cloglog$linear.predictors
 pred.cloglog <- 1- exp(-exp(L.hat) )
```

(b) Which variables are significant? *The same variables are significant in all models: intercept, age, hisp, and race.*

13.9 Using the NHIS data set, fit a classification tree for the response (resp) variable using the covariates age, sex, hisp, race, parents, and educ. Require that a minimum of 50 cases be assigned to each node. Describe the composition of each node in words and draw a picture of the tree. Compute the unweighted response rates in each of the nodes that are formed.

```
require(rpart)
set.seed(15097)
nhis <- data.frame(nhis)
t1 <- rpart(resp ~age + sex + hisp + race + parents + educ,
      method = "class",
      control = rpart.control(minbucket = 50, cp=0),
      data = nhis)
print(t1, digits=2)

par(mfrow=c(1,1))
plot(t1, uniform=TRUE, compress=TRUE, margin = 0.1)
text(t1, use.n=TRUE, all=TRUE,
     digits=4,
     cex=1,
     pretty=1,
     fancy=TRUE,
     xpd = TRUE,
     font = 3)
title("Tree for identifying nonresponse adjustment cells
          in the NHIS data set")

n= 3911
node), split, n, loss, yval, (yprob)
      * denotes terminal node
 1) root 3911 1200 1 (0.31 0.69)
   2) educ$<$ 4 1964   690 1 (0.35 0.65)
     4) age$>$=56 588   240 1 (0.41 0.59)
       8) hisp$<$ 1.5 88    44 0 (0.50 0.50) *
       9) hisp$>$=1.5 500   200 1 (0.39 0.61)
        18) age$>$=82 73    36 1 (0.49 0.51) *
        19) age$<$ 82 427   160 1 (0.37 0.63)
           38) educ$>$$=2 261   100 1 (0.39 0.61)
              76) age$<$ 60 52    25 0 (0.52 0.48) *
              77) age$>$=60 209    76 1 (0.36 0.64) *
           39) educ$<$ 2 166    56 1 (0.34 0.66) *
     5) age$<$ 56 1376   450 1 (0.33 0.67)
      10) parents$<$ 3.5 277   110 1 (0.40 0.60)
        20) age$>$=32 67    31 0 (0.54 0.46) *
        21) age$<$ 32 210    75 1 (0.36 0.64) *
      11) parents$>$=3.5 1099   340 1 (0.31 0.69) *
   3) educ$>$$=4 1947   520 1 (0.27 0.73) *
```

Tree for identifying nonresponse adjustment cells in the NHIS data set
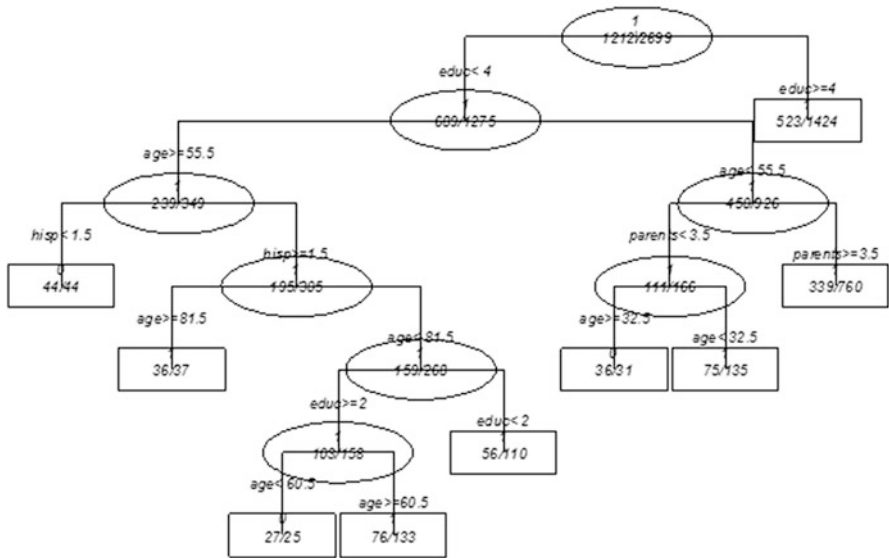
Chapter 14

14.1 Use the `smho.N874` data set to complete this exercise on poststratification.

(a) What are the means of expenditures in the five hospital types in the population? What should you look for in order for poststratification to be worth considering?

(b) Compute the population counts of facilities by hospital type, treating the smho98 data set as the full population. Compute the unweighted sample counts by hospital type to verify that each type is represented in the sample. If one of the hospital types was not represented in the sample, what would be the practical and theoretical implications? Discuss this in the context of design-based and model-based inference.

(c) Calculate the set of poststratified weights for the sample using hospital type as the poststratification variable. What do the weights sum to before and after poststratification? Is this what you expect?

(d) Verify that the calibration controls are met by the set of poststratified weights.

(e) Estimate the population total of expenditures and its standard error for the expansion estimator under the *srswor* design and for the poststratified estimator. Be sure and incorporate a finite population correction factor into the variance estimates. Discuss any similarities or differences in the estimated totals and SEs.

```
attach("C:\\Data\\smho98sub.RData", pos=2)
require(sampling)
require(doBy)
set.seed(-530049348)
smho <- smho98sub
    # (a) population means of expenditures by hospital type
 summaryBy(EXPTOTAL ~ hosp.type, data = smho98sub, fun = mean)
   hosp.type EXPTOTAL.mean
 1         1      21240408
 2         2      10852136
 3         3       4913008
 4         4       6118415
 5         5      12041188
```

*Poststrata will be effective if the PS have different means, which they do in
this case.*

```
    #   Select an srswor and poststratify
 n <- 80
 N <- nrow(smho)
    # select  srswor of size n
 sam <- sample(1:N, n)
 samdat <- smho[sam, ]
```

```
# (b) Population and sample counts by hospital type
 table(smho[, "hosp.type"])
   1   2   3   4   5
 215 115 252 149 143
 table(samdat[, "hosp.type"])
  1  2  3  4  5
 17 13 23 15 12
```

*If one of the poststrata was not represented in the population, the practi-
cal implication is that it would have to be collapsed with one of the other
poststrata in order to compute an estimate. This creates a kind of adaptive
procedure for which the design-based theory for the poststratified estimator
does not apply. The usual assumption there is that every poststratum is in
the sample. If a different model applies in each poststratum, e.g., each PS
has a different mean, then the PS estimator is not model unbiased for that
configuration of sample units.*

```
# (c) poststratified weights, srs weights
 d <- rep(N/n, n)
 f1 <- rep(n/N, n)
 N.hosp <- table(smho[, "hosp.type"])
 require(survey)
 smho.dsgn <- svydesign(ids = ~0,     # no clusters
          strata = NULL,    # no strata
          fpc = ~f1,
          data = data.frame(samdat),
```

```
            weights = ~d)

    # form vector of pop totals and and poststratify
 pop.tots <- c('(Intercept)'= 874, Ng = N.hosp[-1])
 ps.calib <- calibrate(design = smho.dsgn,
     formula = ~ as.factor(hosp.type),
     population = pop.tots,
     bounds = c(-Inf,Inf),
     calfun = c("linear"),
     )
    # sum of weights before and after PS
 sum(weights(smho.dsgn))
 [1] 874
 sum(weights(ps.calib))
 [1] 874
```

Both sets of weights sum to the population size N=874 as they should.

```
# (d) Verify that calibration controls are met
 svytotal(~ as.factor(hosp.type), ps.calib)
                        total        SE
 as.factor(hosp.type)1   215 1.829e-14
 as.factor(hosp.type)2   115 2.974e-15
 as.factor(hosp.type)3   252 9.790e-15
 as.factor(hosp.type)4   149 3.897e-15
 as.factor(hosp.type)5   143 6.476e-15

# (e) Estimate population total of expenditures and SEs
    # PS standard error and cv
 svytotal(~ EXPTOTAL, ps.calib)
            total         SE
 EXPTOTAL 9406934020 1323048236
 cv(svytotal(~ EXPTOTAL, ps.calib))
  EXPTOTAL
 0.1406461
    # srs standard error and cv
 svytotal(~ EXPTOTAL, smho.dsgn)
              total         SE
 EXPTOTAL 9085181570 1363966973
 cv(svytotal(~ EXPTOTAL, smho.dsgn))
  EXPTOTAL
 0.1501310
```

The actual total in the population is 9,686,295,207, so the PS estimate is closer to the population total in this particular sample. The SE of the PS estimator is slightly lower, but poststratifying has not improved the precision of the estimated total much. Of course, one sample does not tell us anything about the long-run performance.

14.6 Using the random seed value of 15097 in R, select a sample of $n=50$ hospitals from the data file Hospital_pop.txt with probabilities proportional to the square root of the number of BEDS, i.e., *pps* $\left(x^{1/2}\right)$.

```
   # load sampling package
require(sampling)
   # attach Hospitals pop
attach("C:\\Data\\hospital.RData")
   #Random seed for sample selection
set.seed(15097)
   # Calculate 1-draw selection probabilities - pps
mos <- sqrt(hospital$x)
   #Calculate 1-draw selection probabilities
hospital$prbs.1d <- mos / sum(mos)
summary(hospital$prbs.1d)
     Min.   1st Qu.   Median      Mean   3rd Qu.      Max.
0.0005277 0.0016850 0.0025470 0.0025450 0.0033080 0.0052400

   # Select sample - pps
   #Define size of sample
n <- 50
   # probabilities for selecting a sample of n
pk <- n * hospital$prbs.1d
   #PPS sample
sam <- UPrandomsystematic(pk)
sam <- sam==1
sam.dat <- hospital[sam, ]
   #Design weights
dsgn.wts <- 1/pk[sam]
sum(dsgn.wts)
[1] 393.8783
summary(dsgn.wts)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  3.955   5.383   6.837   7.878   9.746  27.490
```

(a) Calculate the estimated design effects using Spencer's formula and Kish's approximation.

```
     #Calculate WLS values
   sam.wls <- lm(y ~ prbs.1d, data=sam.dat, weights=dsgn.wts)
     #DEFF component - var of y
   sam.mean.y <- sum(sam.dat$y * dsgn.wts) / sum(dsgn.wts)
   sam.mean.y
   [1] 805.9594
   sam.var.y  <- sum(dsgn.wts * (sam.dat$y - sam.mean.y)^2) /
                     sum(dsgn.wts)
   sam.var.y
   [1] 263510.3
     #DEFF component - alpha squared
   sam.alpha2 <- sam.wls$coefficients[1] ^2
   sam.alpha2
   (Intercept)
      141821.9
     #DEFF component - squared correlation
   sam.rho2.yP <- summary(sam.wls)$r.squared
   sam.rho2.yP
   [1] 0.8261859
     #DEFF component - Kish
```

```
    kish.deff <- n*sum(dsgn.wts^2) / (sum(dsgn.wts)^2)
    kish.deff
    [1] 1.231421
        #Spencer's DEFF
    spencers.deff <- as.numeric((1 - sam.rho2.yP)*kish.deff +
                      (sam.alpha2/sam.var.y)*(kish.deff-1))
    spencers.deff
    [1] 0.3385895
```

(b) Describe the estimators of the population total to which the Kish and
    Spencer *deff*'s refer. Why do the computed values differ? Which do you
    think is the most relevant here? Why?
    The Kish deff is 1.23; Spencer's is 0.34. If the goal is to estimate the total
    of discharges (y), then Spencer's is more appropriate.
(c) Estimate the total of discharges (y) in the population using the π-
    estimator along with its SE and *CV*. How does this compare to the
    estimate of the variance of the total from a simple random sample of
    n=50. Estimate the *srswor* variance from the sample of 50 selected for
    this problem.

```
 h.dsgn <- svydesign(ids = ~0,
              strata = NULL,
              data = data.frame(sam.dat),
              weights = ~dsgn.wts)
 svytotal(~y, h.dsgn)
    total    SE
 y 317450 14682
 cv(svytotal(~y, h.dsgn))
          y
 0.04624832
     # estimate the SE if an srswor had been selected
 w <- dsgn.wts
 y <- sam.dat$y
 wm <- weighted.mean(x=y, w=w)
 sig2 <- (n/(n-1) * sum(w*(y - wm)^2) / (sum(w)-1))
 SE.srs <- sqrt(N*(N/n - 1)* sig2)
 SE.srs
 [1] 51365.75
```

The SE is much smaller for the π-estimator in pps sampling than in srswor
(14682 vs. 51365.75). This is because discharges (y) is related to beds (x) and
square root of beds. Spencer's deff reflects this fact but Kish's does not.

14.8 Using the data file smho.N874, (a) calculate the probabilities for all
population units in a sample of 50 selected with probabilities proportional to
the following measure of size (MOS): EXPTOTAL.

(a) Select a sample of size 50 using the probabilities computed in (a).

```
    # set mos = expenditures
mos <- smho$EXPTOTAL
n <- 50
N <- nrow(smho)
cert <- mos N * mean(mos)/n
certs1 <- (1:N)[cert]
certs1
[1] 161

set.seed(429336912)
n.nc <- n - length(certs1)
pk <- n.nc * mos / sum(mos[-certs1])
pk[certs1] <- 1
sam <- UPrandomsystematic(pk[-certs1])
nc.units <- (1:N)[-certs1]
    # Sample units are:
noncerts <- nc.units[sam == 1]
sam.units <- sort(c(certs1, nc.units[sam == 1]))
sam.units
 [1]   18   21   28   61   68   79   81   84   93  152  155  156  159
 161  162  163  168  171  189  190  193  199  204  207  221  246  250
 286  315  386  506  515  539  557  610  628  660  666  674  679  695
 696  713  722  728  754  802  819  822  864
sam.dat <- smho[sam.units, ]
wk <- 1/pk[sam.units]
summary(wk)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   4.425   7.971  19.820  19.650 206.600
```

(b) Compute Kish's and Spencer's *deff* for this sample. In the case of Spencer's *deff* use the variable SEENCNT as y.

```
p.d1 <- pk/n.nc
p.d1[certs1] <- mos[certs1]/sum(mos)
sam.dat$prbs.1d <- p.d1[sam.units]
    #Calculate WLS values
sam.wls <- lm(SEENCNT ~ prbs.1d, data=sam.dat, weights=wk)
    #DEFF component - var of y
sam.mean.y <- sum(sam.dat$SEENCNT * wk) / sum(wk)
sam.mean.y
[1] 2306.742
sam.var.y  <- sum(wk*(sam.dat$SEENCNT-sam.mean.y)^2)/sum(wk)
sam.var.y
[1] 6359677
    #DEFF component - alpha squared
sam.alpha2 <- sam.wls$coefficients[1] ^2
sam.alpha2
(Intercept)
   3352160
    #DEFF component - squared correlation
sam.rho2.yP <- summary(sam.wls)$r.squared
sam.rho2.yP
[1] 0.09294825
    #DEFF component - Kish
```

```
kish.deff <- n*sum(wk^2) / (sum(wk)^2)
kish.deff
[1] 4.268636
            #Spencer's DEFF
spencers.deff <- as.numeric((1 - sam.rho2.yP)*kish.deff +
                   (sam.alpha2/sam.var.y)*(kish.deff-1))
spencers.deff
[1] 5.59476
```

(c) Explain in words the meaning of the value you obtained in (c) for $1+L$?
    What should be considered in determining whether the value is exces-
    sively large or not? How do Kish's and Spencer's measures compare in
    this problem?
    *A value of 4.27 means that the variance of a mean is 4.27 times larger
    than it would be if equal weighting were optimal. However, pps sampling
    with probabilities proportional to EXPTOTAL may be very efficient for
    some estimates. The estimands that are important in the sample must be
    considered to decide whether 4.27 is a problem or not. In this case, Kish
    and Spencer's deff's are both large because SEENCNT is only weakly
    related to the MOS, EXPTOTAL. Both are saying that EXPTOTAL is
    not a good MOS if SEENCNT is the most important analysis variable.*

(d) Repeat parts (a)–(d) using BEDS as the MOS. Set the MOS for any unit
    with BEDS = 0 to the minimum value of BEDS for those with non-zero
    BEDS.

```
# (a) Compute MOS
    # set mos = BEDS
 mos <- smho$BEDS
 mos[mos == 0] <- min(mos[mos 0])
 n <- 50
 N <- nrow(smho)
 cert <- mos N * mean(mos)/n
 sum(cert)
 [1] 0
 certs1 <- (1:N)[cert]
 certs1      # no certs in this case
 integer(0)

# (b) Select a sample of size 50.
 set.seed(429336912)
 pk <- n * mos / sum(mos)
 sam <- UPrandomsystematic(pk)
    # Sample units are:
 sam.units <- (1:N)[sam == 1]
 sam.units
  [1]    6    9   33   49   77   82 106 111 129 154 157 163
  167 179 181 190 193 197 207 210 233 242 246 265 268
  271 288 334 338 352 360 384 393 403 416 481 499 500
  513 549 614 742 762 770 782 791 822 823 850 852
 sam.dat <- smho[sam.units, ]
 wk <- 1/pk[sam.units]
```

```
   summary(wk)
       Min.  1st Qu.   Median    Mean 3rd Qu.      Max.
      1.049    4.198   10.630   45.980  19.380 1424.000

# (c) Compute Kish's and Spencer's deffs for this sample.
 p.d1 <- pk/n
 sam.dat$prbs.1d <- p.d1[sam.units]
    #Calculate WLS values
 sam.wls <- lm(EXPTOTAL ~ prbs.1d, data=sam.dat, weights=wk)

    #DEFF component - var of y
 sam.mean.y <- sum(sam.dat$EXPTOTAL * wk) / sum(wk)
 sam.mean.y
 [1] 4664970
 sam.var.y  <- sum(wk * (sam.dat$EXPTOTAL -
                        sam.mean.y)^2) / sum(wk)
 sam.var.y
 [1] 8.229275e+13

    #DEFF component - alpha squared
 sam.alpha2 <- sam.wls$coefficients[1] ^2
 sam.alpha2
  (Intercept)
 4.167288e+12

    #DEFF component - squared correlation
 sam.rho2.yP <- summary(sam.wls)$r.squared
 sam.rho2.yP
 [1] 0.6500734

    #DEFF component - Kish
 kish.deff <- n*sum(wk^2) / (sum(wk)^2)
 kish.deff
 [1] 19.69660

    #Spencer's DEFF
 spencers.deff <- as.numeric((1-sam.rho2.yP)*kish.deff +
                  (sam.alpha2/sam.var.y)*(kish.deff-1))
 spencers.deff
 [1] 7.839158
```

The Kish value of 19.7 is extremely large, but so is the Spencer deff of 7.8. The weight summary shows that the largest weight is 1,424 which corresponds to a unit whose MOS was recoded to 1 from 0. In fact, the next smallest weight is 178.0. This appears to be a case where it would be advisable to do either (i) use a different recoding for the MOS, e.g., make the minimal value 5 or 10 rather than 1 or (ii) bound the weights. Quadratic programming may be a good choice to do this.

14.10

(a) Report the summary for the resulting weights, i.e., the min, max, quartiles, and the mean. Do any units have weights that seem to be of concern?

```
smho.N874 <- read.csv("C:\\Data\\smho.N874.csv",
                row.names = 1)
smho <- smho.N874
    # set mos = BEDS
mos <- smho$BEDS
mos[mos == 0] <- min(mos[mos 0])
n <- 50
N <- nrow(smho)
set.seed(429336912)
pk <- n * mos / sum(mos)
sam <- UPrandomsystematic(pk)
    # Sample units are:
sam.units <- (1:N)[sam == 1]
sam.units
 [1]   6   9  33  49  77  82 106 111 129 154 157 163 167
 179 181 190 193 197 207 210 233 242 246 265 268 271 288
 334 338 352 360 384 393 403 416 481 499 500 513 549 614
 742 762 770 782 791 822 823 850 852
sam.dat <- smho[sam.units, ]
d <- 1/pk[sam.units]
summary(d)
    Min.  1st Qu.   Median    Mean  3rd Qu.     Max.
   1.049    4.198   10.630   45.980  19.380 1424.000
sort(d)
 [1]    1.049462    1.200776    1.206881    1.382641
 1.849506 2.401551    2.401551    2.579928    2.738692
 2.831252 3.901699    3.966908    4.045795    4.653987
 5.353835 5.394394    5.606772    6.593148    6.593148
 6.846731 7.697946    9.494133    9.889722   10.029014
 10.549037 10.707669   13.309533   13.309533   13.435094
 14.241200 14.241200   14.531837   14.990737   15.479565
 16.001348 17.581728   18.988267   19.508493   19.779444
 20.344571 20.942941   21.577576   22.969677   37.476842
 44.503750 59.338333   61.918261   71.206000  178.015000
 1424.120000
```

The largest weight of 1,424 is far bigger than any others. This is not likely to be efficient.

(b) Use quadratic programming to bound the weights in the range [1, 50]. Plot the resulting weights versus the design weights. What was the effect of the bounding? Is quadratic programming an effective way of bounding the weights here?

```
    # Tabulate pop totals for constraints
x.beds <- sum(smho$BEDS)
x.seen <- sum(smho[,"SEENCNT"])
x.eoy <- sum(smho[,"EOYCNT"])
X.hosp <- model.matrix(~ 0 + as.factor(hosp.type):BEDS,
        data = sam.dat)
X <- rbind(sam.dat[, "BEDS"],
        sam.dat[, "SEENCNT"],
        sam.dat[, "EOYCNT"]
    )
```

```
c0a <- c(x.beds, x.seen, x.eoy)
   # Compute full sample weights via quadratic programming
In <- diag(nrow = n)
L <- 1
U <- 50
one <- rep(1, n)
c0b <- c( L * one,
         -U * one)
Cmat <- rbind(X, In, -In)
fs.wts <- solve.QP(Dmat = diag(1/d),
                   dvec = 2 * one,
                   Amat = t(Cmat),
                   bvec = c(c0a, c0b),
                   meq = 3    # 1st 3 are equality constraints
         )
sort(fs.wts$solution)
[1]   1.000000   1.000000   1.000000   1.000000   1.000000 1.000000
1.000000   1.000000   1.000000   1.000000   1.000000   1.000000
1.000000   1.000000   1.000000   1.000000   1.000000   2.821197
2.982887   3.181804   5.488968   9.867670 12.623416 14.049774
14.235627 14.577856 14.728491 17.755924 18.204255 19.235434
19.393679 19.535066 22.993572 24.105195 27.706496 28.908379
29.005913 31.769392 32.262358 33.271899 34.805294 40.470829
50.000000 50.000000 50.000000 50.000000 50.000000 50.000000
50.000000 50.000000
plot(d,fs.wts$solution)
abline(0,1)
```
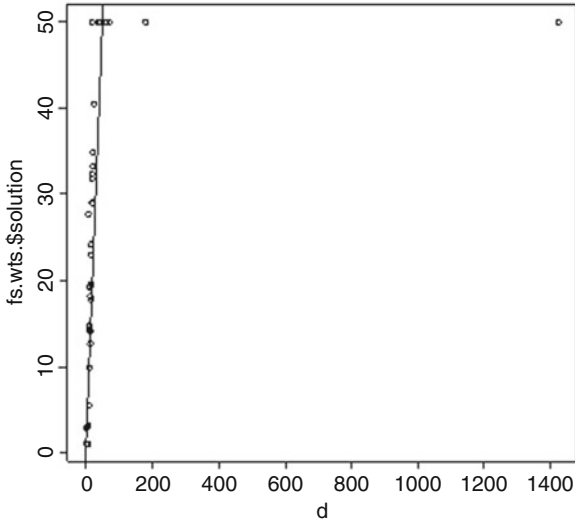

There are eight weights that were trimmed back to 50; otherwise, most weights were not modified too much.

(c) Re-do parts (a) and (b) but recode any unit with BEDS = 0 to BEDS=10. Discuss your results. Are the weight adjustments as extreme as in (b)?

```
mos <- smho$BEDS
mos[mos $<$ 10] <- 10
n <- 50
N <- nrow(smho)
set.seed(429336912)
pk <- n * mos / sum(mos)
sam <- UPrandomsystematic(pk)
   # Sample units are:
sam.units <- (1:N)[sam == 1]
sam.units
 [1]    6   26   49   52   53   77   82   94  111  116  129  136  154
 155  157  163  167  179  181  183  189  190  193  197  233  271  288
 352  360  481  499  500  513  535  721  742  769  782  788  791  802
 811  823  826  832  838  852  864  865  872

sam.dat <- smho[sam.units, ]
d <- 1/pk[sam.units]
summary(d)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.078   4.301  10.720  17.460  17.400 146.300
   # Tabulate pop totals for constraints
x.beds <- sum(smho$BEDS)
x.seen <- sum(smho[,"SEENCNT"])
x.eoy <- sum(smho[,"EOYCNT"])
X.hosp <- model.matrix(~ 0 + as.factor(hosp.type):BEDS,
            data = sam.dat)
X <- rbind(sam.dat[, "BEDS"],
           sam.dat[, "SEENCNT"],
           sam.dat[, "EOYCNT"]
     )
c0a <- c(x.beds, x.seen, x.eoy)
   # Compute full sample weights via quadratic programming
In <- diag(nrow = n)
L <- 1
U <- 50
one <- rep(1, n)
c0b <- c( L * one,
         -U * one)
Cmat <- rbind(X, In, -In)
fs.wts <- solve.QP(Dmat = diag(1/d),
                   dvec = 2 * one,
                   Amat = t(Cmat),
                   bvec = c(c0a, c0b),
                   meq = 3   # 1st 3 are equality constraints
     )
sort(fs.wts$solution)
 [1] 1.000000  1.000000  1.000000  1.000000  1.000000
 1.000000 1.000000  1.000000  1.000000  1.000000  1.000000
 1.000000 1.000000  1.000000  1.000000  1.000000  1.000000
 1.000000 1.000000  3.043553  3.839364  4.839955  6.087290
 7.387588 8.230181 13.038428 14.421487 16.876299 17.517496
 18.532730 20.903125 22.139108 22.909465 23.599381 23.914943
 25.941203 28.815430 29.358972 29.982384 32.729851 32.849170
 37.342522 37.523764 38.214365 45.850978 50.000000 50.000000
```

```
    50.000000 50.000000 50.000000

plot(d,fs.wts$solution)
abline(0,1)
```
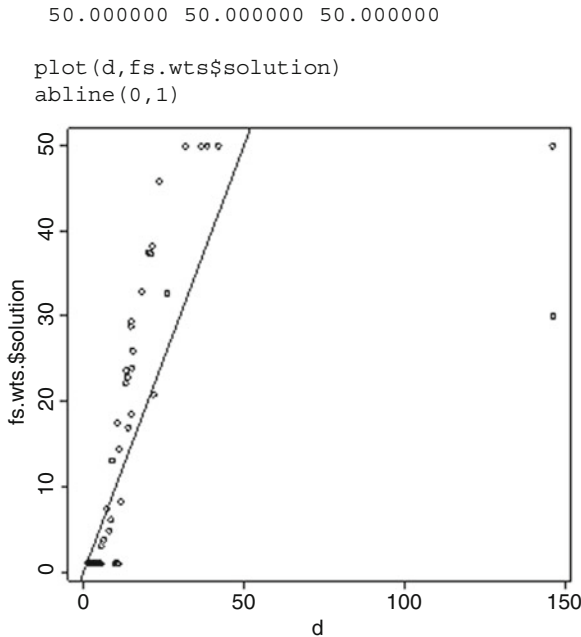

The initial range of the weights (1.078, 146.3) is much smaller here because of the recoding of the MOS. This may be efficient, but specific analysis variables would have to be examined to be sure.

Chapter 15

15.2 The following data were collected from a sample of two PSUs selected from each of two strata.

h	PSU	Y_{hi}
1	1	5
1	2	6
2	1	10
2	2	4
Total		25

Y_{hi} is the weighted PSU total observed for PSU i in stratum h.

(a) Compute the balanced repeated replication (BRR) variance estimator for the estimated total $\hat{y} = \sum_{h=1}^{2} \sum_{i=1}^{2} Y_{hi}$. Specify which form of the BRR estimator you are using. Use the following orthogonal matrix where rows designate the strata and columns the replicates:

$$A = \begin{bmatrix} + & + & + & + \\ + & - & + & - \\ + & + & - & - \\ + & - & - & + \end{bmatrix}.$$

A balanced set of replicates has a number of replicates equal to the smallest number greater than or equal to the number of strata—4 in this case. We can use any two rows of the matrix. Use the last two rows above to denote strata since this will give 4 different estimates. We could use rows 1 and 2 but this will give only two different estimates. However, in the 2 stratum case, the standard BRR variance estimate will be the same whether we use rows 1–2 or 3–4. Using rows 3–4:

Replicate 1: 2*5 + 2*10=30; Replicate 2: 2*5 + 2*4=18; Replicate 3: 2*6 + 2*4=20; Replicate 4: 2*6 + 2*10=32

$$v_B = \tfrac{1}{4}\left[(30-25)^2 + (18-25)^2 + (20-25)^2 + (32-25)^2\right]$$
$$= \tfrac{1}{4}[25 + 49 + 25 + 49] = 37$$

(b) What is the variance formula for the estimated total $\hat{y}$ if PSUs are assumed to be selected with replacement? Evaluate this formula using the data in the table above. How does it compare with your answer in part (a)?

The variance formula is $v_{WR}=\sum_h \frac{n_h}{n_h-1} \sum_{s_h}\left(\hat{Y}_{hi} - \hat{\bar{Y}}_h\right)^2$

.

15.6 Use the `nhis.large` file as a population and select a simple random sample of size $n = 500$. If you are using R, use a random number seed of 428274453. Poststratify the sample to population counts for `age.grp`. (a) Compute the estimated proportion of the population who reported a doctor visit (`doc.visit`) in the 2 weeks prior to the interview. (b) Calculate the SEs using the linearization method and JKn. What would be the effect on estimated SEs of ignoring the poststratification? (c) Estimate the proportions and SEs of the population who reported a doctor visit in a table defined by Hispanic ethnicity (`hisp`). Combine categories 3 and 4 of hisp together. What would be the effect of ignoring the poststratification for these estimates?

```
attach("C:\\Data\\nhis.large.RData", pos=2)
require(sampling)
require(survey)
    # collapse hisp = 3,4
hisp.r <- nhis.large$hisp
hisp.r[nhis.large$hisp ==4] <- 3
table(hisp.r)
    1     2     3
 5031 12637  3920

nhis.large1 <- data.frame(nhis.large, hisp.r)
t1 <- table(nhis.large$doc.visit, nhis.large1$hisp.r)
```

```
100*round(prop.table(t1,2),3);
       1    2    3
  1 12.0 17.2 14.7
  2 88.0 82.8 85.3

nhis.large1$PS <- nhis.large1$age.grp
N.PS <- table(PS = nhis.large1$PS)
N.PS
PS
   1    2    3    4    5
5991 2014 6124 5011 2448
   # select srswor of size n
set.seed(428274453)
n <- 500
N <- nrow(nhis.large1)
sam <- sample(1:N, n)
samdat <- nhis.large1[sam, ]
n.PS <- table(samdat[, "age.grp"])
as.vector(n.PS)
[1] 155  46 128 107  64

   # compute srs weights and sampling fraction
d <- rep(N/n, n)
   # srswor design object
nhis.dsgn <- svydesign(ids = ~0,
          strata = NULL,
          data = data.frame(samdat),
          weights = ~d)

   #  Linearization variances
   # poststratified design object
ps.dsgn <- postStratify(design = nhis.dsgn,
                        strata = ~PS,
                        population = N.PS)
   # Check that weights are calibrated for x's
svytotal(~ as.factor(PS), ps.dsgn)
              total SE
as.factor(PS)1  5991  0
as.factor(PS)2  2014  0
as.factor(PS)3  6124  0
as.factor(PS)4  5011  0
as.factor(PS)5  2448  0

   # PS linearization standard errors and cv's
a1.lin <- round(svymean(~ as.factor(doc.visit),
          ps.dsgn, na.rm=TRUE), 4)
a2.lin <- round(cv(svymean(~ as.factor(doc.visit),
          ps.dsgn, na.rm=TRUE)), 4)
   # crosstab: age group x hispanic
b1.lin <- round(svyby(~as.factor(doc.visit),
               by = ~hisp.r, design = ps.dsgn,
               svymean, na.rm=TRUE), 4)
b2.lin <- round(cv(svyby(~as.factor(doc.visit),
               by = ~hisp.r, design = ps.dsgn,
```

```
                    svymean, na.rm=TRUE)), 4)

   # linearization standard errors and cv's ignoring
   # poststratification
wts <- weights(ps.dsgn)
   # design object ignoring PS
noPS.dsgn <- svydesign(ids = ~0,
          strata = NULL,
          data = data.frame(samdat),
          weights = ~wts)
a1.noPS <- round(svymean(~ as.factor(doc.visit),
              noPS.dsgn, na.rm=TRUE), 4)
a2.noPS <- round(cv(svymean(~ as.factor(doc.visit),
              noPS.dsgn, na.rm=TRUE)), 4)
b1.noPS <- round(svyby(~as.factor(doc.visit), by = ~hisp.r,
              design = noPS.dsgn, svymean, na.rm=TRUE), 4)
b2.noPS <- round(cv(svyby(~as.factor(doc.visit), by = ~hisp.r,
              design = noPS.dsgn, svymean, na.rm=TRUE)),

   #  Jackknife variances
jk1.dsgn <- as.svrepdesign(design = nhis.dsgn, type = "JK1")
   # poststratified design object
jk1.ps.dsgn <- postStratify(design = jk1.dsgn,
                       strata = ~PS,
                       population = N.PS)
   # PS JK1 standard errors and cv's
a1.jk <- round(svymean(~ as.factor(doc.visit), jk1.ps.dsgn,
              na.rm=TRUE), 4)
a1.jk
                        mean      SE
as.factor(doc.visit)1 0.1602 0.0162
as.factor(doc.visit)2 0.8398 0.0162

a2.jk <- round(cv(svymean(~ as.factor(doc.visit), jk1.ps.dsgn,
          na.rm=TRUE)), 4)
a2.jk
as.factor(doc.visit)1 as.factor(doc.visit)2
              0.1008                  0.0192

   # crosstab: age group x hispanic
b1.jk <- round(svyby(~as.factor(doc.visit), by = ~hisp.r,
              design = jk1.ps.dsgn, svymean, na.rm=TRUE), 4)
b2.jk <- round(cv(svyby(~as.factor(doc.visit), by = ~hisp.r,
              design = jk1.ps.dsgn, svymean, na.rm=TRUE)), 4)
SEs <- cbind(b1.noPS, b1.lin, b1.jk)
SEs <- SEs[, -c(1,6,11)] # remove columns of hisp IDs
pt.ests <- SEs[, c(1,5,9)] # keep pt. ests of prop. with doc
                           # visit
SEs <- SEs[, c(3,7,11)]
dimnames(pt.ests)[[1]] <-
    dimnames(SEs)[[1]] <- c("Hispanic", "non-Hisp white",
                            "non-Hisp Black & Other")
dimnames(pt.ests)[[2]] <- c("noPS doc=1","lin PS doc=1",
                            "jk doc=1")
```

```
 pt.ests
                       noPS doc=1 lin PS doc=1 jk doc=1
Hispanic                   0.0785         0.0785    0.0785
non-Hisp white             0.2077         0.2077    0.2077
non-Hisp Black & Other     0.1044         0.1044    0.1044
 dimnames(SEs)[[2]] <- c("noPS SE doc=1","lin SE doc=1",
                       "jk SE doc=1")

# SEs on estimated proportions of persons with doctor visits
# by the Hispanic variable.
 SEs
                       noPS SE doc=1 lin SE doc=1 jk SE doc=1
Hispanic                      0.0253       0.0252      0.0255
non-Hisp white                0.0235       0.0228      0.0230
non-Hisp Black & Other        0.0331       0.0330      0.0336

 CVs <- cbind(b2.noPS, b2.lin, b2.jk)
 CVs <- CVs[, c(1,3,5)]  # keep CV ests of proportion with doc
                       visit

 dimnames(CVs)[[1]] <- c("Hispanic", "non-Hisp white",
                       "non-Hisp Black & Other")
 dimnames(CVs)[[2]] <- c("noPS CV doc=1","linCV doc=1",
                       "jkCV doc=1")
# CVs on estimated proportions of persons with doctor visits
# by the Hispanic variable.
 CVs
                       noPS CV doc=1 linCV doc=1 jkCV doc=1
Hispanic                      0.3222      0.3213     0.3253
non-Hisp white                0.1132      0.1098     0.1109
non-Hisp Black & Other        0.3173      0.3164     0.3215
```

Discussion: The point estimates of the proportions with doctor visits are not affected by the choice of variance estimation method. In this example, there is very little difference in the SE and CV estimates whether poststratification is accounted for or not. Linearization and jackknife SEs and CVs are very similar.

15.8 Repeat Exercise 15.6 using the bootstrap method with 500 replicates. If you are using R, use a random number seed of -711384152. How do your estimates of standard errors and *CV*s compare to the linearization and jackknife estimates in Exercise 15.6?

```
attach("C:\\Data\\nhis.large.RData", pos=2)
 require(sampling)
 require(survey)
    # collapse hisp = 3,4
 hisp.r <- nhis.large$hisp
 hisp.r[nhis.large$hisp ==4] <- 3
 table(hisp.r)
 nhis.large1 <- data.frame(nhis.large, hisp.r)
 nhis.large1$PS <- nhis.large1$age.grp
 N.PS <- table(PS = nhis.large1$PS)
```

```
   # select  srswor of size n
set.seed(-711384152)
n <- 500
N <- nrow(nhis.large1)
sam <- sample(1:N, n)
samdat <- nhis.large1[sam, ]
n.PS <- table(samdat[, "age.grp"])
as.vector(n.PS)

   # compute srs weights and sampling fraction
d <- rep(N/n, n)
   # srswor design object
nhis.dsgn <- svydesign(ids = ~0,
         strata = NULL,
         data = data.frame(samdat),
         weights = ~d)
   # create design with boostrap wts.
   # Rao-Wu version used with mh = nh-1
nhis.boot <- as.svrepdesign(design = nhis.dsgn,
                            type = "subbootstrap",
                            replicates = 500)
   # poststratified design object
boot.ps <- postStratify(design = nhis.boot,
                        strata = ~PS,
                        population = N.PS)

   # PS boot standard errors and cv's
a1.boot <- round(svymean(~ as.factor(doc.visit),
            boot.ps, na.rm=TRUE), 4)
a2.boot <- round(cv(svymean(~ as.factor(doc.visit),
            boot.ps, na.rm=TRUE)), 4)
   # crosstab: age group x hispanic
b1.boot <- round(svyby(~as.factor(doc.visit), by = ~hisp.r,
         design = boot.ps, svymean, na.rm=TRUE), 4)
b2.boot <- round(cv(svyby(~as.factor(doc.visit), by = ~hisp.r,
         design = boot.ps, svymean, na.rm=TRUE)), 4)

ests <- cbind(b1.boot, b2.boot)
ests <- ests[, c(2,4,6)] # keep ests for prop. with doc visit
dimnames(ests)[[1]] <- c("Hispanic", "non-Hisp white",
                     "non-Hisp Black & Other")
dimnames(ests)[[2]] <- c("doc=1", "SE","CV")
ests
                         doc=1     SE     CV
Hispanic                0.1613 0.0338 0.2096
non-Hisp white          0.2032 0.0232 0.1140
non-Hisp Black & Other  0.1097 0.0314 0.2863
```

Author Index

Subject Index