Survey Weights: A Step-by-Step Guide to Calculation

RICHARD VALLIANT
Universities of Michigan & Maryland

JILL A. DEVER
RTI International (Washington, DC)

A Stata Press Publication
StataCorp LLC
College Station, Texas

Published by Stata Press, 4905 Lakeway Drive, College Station, Texas 77845
Typeset in LaTeX 2ε
Printed in the United States of America
10 9 8 7 6 5 4 3 2 1

Print ISBN-10: 1-59718-260-5
Print ISBN-13: 978-1-59718-260-7
ePub ISBN-10: 1-59718-261-3
ePub ISBN-13: 978-1-59718-261-4
Mobi ISBN-10: 1-59718-262-1
Mobi ISBN-13: 978-1-59718-262-1

Library of Congress Control Number: 2017960405

Acknowledgments

We are indebted to several people who have answered questions and encouraged us in the writing of this book. Jeff Pitblado of StataCorp programmed `svycal`, which is a new Stata procedure that can handle raking, poststratification, general regression, and more general calibration estimation. He also answered many specific Stata questions. This book would not have been possible without him.

Matthias Schonlau at the University of Waterloo provided valuable assistance on how to use his `boost` plug-in and how to tune parameters in boosting. Nicholas Winter helped us several times with questions about his `survwgt` package, which seems to get far less publicity than it deserves. Stas Kolenikov advised us on Stata's general capabilities and on his `ipfraking` raking procedure, which is also a useful tool for computing survey weights.

We thank Frauke Kreuter for many things. Her boundless energy and endless fount of ideas have pushed us along for years. Finally, we thank our spouses, Carla Maffeo and Vince Iannacchione, for their support throughout this several-year project.

Richard Valliant
Jill A. Dever

November 2017

Contents

(Pages omitted)

Preface

Many data analysts use survey data and understand the general purpose of survey weights. However, they may not have studied the details of how weights are computed, nor do they understand the purpose of different steps used in weighting. *Survey Weights: A Step-by-step Guide to Calculation* is intended to fill these gaps in understanding. Throughout the book, we explain the theoretical rationale for why steps are done. Plus, we include many examples that give analysts tools for actually computing weights themselves in Stata.

We assume that the reader is familiar with Stata. If not, Kohler and Kreuter (2012) provide a good introduction.

Finally, we also assume that the reader has some applied sampling experience and knowledge of "lite" theory. Concepts of with-replacement versus without-replacement sampling and single- versus multistage designs should be familiar. Sources for sampling theory and associated applications abound, including Valliant, Dever, and Kreuter (2013), Lohr (2010), and Särndal, Swensson, and Wretman (1992), to name just a few.

Structure of the book

When faced with a new dataset, it is good practice to ask yourself a few questions before analyzing the data. For example,

- Am I dealing with a sample, or does the dataset contain a whole population?
- If it is a sample, how was it selected?
- What is my goal for the analysis? Am I trying to draw inference to the population?
- Do I need to weight my sample to project it to the population?
- Do I need to weight my data to compensate for the fact that the sample does not correctly cover the desired population?

Some datasets you encounter might already contain weights, and it is useful to understand how they were constructed. If you collect data yourself, you might need to construct weights on your own. In both cases, this book will give useful guidance, both for the construction and for the use of survey weights. This book can be read straight through but can also serve as a reference for specific procedures you may need to understand. You can skip around to particular topics and look at the examples for useful code.

We start our book with a general introduction to survey weighting in chapter 1. Weights are intended to project a sample to some larger population. The steps in weight calculation can be justified in different ways, depending on whether a probability or nonprobability sample is used. An overview of the typical steps is given in this chapter, including a flowchart of the steps.

Chapter 2 covers the initial weighting steps in probability samples. The first step is to compute base weights calculated as the inverse of selection probabilities. In some applications, because of inadequate information, it is unclear whether some sample units are actually eligible for the survey. Adjustments can be made to the known eligible units to account for those with an unknown status.

Most surveys suffer from some degree of nonresponse. Chapter 3 reviews methods of nonresponse adjustment. A typical approach is to put sample units into groups (cells) based on characteristics of the units or estimates of the probabilities that units respond to the survey. This chapter also covers another option for cell creation—using machine learning algorithms like CART, random forests, or boosting to classify units.

Chapter 4 covers calibration or adjusting weights so that sample estimates of totals for a set of variables equal their corresponding population totals. Calibration is an important step in correcting coverage problems and nonresponse and, in addition, can also reduce variances.

Chapter 5 discusses options for variance estimation, including exact formulas, linearization, and replication. Using multiple adjustments in weight calculation, as described in the previous chapters, does affect the variance of point estimates of descriptive quantities like means and totals. We illustrate how these multiple effects can be reflected using replication variances.

Not all sets of survey data are selected via probability samples. Even if the initial sample is probability, an investigator often loses control over which units actually provide data. This is especially true in the current climate, in which people, businesses, and institutions are progressively becoming more resistant to cooperating. Chapter 6 describes methods to weight nonprobability samples. The general thinking about estimating propensities of cooperation and using calibration models, covered in chapters 3 and 4, can be adapted to the nonprobability situation.

Chapter 7 covers a few special situations. Normalized weights are scaled so that they sum to the number of units in the sample—not to an estimate of the population size. Although we do not recommend them, normalized weights are used in some applications, particularly in public opinion surveys. Other topics in this chapter include datasets with multiple weights, two-phase sampling, and weights for composite estimation. Some survey datasets come with more than one weight for each case, especially when subsamples of units are selected for different purposes. Two-phase sampling is often used when more intensive efforts are made to convert nonrespondents for a subsample of cases. Composite weighting is used to combine different samples from different frames such as persons with landline telephones and persons with cell phones. This chapter also covers

whether to use survey weights when fitting models. We describe the issues that need to be considered and give some analyses that can be done when deciding whether to use weights in fitting linear and nonlinear models from survey data.

Chapter 8 covers the unexciting but essential procedures needed for quality control when computing survey weights. An orderly system needs to be laid out in advance to guide the sequence of weighting steps, to list quality checks that will be made at every step, and to document the entire process.

Data files and programs for this book

The data and program files used in the examples are available on the Internet. You can access these files from within Stata or by downloading a zip archive. For either method, we suggest that you create a new directory and download the materials there.

- If the machine you are using to run Stata is connected to the Internet, you can download the files from within Stata. To do this, type the following commands in the Stata Command window:

```
. net from http://www.stata-press.com/data/svywt/
. net describe svywt
. net install svywt
. net get svywt
```

  Notice that the statements above are prefaced by "." as in the Stata Results window. We use this convention throughout the book.

- The files are also stored as a zip archive, which you can download by pointing your browser to http://www.stata-press.com/data/svywt/svywt.zip.

  To extract the file `svywt.zip`, create a new folder, for example, `svywt`, copy `svywt.zip` into this folder, and unzip the file `svywt.zip` using any program that can extract zip archives. Make sure to preserve the subdirectory structure contained in the zip file.

Throughout the book, we assume that your current working directory (folder) is the directory where you have stored our files. This is important if you want to reproduce our examples.

Ensure that you do not replace our files with a modified version of the same file; avoid using the command `save, replace` while working with our files.

(Pages omitted)

1 Overview of weighting

Survey datasets, released to the public through a general- or restricted-use agreement, usually come with at least one analysis weight for each respondent record in the sample. (In some applications, more than one weight may be provided for each record for special purposes discussed in chapter 7.) Analysts interested in calculating population estimates are told to use the same set of weights for all analyses—means, totals, linear and nonlinear models, etc. The benefits and drawbacks of a single analysis weight compared with multiple weights for tailored analytic objectives is reviewed in section 1.3.

Analysis weights are designed to

1. account for the probabilities used to select units (in cases where random sampling is used);
2. adjust in cases where it cannot be determined whether some sample units are members of the population under study;
3. adjust for eligible units that do not respond to the survey to limit the effects of nonresponse bias; and
4. incorporate external data to reduce standard errors of estimates and to compensate when the sample does not correctly cover the desired population.

However, unless you are the developer of the weights, the datasets typically contain the final analysis weights and not the adjustments for the above conditions.

Survey statisticians usually think of weighting in the context of probability samples, where units are selected by some random means from a well-defined population. All four steps above can be applied to probability samples. However, because of the current popularity of volunteer web panels and other kinds of "found" data, how to weight nonprobability samples is also worth considering. For those samples, steps 3 and 4 can be used (see chapter 6).

This chapter gives an overview of the purposes of weighting, underlying theory and sampling methods, and some problems that are considered when constructing a set of weights. The information in this chapter forms the basis for our discussion in this book. Specifically, the last section of this chapter contains an overview of weighting procedures and serves as an important reference for the remaining chapters.

1.1 Reasons for weighting

The fundamental reason for using weights when analyzing survey data is to produce estimates for some larger target population, that is, population inference. Ideally, the estimates will a) be unbiased or consistent in a sense described later, b) have standard errors that are as small as is feasible given the sample size and sample design, and c) correct for deficiencies in how the sample covers the desired population. Depending on the type of analysis being done, the population may be some well-defined finite population, like all adults aged 18 years and older in a country. The goal when making other estimates, like those of parameters in a regression model, may be to represent some population that, at least conceptually, is broader than any given finite population.

A finite population is a collection of units (also referred to as elements or cases) that could, in principle, be completely listed so that a census could be conducted to collect data from each unit. Examples, in addition to the adult population mentioned above, are elementary schools in a county, hospitals in a state, registered voters in a city, and retail business establishments in a province.

Defining the units that are members of a finite population (that is, eligible units) may require some thought, depending on the type of population. Whether a person is age 18 or older (and eligible to vote in the United States) seems straightforward, but defining what constitutes a business establishment is more difficult. Often, the composition of a population can change over time so that a specific time period must be part of the definition of the population. For example, a finite population of registered voters might be defined as those persons who are registered as of the date an election is to be held. The January labor force in a country may be defined as all persons who are employed or unemployed but seeking a job during the second week of that month.

Target populations and sampling frames

Understanding the distinction between a target population (also referred to as the universe of all population members or just universe) and a sampling frame is important when assessing the strengths and weaknesses of a sample. The target population is the population for which inferences or estimates are desired. The sampling frame is the set of units from which the sample is selected. Ideally, the sampling frame and the target population are the same. In that case, we say that the sampling frame completely covers the target population. However, there are many instances where the two do not coincide.

Figure 1.1 is a diagram of how the universe U, the sampling frame F, the sample s, and the complement of the sample within U, s^c, might be related. The frame F can omit some eligible units (undercoverage) and include other ineligible units (overcoverage). The eligibles in the frame in figure 1.1 are denoted by the intersection of U and F, $U \cap F$, while the ineligibles in the frame are denoted by those not included in U, $F - U$. The sample s can include both eligible units in $s \cap U$ and ineligible units in $s \cap (F - U)$. The latter condition occurs if the true eligibility of the units on the frame is unknown

when the sample is selected. In the figure, the eligible units that are not in the frame or sample are denoted by $U - F$. In the ideal situation, the frame completely covers the population so that $F = U$. The purpose of weights is to project the eligible sample, $s \cap U$, to the full universe, U. As is apparent from the figure, this will require eliminating the ineligible units from the sample (or at least those known to be ineligible) if such information is not available to remove them initially from the frame. We also hope to use the sample to represent the units in the universe that were not in the frame, $U - F$, and consequently had no chance of being selected for the sample. One of the functions of weighting is to attempt to correct for such coverage errors.

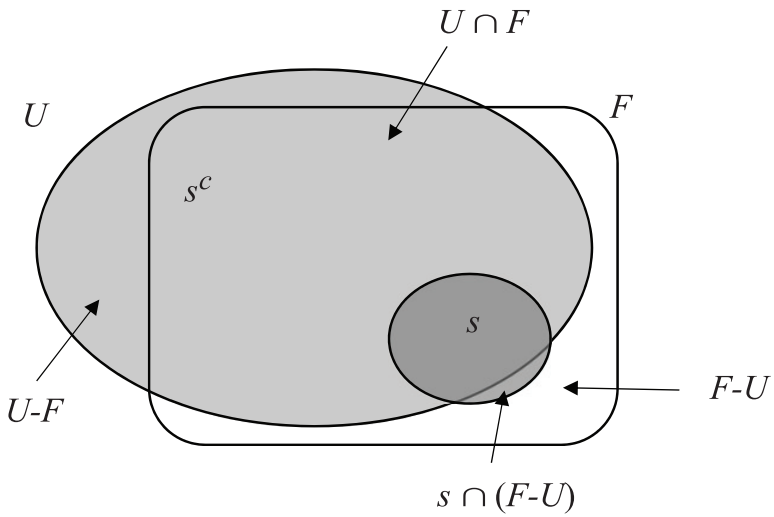

Figure 1.1. Illustration of sampling frame with over- and undercoverage of target population

The most straightforward case of a sampling frame is a list of every unit in the target population. For example, if we want to survey the members of some professional organization like the Royal Statistical Society (target population), a current membership list (sampling frame) may be available from which the sample can be selected. However, if the list was somewhat outdated because it omits people who became members in the last month, or it still contains some deceased members, the frame would have coverage errors. Current members not covered by the list cannot be sampled, although they would be eligible for the study. Past members covered by the list can be sampled, although they are ineligible for the study.

A complete list of the members of the target population is not always available, but it may be possible to construct a frame that does cover the whole population. For example, in household surveys, a list of all households or people who live in them is not available in many countries. Even if a government agency has such a list, it may not be accessible to private survey organizations. Standard practice is to compile a frame in

stages. For example, a sample of geographic areas is selected, perhaps in several stages, and a list of households is compiled only within the sample areas. When executed properly, this technique will provide virtually complete coverage. However, in practice, achieving complete coverage of a household population is difficult or impossible. Even the Current Population Survey in the United States, which is quite well conducted, had about 15% undercoverage of persons in 2013 (U.S. Census Bureau 2013).

Types of statistics

Descriptive statistics, like means or totals, are usually thought of as estimates of the quantities that would be obtained if a census were conducted of a finite population. For example, if the estimate is for the mean salary and wage income per person in a particular calendar year, the target for the sample estimate is the mean that would be obtained if all persons in the finite population were enumerated and the income collected for each. A population total is another example of a descriptive statistic. The finite population total itself is $t_y = \sum_{i \in U} y_i$, where U is the set of all N units in the population. Suppose a sample of n units is selected from the population. An estimated total often has the form $\widehat{t}_y = \sum_{i \in s} w_i y_i$, where i denotes a unit, s is the set of n units in the sample, w_i is a weight assigned to unit i, and y_i is the value of a data item collected for unit i. Weights that are appropriate for estimating totals are generally larger than or equal to 1 because $n < N$, and the weights need to inflate the sample to the larger population. In fact, for $y_i = 1$ for all units in the sample, $\sum_{i \in s} w_i = \widehat{N}$, is an estimate of the finite population size. Note that we use "hat notation" to signify estimates such as $\widehat{N}$, the estimate of the true population size, N.

Survey weights can also be used to estimate more complicated quantities like model parameters. For example, consider the simple linear regression model $y_i = \alpha + \beta x_i + \varepsilon_i$, where α and β are parameters, and the ε_i's are errors that are independent under the model with mean 0 and variance σ^2. The survey-weighted estimate of the slope computed by Stata and other software that handle survey data is

$$\widehat{\beta} = \frac{\sum_{i \in s} w_i \left(y_i - \overline{y}_w\right)\left(x_i - \overline{x}_w\right)}{\sum_{i \in s} w_i \left(x_i - \overline{x}_w\right)^2} = \frac{\sum_{i \in s} w_i x_i y_i - \left(\sum_{i \in s} w_i\right) \overline{x}_w \overline{y}_w}{\sum_{i \in s} w_i x_i^2 - \left(\sum_{i \in s} w_i\right) \overline{x}_w^2}$$

with $\overline{y}_w = \sum_{i \in s} w_i y_i / \sum_{i \in s} w_i$ and $\overline{x}_w$ defined similarly. As the second expression for $\widehat{\beta}$ shows, the estimated slope is a combination of several different estimated totals. Thus, estimated totals are frequently the building blocks for calculating quantities that are more complicated.

Estimates of model parameters can be interpreted in one of two ways. The first is the same as for descriptive statistics: $\widehat{\beta}$ estimates the value that would be obtained if a census were done and the model fit via ordinary least squares (that is, without weights) for the full, finite population. The second interpretation is, perhaps, more subtle: $\widehat{\beta}$ estimates a model parameter that applies to units beyond those in the fixed, finite population from which the sample was drawn. For example, suppose a sample of persons is selected in April 2015, and an analyst regresses personal income on years of

education. The analyst is probably interested in making a statement about the effect of education on income not just in April 2015 but also without regard to the month when the survey happened to have been done. This also raises the question of whether the survey weights should be used at all in model fitting—a topic we address in more detail in chapter 7.

1.2 Probability sampling versus nonprobability sampling

Survey samples can be selected in one of two ways. The first is through a defined probabilistic method that is reproducible and is labeled as probability sampling. The second is by way of an undefined sampling mechanism that is not exactly reproducible, known in the survey world most recently as nonprobability sampling. The method that is used affects how weights are calculated.

Probability sampling means that units are selected from the finite population in some random manner. Probability sampling has a very specific, technical definition given in Särndal, Swensson, and Wretman (1992) and other books on sampling theory. Four conditions must be satisfied for a sample to be a probability sample:

1. The set of all samples that are possible to obtain with the specified sampling procedure can (in principle) be enumerated.
2. Each possible sample s has a known probability of selection, $p(s)$.
3. Every unit in the target population has a knowable, nonzero probability of selection.
4. One set of sample units is selected with the probability associated with the set.

If a probability sample is selected, the first step in weighting is to compute an initial or base weight for each unit, which is the inverse of its selection probability. Base weights are mentioned in section 1.7 and described further in chapter 2.

Although the requirements above seem to imply that every possible sample would have to be identified, a probability sample can be selected in a way that does not require listing all the possibilities. Standard procedures also require only that the probabilities of selection of individual units be tracked—values of $p(s)$ are unnecessary.

Probability samples are the standard for governmental surveys that publish official statistics, like the unemployment rate, the inflation rate, and statistics on the health of a population. If time and budget allow, other surveys like pre-election polls may also select probability samples. This method of sampling provides one mathematical basis for making estimates, as discussed in section 1.3. It also adds a degree of face validity to the results. A survey designer cannot be accused of injecting conscious or unconscious biases into the selection of units when a random mechanism is used to decide which units are picked for the sample. Because every element in the population has a chance of being selected for a sample, the sample covers the entire population. If

enough information is available on the frame in advance of sampling, a survey designer can also control the distribution of the sample among various subgroups.

On the other hand, it may be cheaper and quicker, or only feasible, to acquire sample cases without a defined probability method (that is, by using nonprobability methods). Characteristics of interest may be time sensitive, and sampling may have to be done in the field by data collectors. Asking visitors to a website to participate in a survey voluntarily is one way that is currently being used to collect sample data. For example, a survey sponsor can inexpensively accumulate a huge number of persons this way and request that they become part of a panel that will cooperate in future surveys. One obvious criticism of this approach is that only a selective group of persons may visit the website used for recruiting. The persons who volunteer may be a poor cross-section of the population at large; that is, the sample may be subject to severe coverage error. Of course, this sort of criticism can be levied against any sample where there is no control or limited control over which sample units actually participate. A committee of the American Association for Public Opinion Research (AAPOR) conducted an extensive review of nonprobability samples (Baker et al. 2013b). Elliott and Valliant (2017) review the theoretical issues with inference from nonprobability samples and some of the methods that have been proposed for estimation. We investigate weighting for nonprobability surveys in detail in chapter 6.

Samples often live in some fuzzy netherworld between probability and nonprobability. A sample may begin as a probability sample but then suffer from a high rate of nonresponse. Because the survey designer cannot completely control which units respond, the set of units that ultimately respond may not reflect the intended probability sample. Nevertheless, starting with a probability sample selected from a high-quality frame provides some degree of comfort that a sample will have limited coverage errors.

A web panel of persons is a case in point. One approach to forming a web panel is to select a large telephone sample of households and request the cooperation of all persons over a certain age. The initial sample may be a probability sample of all telephone numbers known to be in use, but the resulting panel can suffer from at least two problems. If any phone numbers are omitted from the sampling frame, an undercoverage problem may result if the omitted portion differs from those on the frame. For example, if a frame uses only landline phones, then households with only cell phones cannot be selected. Telephone surveys also often have poor response rates—30% or less is common in the United States. If the respondents are not randomly spread over the initial sample, then there may be nonresponse bias, another source of potential undercoverage.

As discussed in chapters 3 and 4, weights can be constructed that attempt to adjust for both coverage and nonresponse error. The success of these adjustments depends on strong assumptions that are described there.

(Pages omitted)

4 Calibration and other uses of auxiliary data in weighting

Calibration is usually the last step in weighting and is extremely important in many surveys. Auxiliary data are used to reduce standard errors and to correct coverage problems introduced through the frame or through nonresponse not corrected with a prior adjustment. By auxiliary data, we mean information that is available for the entire frame or target population, either for each individual population unit or in aggregate form. These may be obtainable because a frame of all units in the population is available that has the auxiliary data on each unit. Surveys of business establishments or institutions may have such frames.

The standard application of calibration uses (aggregated) population totals of the auxiliaries to compute weights. Population totals for some variables may be available from a source separate from the survey, like a census. In a business survey, the frame might have the number of employees from an earlier time period for each establishment. In a household survey, counts of persons in groups defined by age, race, and gender may be published from a census, from population projections, or from a separate survey(s). In this chapter, we review several ways to use those data.

Figure 1.1 showed an example of a sample that had both under- and overcoverage. A common situation in household surveys is that some groups of persons are not covered as well as others by the sample design. In the United States, for example, estimates of the numbers of persons from most household surveys are less than counts from the most recent census, with the undercoverage being especially severe for young black or Hispanic males (for example, see U.S. Census Bureau [2014, table 2]). By computing weights that sum to census counts for male blacks and Hispanics, the undercoverage is corrected, at least, in the sense that estimates of population counts match the census counts. Details of some of the techniques for doing this are given in sections 4.1 and 4.3.

Inherent in this discussion is the need for the survey and the control total source(s) to use consistent data collection methods. Both should capture the data in the same way—identical questions and mode of data collection—from the same inferential population. In practice, researchers may have to decide if the level of consistency is close enough. If, however, the frame population differs greatly from the inferential population captured in the control total source(s), then strong assumptions are needed to proceed with weight calibration.

The estimators we cover in this chapter all have a superpopulation model that supports them.[1] Taking a model-based view of inference as discussed in section 1.3, the estimators will be approximately or exactly unbiased for an estimated total of y if the model holds for that y in the population. From a design-based point of view, the estimator will be approximately unbiased in repeated sampling and more efficient than one that does not use the auxiliary data.[2] (Note that if the sample is not a probability sample, as covered in chapter 6, then some type of model-based estimation is the only option for analysis.) We point out the appropriate model for each technique as we go along. Model fitting and checking should be an important step in deciding what form of estimator to use, although practitioners often ignore this step.

4.1 Poststratified estimators

Poststratified and raking estimators are two of the most commonly used calibration estimators. They are especially popular in household surveys of persons where the auxiliary variables are indicators for demographic groups. For example, persons may be classified by age group, gender, and race. Poststratification is implemented within calibration weighting classes formed by crossing all categories of the qualitative variables and constructing weights that reproduce the class-specific population counts in the weighted estimates. Poststratification can also be done using a single variable like age group. The poststratified estimator of a total is defined as

$$\widehat{t}_{y\mathrm{PS}} = \sum_{\gamma=1}^{G} N_\gamma \left(\widehat{t}_{y\gamma} \Big/ \widehat{N}_\gamma \right) \tag{4.1}$$

where $\widehat{t}_{y\gamma} = \sum_{s_\gamma} d_i y_i$ is the estimated total of y in weighting class (or poststratum) γ based on the input weights, s_γ is the set of responding sample units in poststratum γ, $\widehat{N}_\gamma = \sum_{s_\gamma} d_i$ is the estimated population size of poststratum γ based on the input weights, N_γ is the population count (also known as a control or control total) for the poststratum γ, and G is the total number of poststrata. The input weights, d_i, can be base weights or weights that have been adjusted for unknown eligibility or nonresponse. The implied final weight for unit i in poststratum γ is

$$w_i = d_i \frac{N_\gamma}{\widehat{N}_\gamma} \tag{4.2}$$

where $N_\gamma \Big/ \widehat{N}_\gamma$ is the poststratification adjustment (factor). With that definition of the weight, we can write the estimator as $\widehat{t}_{y\mathrm{PS}} = \sum_{i \in s} w_i y_i$, that is, a weighted sum of the data values.

1. The term "superpopulation model" means a statistical model that appears to agree with the structure of the data observed in the finite population.
2. This design-based property is contingent on any nonresponse bias having been corrected using the methods in chapter 3. Consequently, a more honest term might be "pseudodesign based", because which units respond is almost never under the survey designer's control.

The weighting classes are called poststrata because they are applied after the sample is selected and data are collected. This gives flexibility in which variables are used to define the poststrata because they do not have to be available when the sample is designed. Consequently, poststratification is a good way to use auxiliaries that you think are effective predictors of important variables collected in the survey but cannot be easily used for sample selection or for a nonresponse adjustment. For example, in a household survey, many countries do not have a frame of persons that includes race and educational attainment, even though those variables may be correlates of many analysis variables.

In the model that supports poststratification, each unit in poststratum γ has the same mean

$$E_M(y_i) = \mu_\gamma$$

for all units i in group γ. The units in the sample can be independent under the model, or the model can reflect clustering in the sample (and population) so that elements within clusters are correlated under the model.

Example 4.1: Poststratification. The example below reads a sample of 2,000 persons selected from the U.S. National Health Interview (NHIS, `nhis.large.dta`) population in `PracTools`, which has $N = 21588$ persons and poststratifies it using a 5-category age-group variable. The full code is in `ex.4.1_poststrat.do`. The `svyset` statement specifies the design weight as `wt`. Two options are used: `poststrata()`, which names the variable that holds the poststratum values (`age_grp` in this case), and `postweight()`, which gives the variable that contains the poststratification control totals (`poptots`). Each record on the file that is in a given poststratum should have the same value in the `poptots` field. As a check, we run a table of the weighted population estimates for age group. Each estimated count is equal to the control total in each group, plus their standard errors (SEs) are 0. The estimated SEs are 0 for any control total used in a calibration adjustment because the algorithm forces the estimates to be the same in every sample.

```
. use http://www.stata-press.com/data/svywt/nhis_sam.dta
. label define age_lab 1 "<18" 2 "18-24" 3 "25-44" 4 "45-64" 5 "65+"
. label values age_grp age_lab
. svyset [pweight=wt], poststrata(age_grp) postweight(poptots)
      pweight: wt
          VCE: linearized
   Poststrata: age_grp
   Postweight: poptots
  Single unit: missing
     Strata 1: <one>
         SU 1: <observations>
        FPC 1: <zero>
. svy: tabulate age_grp, count se
(running tabulate on estimation sample)
Number of strata   =         1                  Number of obs     =      2,000
Number of PSUs     =     2,000                  Population size   =     21,588
N. of poststrata   =         5                  Design df         =      1,999
```

age.grp	count	se
<18	5991	0
18-24	2014	0
25-44	6124	0
45-64	5011	0
65+	2448	0
Total	2.2e+04	

```
  Key:  count     =  weighted count
        se        =  linearized standard error of weighted count
```

A shortcoming of the Stata code above is that the poststratified weights in (4.2) are not saved as a separate variable and, consequently, cannot be stored for later use. However, using `svycal`, the code below does compute and save the weights. These weights will be saved if the data file is saved. The population controls are stored in a matrix called `poptotals`. Notice that the matrix contains six positions, with the first holding the sum of the age group control totals that are in positions 2–6 corresponding to the five levels of `age_grp`. The first position is the control total for the constant in the poststratification model, which is the population size. The `gen(ps_wt)` option of `svycal` causes the poststratified weights to be saved in `ps_wt`. `svycal` is a general command that will be used for more elaborate applications of calibration in later sections.

```
. matrix poptotals = 21588, 5991, 2014, 6124, 5011, 2448
. matrix colnames poptotals = _cons 1.age_grp 2.age_grp 3.age_grp 4.age_grp
> 5.age_grp
. svycal regress i.age_grp [pw=wt], gen(ps_wt) totals(poptotals)
```

The poststratified weights can be used in analysis, for example, to estimate the proportion of persons who receive Medicaid (a medical assistance program for the poor) in the United States:

```
. generate medicaid1 = abs(medicaid - 2)
. svyset [pweight=wt], poststrata(age_grp) postweight(poptots)

. svy: mean medicaid1
  (output omitted)
```

	Mean	Linearized Std. Err.	[95% Conf.	Interval]
medicaid1	.1052065	.0070797	.0913219	.119091

■

Warning. Be wary of specifying the survey design in a way that does not recognize that poststratification has been used. For example, if `svyset` is issued as

```
svyset [pweight=ps_wt]
```

Stata will use the wrong variance estimation formula. SEs of the estimated age group population counts will be nonzero because we do not give ourselves proper "credit" for poststratifying. SEs may be overestimated for variables for which poststratification does result in precision gains.

If you are the database constructor and are supplying a file with poststratified weights to other users, they need to know both definitions of the poststrata and their control totals to correctly compute SEs. This proviso applies to all other calibrated weights, including raking and general regression that will be covered later in this chapter.

Recovering control totals from public-use files

Public-use files are often supplied with weights computed via poststratification or another calibration procedure. To correctly compute SEs, Stata needs to be informed that poststratification was used. This requires knowing both the variable(s) used to define poststrata and the population (control) totals. As a data user, you may have to rely on the survey documentation to try to reconstruct both. For example, suppose the documentation says that poststrata were based on age × gender. A survey-weighted tabulation of the estimated population counts in the age × gender table will reproduce the population controls. If the documentation does not clearly list the age categories,

you may have to contact the organization that provided the database or make an educated guess as to what they were.

Given your own construction of a poststratum variable and the recovered population totals, `svyset` can be specified as in example 4.1. Expending this effort is worthwhile to get more honest SEs. These better SEs could well be smaller than if you treated the weights as inverse selection probability weights like in the warning above.

Example 4.2: Recovering poststratification totals. We use the same sample as in example 4.1 but now use `pswt` in the dataset as the weight. (The code is in `ex.4.2_recover.poptots.do`.) If the poststratified weights, `pswt`, are treated as inverse selection probabilities, and the total number of persons who are not covered by any kind of medical insurance is estimated, we get these results:

```
. use http://www.stata-press.com/data/svywt/nhis_sam, clear
. label define age_lab 1 "<18" 2 "18-24" 3 "25-44" 4 "45-64" 5 "65+"
. label values age_grp age_lab
. generate notcov1 = abs(notcov - 2)
. svyset [pweight = pswt]

. svy: total notcov1
  (output omitted)
```

	Total	Linearized Std. Err.	[95% Conf.	Interval]
notcov1	4045.863	195.5004	3662.454	4429.273

```
. svy: total notcov1, over(age_grp)
  (output omitted)
    _subpop_1: age_grp = <18
    _subpop_2: age_grp = 18-24
    _subpop_3: age_grp = 25-44
    _subpop_4: age_grp = 45-64
    _subpop_5: age_grp = 65+
```

Over	Total	Linearized Std. Err.	[95% Conf.	Interval]
notcov1				
_subpop_1	875.2284	102.0084	675.1727	1075.284
_subpop_2	689.9815	112.391	469.5637	910.3993
_subpop_3	1653.662	116.8108	1424.576	1882.747
_subpop_4	775.1823	82.7784	612.8399	937.5248
_subpop_5	51.80952	25.88503	1.044608	102.5744

Next, suppose that the weights on the file were created by poststratifying by the five age groups. We can recover the population totals with a tabulation.

```
. svy: tabulate age_grp, count format(%12.0f)
```
(*output omitted*)

age.grp	count
<18	5991
18-24	2014
25-44	6124
45-64	5011
65+	2448
Total	21588

Key: count = weighted count

Then, the poststratification control counts are appended to the dataset with the following:

```
. generate pstot = 5991
. replace pstot = 2014 if age_grp == 2
. replace pstot = 6124 if age_grp == 3
. replace pstot =  5011 if age_grp == 4
. replace pstot =  2448 if age_grp == 5
```

Retabulating the estimated total of persons not covered by medical insurance and accounting for the poststratification gives

```
. svyset [pweight=wt], poststrata(age_grp) postweight(pstot)

.     * nocov1 totals overall and by age_grp
. svy: total notcov1
```
(*output omitted*)

	Total	Linearized Std. Err.	[95% Conf.	Interval]
notcov1	4100.936	190.2382	3727.847	4474.025

```
. svy: total notcov1, over(age_grp)
```
(*output omitted*)

Over	Total	Linearized Std. Err.	[95% Conf.	Interval]
notcov1				
_subpop_1	882.4917	96.73621	692.7757	1072.208
_subpop_2	703	93.26878	520.0842	885.9158
_subpop_3	1671.016	105.6485	1463.822	1878.211
_subpop_4	791.2105	79.24024	635.807	946.614
_subpop_5	53.21739	26.32456	1.590482	104.8443

Note that we do not have to compute new poststratified weights using `svycal` as was done in example 4.1 because `wt` is already poststratified. The SE on the full population estimate is reduced by about 2.7% $(1 - 190.2382/195.5004)$ when the poststratification is accounted for. There are also reductions in the SEs of estimated totals for individual age groups. The reason that SEs are smaller when poststratification is properly credited is that there are substantial differences among age groups in the proportions not covered by medical insurance as is evident from the table below. In other words, there is an association between the poststrata and the variable of interest.

```
. svy: mean notcov1, over(age_grp)
  (output omitted)
```

Over	Mean	Linearized Std. Err.	[95% Conf.	Interval]
notcov1				
_subpop_1	.1473029	.0161469	.1156361	.1789697
_subpop_2	.3490566	.0463102	.2582344	.4398788
_subpop_3	.2728636	.0172515	.2390304	.3066968
_subpop_4	.1578947	.0158133	.1268823	.1889072
_subpop_5	.0217391	.0107535	.0006497	.0428286

Finally, we should note that any reduction in SEs will be most important for estimated totals. Poststratification may not improve the precision of estimated proportions (or any ratio estimator where weights are included in the numerator and denominator) much. ■

Control totals, however, are not always exactly recoverable from public-use files. As discussed in Kim, Li, and Valliant (2007) some weighting cells, but not an entire level of a complicated set of poststrata, may need to be collapsed because of small or zero sample size. For example, say that poststrata are constructed by gender × race (Hispanic, non-Hispanic [NH], white, NH black, NH Asian, and NH other). If there are few respondents in the female NH Asian weighting cell (say, 10) but plenty of NH Asian males, then the statistician may choose to collapse NH Asian females with NH other females but leave the category alone for males. The associated documentation may not include such details. Therefore, we recommend checking the respondent sample size for the poststrata; if counts are small, then producing population control totals for your own version of collapsed cells or for a raking estimator (discussed in section 4.2) may be preferred.

4.2 Raking estimators

Raking is another commonly used method of adjusting weights to control totals. In this method, marginal population controls can be used for two or more variables. Raking is often used when several variables are predictive of either coverage or the analysis variables (or both) but the sample sizes in some cells would be small if all variables

were fully crossed as they would be for poststratification based on multiple variables. Another reason to use raking is when control totals are only available at the margins, say, in a published report but not for the corresponding poststrata. Like poststratification, raking has an associated linear model. Taking the case of two raking dimensions, the model mean for unit i in level j of the first variable and level k of the second is

$$E_M(y_i) = \mu + \alpha_j + \beta_k$$

where α_j and β_k are main effects. Even with two dimensions several variables can be used because a dimension can be a cross-classification. For example, one dimension might be (age group) $\times$ education and the other race $\times$ (income group).

Example 4.3: Raking with svycal. The code below (also in `ex.4.3_rake.do`) uses `svycal` to rake to control totals for age group and Hispanic using the same `nhis_sam` sample as above. The variable `hisp` is recoded from 4 categories to 3 in `hispr` because the sample size in category 4 (non-Hispanic all other race groups) is small. The `gen(rake_wt)` option saves the weights in a variable called `rake_wt`. The `totals()` option lists the control totals in order for the levels of `age_grp` and `hispr`.

```
. use http://www.stata-press.com/data/svywt/nhis_sam, clear
. label define age_lab 1 "<18" 2 "18-24" 3 "25-44" 4 "45-64" 5 "65+"
. label values age_grp age_lab
. recode hisp (1=1) (2=2) (3=3) (4=3), generate(hispr)
. svycal rake i.age_grp i.hispr [pw=wt], gen(rake_wt)
> totals(_cons=21588 1.age_grp=5991 2.age_grp=2014 3.age_grp=6124
> 4.age_grp=5011 5.age_grp=2448 1.hispr=5031 2.hispr=12637 3.hispr=3920)

. summarize(rake_wt)
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
     rake_wt |      2,000      10.794    2.478362   8.909467   20.06054
```

■

Using the raked weights to estimate the proportion of persons receiving Medicaid, we get

```
. svyset [pweight=wt], rake(i.age_grp i.hispr,
> totals(_cons=21588 1.age_grp=5991 2.age_grp=2014 3.age_grp=6124
> 4.age_grp=5011 5.age_grp=2448 1.hispr=5031 2.hispr=12637 3.hispr=3920))
. generate medicaid1 = abs(medicaid - 2)

. svy: mean medicaid1
  (output omitted)
```

	Mean	Linearized Std. Err.	[95% Conf.	Interval]
medicaid1	.1067731	.0070633	.0929209	.1206254

which is similar to the estimated mean and SE in poststratification where only age group was used. Notice that in the `svyset` statement above, the specification of `totals` in `svycal` includes marginal population counts for both age group and Hispanic.

The `rake()` option also allows bounds to be set on the relative change of the raked weights to the input weights. This code sets lower and upper bounds on the ratio of `rake_wt`/`wt` using the `ll()` and `ul()` options:

```
. svycal rake i.age_grp i.hispr [pw=wt], gen(rake_wtB) totals(_cons=21588
> 1.age_grp=5991 2.age_grp=2014 3.age_grp=6124 4.age_grp=5011 5.age_grp=2448
> 1.hispr=5031 2.hispr=12637 3.hispr=3920) ll(0.8) ul(1.2)
```

This option can be used if unbounded raking makes some extremely large adjustments to the input weights that the analyst feels are untrustworthy. The `ll()` and `ul()` options can also be used in `svycal regress` discussed below. A key point to remember is that these options put a bound on the weight adjustments not on the weights themselves. The formal statement of the constraint that is put on the final weights w_i is

$$L \leq w_i/d_i \leq U$$

where d_i is the input weight for unit i and L and U are, respectively, the lower and upper bounds on the weight ratio.

The sizes of L and U will depend in part on the quality of the input weights. L can be set so that no final weight is less than or equal to 0 or allowed to be less than 1 on the premise that each unit should at least represent itself. U will depend on how large the adjustments need to be to compensate for nonresponse and noncoverage. A sample with very low response or poor coverage of the target population will need larger upward adjustments to the input weights than a sample with high response and near complete coverage.

Another versatile command for raking is `ipfraking` (Kolenikov 2014) (install by typing `net install http://www.stata-journal.com/software/sj14-1/st0323`). It has several useful features, including weight trimming and diagnostics. The code below repeats the example above, in which the NHIS sample is raked to margins for age groups and Hispanicity. Some preliminary coding is needed in addition to the labeling and recoding of the `hisp` variable, which was also done above. The vectors of control totals must be stored as matrices with three requirements:

1. Each matrix must be a $1 \times c$ row vector, where c is the number of control totals in the matrix.
2. Each matrix must have column names in Stata estimation results format, that is, *varname*`:` #.
3. Each matrix must have a row name that contains the categorical variable for which the totals were computed.

Because of requirement 2 above, the command `generate _one = 1` is issued below. This creates the variable `_one`, which is also set as the prefix to the column names in the