Valliant, Dever, Kreuter

Practical Tools for Designing and Weighting Survey Samples

-subtitle-

April 23, 2012

Springer

Contents

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 is defined here that will come in handy throughout the book.

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, population statistics for which 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*, by Groves, Fowler, Couper, Lepkowski, Singer, and Tourangeau (2004) (rob:BIB) or *Introduction to Survey Quality* by Biemer and Lyberg (2003) (robp:BIB) cover these issues nicely and

are strongly recommended as supplements to the material presented in this book. The primary focus here is the sixth step and, thus we will only briefly comment on the other five, to the extent 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 U.S. 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 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 U.S. 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?

---

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

[2] `http://www40.statcan.gc.ca/l01/cst01/other/lfs/lfsintro-eng.htm`

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

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). 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 U.S. 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 hardcopy 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).

At this writing, a fairly new sampling frame exists in the U.S. that is based on the U.S. Postal Service (USPS) Delivery Sequence File (DSF) (Iannacchione et al. 2003; Iannacchione 2011; Link et al. 2008)(robp:BIB). 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, than will 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)(robp:BIB). 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. 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*.

As apparent from the definition, the *deff* is specific to a particular estimator, 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 random 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 available 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 recontacted using a different mode of data collection, or a study of families with children under the age of 5 using a sample of the general population).

Each of the three general designs above usually involves probability sampling. Srndal, Swensson, and Wretman (1992, sec. 1.3)(robp:BIB) 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 non-zero probability of selection. (iv) One set of sample elements is selected with the probability associated with the set. Weights for sample units 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—*indirect* and *direct*. 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. For example, in a survey of households, a common practice is to select a sample of geographic areas, called primary sampling units (PSUs) first. Within the sample PSUs, households may be selected from (*i*) lists compiled by research personnel (sometimes referred to as listers) who canvass the area (in a process known as *counting and listing*), or (*ii*) lists maintained by organizations such as the U.S. Postal Service (USPS) delivery sequence file.

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 U.S. 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 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) (robp:BIB).

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 (2000, ch. 16)(robp:BIB) 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 (Bureau of the Census 2002)(robp:BIB). 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. In urban areas, it may also be difficult to identify all

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

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.

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.

1.2 Chapter Guide

The book is divided into three (robp: I count 4 parts) 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(robp:REF)) provide 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(robp:REF) 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 pre-specified difference in means, proportions, etc. between some subgroups.

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 socio-economic 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 *multi-criteria 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. 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 Chapter 6(robp:REF), 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 Chapter 8(robp:REF) to design an area sample and allocate units to geographic clusters in such a way that the size of the samples of persons are controlled for some important demographic groups. Chapters 9(robp:REF) and 10(robp:REF) cover the design of samples of those geographic clusters. The U.S. National Health and Nutrition Examination Survey (NHANES; CDC 2009)(robp:BIB?) 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 has 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 Chapter 12(robp:REF) on calculating weights for a personnel survey, like the one designed in Project 1(robp:REF?). Chapters 13(robp:REF) and 14(robp:REF) 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(robp:REF) 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 Chapter 14(robp:REF). Methods of weight trimming using quadratic programming and other, more *ad hoc* methods are also dealt with in that chapter.

Chapter 15(robp:REF) 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 Chapter 15(robp:REF).

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 people in these subgroups is not available in advance of sampling, the technique of *multiphase sampling* can be used as described in Chapter 17(robp:REF). A large initial sample is selected, and group identity determined for each person 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(robp:REF) 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 re-do 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 Development Core Team 2005)(robp:BIB), which is available for free. Additional materials are provided in the Appendices. *Appendix A*(robp:REF) contains a primer on the R programming language (R Development Core Team 2005)(robp:BIB) and functions developed for Chapter examples. Datasets used in many of the examples are described in *Appendix B*(robp:REF); small datasets are provided within these pages while larger files are available through the book's Web address. Details are given in *Appendix C*(robp:REF) on several group projects developed for a course based on this text.

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; and
- 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.

2.1 Specification for the Study

The Verkeer NetUltraValid (VNUV) International Corporation is preparing to conduct Cycle 5 of its yearly 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. 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 lunch-time 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 2–3 day difference 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 2.6 (robp:REF) below.*
6. Are there additional requirements for the design, such as estimates by gender, by number of dependents, etc. 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 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 Analysis

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 2.1)(robp:REF).

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, 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 2.2)(robp:REF).

(robp:1.table is cycle 5, 2.table corresponds to cycle 4?)

Table 2.1 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

The information in Tables 2.3-2.6(robp:REF) 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.

(robp:break)

2.4 Documentation

With the preliminary analysis complete, the design team began to draft the sampling report to the Senior Council using the following annotated outline:

Table 2.2 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 >

Title = *UNUV Climate Survey Cycle 5 Sample Design Report* (robp:UNUV or VNUV?)

1. Executive Summary

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

   a) Optimization Requirements
      - Optimization details including constraints, budget, etc.
      - Detail the minimum domain sizes and mechanics used to determine the sizes
   b) Optimization Results
      - Results = minimum respondent sample size per stratum
      - Marginal sample sizes for key reporting domains
      - Estimated precision achieved by optimization results
   c) Inflation Adjustments to allocation solution
      - Nonresponse adjustments
      - Adjustments for ineligible sample members
   d) Final Sample Allocation
      - Marginal sample sizes for key reporting domains
   e) Sensitivity Analysis

Table 2.3 Distribution of Response Status by Business Unit, Salary Grade, and Tenure: VNUV Climate Survey Cycle 4, Survey Division

			Total			Eligible				
Business	Salary		Sample	Ineligible[a]		Total	Respondent		Nonrespondent	
Unit	Grade	Tenure	n	n	pct[b]	n	n	pct[c]	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 that 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).

- 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

Table 2.4 Estimates for Four Key Questions by Business Unit, Salary Grade, and Tenure: VNUV Climate Survey Cycle 4, Survey Division

Business	Salary		Proportion (Strongly) Agree			Avg Number of Training Classes	
Unit	Grade	Tenure	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.

2.5 Next Steps

The optimization problem and a proposed solution to the sampling design task discussed in this chapter will be revealed in Chapter 7(robp:REF). 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 Chapter 7(robp:REF) to provide insight into the design team's decisions and procedures.

Table 2.5 Estimates By Business Unit, Salary Grade, and Tenure:VNUV Climate Survey Cycle 4, Survey Division

Business Unit	Salary Grade	Tenure	Proportion (Strongly) Agree Q5	Q12	Q15	Avg Number of Training Classes 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

Table 2.6 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.

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 estimates discussed in Sect. 3.5. The chapter concludes with some discussion of design effects and software for sample selection in Sect. 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 studied in Chapter 5(robp:REF).

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 income 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 Chapter 2(robp:REF). Chapter 5(robp:REF) covers 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 Chapter 9(robp:REF).

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 purpose 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.
- The *population (or unit) coefficient of variation* of Y is $CV_U = S / \bar{y}_U$. The square of the population CV, $S^2 / \bar{y}_U^2$, 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 $cv\left(\hat{\theta}\right) = \sqrt{v\left(\hat{\theta}\right)}\Big/\hat{\theta}$, 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 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.

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 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. We discuss several 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, etc.).
- Coefficient of variation—CV's 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."

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

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

where N is the number of units in the target population on the sampling frame and $S^2 = \sum_{i=1}^{N}\left(y_i - \bar{y}_U\right)^2/(N-1)$ 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

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

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

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

We can set the squared CV or relvariance in (3.3) to some desired value, CV_0^2, 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 (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 (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, equation (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, `n.cont`, will compute a sample size using either CV_0 or V_0 as input. The parameters used by the function are shown below:

```
n.cont(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 A(robp:REF), which shows the code for the functions and detailed explanations of the parameters. The functions can be used after pasting the code into an RGui session.

Example 3.1. 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 `n.cont(CV0=0.05, CVpop=2)`. The resulting sample size is 1,600. If the population size is $N = 500$, then `n.cont(CV0=0.05, CVpop=2, N=500)`

results in a (rounded) sample size of 381. The finite population correction 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 U.S. National Center for Health Statistics flags any estimate it publishes that has a CV of 30% or more and labels it as "unreliable." 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 that level of precision was satisfactory. In that case, the same sample design and allocation could be used again. Some values of CV's from government-sponsored surveys in the U.S. are listed in Table 3.1(robp:REF). These obviously have quite a large range. CV's 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.

(robp:FLOAT TABLE 3.1!)

In some instances, a precision target may be set by an administrative group. For example, the Council of the European Union (1998)(robp:BIB) 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)(robp:BIB) for income and living conditions estimates. An effective sample size, n_{eff}, was defined in Chapter 1(robp:REF) and is equal to the number of sample elementary 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 Chapters 9 (robp:REF) and 10(robp:REF).

Example 3.2. The U.S. 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 of all capital assets that can be depreciated on a 5-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

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 U.S. 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*
Hospital Discharge Survey 2005[e]	Total days of hospital care for heart disease	21.3% *CV*

[a] BLS (2006)(robp:BIB)
[b] BLS (2009)(robp:BIB)
[c] Thayer and Diamond (2002)(robp:BIB)
[d] Deak, et al. (2002, Table 28a.1)(robp:BIB)
[e] CDC (2005, Tables I, II)(robp:BIB)

confidence interval is used, the requirement is(robp:Wirklich ein kleines e?) $e = 1.645 \times CV\left(\hat{T}\right) \leq 0.10$. 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.(robp:2seitige fussnote)[1] Since $CV(\bar{y}_s) = CV\left(\hat{T}\right)$ under simple random sampling, the IRS bound is equivalent to $CV\left(\hat{T}\right) \leq 0.10\,/1.645$. If the population CV is 1, the sample size that would meet the IRS requirement is 271, which is obtained via `n.cont(CV0=0.10/1.645, CVpop=1)`.

[1] See, e.g., Internal Revenue Service *Cost Segregation Audit Techniques Guide*, December 2007, `http://www.irs.gov/businesses/article/0,,id=134672,00.html`; Internal Revenue Bulletin: 2007-23, §199 Deduction, Appendix A, June 4, 2007; Internal Revenue Bulletin: 2004-20, Meals and Entertainment Expenses, `http://www.irs.gov/irb/2007-23_IRB/ar10.html`

Example 3.3. Revisiting the data gathered for the VNUV Climate Survey (Project 1 in Chapter 2(robp:REF)), 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 $CV^2(\bar{y}_s) = \left(n^{-1} - N^{-1}\right) CV_U^2$ where $CV^2(\bar{y}_s)$ is from the previous survey, the population (unit) CV within each stratum can be computed as $CV_U^2 = CV^2(\bar{y}_s) \big/ \left(n^{-1} - N^{-1}\right)$. Information for the SR business unit, key to calculating the sample sizes, includes: (robp:enhance table)

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 2.1.(robp:REF)

[b] Estimated means and standard errors were obtained from a prior survey and are shown in Table 2.5.(robp:REF)

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:

```
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
n.cont(CV0=0.05, CVpop = CVpoph, N=Nh)
        # SR business unit sample
n.cont(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 0.06 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 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	28.2
	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(robp:REF?), 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

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

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

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

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

Based on equation (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 the exercises (robp:REF).

(robp:ACHTUNGFUSSNOTEaufrichtigerSeite)

Whether the sample size should be computed via the formula given in (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 (3.8) needed for CV's 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, `n.prop`, will compute the sample size using (3.8), assuming that a target CV_0 is set, or using (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: `n.prop(CV0=NULL, V0=NULL, pU=NULL, N=Inf)`

Example 3.4. 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 `n.prop(V0=0.0005^2, N=Inf, pU=0.01)` or `n.prop(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 one percent. This would, at least, confirm any suspicion that the proportion is quite small. If the 1/2 of one percent goal is translated to mean that a 95% confidence interval should have a half-width of 0.005, this means that

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

[3] The population size is assumed to be large so that the finite population correction is one.

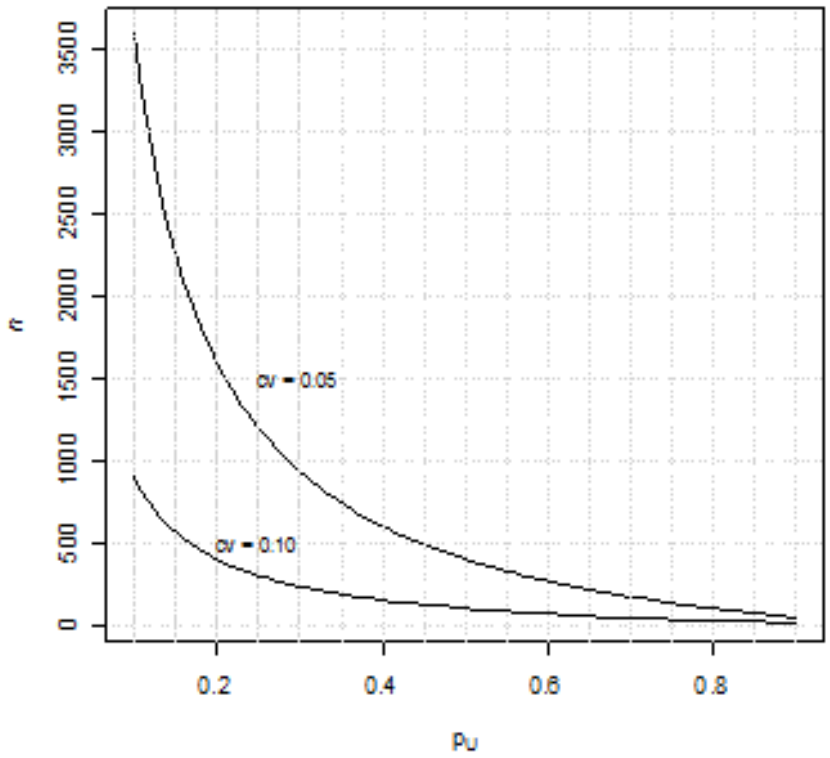

Fig. 3.1 Approximate sample sizes from equation (3.8) required to achieve CV's of 0.05 and 0.10 for population proportions ranging from 0.10 to 0.90.[3]

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 `n.prop` to compute this is `n.prop(V0=(0.005/1.96)^2, N=Inf, pU=0.01)`.

The function `n.prop` 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 `n.prop(CV0=0.05, N=Inf, pU=c(0.01, 0.05, 0.10))` with results, $n =$ 39600, 7600, and 3600.

Example 3.5. Returning again to Project 1 (Chapter 2(robp:REF)), 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:

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

[a] Counts of employees shown in Table 2.1.(robp:REF)
[b] Estimated proportion of employees who strongly agree with the statement in question 5.

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 `n.prop` R 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., `n.prop(CV0=0.06, N=Inf, pU=0.82)` for salary grades A1-A3, which gives $n = 61$, compared with `n.prop(CV0=0.06, N=68, pU=0.82)`, which yields $n = 34$. 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 CV's. If the tolerance (sometimes called the *margin of error*) is e and the goal is to be within e of the population mean with probability $1 - \alpha$, this translates to

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

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

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

this corresponds to setting $e = z_{1-\alpha/2}\, CV(\bar{y}_s)$. (See the exercise 3.4.) If we set the margin of error to e_0, then (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 margin of error in (3.11) is set to e_0, we obtain

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

In the particular case of estimating a proportion, we set $S^2 = Np_Uq_U/(N-1)$ in (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 (3.9) once we note that $V_0 = e^2 \Big/ z_{1-\alpha/2}^2$. Either (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 $S^2 \big/ \bar{y}_U^2 = Nq_U/[(N-1)\,p_U]$ in (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 (3.8). The R function, `n.prop.moe`, will calculate sample sizes using (3.14) or (3.15), corresponding to whether we set the margin of error in terms of (3.10) or (3.11). The type of margin of error is selected by the parameter `moe.sw` where `moe.sw=1` invokes equation (3.14), i.e., $e = z_{1-\alpha/2}\sqrt{V(p_s)}$ and `moe.sw=2` invokes equation (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.

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

Example 3.6. Suppose that we want to estimate a proportion for a characteristic where $p_U = 0.5$ with a margin of error 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 (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

```
n.prop.moe(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)(robp:BIB). Brown et al (2001) and Newcombe (1998)(robp:BIB) 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 margin of error. 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 `n.wilson` will calculate a sample size using inputs for p_U and e. As in `n.prop.moe`, the desired margin of error 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 `n.wilson(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. Suppose that $p_U = 0.04$ and the desired half-width of the CI is 0.01. The function call and output are:

```
n.wilson(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 margin-of-error computation using `n.prop.moe` is

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

[1] 1475.120
```

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 $\log(p_s/q_s)$ with $q_s = 1 - p_s$. A linear approximation to the log-odds is

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

The approximate variance of $\log(p_s/q_s)$ is then estimated as

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

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

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

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

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

where

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

and $k = q_U/p_U$. The R function, `n.log.odds`, will evaluate the sample size in (3.17). The function accepts the same five parameters as `n.prop.moe`. The desired margin of error 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:

```
n.log.odds(moe.sw, e, alpha=0.05, pU, N=Inf)
```

Another transformation that is sometimes used when calculating a CI for a proportion is $\arcsin\left(\sqrt{p_s}\right)$. This transformation is not included here because

it does not appear amenable to sample size calculation when setting a margin of error.

Example 3.8. 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 from our three functions for computing samples sizes are:

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

[1] 1500.460

n.wilson(moe.sw=1, pU=0.04, e=0.01)$n.sam

[1] 1492.151

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

[1] 1475.120
```

The sample sizes are within about 2 percent 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 U.S. Bureau of Labor Statistics[4] publishes estimated percentages of workers that receive different benefits from their employers, the National Center for Health Statistics[5] 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[6] 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 micro-data sets for earlier samples will be available. For instance, the Common Core of Data (CCD)[7] from NCES contains population data files of elementary and

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

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

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

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

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. 3.7. Note that the design team for Project 1(robp:REF?) had direct access to the relevant data sources and could therefore produce the estimates provided in Tables 1-6 in Chapter 2(robp:REF).

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;
- $S_h^2 = \sum_{i=1}^{N_h} \left(y_{hi} - \bar{y}_{U_h}\right)^2 / (N_h - 1)$, the population variance in stratum h;
- U_h = set of all units in the population from stratum h; and
- 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:

$$\bar{y}_{st} = \sum_{h=1}^{H} W_h \bar{y}_{s_h} \tag{3.18}$$

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

$$p_{st} = \sum_{h=1}^{H} W_h p_{s_h} \tag{3.19}$$

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 section 3.4.1(robp:REF es gibt keine Section 3.4.1 ?) 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); and
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 revenues. To determine a good measure of size, regression modeling should be done assuming that some relevant data are available. This method of stratification is closely related to probability proportional to size 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 4 and 5, the proportion of the sample allocated to a stratum is the same and is given in (3.23). The two methods lead to different total sample sizes as shown in (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

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

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

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

The proportion of the sample allocated to stratum h is

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

As is apparent from the expression given in (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

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

If the CV of $\bar{y}_{st}$ is fixed at CV_0, this implies that V_0 in (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 (3.23) but the total sample size is

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

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 (3.21) and (3.24) are:

(i) Verify that $\sum_h n_h c_h$ approximately respects the cost constraint, and
(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 (3.21) is probably the one used more often. The usual situation is that an investigator has a pre-determined amount of money to spend. Any sample that is selected must fit within that budget. Another standard occurrence is that part-way through a study the budget is changed—usually cut—or that the unit costs c_h are higher than expected. Consequently, mid-course 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 (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 (3.20) could be specialized using the formulas for allocations 1–5, this is unnecessary and, in fact, undesirable for computer programming. When evaluating (3.20) from a sample, the population variance, S_h^2, can be estimated as described in Sect. 3.4.

The R function, `str.alloc`, 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. Table 3.1 gives stratum population counts and standard deviations of total expenditures based on the 1998 Survey of Mental Health Organizations (SMHO)[8]. The survey data set is treated as the population (smho98) for this example. 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 3.2 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 multi-service or substance abuse hospitals. After considering cost, these two strata account for only 60% of the sample (a 13 percentage point reduction) because the cost per organization is higher than for other strata.

(robp:center n_h in table) (robp:center cost c_h in table)

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 (3.25) is

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

Table 3.1 Statistics on total expenditures for a population of mental health organizations.

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 Outpatient	149	6,118,415	9,817,762	1.605
6	Multi-service or Substance Abuse	144	15,567,731	44,553,355	2.862
Total		875	11,664,181		

Table 3.2 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 $\frac{n_h}{n} = \frac{W_h S_h/\sqrt{c_h}}{\sum_{h=1}^{H}(W_h S_h/\sqrt{c_h})}$
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	Multi-service or Substance Abuse	1,000	0.386	0.339
Total			1.000	1.000

$(CV_0 \times \bar{y}_U)^2$. Using this to evaluate (3.25) gives sample sizes of 406 and 198 for CV targets of 0.05 and 0.10.

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

(robp:first line 2 line headers) 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)
```

```
str.alloc(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)
str.alloc(Nh = Nh, Sh = Sh, cost = 100000, ch = ch,
          alloc = "totcost")
str.alloc(Nh = Nh, Sh = Sh, cost = 200000, ch = ch,
          alloc = "totcost")
```

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

```
str.alloc(Nh = Nh, Sh = Sh, CV0 = 0.05, ch = ch,
          ybarU = 11664181, alloc = "totvar")
str.alloc(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 <- str.alloc(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 (3.26).

Table 3.3 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 Outpatient	0.089	0.031
6	Multi-service or Substance Abuse	0.406	0.644
Total		1.000	1.000

Example 3.10. Continuing with the previous example, the results of calculating the allocations for the mental health organizations based on the criteria in (3.26) and (3.27) are shown in Table 3.3. These allocations are both more extreme than those in Table 3.2 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 for estimating 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(robp:REF) will cover sample allocation tasks using more than one analysis variable.

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 involves 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 probability proportional to size sampling to models and the use of regression estimators of means and totals. Chapter 14(robp:REF) 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 measure of size (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 probability proportional to size 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(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,

$$\begin{aligned} E_M(y_i) &= \beta x_i \; , \\ V_M(y_i) &= v_i \end{aligned} \tag{3.29}$$

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 (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:

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

(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 Chapter 18(robp:REF).

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 probability proportional to size (Chromy 1979)(robp:BIB). Section 3.7 covers some of the software packages available for selecting samples with varying probabilities.

Determining a Measure of Size

(robp:BOLD VECTORS!)

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 Chapters 9(robp:REF) and 10(robp:REF), 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)(robp:BIB). 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 $E_M\left(y_i\right) = x_i^T \beta$ and $V_M\left(y_i\right) = v_i$ with x_i defined as a vector of x's, β defined as a vector of regression slopes of the same dimension as x_i, and x_i^T is the transpose of the 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 Chapter 14(robp:REF).

A model that may fit some establishment or institutional populations reasonably well has a variance with the form, $V_M\left(y_i\right) = \sigma^2 x_i^\gamma$, 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, $E_M\left(y_i\right)$, 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 $V_M\left(y_i\right)$, regardless of its form. When $V_M\left(y_i\right) = \sigma^2 x_i^\gamma$, the slope in a regression of $\ln\left(e_i^2\right)$ on $\ln\left(x_i\right)$ is an approximate estimate of γ. Henry and

Valliant (2009)(robp:BIB) give more detail along with applications. Two R functions that will iteratively estimate γ are `gamma.fit` along with `gam.est` given in Appendix A(robp:REF). Note that `gam.est` is set up for a regression without an intercept. If an intercept is desired, the matrix $\mathbf{X}$, which is an input to `gamma.fit` must be defined to include a column of 1's. The parameters used by `gamma.fit` 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. 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 y to be the vectors of the 10 values for these units. The matrix $\mathbf{X}$ contains columns for $\sqrt{x}$ and x. To estimate γ, the call to `gamma.fit` and its output are:

```
X <- cbind(sqrt(x), x)
gamma.fit(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 MOS's 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 `gamma.fit` 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 (3.31) looks just like $\hat{\bar{y}}_\pi$ above, the selection probability of unit i is not np_i in with

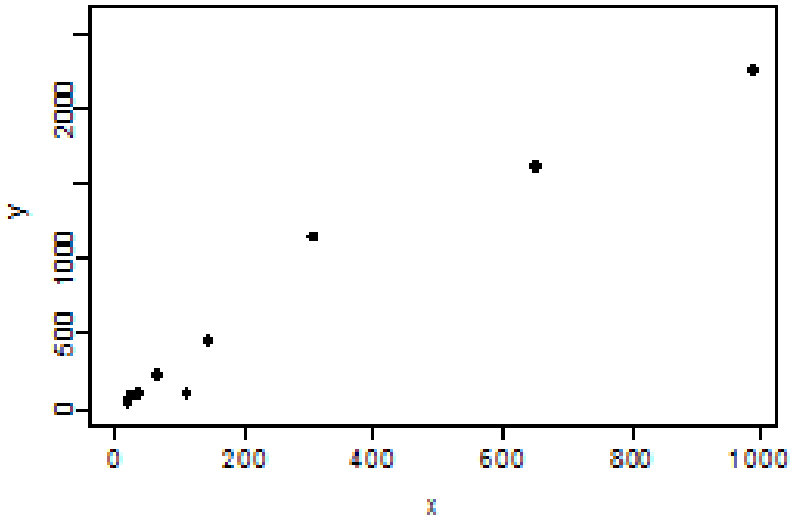

Fig. 3.2 Scatterplot of a sample of $n = 10$ sample units from the hospital population.

replacement sampling; it is actually $1 - (1 - p_i)^n$. The variance of $\hat{\bar{y}}_{pwr}$ in *ppswr* sampling is

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

where T is the population total of y. The obvious advantage of (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, (3.32) can be solved to give the sample size as

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

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's.

Example 3.12. Figure 3.3 plots total expenditures by the number of beds for the 671 organizations in the SMHO population that have non-zero 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 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 (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 (3.32) refers only to the subuniverse of $N-n_t$ non-(take-alls). The result is $V_1 = 9.53703\text{e+}19$. The sample size formula (3.33) is then

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

with $\bar{y}_U = 13,667,706$. For $CV_0 = 0.15$, (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 (3.34) recomputed for the remaining units. A few iterations may be needed to identify all of the take-alls. Alternatively, a cutoff like $x_i \geq 0.8N\bar{x}_U/n$ 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(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(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 Chapter 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(robp:REF) 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

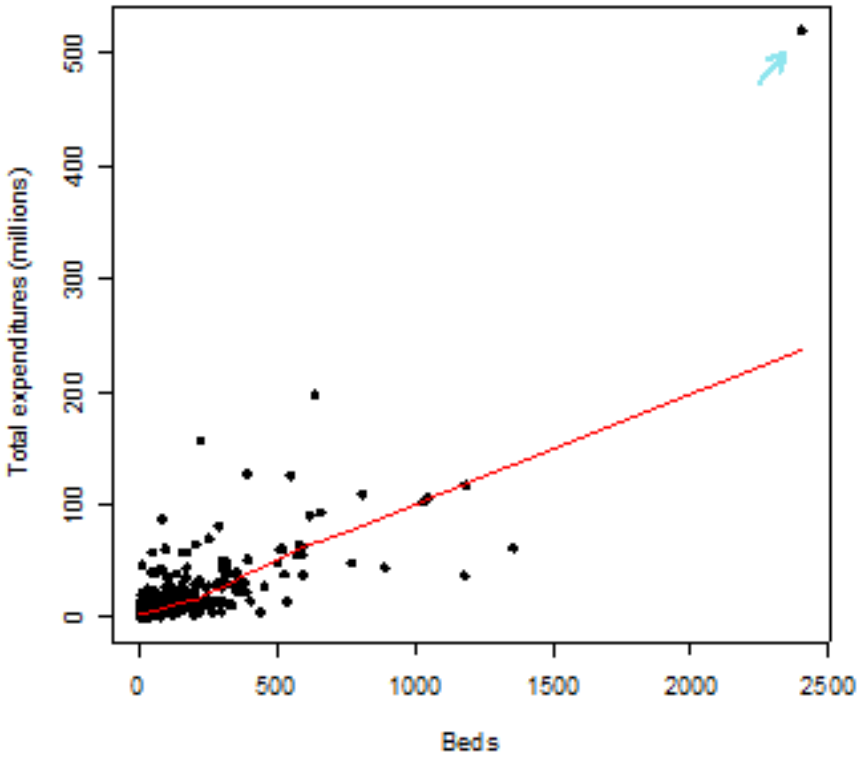

Fig. 3.3 Plot of total expenditures vs. number of beds for the SMHO population. The gray line is a nonparametric smoother (lowess).(robp:Ich seh nur ne rote Linie)

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 (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 MOS's 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 measure of size). 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:

$$\begin{aligned} E_M(y_i) &= \sum_{j=1}^{p} \beta_j x_{ji} \,, \\ V_M(y_i) &= \sigma^2 v_i \end{aligned} \tag{3.35}$$

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 j^{th} 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(robp:REF?). 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 probability proportional to size 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 $\left[E_M\left(\bar{y}_U\right)\right]^2$, we get a kind of relvariance. Setting the result equal to CV_0^2 and solving for n leads to

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

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

$$\begin{aligned} E_M\left(y_i\right) &= \beta_1 \sqrt{x_i} + \beta_2 x_i\ , \\ V_M\left(y_i\right) &= \sigma^2 x_i\ , \end{aligned} \tag{3.38}$$

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 $E_M\left(y_i\right)$. 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, Thm. 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 (3.37). The next example illustrates the calculation with the smho98 population.

Example 3.14. 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 (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 (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

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

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

Example 3.15. 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))
n.cont(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 $k^{th}(k = 10$ 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 chapter 14(robp:REF)).

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 *srswor*, *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 U.S. Internal Revenue Service received over 130 million tax returns for individuals and selected a sample of about 200,000 returns using Poisson sampling (Henry, et al. 2008)(robp:BIB). 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 Sect. 3.4 all involve population parameters. These must be estimated from a previous sample or from a secondary dataset. 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

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

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

In the special case of a binary variable, $\bar{y}_{s_0}$ reduces to the sample proportion p_0 and $\hat{S}^2 = n_0 p_0 \left(1 - p_0\right) / (n_0 - 1)$. 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 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

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

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

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 $s_h^2 = n_{0h} p_{0h} \left(1 - p_{0h}\right) / (n_{0h} - 1)$.

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 (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 (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 $v\left(\hat{t}_\pi\right) = \hat{V}_1/\left(N^2 n\right)$ with $\hat{t}_\pi = n^{-1}\sum_{s_0} y_i/p_i$. The third term on the right-hand side of (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 (3.43) is that it can be negative, which is, of course, impossible for a variance parameter. 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:

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

where $\bar{y}_w = \sum_{s_0} w_i y_i / \sum_{s_0} w_i$. This form of estimator 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. 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 3.4. We compute $\hat{V}\left(\hat{\bar{y}}_{pwr}\right) \equiv \hat{V}_1/\left(N^2 n\right)$ 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* estimate, we first calculate $\hat{V}_1$ in (3.41) as $\hat{V}_1 = 11,001,669,955$. Substituting this value into the $\hat{V}\left(\hat{\bar{y}}_{pwr}\right)$ formula (3.32), we have

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

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 $p_i = x_i\,/\sum_U x_i\,.$

$$\begin{aligned}\hat{V}_1 &= \frac{1}{n}\left(\sum_U \sqrt{x_i}\right)\sum_{s_0}\frac{y_i^2}{\sqrt{x_i}p_i} - \left(\frac{1}{n}\sum_{s_0}\frac{y_i}{p_i}\right)^2 + v\left(\hat{\bar{y}}_{pwr}\right)\\ &= \frac{5992.3}{20}410,727,850 - 319,545^2 + 3,561.6^2\end{aligned}$$

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 3.4 Sample data for 10 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	1239	472
155	402	160	321	1258	474
192	732	227	329	1657	498
200	925	235	354	2116	562
228	632	275	360	1326	584
243	557	300	369	1606	635
253	1226	310	373	1707	670
289	896	378	376	2089	712
297	2190	400	378	1283	760
315	1948	461	381	1239	816

Example 3.17. 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 (3.3). Thus, we need to estimate the unit variance S^2 using (3.44). Evaluating this with the data for the 10 sample hospitals in Table 3.4, 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 section 3.5.2(robp:REF, section?), 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. One of the exercises (robp:REF?) 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 that have the 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

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

(The inequality reverses since $\log(1-p_U)$ is negative.) The table 3.5 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 3.5 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 distribution. 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)(robp:BIB) cover an interesting method known as the "rule-of-three" which derives from the formula $(1-p_U)^n \leq \alpha$ that led to (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 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)(robp:BIB) 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 Chapter 17(robp:REF).

3.5.2 Domain Estimates

Most multipurpose surveys make separate estimates for domains or subpopulations. Kish (1987)(robp:BIB) 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 (e.g., African-Americans over the age of 50 in the U.S.);
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 survey in the U.S., 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. 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 $\times$ 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 five 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 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 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.[9] 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,

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

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 Chapter 11(robp:REF). 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. NHANES III is an excellent example of two-phase sampling.[10] In this national study, conducted 1988-1994, analytic domains were based on race/ethnicity (Black, White and All Other, Mexican-Americans) and age groups that varied depending on the race/ethnicity 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

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

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 (1953, Sect. 4.10), Cochran (1977, Sect. 2.11)). Using this version of S^2, the relvariance of $\hat{t}_d$ is

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

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

$$n = \frac{CV_d^2 + Q_d}{P_dCV_0^2 + \frac{CV_d^2 + Q_d}{N}} \doteq \frac{CV_d^2 + Q_d}{P_dCV_0^2} \tag{3.47}$$

[10] http://www.cdc.gov/nchs/data/series/sr_02/sr02_113.pdf

The second line comes from assuming that the population size N is large. Notice that (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 $\hat{\bar{y}}_d = \hat{t}_d \Big/ \hat{N}_d$ 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 e with $e_i = \delta_i (y_i - \bar{y}_{Ud})$. The approximate variance is then

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

with $S_e^2 = (N-1)^{-1} \sum_U e_i^2$. 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

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

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 (3.47) for estimating the domain total as illustrated in Table 3.6. 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.

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

Table 3.6 Sizes of simple random samples required to achieve a CV of 0.05 for estimated domain totals and means for different sizes of domains. The population size is assumed to be large; domain relvariance is $CV_d^2 = 1$.

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

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

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

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

(robp:break frac after $s_d^2 h$) 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 $S_{dh}^2 = \sum_{i \in U_{dh}} \left(y_{hi} - \bar{y}_{U_{dh}}\right)^2 / (N_{dh} - 1)$ 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

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

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, (3.50) is difficult to use. If $n_{dh} \doteq n_h P_{dh}$, this can be substituted in (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 $S_h^{*2} = \frac{P_{dh}}{P_d^2}\left[S_{dh}^2 + Q_{dh}\left(\bar{y}_{U_{dh}} - \bar{y}_{U_d}\right)^2\right]$. 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 (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 (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 (Research Triangle Institute 2008)(robp:BIB?) 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, survey organizations had to rely on computer programs developed by their own staffs 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. Both the `pps` and `sampling` 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 Website. User-defined functions are easily created as discussed in this and other chapters—see Appendix A(robp:REF) for a complete list of author-defined R functions used in this text.

Example 3.18. We wish to select 10 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`. Because R code is executed line-by-line, the resulting output is shown below following the code executed after the command prompt symbol >.

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]
   ORGID IPCCCENS IPCCMALE IPCCFEML IPCCUNKG
1 010345       28       15       13        0
2 020025       28       15       13        0
3 040061       30       10       20        0
4 040087       40       21       19        0
5 050003       30       16       14        0

#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
```

Example 3.19. 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 probability proportional to size sample within

five design strata with a size measure defined as the square root of the bed size, i.e., $pp(\sqrt{x})$ discussed in Sect. 3.5.2. We will use the R function `ppssstrat` from the `pps` package to draw an (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`[11] 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

[11] http://support.sas.com/documentation/cdl/en/statug/59654/HTML/default/surveyselect_toc.htm

be selected with METHOD=SYS. Probability proportional to size samples are 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)(robp:BIB) 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 Chapters 9(robp:REF) and 10(robp:REF).

Example 3.20. 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 input data file, SMHO98, and two additional variables:

1. SelectionProb – the probability of selection into the sample; and
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 Chapters 14(robp:REF) and 15(robp:REF).

Example 3.21. A $pp(\sqrt{x})$ 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.

Stratum5	SampleSize
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 U.S. Bureau of Labor Statistics, 71 percent of all workers in private industry had access to employer-sponsored medical care plans, 52 percent of all workers participated in medical care plans in March 2006, and 7 percent 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 percentage points.
(c) Margin of error of 3 percentage 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 $\Pr\left(\left|\frac{\bar{y}_s - \bar{y}_U}{\bar{y}_U}\right| \leq e\right) = 1 - \alpha$, 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 margin of error e is set for estimating a proportion.

3.6. Verify formula (3.17) for the required sample size derived from the margin of error 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 margin of error 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 and
(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 3.1. 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 inpatient visits 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 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?
(d) Explain any differences in the results for parts (a) and (c).

3.11. Show that (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, $\hat{V}_1 \big/ N^2 n = [n(n-1)]^{-1}\sum_{s_0}(y_i - \bar{y}_{s_0})^2$.

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 measure of size. 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_pop.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 10 samples of size 20 from the hospital population using probability proportional to the number of beds as in Example 3.15. Calculate the estimate $\hat{V}_1$ in (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 1000 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 non-zero value. For example, the U.S. 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)[12]. Some employees are engaged in qualified research for some percentage of their time (the non-zeros); 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

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

where N_1 is the number of elements with non-zero values, $P = N_1 / N$ is the proportion of elements with non-zero values, $Q = 1 - P$, $\bar{y}_{U1}$ is the mean for elements with non-zero 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 non-zero values. In the example, N_1 would be the number of employees 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 non-zero elements equal to 1, 2, and 4.

3.16. Consider two different sample designs for the smho.N874 population (`smho.N874.RData`). 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 probability

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

proportional to size 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 measure of size 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

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

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 percent 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 $\hat{N}_d = Nn_d / n$.

(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, Hosmer, Klar, and Lwanga (1990), Schlesselman (1982), and Woodward (1992) (robp:BIBTEX).

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 and SAS.

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, *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

State of nature	Decision: Do not reject H_0	Decision: Reject H_0
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 in either direction from $H_0 : \mu = 3$. 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 degrees of freedom for the t are usually based on some rules-of-thumb. One that is often used is

[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 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.

$$df = \text{number of PSUs} - \text{number of strata} . \tag{4.3}$$

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 (e.g., see Rust 1984, 1985; Valliant and Rust 2010). (robp:BIBTEX)

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 this 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*.

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 as case for a 2-sided test is

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

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

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

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:

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

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

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

Notice that we could use the more elaborate notation β_δ since the power depends on the specific value of the alternative.

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 ones that are close. Three things are needed for a sample size calculation based on power:

1. Value of δ;
2. Desired probability $1 - \beta$ of obtaining a significant test result when the true difference is δ; and
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

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

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

$$\begin{aligned} &\Pr\left(t > t_{1-\alpha}(df) \,|\mu = \mu_0 + \delta\right) \qquad (4.4)\\ \doteq &\Pr\left(\frac{\hat{\bar{y}} - \mu_0}{\sqrt{V\left(\hat{\bar{y}}\right)}} - \frac{\delta}{\sqrt{V\left(\hat{\bar{y}}\right)}} > z_{1-\alpha} - \frac{\delta}{\sqrt{V\left(\hat{\bar{y}}\right)}} \middle| \mu = \mu_0 + \delta\right)\\ = &\Pr\left(Z > z_{1-\alpha} - \frac{\delta}{\sqrt{V\left(\hat{\bar{y}}\right)}}\right) \end{aligned}$$

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 $\delta \Big/ \sqrt{V\left(\hat{\bar{y}}\right)}$ 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. 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.

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 $\left(\hat{\bar{y}} - \mu_0\right) \Big/ \sqrt{v\left(\hat{\bar{y}}\right)}$ will have a noncentral t-distribution with noncentrality parameter $\delta \Big/ \sqrt{V\left(\hat{\bar{y}}\right)}$ 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*

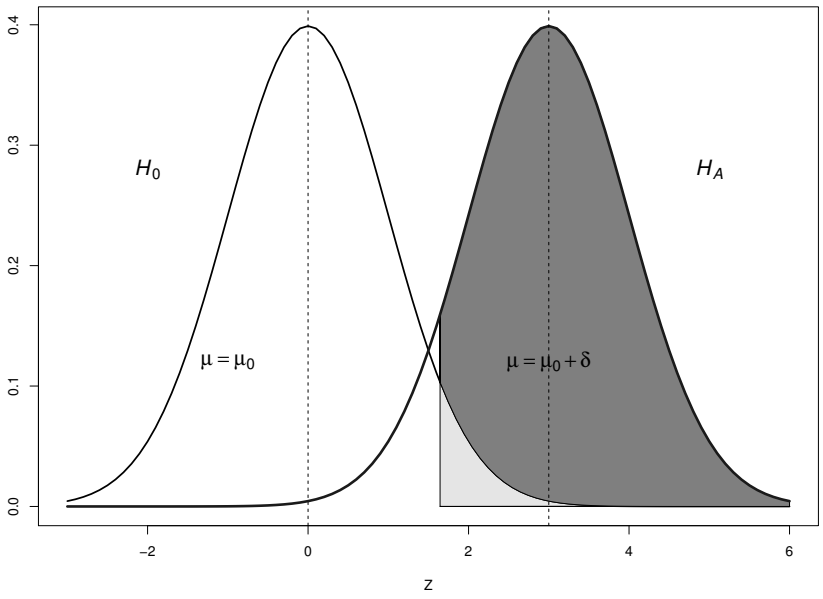

Fig. 4.1 Normal densities of test statistics under H_0 and H_A. $\delta \Big/ \sqrt{V\left(\hat{\bar{y}}\right)}$ is set equal to 3 in this illustration so that $E\{t \,|H_A \text{ is true}\} = 3$. A 1-sided test is conducted at the 0.05 level.

degrees of freedom is greater than $t_{1-\alpha}\,(df)$. This is the method used by the R function `power.t.test` 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 $z_\beta\,(= -z_{1-\beta})$ and solve for the sample size n to obtain

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

For designs other than *srswor*, $V\left(\hat{\bar{y}}\right) = \sigma_y^2/n$ does not hold. A common work-around is to set the solution to (4.5) equal to the effective sample size, defined as $n_{\mathit{eff}} = n/\mathit{deff}$ where $\mathit{deff} = V\left(\hat{\bar{y}}\right)/V_{SRS}\left(\hat{\bar{y}}\right)$, 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. 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 $\mu = \mu_0 + \delta$, 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

$$
\begin{aligned}
&\Pr\left(|t| \le t_{1-\alpha/2}\left(d\hat{f}\right) | \mu = \mu_0 + \delta\right) \\
\doteq &\Pr\left(-z_{1-\alpha/2} \le \frac{\hat{\bar{y}} - \mu_0}{\sqrt{V\left(\hat{\bar{y}}\right)}} < z_{1-\alpha/2} \middle| \mu = \mu_0 + \delta\right) \\
= &\Pr\left(-z_{1-\alpha/2} - \frac{\delta}{\sqrt{V\left(\hat{\bar{y}}\right)}} \le \frac{\hat{\bar{y}} - \mu_0}{\sqrt{V\left(\hat{\bar{y}}\right)}} - \frac{\delta}{\sqrt{V\left(\hat{\bar{y}}\right)}} < z_{1-\alpha/2} - \frac{\delta}{\sqrt{V\left(\hat{\bar{y}}\right)}} \middle| \mu = \mu_0 + \delta\right) \\
= &\Pr\left(-z_{1-\alpha/2} - \frac{\delta}{\sqrt{V\left(\hat{\bar{y}}\right)}} \le Z < z_{1-\alpha/2} - \frac{\delta}{\sqrt{V\left(\hat{\bar{y}}\right)}}\right) \\
= &\Pr\left(Z \le z_{1-\alpha/2} - \frac{\delta}{\sqrt{V\left(\hat{\bar{y}}\right)}}\right) - \Pr\left(Z \le -z_{1-\alpha/2} - \frac{\delta}{\sqrt{V\left(\hat{\bar{y}}\right)}}\right)
\end{aligned}
$$

The power of the test against the alternative is then

$$
\begin{aligned}
&\Pr\left(|t| > t_{1-\alpha/2}\left(d\hat{f}\right) | \mu = \mu_0 + \delta\right) \qquad (4.6) \\
\doteq &1 - \Pr\left(Z \le z_{1-\alpha/2} - \frac{\delta}{\sqrt{V\left(\hat{\bar{y}}\right)}}\right) + \Pr\left(Z \le -z_{1-\alpha/2} - \frac{\delta}{\sqrt{V\left(\hat{\bar{y}}\right)}}\right)
\end{aligned}
$$

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

$$
\begin{aligned}
&\Pr\left(|t| > t_{1-\alpha/2}\left(d\hat{f}\right) | \mu = \mu_0 - \delta\right) \qquad (4.7) \\
\doteq &1 - \Pr\left(Z \le z_{1-\alpha/2} + \frac{\delta}{\sqrt{V\left(\hat{\bar{y}}\right)}}\right) + \Pr\left(Z \le -z_{1-\alpha/2} + \frac{\delta}{\sqrt{V\left(\hat{\bar{y}}\right)}}\right)
\end{aligned}
$$

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

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

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

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

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

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

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

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. Some examples using R are given in Sect. 4.4.

Example 4.3. 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. Figure 4.2 shows screenshots of a spreadsheet that will compute the sample sizes in Examples 4.2 and 4.3. The lower part of the figure shows the formulas while the upper gives numerical results that match those in the examples. The spreadsheet is also available on the book's website. Another excellent reference that combines R and Excel is Heiberger and Neuwirth (2009). (robp:BIBTEX)

	A	B	C	D	E
1		**Example 4.2**		**Example 4.3**	
2		**(a)**	**(b)**	**(a)**	**(b)**
3	?-sided test?	1	1	2	2
4	α	0.05	0.05	0.05	0.05
5	$z_{(1-\alpha/2)}$	1.645	1.645	1.960	1.960
6	β	0.200	0.200	0.200	0.200
7	1-β	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

	A	B	C	D	E
1		**Example 4.2**		**Example 4.3**	
2		**(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. 4.2 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

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

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

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

with $\hat{\bar{d}} = \hat{\bar{x}} - \hat{\bar{y}}$, $v\left(\hat{\bar{d}}\right) = v\left(\hat{\bar{x}}\right) + v\left(\hat{\bar{y}}\right) - 2cov\left(\hat{\bar{x}}, \hat{\bar{y}}\right)$ 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 non-overlapping 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 non-overlapping 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 ($H_0 : \mu_X \leq \mu_Y$) 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 $\delta \Big/ \sqrt{V\left(\hat{\bar{d}}\right)}$ instead of 0. Letting $\mu_D = \mu_X - \mu_Y$, the probability that t_d is in the rejection region is then

$$\Pr\left\{t_d > z_{1-\alpha} \,|\mu_D = \delta\right\} = \Pr\left(t_d - \frac{\delta}{\sqrt{V\left(\hat{\bar{d}}\right)}} > z_{1-\alpha} - \frac{\delta}{\sqrt{V\left(\hat{\bar{d}}\right)}} \middle| \mu_D = \delta\right) \tag{4.10}$$

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

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 $V\left(\hat{\bar{d}}\right) = \sigma_d^2/n$ 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

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

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 $\sigma_d^2 = 2\sigma_0^2\left(1 - \rho\right)$, 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

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

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 4.3 graphs the power in a two-sided test of $H_0 : \mu_D = 0$ vs. $H_A : |\mu_D| = \delta$ 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. 4.3.

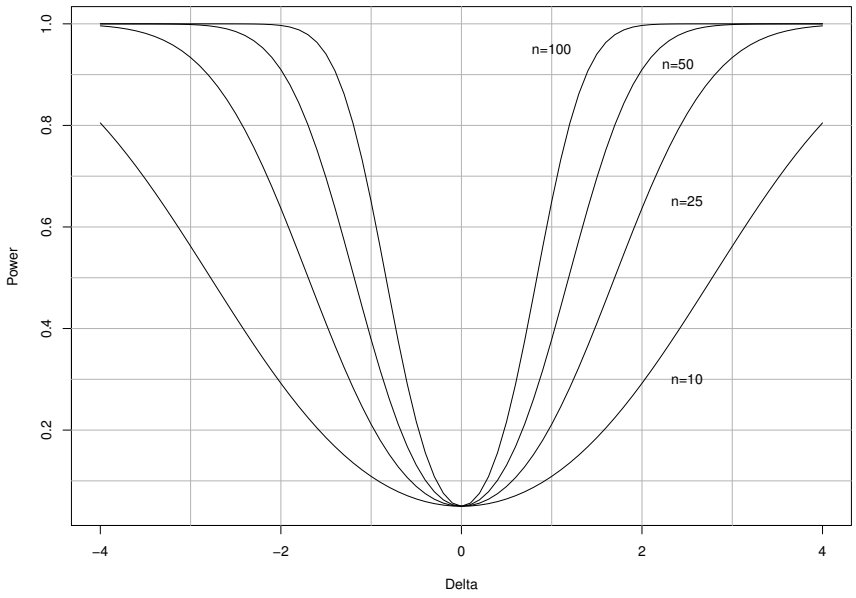

Fig. 4.3 Power for sample sizes of $n = 10, 25, 50, 100$ in a two-sided test of $H_0 : \mu_D = 0$ vs. $H_A : |\mu_D| = \delta$ ($\alpha = 0.05$, $\sigma_d = 3$).

Partially Overlapping Samples

The case of partially overlapping samples can also be handled (e.g., see Woodward 1992) (robp:BIB). 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 or planned sample rotation. 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 $s_2 \cup s_{12}$. 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

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

After some calculation the variance can be expressed as

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

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

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

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

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

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

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

Note that if $n_1 = n_2$, then $r = 1$ and (4.15) equals (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

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

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

$$\begin{aligned}
&\Pr\left(t_{\Delta p} > z_{1-\alpha} \,|P_2 - P_1 = \delta\right) \qquad (4.17)\\
&= \Pr\left(p_1 - p_2 > z_{1-\alpha}\sqrt{2\bar{p}\left(1 - \bar{p}\right)/n}\,\middle|\, P_2 - P_1 = \delta\right)\\
&\doteq \Pr\left(\frac{p_1 - p_2}{\sqrt{\left(P_1 Q_1 + P_2 Q_2\right)/n}} - \frac{\delta}{\sqrt{\left(P_1 Q_1 + P_2 Q_2\right)/n}} > \frac{z_{1-\alpha}\sqrt{2\bar{P}\left(1 - \bar{P}\right)/n} - \delta}{\sqrt{\left(P_1 Q_1 + P_2 Q_2\right)/n}}\,\middle|\, P_2 - P_1 = \delta\right)\\
&= \Pr\left(Z > \frac{z_{1-\alpha}\sqrt{2\bar{P}\left(1 - \bar{P}\right)/n} - \delta}{\sqrt{\left(P_1 Q_1 + P_2 Q_2\right)/n}}\right) .
\end{aligned}$$

(robp:break third line after ¿) 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`, 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

$$\rho = \left(P_{xy} - P_x P_y\right) \Big/ \left[P_x\left(1 - P_x\right) P_y\left(1 - P_y\right)\right]^{1/2} \quad .$$

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 $P_{xy} \leq \min\left(P_x, P_y\right)$. 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 $\Delta = \left[P_x\left(1 - P_x\right) P_y\left(1 - P_y\right)\right]^{1/2}$.

Arcsine square root transformation

When a characteristic is extremely rare or prevalent, the normal approximation for (4.16) can be poor. One rule-of-thumb is that np and $n\left(1 - p\right)$ 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)(robp:BIB), 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

$$\begin{aligned} &\Pr\left(t_\phi > z_{1-\alpha} \,|P_1 - P_2 = \delta\right) \qquad (4.20)\\ &= \Pr\left(t_\phi - \frac{\delta_\phi}{\sqrt{V(\phi_1 - \phi_2)}} > z_{1-\alpha} - \frac{\delta_\phi}{\sqrt{V(\phi_1 - \phi_2)}} \middle| P_1 - P_2 = \delta\right) \\ &\doteq \Pr\left(Z > z_{1-\alpha} - \delta_\phi \sqrt{2n}\right) \end{aligned}$$

(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

$$V(\phi_1 - \phi_2) = \frac{2}{n}\frac{1}{\bar{P}\bar{Q}} \text{ under } H_0 \text{ and}$$

$$V(\phi_1 - \phi_2) = \frac{1}{n}\left(\frac{1}{P_1Q_1} + \frac{1}{P_2Q_2}\right) \text{ under } H_A,$$

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

$$\Pr\left(t_\phi > z_{1-\alpha} \,|P_2 - P_1 = \delta\,\right) \doteq \Pr\left(Z > \frac{z_{1-\alpha}\sqrt{2\left[n\bar{P}\left(1-\bar{P}\right)\right]^{-1}} - \delta_\phi}{\sqrt{n^{-1}\left[\left(P_1Q_1\right)^{-1} + \left(P_2Q_2\right)^{-1}\right]}}\right)$$

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 gives

$$n = \left(\frac{z_{1-\alpha}\sqrt{2V_0} - z_\beta\sqrt{V_A}}{\delta_\phi}\right)^2 \tag{4.23}$$

where $V_0 = \left[\bar{P}\left(1-\bar{P}\right)\right]^{-1}$ and $V_A = \left(P_1Q_1\right)^{-1} + \left(P_2Q_2\right)^{-1}$. 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

$$P_1 - P_2 = P_2 (R - 1) \ .$$

Consequently, if a sample size is desired to detect a relative risk of R^*, this corresponds to detecting a difference of $\delta = P_2 (R^* - 1)$. 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

$$\hat{\delta}_E = \frac{\bar{x}_1 - \bar{x}_2}{s} \ . \tag{4.25}$$

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) (robp:BIB). The general idea of effect size is due to Cohen (1988) (robp:BIB). 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 3300*sqrt(1000)/sqrt(1000) = 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 non-central 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):

$$n = \frac{400\,(1.645 + 0.84)^2}{5^2}$$

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` developed for the book and shown in Appendix A(robp:REF). 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 A(robp:REF) 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 Power	N Per Group
800	1892

Exercises

4.1. The average disposable employment income per worker in Mexico in 2002 was approximately \$6100 U.S. 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 \$7000. 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 percent 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 percent response rate for females and a 60 percent 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 percent 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 is 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 are 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 a 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 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, $\hat{\bar{d}} = \hat{\bar{x}} - \hat{\bar{y}}$, is $V\left(\hat{\bar{d}}\right) = \frac{\sigma_x^2}{n_1} + \frac{\sigma_y^2}{n_2} - 2\sigma_{xy}\frac{n_{12}}{n_1 n_2}$ as shown in (4.12). When $n_{12} = \gamma n_1$ and $r = n_1/n_2$, show that this reduces to $V\left(\hat{\bar{d}}\right) = n_1^{-1}\left[\sigma_x^2 + r\sigma_y^2 - 2\gamma r\sigma_{xy}\right]$. When $\sigma_x^2 = \sigma_y^2 = \sigma_0^2$, show that $V\left(\hat{\bar{d}}\right) = \sigma_0^2 n_1^{-1}\left[1 + r\left(1 - 2\gamma\rho\right)\right]$.

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 over-runs 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 U.S. military to refer to proposed training and logistics schedules (Freund 1994, Dantzig 1963) (robp:BIB). 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 CV's 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; and
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 function `constrOptim.nl`, 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. 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 Chapter 3, 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 CV's of the estimated means for variable j for stratum h (a design domain as described in Chapter 3) and cross-strata domains (called a cross class in Chapter 3)

1. a. 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 $relvar\left(\hat{\bar{y}}_j\right) = V\left(\hat{\bar{y}}_j\right)/\bar{y}_{Uj}^2$.

1. a. Subject to the constraints:
2. $n_h \leq N_h$ for all h;
3. $n_h \geq n_{\min}$, a minimum sample size in every stratum ($n_{\min} \geq 2$ in general);
4. $\left[CV\left(\bar{y}_{j,sh}\right)\right]^2 \leq \left(CV_{0jh}\right)^2$ for certain strata and variables;
5. $\left[CV\left(\hat{\bar{y}}_{dj}\right)\right]^2 \leq \left(CV_{0dj}\right)^2$ for certain domains and variables; and
6. $C = C_0 + \sum_{h=1}^{H} c_h n_h$

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 $[CV(\bar{y}_{j,sh})]^2$ and $\left[CV\left(\hat{\bar{y}}_{dj}\right)\right]^2$. In almost all nonlinear problems, there are no closed-form, exact solutions like the ones we noted for stratified sampling in section 3.4.2. 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.

1. a. **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 Table 5.1 gives stratum means, standard deviations, and proportions for an artificial population of business establishments. The U.S. 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:

1. Budget on variable costs = \$300,000 U.S.;
2. $CV \leq 0.05$ on estimated total number of employees;
3. At least 100 establishments are sampled in each sector, $n_h \geq 100$;
4. The number sampled in each stratum is less than the population count, $n_h \leq N_h$;
5. $CV \leq 0.03$ on estimated total number of establishments claiming the research tax credit; and
6. $CV \leq 0.03$ 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 to estimated totals and means will be the same (as discussed in section 3.4.1). Recall that the relvariance of an estimated total in an *stsrswor* is

$$relvar\left(\hat{T}\right) = T^{-2} \sum_h N_h \left(\frac{N_h}{n_h} - 1\right) S_h^2$$

[1] $http://www.solver.com/excel-solver.htm$

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 website. A screenshot of the spreadsheet showing row and column labels is in Figure 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 CV's 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 Figure 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 <= (**??**)^2 (relvariance of estimated total employees)
$N$11 <= (**??**)^2 (relvariance of estimated total number of establishments claiming the research tax credit)
$O$11 <= (**??**)^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 (Figure 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 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 Example 5.2.

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, 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 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.

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	Establishments N_h	ε_h	Revenue (millions)	Employees	Revenue (millions)	Employees	Claimed research credit	Had offshore affiliates
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

Figure 5.1 Excel spreadsheet set up for use with Solver.

	A	B	C	D	E	F	G	H	I	J	K	L
1					Population means		Population standard dev.		Population proportion			
2	h	Sector	Establishments Nh	ch	Revenue (millions)	Employees	Revenue (millions)	Employees	Claimed research credit	Had offshore affiliates	nh	Revenue Nh*(Nh/nh - 1) * Sh^2
3	1	Manufacturing	8,221	120	85	511	170.0	255.50	0.8	0.86	414	2,524,570,862
4	2	Retail	11,738	80	11	21	8.8	5.25	0.2	0.83	317	32,700,936
5	3	Wholesale	4,333	80	23	70	23.0	35.00	0.5	0.83	124	78,108,889
6	4	Service	22,869	90	17	32	25.5	32.00	0.3	0.21	1,395	227,670,396
7	5	Finance	5,467	150	126	157	315.0	471.00	0.9	0.77	597	4,425,615,684
8												
9	Pop Total		58,668		1,834,157	5,316,946			21,253.90	9,664.87	2,847	7,289,665,967
10	Pop Mean				36	105			0.420	0.195		
11										variance of t.hat		0.00219668
12	Budget			$ 300,000						CV of t.hat		0.84665
13	Total sample cost			$ 300,000								

Figure 5.2 Screenshot of the Excel Solver dialogue screen.

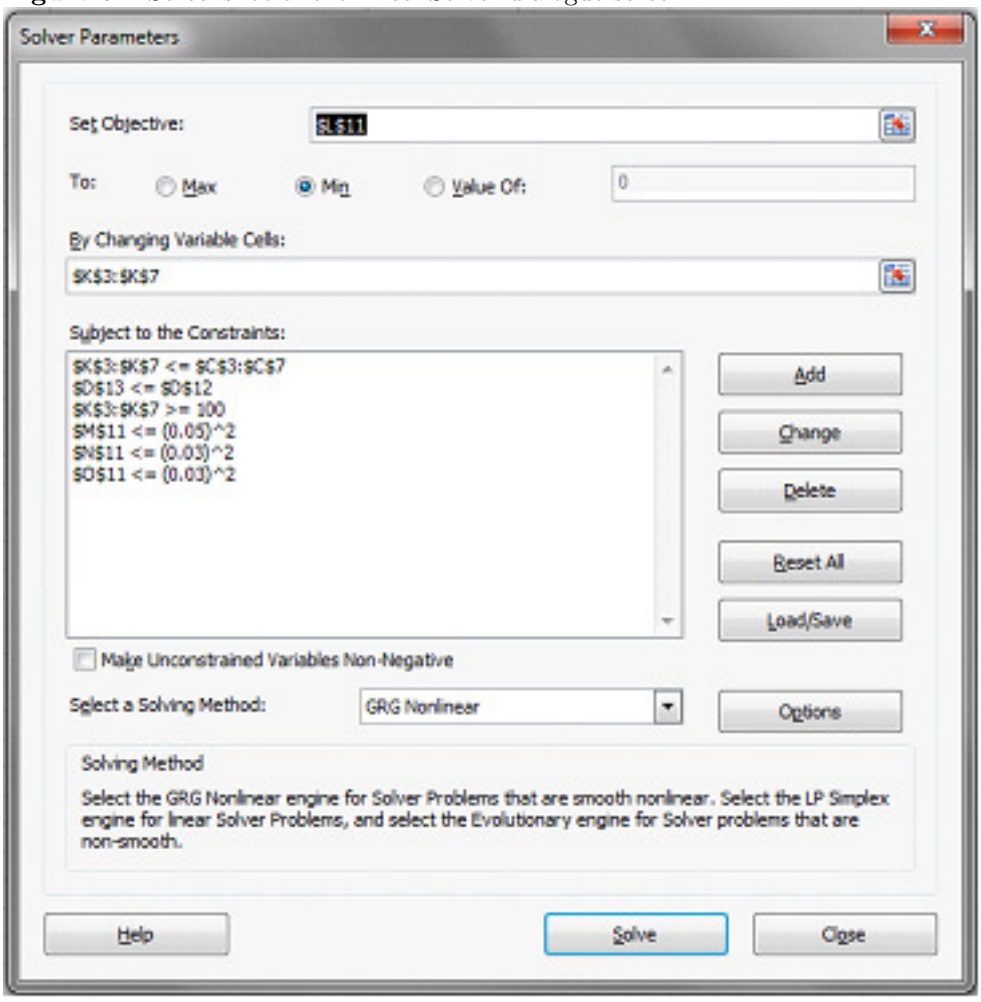

Figure 5.3 Screenshot of the Change Constraint dialogue screen.

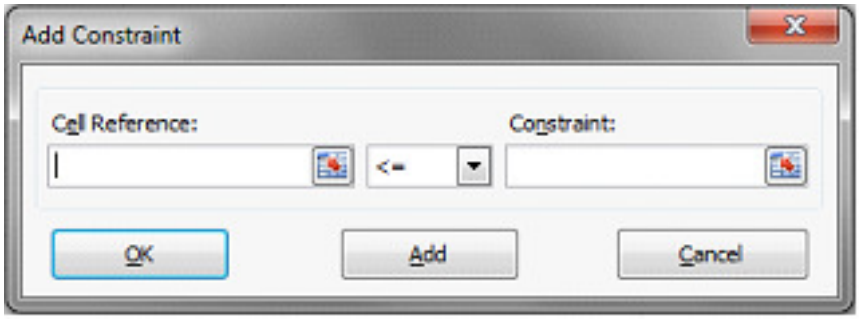
Add Constraint
Cell Reference:
<=
Constraint:
OK
Add
Cancel

Figure 5.4 Solver Options window where tuning parameters can be set and models saved.

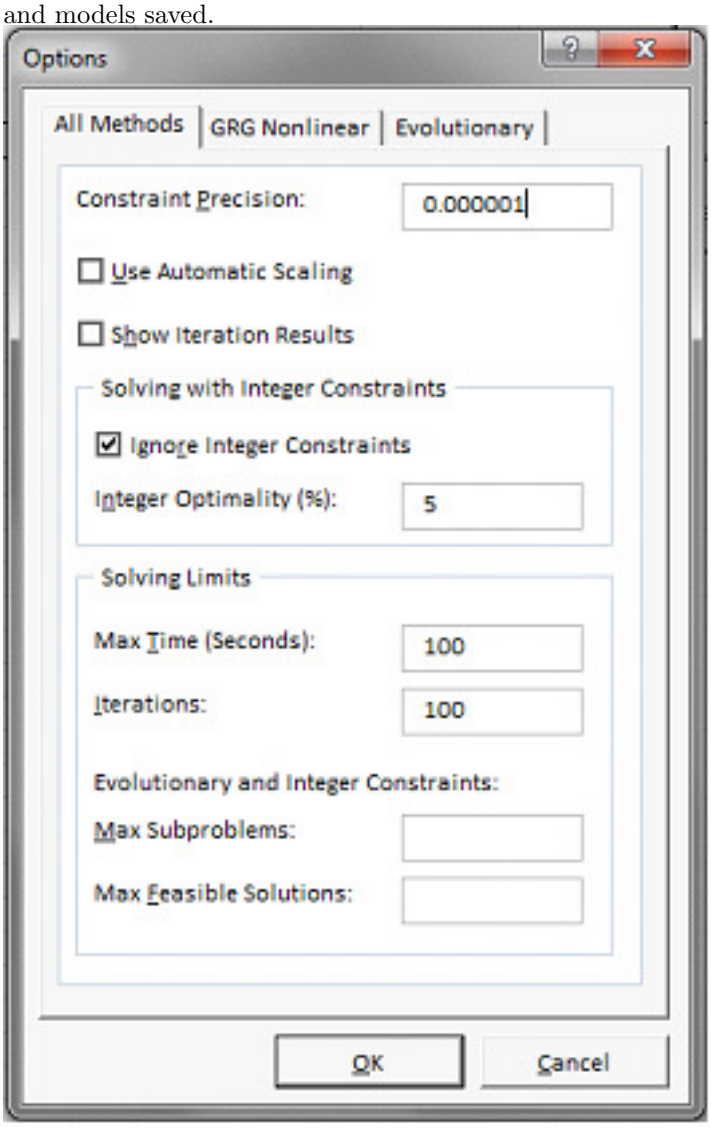

Table 5.2 Solution to the optimization problem in Example 5.2.

Stratum	Sector	n_h
1	Manufacturing	414
2	Retail	317
3	Wholesale	124
4	Service	1,395
5	Finance	597
Total		2,847

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 Example 5.2. First, the original and final values of the objective function are listed. Initial values for the n_h's are needed to get the algorithm started; n_h=500 was used for all strata in this case. In 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 Figure 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 non-binding at the solution; and the slack value. The Name column is the combination of the row and column label for the constraint, e.g., "relvariance of t.hat Offshore Nh*(Nh/nh - 1) * Shˆ2". 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 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 Figure 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*. In Example 5.2, 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 Example 5.2 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 -1.644E-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(-1.644E-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)(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 Figure 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 Example 5.2, 10 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 optimization problems that may be tried. As in all applications, good bookkeeping is a critical part of good organization.

Figure 5.5 Solver's Answer Report for Example 5.2

Figure 5.6 Solver's Sensitivity Report for Example 5.2

Starting Values. Finally, we note that the solution may be sensitive to the starting values of the decision variables. In Example 5.2, 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 Chapter 1. It is advisable to find solutions 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. 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 workaround is set one cell equal to $\max_h (n_h/N_h)$ and 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.

1. a. **SAS PROC NLP**

Multicriteria optimization in SAS can be conducted using the procedures proc nlp (nonlinear programming) 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

[2] *More detailed help is available for this and all other options at www.solver.com/excel2010/solverhelp.htm.*

$$\begin{array}{ll} c_i(x) = 0 & i = 1, \ldots, m_1 \\ c_i(x) \geq 0 & i = m_1, \ldots, m_1 + m_2 \\ \ell_j \leq x_j \leq u_j & j = 1, \ldots, p \end{array}$$

The vector $\mathbf{x}$ contains the decision variables; the $c_i(x)$ 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 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:

1. a. there are no specific limits on numbers of decision variables and constraints other than those imposed by computer memory and hard drive size;
   b. detailed documentation is produced in a SAS log file; and
   c. 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 SASprogram. Detailed information on more advanced techniques in proc nlp (and other procedures) may be obtained from the SAS OnlineDoc website[3]. Once at the SAS OnlineDoc website, 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 user's guide. For your own special purpose programs, choice (i) should be sufficient. The comments should include a header giving:

1. a. Purpose of the program
   b. Name of programmer
   c. Date written
   d. Date(s) revised and changes made in each revision.

Choices (ii) and (iii) are used by R, SAS, Stata, and other multipurpose packages. We discuss program documentation in more detail in the quality control chapter (Chapter 18).

Example 5.3 The SAS 9.2 proc nlp code, program log, and output file used in this example are located in the *Example 5.3 (NLP)* files (*.sas*, *.log*, and *.lst*

[3] *http://support.sas.com/documentation/*

files, respectively) on the book's website. The code is also shown in Figure 5.7.

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] array), 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 "hardcoded" 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], array 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 $\sum_{i=1}^{4} cost[\mathrm{i}] \times n[\mathrm{i}]$.

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 section 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 multi-component 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 website for more details). We chose the Nelder-Mead simplex technique (TECH=nmsimp) because of the nonlinear constraints we are required to impose (see, e.g., constraint (iii) in section 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 5.3 (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) and sum to an overall sample size of 2,848 after rounding up each value. This sample allocation results in a study budget slightly more than the specified $300,000 (ACT 2.2737E-12). The linear budget constraint is called active (ACT) because it is violated, but the amount of the violation is trivial. There is also a minor violation in the CV for the fourth variable, offshore affiliates

(relvar4_L 0.000900 -223E-19 Active NLIC). We also note that the estimated relvariance for the total amount of revenue is given by the objective function (Value of Objective Function = 0.0021705236). Taking the square root gives the CV of total revenues of about 4.7%, which is larger than that of the other variables.

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 (**??**) greater than 100
Upper Bounds	5	Sample sizes (**??**) less than pop sizes
Linear Constraints	1	Cost model
Nonlinear Constraints	3	Constraints on three CVs

Figure 5.7 SAS 9.2 proc nlp code for the optimization problem in Example 5.3.

Table 5.4 Summary of results for Solver, proc nlp, proc optmodel, and constrOptim.nl optimization solutions.

h		Excel Solver (init=500)	SAS NLP[1] (init=500)		SAS OPTMODEL[1] with SQP (init=100)		constrOptim.nl.R (init=1100)	
		nh	**nh**	***Diff***	**nh**	***Diff***	**nh**	***Diff***
1	**Sector Manufacturing**	414	413	*-1*	369	*-45*	430	*16*
2	**Retail**	317	318	*1*	367	*49*	233	*-84*
3	**Wholesale**	124	124	*0*	100	*-23*	114	*-10*
4	**Service**	1,395	1,397	*2*	1,386	*-9*	1,535	*140*
5	**Finance**	597	596	*-1*	625	*28*	550	*-47*
		2,847	2,848	*1*	2,846	*-1*	2,862	*15*
		CV	**CV**	***% RelDiff***	**CV**	***% RelDiff***	**CV**	***% RelDiff***
1	**Revenue (millions)** [2]	4.65%	4.66%	0.09%	4.69%	*0.9%*	4.75%	2.2%
2	**Employees**	2.39%	2.39%	0.07%	2.41%	*0.9%*	2.44%	2.1%
3	**Research credit**	2.09%	2.08%	-0.11%	2.09%	*0.4%*	2.19%	4.8%
4	**Offshore affiliates**	3.00%					3.00%	0%
	Objective function	0.217%	0.217%	*0.2%*	0.220%	*1.7%*	0.226%	2.2%

[1] *The procedures were implemented in SAS 9.2.*
[2] *Minimized in the optimization*

1. a. **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 non-linear optimization. The optmodel procedure has many of the same advantages noted for proc nlp. This newer SAS procedure uses "optmodel language" which is advertised as enabling a quick translation of an optimization "word problem" into executable program code. However, the non-linear optimization techniques currently listed for this procedure are fewer than those specified for proc nlp.

Example 5.4 We recast the proc nlp code presented in Example 5.3 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 5.4 OptModel* files on the book's website. The code is also shown in Figure 5.8.

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_54 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 non-linear 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, initializes 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 optimization 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 1.7% higher 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 a difference exists 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., sensitivity analysis) is always a useful practice.

The last section of 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, and then to randomly select the cases from the sampling frame. Without this text file, statisticians must, for example, cut-and-paste the optimization results into the sampling program—a problem when the optimization 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.

Figure 5.8 SAS proc optmodel code for the optimization problem in Example 5.4.

```
/*****************************************************************/
/* Program: Example 5.4 (OptModel).sas */
/* Date: 10/17/10 */
/* Author: J.Dever */
/* Purpose: Solve example optimization problem. */
/*****************************************************************/
options nocenter orientation=portrait;

TITLE1 "Example 5.4";

*****************************************************************;
Title2 "Load Information";
*****************************************************************;
DATA Example_54;
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_54;
SET Example_54;
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_54 UNIFORM NOOBS; RUN;

*****************************************************************;
Title2 "Sample Allocation - Initial Solution";
*****************************************************************;
PROC OPTMODEL;

*_____ LOAD PARAMETERS _____*;
*Stratum frame counts;
NUMBER Nh{1..5};
```

1. a. **R constrOptim.nl**

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, Nielsen, and Tingleff 2004). This algorithm is different from the ones in Excel Solver and SAS. Figure 5.9 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

Figure 5.9 constrOptim.nl code for the optimization problem in Example 5.2.

```
#*****************************************************************************
# FILE: constrOptim.example.R
# PURPOSE: Use constrOptim.nl to solve allocation problem in Example 5.2
# DATE: 9/14/09
# AUTHOR: R. Valliant
#*****************************************************************************
```

```
require(alabama)
require(numDeriv) # need to have "numDeriv" package installed
# 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 varsOptim.nl
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[2*length(nh) + 4] <- budget - sum(nh * ch)
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, # default is TRUE, using FALSE is worse
method = "BFGS" # default
# method = "Nelder-Mead" # worse objective value than BFGS
# method = "CG" # objective value about same as BFGS but slower
execution
)
)
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 Figure 5.9. 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 Example 5.2 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 percent higher than the 0.00216695obtained with Solver and proc nlp.

1. a. 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 tradeoffs to clients. We recommend that researchers establish and maintain an accounting system to document:

1. initial values set for the optimization problem;
2. optimization results such as attained constraints and decision-variable values;

3. reasons for changing the optimization components; and
4. 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

1. a. A researcher would like to survey the mathematics teachers in the elementary and secondary schools in CityMontgomery, Howard, and CityPrince George's counties in the state of placeStateMaryland. 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 percentage 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?

1. a. 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 CV's 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?

1. a. Re-solve Example 5.2 with the following constraints:
2. Budget on variable costs = $300,000;
3. $CV \leq 0.05$ on estimated total number of employees;
4. At least 100 establishments are sampled in each sector;
5. The number sampled in each stratum is less than the population count, $n_h \leq N_h$;
6. $CV \leq 0.03$ on estimated total number of establishments claiming the research tax credit; and

7. $CV \leq 0.05$ 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.

1. a. Re-solve 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.

1. a. Determine the allocation to strata in Example 5.2 based on the following set up. Minimize $\Phi = 0.75 relvar\left(\hat{T}_{rev}\right) + 0.25 relvar\left(\hat{T}_{emp}\right)$, 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
   b. Sample at least 200 establishments in each stratum;
   c. The number sampled from a stratum should be less than 20% of the stratum population;
   d. The CV's on the estimated total numbers of establishments claiming the research credit and having offshore affiliates should be at most 0.02.
   e. The budget is $450,000.

Part II
Multi-stage Designs

Part III
Survey Weights and Analyses

Part IV
Other Topics

Appendix A
R Functions

This is just a test.

References

Brown L, Cai T, Das Gupta A (2001) Interval estimation for a binomial proportion. Statistical Science 16:101–133

Cochran WG (1977) Sampling Techniques. John Wiley & Sons, Inc., New York

Hansen MH, Hurwitz WH, Madow WG (1953) Sample Survey Methods and Theory, Volume I. John Wiley & Sons, Inc., New York

Isaki CT, Fuller WA (1982) Survey design under the regression superpopulation model. Journal of the American Statistical Association 77(377):89–96

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

Kott PS (1988) Model-based finite population correction for the horvitz-thompson estimator. Biometrika 75:797–799

Lohr SL (1999) Sampling: Design and Analysis. Duxbury Press, New York

Royall RM (1986) The effect of sample size on the meaning of significance tests. The American Statistician 40:313–315

Särndal CE, Swensson B, Wretman J (1992) Model Assisted Survey Sampling. Springer-Verlag, Inc., New York

Valliant R, Dorfman AH, Royall RM (2000) Finite Population Sampling and Inference: A Prediction Approach. John Wiley & Sons, Inc., New York