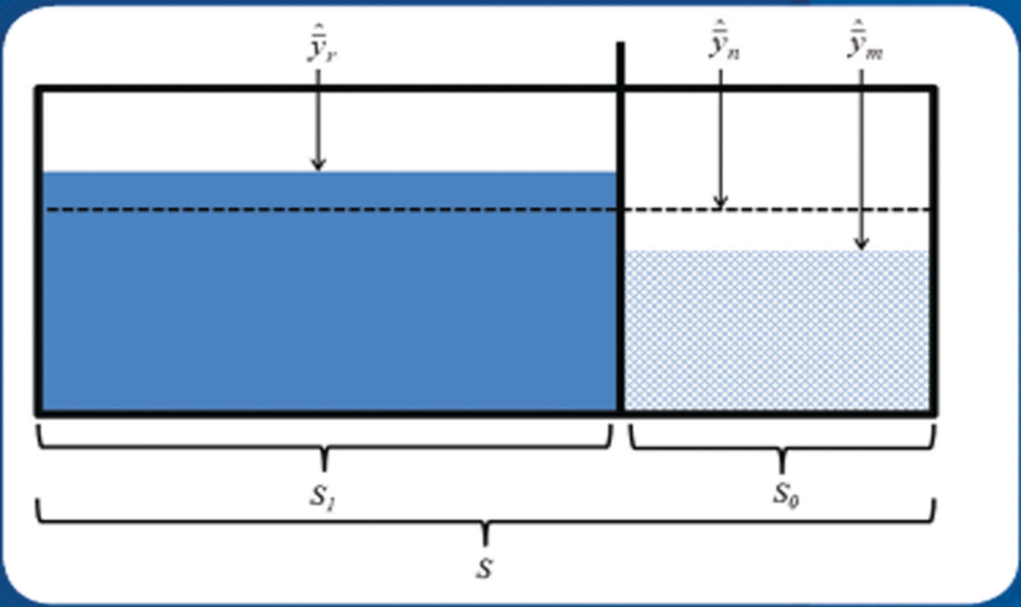
COMPLEX SURVEY DATA ANALYSIS WITH SAS®



Taylor H. Lewis

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

COMPLEX SURVEY DATA ANALYSIS WITH SAS®

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

COMPLEX SURVEY DATA ANALYSIS WITH SAS®

Taylor H. Lewis
Department of Statistics
George Mason University
Fairfax, Virginia, USA

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business
A CHAPMAN & HALL BOOK

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Printed on acid-free paper
Version Date: 20160427

International Standard Book Number-13: 978-1-4987-7677-6 (Hardback)

Library of Congress Cataloging-in-Publication Data

Names: Lewis, Taylor H.
Title: Complex survey data analysis with SAS / Taylor H. Lewis.
Description: Boca Raton : Taylor & Francis, 2016. | “A CRC title.” | Includes
bibliographical references and index.
Identifiers: LCCN 2016008468 | ISBN 9781498776776
Subjects: LCSH: Multivariate analysis--Data processing. | Regression
analysis--Data processing. | Sampling (Statistics) | SAS (Computer file) |
Surveys.
Classification: LCC QA278 .L49 2016 | DDC 519.5/30285555--dc23
LC record available at https://lccn.loc.gov/2016008468

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

To my three beautiful girls: Katie, Tessa, and Willow

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Contents

Preface

I wrote this book to serve as a handy desk-side reference for students, researchers, or any other kind of data analysts who have an intermediate to advanced grasp on fundamental statistical techniques, but who may not be familiar with the nuances introduced by complex survey data. As I embarked on this project, my intent was never to supplant or even compete with any of the established textbooks on the subject, such as Kish (1965), Cochran (1977), Lohr (2009), or Heeringa et al. (2010), for that would have surely been a futile endeavor. Rather, my aim was to demonstrate easy-to-follow applications of survey data analysis for researchers like myself who rely on SAS primarily, if not exclusively, for conducting statistical analyses.

All material in this book was developed using SAS Version 9.4 and SAS/STAT 13.1. It features the SURVEY family of SAS/STAT procedures, of which there are six at the time of this writing: PROC SURVEYSELECT, PROC SURVEYMEANS, PROC SURVEYFREQ, PROC SURVEYREG, PROC SURVEYLOGISTIC, and PROC SURVEYPHREG. The last five listed are companions to PROC MEANS, PROC FREQ, PROC REG, PROC LOGISTIC, and PROC PHREG, respectively. As will be explained in great detail over the course of this book, you should use one of the SURVEY procedures when analyzing data from a complex survey design. Using one of the non-SURVEY procedures opens the door to making false inferences due to biased point estimates, biased measures of variability, or both.

The book is structured as follows. Chapter 1 is a brief introduction to the practice of applied survey research. Key terminology and notation associated with the process of estimating finite population parameters from a sample are defined, as are the four features of complex survey data. Chapter 2 covers PROC SURVEYSELECT, which boasts numerous useful routines to facilitate the task of selecting probability samples. Descriptive, univariate analyses of continuous variables are covered in Chapter 3. Categorical variable analysis, both univariate and multivariate, is discussed in Chapter 4. Chapters 5 and 6 deal with the analytic task of fitting a linear model to survey data using PROC SURVEYREG or PROC SURVEYLOGISTIC, respectively. Both chapters begin with a section reviewing the assumptions, interpretations, and key formulas within the milieu of data collected via a simple random sample and then segue into the conceptual and formulaic differences within the realm of complex survey data. Chapter 7 is a foray into survival analysis, highlighting in large part the newest addition to the SURVEY family of SAS procedures, PROC SURVEYPHREG, which was developed for fitting Cox proportional hazard regression models. Chapter 8 delves into the concept of domain estimation, referring to any analysis targeting a subset of the target population. We will see why the DOMAIN statement should be used instead of subsetting a complex survey data set. Chapter 9 touches on the increasingly popular class of variance estimators known as replication techniques, which offer analysts a flexible alternative to Taylor series linearization, the default variance estimation method used in all SURVEY procedures. Because missing data is a ubiquitous problem in applied survey research, in my humble opinion, any book would be remiss without some material discussing approaches used to compensate for it. To that end, Chapter 10 demonstrates methods to adjust the weights of responding cases to better reflect known characteristics of the sample (or population), and Chapter 11 considers methods to impute, or fill in, plausible values for the missing data.

Each chapter was designed to be read on its own and not necessarily in sequence. As each chapter progresses, the material tends to build on itself and become more complex. It is my belief that oscillating between numerous survey data sets can only serve to detract from the reader's ability to compare and contrast the current syntax example with those preceding it in the chapter. As such, examples in most chapters are motivated within the context of a single complex survey data set.

Many people contributed to this effort and are owed my sincere thanks. First and foremost, I am especially grateful for the encouragement I received from Shelley Sessoms at SAS, who believed in me and the vision I had for this book and on several occasions helped me overcome my self-doubt about whether I could complete this colossal task. I thank Rob Calver for resuscitating this project, and for the professionalism he and his entire team at CRC Press exhibited. I am also very grateful for the encouragement I received on both professional and personal levels from Frauke Kreuter, Partha Lahiri, Brady West, Roberto Valverde, Abera Wouhib, Sidney Fisher, Mircea Marcu, Karl Hess, Kimya Lee, Chris Daman, Glenn White, Eugene Pangalos, Graham Evans, Ryan Fulcher, Matt Fuller, and Doug Drewry. Richard Valliant and Richard Sigman were kind enough to endure several lengthy conversations during which I solicited their advice on certain content issues that surfaced as the book took shape. I also thank Richard Sigman for agreeing to serve as a reviewer. He, along with Patricia Berglund, Peter Timusk, Donna Brogan, and several anonymous reviewers at SAS Institute, provided valuable feedback that greatly improved the overall quality of the book. The same can be said for Casey Copen and Kimberly Daniels at the National Center for Health Statistics, who carefully reviewed all discussion and examples pertaining to the National Survey of Family Growth. I claim full responsibility for any and all errors that remain. Lastly, and most importantly, I thank my wife, Katie, for being extraordinarily patient and understanding during this process, selflessly tolerating my extended absences from home without (too much) complaint. None of this would have been possible without her love and support.

Taylor Lewis
Arlington, Virginia

Author

Taylor H. Lewis is a PhD graduate of the Joint Program in Survey Methodology at the University of Maryland, College Park, and an adjunct professor in the George Mason University Department of Statistics. An avid SAS user for 15 years, Taylor is a SAS Certified Advanced programmer and a nationally recognized SAS educator who has produced dozens of papers and workshops illustrating how to efficiently and effectively conduct statistical analyses using SAS.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

1

Features and Examples of Complex Surveys

1.1 Introduction

In the era of *Big Data*, the variety of statistics that can be generated is ostensibly limitless. Given the copious and ever-expanding types of data being collected, there are many questions that can be answered from analyzing a data set already in existence or perhaps one even updated in real time. For instance, a credit card issuer seeking to determine the total amount of charges made by its customers on gasoline during a particular year may have this information readily retrievable from one or more databases. If so, a straightforward query can produce the answer. On the other hand, determining the average amount the typical U.S. household spends on gasoline presents a much more complicated estimation problem. Collecting data from all households in the United States would obviously be exceedingly costly, if not an outright logistical impossibility. One could probably make some progress pooling the comprehensive set of credit card issuers' databases and trying to group data into distinct households via the primary account holder's address, but not all households own and use a credit card, and so this would exclude any non-credit-card payment such as one made by cash or check. A survey of the general U.S. population is needed to acquire this kind of information. One such survey is the Consumer Expenditure Survey, sponsored by the Bureau of Labor Statistics, which reported in September 2015 that the average household spent approximately $2500 on gasoline during calendar year 2014 (http://www.bls.gov/news.release/cesan.htm).

It is truly a marvel to consider the breadth of statistics such as these available to answer a myriad of questions posed by researchers, policymakers, and the general public. The legitimacy of these statistics is attributable to the fields of survey sampling and survey research methodology, which together have engendered a wide variety of techniques and approaches to practical data collection problems. Hansen (1987) and Converse (1987) present nice summaries of the two fields' overlapping histories. The practice of modern survey sampling began with the argument that a sample should be *representative* (Kiaer, 1895) and drawn using techniques of randomization

(Bowley, 1906). The seminal paper by Neyman (1934) provided much needed theoretical foundations for the concept. Nowadays, there are textbooks entirely devoted to considerations for designing an efficient sample (Hansen et al., 1953; Kish, 1965; Cochran, 1977; Lohr, 2009; Valliant et al., 2013). Best practices in the art of fielding a survey have emerged more recently, and they continue to evolve in response to changes in information technology, communication patterns of the general public, and other societal norms. There are certainly a number of excellent volumes summarizing the literature on survey methods (Couper, 2008; Dillman et al., 2009; Groves et al., 2009), but they tend to have a shorter shelf life—or at least warrant a new edition more frequently—than those on sampling.

Survey research is the quintessence of an interdisciplinary field. While the opening example was motivated by expenditure data that might be of interest to an economist, there are analogous survey efforts aimed at producing statistics related to agriculture and crop yields, scholastic achievement, and crime, just to name a few. Of course, collecting data comes at a cost. In the United States, for-profit businesses do not generally conduct surveys and release raw data files to the public free of charge. More commonly, surveys are funded by one or more government agencies. These agencies are ideally apolitical and charged solely with the task of impartially collecting and disseminating data. The set of Principal Statistical Agencies listed at http://fedstats.sites.usa.gov/agencies/ more or less fits this description. Aside from preformatted tables and reports, data dissemination often takes the form of a raw or microdata file posted on the survey website for open access. Indeed, many of the examples in this book are drawn from three *real-world* survey data sets sponsored by two of these agencies, the National Center for Health Statistics (NCHS) and the Energy Information Administration. Namely, the National Ambulatory Medical Care Survey (NAMCS), the National Survey of Family Growth (NSFG), and the Commercial Buildings Energy Consumption Survey (CBECS) will be formally introduced in Section 1.5.

As authoritative or *official* as these statistics seem, it is important to bear in mind they are *estimates*. The term *estimate* can sometimes be confused with the very similar term *estimator*, but the two terms have different meanings. An estimate is the value computed from a sample, whereas an estimator is the method or technique used to produce the estimate (see Program 3.7 for a comparison of two unbiased estimators of a total). If the entire survey process were conducted anew, there are a variety of reasons one would expect an estimate to differ somewhat, but this book focuses primarily on quantifying the portion of this variability attributable to *sampling error* or the variability owing to the fact that we have collected data for only a portion of the population. Using formal statistical theory and a single sample data set in hand, however, there are established ways to calculate an unbiased estimate of the sampling error, which can be reported alongside the estimate or used to form a confidence interval or to conduct a hypothesis test. A distinctive aspect of *complex* survey data, the features of which will be detailed in

Section 1.4 and which are all too often overlooked by applied researchers, is that the techniques and formulas for estimating sampling error one learns in general statistics courses or from general statistics textbooks frequently do not carry over intact. The reason is that complex surveys often employ alternative sample designs, either for the purpose of statistical efficiency or out of necessity to control data collection costs. The implied data-generating mechanism in general statistics courses is simple random sampling with replacement (SRSWR) of the ultimate units for which data are measured. In applied survey research, that particular data-generating mechanism is the exception rather than the rule.

Section 1.2 establishes some of the terminology pertaining to applied survey research that will be used throughout this book. Section 1.3 previews the SAS/STAT procedures that have been developed to facilitate complex survey data analysis. These are all prefixed with the word SURVEY (e.g., PROC SURVEYMEANS is the companion procedure to PROC MEANS). Section 1.4 introduces the four features that may be present in a survey data set to justify the qualifier *complex*: (1) finite population corrections (FPCs), (2) stratification, (3) clustering, and (4) unequal weights. This chapter concludes with a discussion of the three real-world complex survey data sets from which many of the book's examples are drawn. There is some brief commentary on the motivation behind each survey effort, the type of sample design employed, the complex survey features present in the data set, and specific examples of estimates produced.

1.2 Definitions and Terminology

Groves et al. (2009, p. 2) define a survey as a "systematic method for gathering information from (a sample of) entities for constructing quantitative descriptors of the attributes of the larger population for which the entities are members." They use the term "entities" to stress the fact that, although the word "survey" often has the connotation of an opinion poll or a battery of questions directed at humans, this is not always the case. Other example entities are farms, businesses, or even events. Parenthetically including the phrase "a sample of" serves to remind us that not all surveys involve sampling. A *census* is the term describing a survey that aims to collect data on or enumerate an entire population.

One of the first stages of any survey sampling effort is defining the *target population*, the "larger population" alluded to in the Groves et al. definition about which inferences are desired. The target population often carries an ambitious, all-encompassing label, such as "the general U.S. population." The next step is to construct a list, or *sampling frame*, from which a random sample of *sampling units* can be drawn. The totality of entities covered by this list is called the *survey population*, which does not always coincide perfectly with the target population. For example, there is no population registry in the United States as there is in many European countries to serve as a sampling frame. There is an important distinction to be made between the sampling units and the *population elements*, or the ultimate analytic units for which measurements are taken and inferences drawn. The two are not always one and the same. Sometimes, the sampling frame consists of clusters of the population elements. Considering the goal of making inferences on the general U.S. population, even if a population registry existed, it might be oriented by household or address instead of by individual. This would present a cluster sampling situation, which is permissible but introduces changes to the more familiar formulas used for estimation. We will discuss cluster sampling more in Section 1.4.4 with the help of a simple example.

A sampling frame's makeup is often influenced by the survey's *mode* or method of data collection (e.g., in-person interview or self-administered paper questionnaire by mail). For example, a popular method for administering surveys by telephone is *random-digit dialing* (RDD), in which the sampling frame consists of a list of landline telephone numbers. A survey opting for this mode may still consider the target population "the general U.S. population," but the survey population is actually the subset of U.S. households with a landline telephone.

Figure 1.1 illustrates how the target population and survey population may not always be one and the same. The area within the target population that does not fall within the survey population area is of most concern. That represents *undercoverage*, meaning population elements that have no chance of being selected into the sample. Continuing with the RDD example, households without a landline telephone fall within this domain. Sometimes, it is possible to supplement one sampling frame with another to capture this group (e.g., by incorporating a sampling frame consisting of cellular telephone numbers), but that can introduce duplicated sampling units (i.e., households with landline *and* cellular numbers, therefore present in both frames), which can be a nuisance to deal with in its own right (Lohr and Rao, 2000). Another remedy often pursued is to conduct weighting adjustment techniques such as post-stratification or raking. These techniques will be discussed in Chapter 10.

There is also an area in Figure 1.1 delineating a portion of the survey population falling outside the bounds of the target population. This represents extraneous, or ineligible, sampling units on the sampling frame that may be selected as part of the sample. With respect to an RDD survey of U.S. households, a few examples are inoperable, unassigned, or business telephone numbers. These are represented by the area to the right of the vertical line in the oval labeled "Sample." An appreciable rate of ineligibility can cause inefficiencies in the sense that these units must be *screened* out where identified, but it is generally easier to handle than undercoverage.

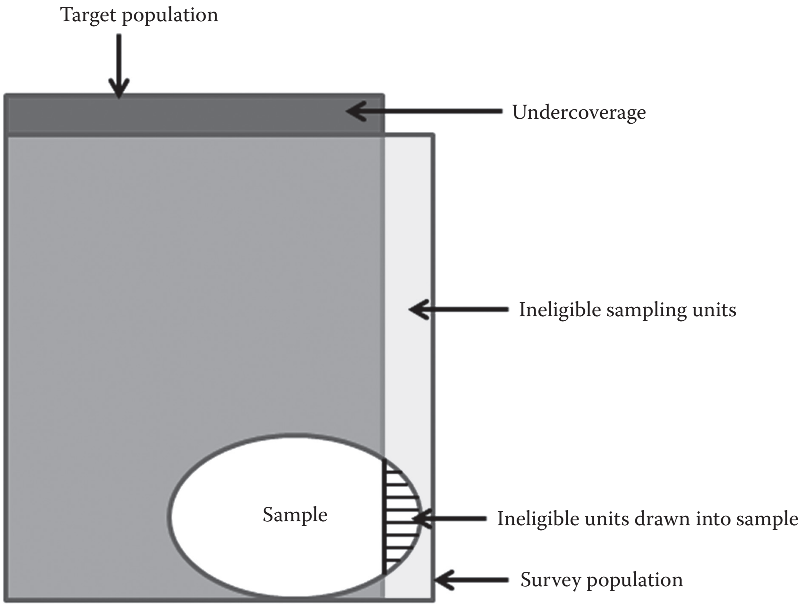


FIGURE 1.1
Visualization of a sample relative to the target population and survey population.

It is not unusual for those unfamiliar with survey sampling theory to be skeptical at how valid statistical inferences can be made using a moderate sample of size n from a comparatively large population of size N. This is a testament to the *central limit theorem*, which states that the distribution of a sequence of means computed from independent, and random samples of size n is always normally distributed as long as n is *sufficiently large*. Despite the vague descriptor *sufficiently large* fostering a certain amount of debate amongst statisticians—some might say 25, others 50, and still others 100—the pervading virtue of the theorem is that it is true regardless of the underlying variable's distribution. In other words, you do not need normally distributed data for this theorem to hold. This assertion is best illustrated by simulation.

Figure 1.2 shows the distribution of three variables from a fabricated finite population of $N=100{,}000$. The first variable is normally distributed with a population mean of $\bar{y}_1 = 5$. The second is right-skewed with a mean of $\bar{y}_2 = 2$, whereas the third variable has a bimodal distribution with a mean of $\bar{y}_3 = 3.75$. Figure 1.3 immediately following displays the result of a simulation that involved drawing 5000 samples of size $n=15$, $n=30$, and $n=100$ from this population and computing the sample mean for each of y_1, y_2, and y_3. That is, the figure is comprised of histograms summarizing the distribution of the three variables' sample means with respect to each of the three sample sizes. As in Figure 1.2, the row factor distinguishes the three variables, while

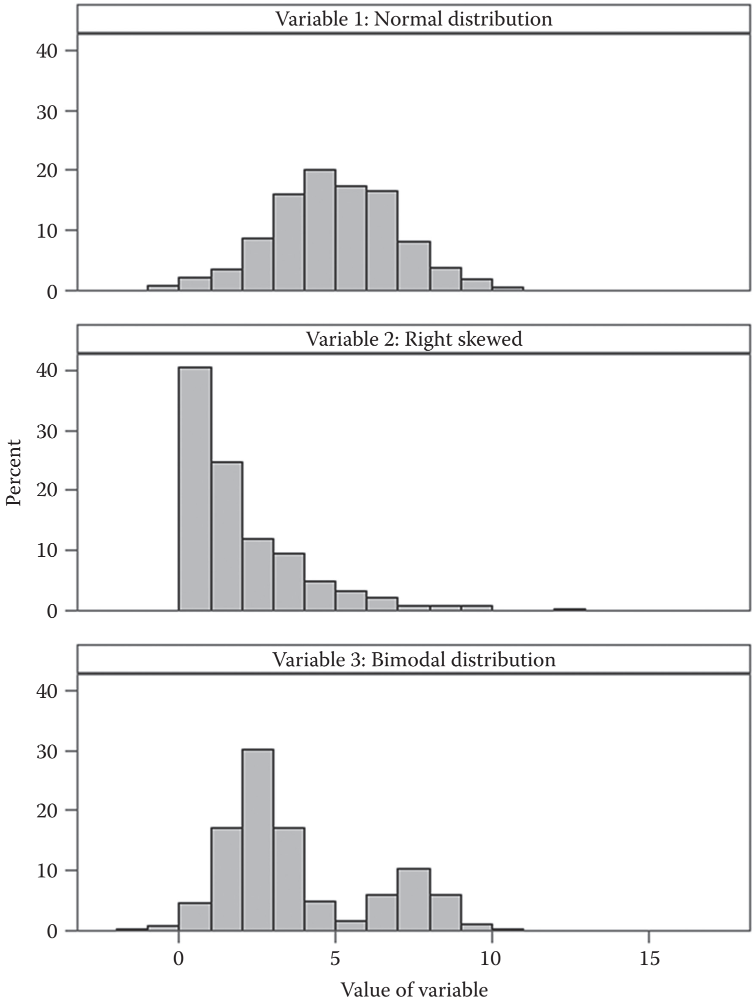


FIGURE 1.2
Distribution of three variables from a finite population of $N = 100{,}000$.

the column factor (moving left to right) distinguishes the increasing sample sizes. There are a few key takeaways from examining Figure 1.3:

- All sample mean distributions closely resemble a normal distribution, which has been superimposed on the histograms. Again, this is true regardless of the underlying distribution.
- The average or *expected* value of the 5000 sample mean values is the population mean, which is to say the sample mean is an unbiased estimate for the mean of the entire population.

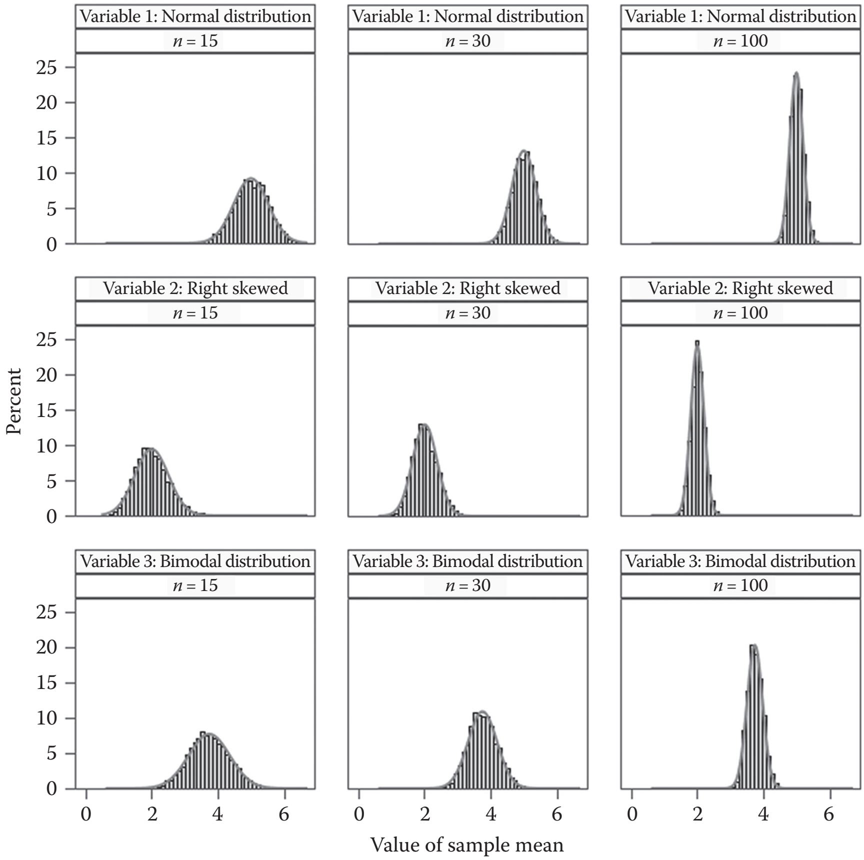


FIGURE 1.3
Sample mean distributions of 5000 samples of size 15, 30, and 100 drawn from the finite population in Figure 1.2.

- The distributions get "skinnier" as the sample size increases. This reflects increased precision or less deviation amongst the means from one sample to the next.

Knowing the sampling distribution of statistics such as the mean is what justifies our ability to make inferences about the larger population (e.g., to form confidence intervals, conduct significance tests, and calculate *p* values). In practice, we do not have a set of 5000 independent samples, but usually a single sample to work with. From this sample, a statistic is produced, which we typically top with a hat to distinguish it from the true population parameter it is estimating. Using general theta notation, we can denote $\hat{\theta}$ as the sample-based estimate or *point estimate* of θ. For example, $\hat{\bar{y}}$ refers to the point estimate of $\bar{y}$ from a particular sample. We acknowledge, however, that this estimate likely does not match the population parameter exactly. A fundamental quantification of the anticipated deviation is the *variance*, which can be expressed as $\text{Var}(\hat{\theta}) = E[(\hat{\theta} - \theta)^2]$. The variance represents the average squared deviation of the sample-based estimates from the true population value over all possible samples. This quantity is rarely known, except perhaps in simulated settings such as the one discussed earlier. But drawing from formal statistical theory, there are established computational methods to formulate an unbiased estimate of it using only the single sample at hand. The sample-based estimate is referred to as the *estimated variance* or *approximated variance* and denoted $\text{var}(\hat{\theta})$ (with a lowercase "v"). Over the course of the book, we will see expressions of variance estimates for a whole host of statistics, along with program examples demonstrating SURVEY procedure syntax to have SAS carry out the computational legwork for us.

Numerous related measures of uncertainty can be derived from the estimated variance. Since variability in squared units can be difficult to interpret, a common transformation is to take the square root of this quantity, which returns the *standard error* of the statistic denoted $\text{se}(\hat{\theta}) = \sqrt{\text{var}(\hat{\theta})}$. Along with the estimate itself, the standard error can be used to form a confidence interval $\hat{\theta} \pm t_{df,\alpha/2}\text{se}(\hat{\theta})$, where $t_{df,\alpha/2}$ represents the 100(1 − α/2)th percentile of a *t* distribution with *df* degrees of freedom. Another popular measure is the relative standard error of the estimate, also known as the *coefficient of variation*, which is defined as $\text{CV}(\hat{\theta}) = \text{se}(\hat{\theta})/\hat{\theta}$. This has an appealing interpretation. A value of 0.1, for example, indicates the standard error of the statistic represents 10% of the magnitude of the statistic itself. A target CV can be used as a sample design criterion before the survey is administered or as a threshold for whether a survey estimate is deemed precise enough to be published. Unlike the variance and standard error, the CV is unitless, thereby permitting precision comparisons regardless of scale (e.g., dollars versus percentages) or type of estimate (e.g., means versus totals).

It should be emphasized that these assertions only apply if the sample is selected randomly. An impressive-sounding sample size of, say, 10,000 means little if it were drawn using a nonrandom process such as convenience sampling, quota sampling, or judgment sampling (see Lohr, 2009, p. 5). The essential requirement is that every sampling unit on the sampling frame has a known, nonzero chance of being selected as part of the sample. The selection probabilities need not be equal, but they should be fixed and nonzero.

1.3 Overview of SAS/STAT Procedures Available to Analyze Survey Data

Table 1.1 previews the SURVEY procedures to be covered in the coming chapters. PROC SURVEYSELECT was developed to facilitate the task of sample

TABLE 1.1

Summary of SURVEY Procedures Currently Available and Primary Chapter(s) in Which They Are Covered

Procedure	Analytic Tools	Chapter(s)
PROC SURVEYSELECT	Variety of built-in routines for selecting random samples; also contains a few methods for allocating a fixed sample size amongst strata and determining a necessary sample size given certain constraints	Chapter 2
PROC SURVEYMEANS	Descriptive statistics such as means, totals, ratios, quantiles, as well as their corresponding measures of uncertainty	Chapter 3
PROC SURVEYFREQ	Tabulations, tests of association, odds ratios, and risk statistics	Chapter 4
PROC SURVEYREG	Regression models where the outcome is a continuous variable	Chapter 5
PROC SURVEYLOGISTIC	Regression models where the outcome is a categorical variable	Chapters 6 and 7
PROC SURVEYPHREG	Cox proportional hazards models for time-to-event data (survival analyses)	Chapter 7

selection, so it is generally more useful at the survey design stage rather than the analysis stage. Once the data have been collected, there are five additional SURVEY procedures available to produce descriptive statistics and conduct more sophisticated analytic tasks such as multivariate modeling. Each procedure is typically covered in a chapter of its own, but later chapters covering cross-cutting topics such as domain estimation (Chapter 8) and replication (Chapter 9) demonstrate more than one specific SURVEY procedure. We will explore alternative (non-SURVEY) SAS/STAT procedures and user-defined macros in the final two chapters, which deal with techniques for handling missing data.

1.4 Four Features of Complex Surveys

1.4.1 A Hypothetical Expenditure Survey

To motivate exposition of the four features of complex survey data, suppose a market research firm has been hired to assess the spending habits of $N = 2000$ adults living in a small town. Two example estimates of interest are the average amount of money an adult spent in the previous year on over-the-counter (OTC) medications and the average amount spent on travel outside the town. The ensuing discussion centers around a few (hypothetically carried out) sample designs to collect this expenditure data for a sample of $n=400$ adults and the particular complex survey features introduced.

1.4.2 Finite Population Corrections

Suppose the first sample design involved compiling the names and contact information for all $N=2000$ people in the town onto a sampling frame and drawing a simple random sample (SRS) of $n=400$ of them to follow up with to collect the expenditure information. From this sample, the estimated average for a given expenditure y would be calculated as $\hat{\bar{y}}=\sum_{i=1}^{n=400} y_i/n$. The research firm might then reference an introductory statistics textbook and calculate an estimated *element variance* of y as $s^2=\sum_{i=1}^{n=400}(y_i-\hat{\bar{y}})^2\big/(n-1)$, an unbiased, sample-based estimate of the population element variance, or $S^2=\sum_{i=1}^{N=2000}(y_i-\bar{y})^2/(N-1)$, and use this quantity to estimate the variance of the sample mean by $\text{var}(\hat{\bar{y}})=s^2/n$. They might then construct a 95% confidence interval as $\hat{\bar{y}}\pm 1.96\sqrt{\text{var}(\hat{\bar{y}})}$, where $\sqrt{\text{var}(\hat{\bar{y}})}=\text{se}(\hat{\bar{y}})$ is the standard error of $\hat{\bar{y}}$. Note how the standard error differs conceptually and computationally from the *standard deviation* of y, which is $S=\sqrt{S^2}$ for the full population and estimated by $s=\sqrt{s^2}$ from the sample.

It turns out the market research firm's calculations would overestimate the variance of the sample mean because whenever the sampling fraction n/N is nonnegligible (as is the case with 400/2000), there is an additional term that enters into variance estimate calculations called the FPC. A more accurate estimate of the variance of the sample mean is $\text{var}(\hat{\bar{y}})=(s^2/n)\big(1-(n/N)\big)$, where the term (1 − ($n/N$)) is the FPC. Notice how the FPC tends to 0 as n approaches N. Since the purpose of the estimated variance is to quantify the sample-to-sample variability of the estimate, an intuitive result is that it decreases as the portion of the population in each respective sample increases. In the extreme case when $n=N$, or when a census is undertaken, the variance accounting for the FPC would be 0. And despite the discussion in this section pertaining strictly to estimating the variance of a sample mean, a comparable variance formula modification occurs for other statistics such as totals and regression coefficients.

The difference between the two variance perspectives is that the traditional formula implicitly assumes data were collected under a SRSWR design, meaning each unit in the population could have been selected into the sample more than one time. Equivalently, the tacit assumption could be that data were collected using simple random sampling without replacement (SRSWOR) from an effectively infinite population, a corollary of which is that the sampling fraction is negligible and can be ignored. Sampling products from an assembly line or trees in a large forest might fit reasonably well within this paradigm. But in contrast, survey research frequently involves sampling from a finite population, such as employees of a company or residents of a municipality, in which case adopting the SRSWOR design formulas is more appropriate.

There are two options available within the SURVEY procedures to account for the FPC. The first is to specify the population total N in the TOTAL = option of the PROC statement. The second is to specify the sampling fraction n/N in the RATE = option of the PROC statement. With respect to the sample design presently considered, specifying TOTAL = 2000 or RATE = 0.20 has the same effect. The syntax to account for the FPC is identical across all SURVEY procedures, and the same is true for the other three features of complex survey data as well.

Suppose the SAS data set SAMPLE_SRSWOR contains the results of this survey of $n = 400$ adults in the town. Program 1.1 consists of two PROC SURVEYMEANS runs on this data set. We will explore the features and capabilities of PROC SURVEYMEANS in more detail in Chapter 3, but for the moment note that we are requesting the sample mean and its estimated variance for the OTC expenditures variable (EXP_OTCMEDS). The first run assumes the sample was selected with replacement. Since there are no complex survey features specified, it produces the same figures that would be generated from PROC MEANS. The second requests the same statistics but specifies TOTAL=2000 in the PROC statement, in effect alerting SAS that sampling was done without replacement and so an FPC should be incorporated. (The SURVEY procedure determines n from the input data set.)

Program 1.1: Illustration of the Effect of an FPC on Measures of Variability

```
title1 ‘Simple Random Sampling without Replacement’;
title2 ‘Estimating a Sample Mean and its Variance Ignoring the
FPC’;
proc surveymeans data=sample_SRSWOR mean var;
  var exp_OTCmeds;
run;

title2 ‘Estimating a Sample Mean and its Variance Accounting
for the FPC’;
proc surveymeans data=sample_SRSWOR total=2000 mean var;
  var exp_OTCmeds;
run;
```

Simple Random Sampling without Replacement
Estimating a Sample Mean and Its Variance Ignoring the FPC

SURVEYMEANS Procedure

Data Summary	
Number of observations	400

Statistics			
Variable	**Mean**	**Std Error of Mean**	**Var of Mean**
exp_OTCmeds	17.645854	0.683045	0.466550

Simple Random Sampling without Replacement
Estimating a Sample Mean and Its Variance Accounting for the FPC

SURVEYMEANS Procedure

Data Summary	
Number of observations	400

Statistics			
Variable	**Mean**	**Std Error of Mean**	**Var of Mean**
exp_OTCmeds	17.645854	0.610934	0.373240

The sample mean is equivalent ($17.65) in both PROC SURVEYMEANS runs, but measures of variability are smaller with the FPC incorporated. Specifically, the estimated variance of the mean has been reduced by a factor of 20% as we can observe that $0.3732 = 0.4666 * (1-(400/2000))$. Since the standard error of the mean is just the square root of the variance, by comparison it has been reduced to $0.6109 = 0.6830^{*}\sqrt{1-\left(400/2000\right)}$.

Where applicable, the FPC is beneficial to incorporate into measures of variability because doing so results in increased precision and, therefore, more statistical power. There are occasions, however, when the FPC is known to exist but intentionally ignored. This is often done when assuming a with-replacement sample design dramatically simplifies the variance estimation task (see discussion regarding the ultimate cluster assumption in Section 1.4.4), especially when there is only a marginal precision gain to be realized from adopting the without-replacement variance formula. Providing a few numbers to consider, with a sampling fraction of 10%, we would anticipate about a 5% reduction in the standard error; if the sampling fraction were 5%, the reduction would be around 3%. While the with-replacement assumption typically imposes an overestimation of variability, the rationale behind this practice is that the computational efficiencies outweigh the minor sacrifice in precision.

1.4.3 Stratification

The second feature of complex survey data is *stratification*, which involves partitioning the sampling frame into H mutually exclusive and exhaustive *strata* (singular: stratum), and then independently drawing a sample within each. There are numerous reasons the technique is used in practice, but a few examples are as follows:

- *Ensure representation of less prevalent subgroups in the population*. If there is a rare subgroup in the population that can be identified on the sampling frame, it can be sequestered into its own stratum to provide greater control over the number of units sampled. In practice, sometimes the given subgroup's stratum is so small that it makes more sense to simply conduct a census of those units rather than select a sample of them.
- *Administer multiple modes of data collection*. To increase representation of the target population, some survey sponsors utilize more than one mode of data collection (de Leeuw, 2005). When the separate modes are pursued via separate sampling frames, these frames can sometimes be treated as strata of a more comprehensive sampling frame.
- *Increase precision of overall estimates*. When strata are constructed homogeneously with respect to the key outcome variable(s), there can be substantial precision gains.

To illustrate how precision can be increased if the stratification scheme is carried out prudently, let us return to the expenditure survey example and consider an alternative sample design. Suppose there is a river evenly dividing the hypothetical town's population into an east and a west side, each with 1000 adults, and that adults living on the west side of the river tend to be more affluent. It is foreseeable that certain spending behaviors could differ markedly between adults on either side of the river. Since the two key outcome variables deal with expenditures, this would be a good candidate stratification variable.

For sake of an example, let us assume that the firm is able to stratify their sampling frame accordingly, allocating the overall sample size of $n=400$ adults into $n_1=200$ adults sampled without replacement from the west side and $n_2=200$ from the east, and that the results have been stored in a data set called SAMPLE_STR_SRSWOR. To account for stratification in the sample design, we specify the stratum identifier variable(s) on the survey data set in the STRATA statement of the SURVEY procedure. For the present example, the variable CITYSIDE defines which of the $H=2$ strata the observation belongs to, a character variable with two possible values: "West" or "East."

Like Program 1.1, Program 1.2 consists of two PROC SURVEYMEANS runs on the survey data set, except this time we are analyzing a measure of travel expenditures (EXP_TRAVEL) instead of OTC medications (EXP_OTCMEDS). The first run ignores the stratification and assumes a sample of size 400 was selected without replacement from the population of 2000. The second run properly accounts for the stratification by placing CITYSIDE in the STRATA statement. Observe how the FPC is supplied by way of a secondary data set called TOTALS in the second run. This is because the FPC is a stratum-specific quantity. When there is no stratification or the stratification is ignored (as in the first run), one number is sufficient, but you can specify stratum-specific population totals, or N_hs, via a supplementary data set containing a like-named and like-formatted stratum variable(s) and the key variable _TOTAL_ (or _RATE_, if you are opting to provide sampling fractions instead).

Program 1.2: Illustration of the Effect of Stratification on Measures of Variability

```
title1 ‘Stratified Simple Random Sampling without
Replacement’;
title2 ‘Estimating a Sample Mean and its Variance Ignoring the
Stratification’;
proc surveymeans data=sample_str_SRSWOR total=2000 mean var;
  var exp_travel;
run;

data totals;
  length cityside $4;
  input cityside _TOTAL_;
datalines;
East 1000
West 1000
;
run;

title2 ‘Estimating a Sample Mean and its Variance Accounting
for the Stratification’;
proc surveymeans data=sample_str_SRSWOR total=totals mean var;
  strata cityside;
  var exp_travel;
run;
```

Stratified Simple Random Sampling without Replacement

Estimating a Sample Mean and Its Variance Ignoring the Stratification

SURVEYMEANS Procedure

Data Summary	
Number of observations	400

Statistics			
Variable	**Mean**	**Std Error of Mean**	**Var of Mean**
exp_travel	1363.179844	92.594306	8573.705490

Stratified Simple Random Sampling without Replacement
Estimating a Sample Mean and Its Variance Accounting for the Stratification

SURVEYMEANS Procedure

Data Summary	
Number of strata	2
Number of observations	400

Statistics			
Variable	**Mean**	**Std Error of Mean**	**Var of Mean**
exp_travel	1363.179844	77.564901	6016.313916

The sample mean reported by PROC SURVEYMEANS is the same ($1363.18) in either case, but accounting for the stratification reduced the variance by almost one-third. Aside from a few rare circumstances, stratification increases the precision of overall estimates. It should be acknowledged, however, that any gains achievable are variable-specific and less pronounced for dichotomous variables (Kish, 1965). For instance, expenditures on OTC medications are likely much less disparate across CITYSIDE as expenditures of this sort seem less influenced by personal wealth than those related to travel.

Because sampling is performed independently within each stratum, we are able to essentially eliminate the between-stratum variability and focus only on the within-stratum variability. To see this, consider how the estimated variance of the overall sample mean under this sample design is given by

$$\text{var}(\hat{\bar{y}}) = \sum_{h=1}^{H=2} \left(\frac{N_h}{N} \right)^2 \frac{s_h^2}{n_h} \left(1 - \frac{n_h}{N_h} \right) = \sum_{h=1}^{H=2} \left(\frac{N_h}{N} \right)^2 \text{var}(\hat{\bar{y}}_h) \tag{1.1}$$

where

N_h is the stratum-specific population size
n_h is the stratum-specific sample size
s_h^2 is the stratum-specific element variance

We can conceptualize this as the variance of a weighted sum of stratum-specific, SRSWOR sample means, where weights are determined by the proportion of the population covered by the given stratum, or N_h/N, where $\sum_{h=1}^{H=2} N_h/N = 1$.

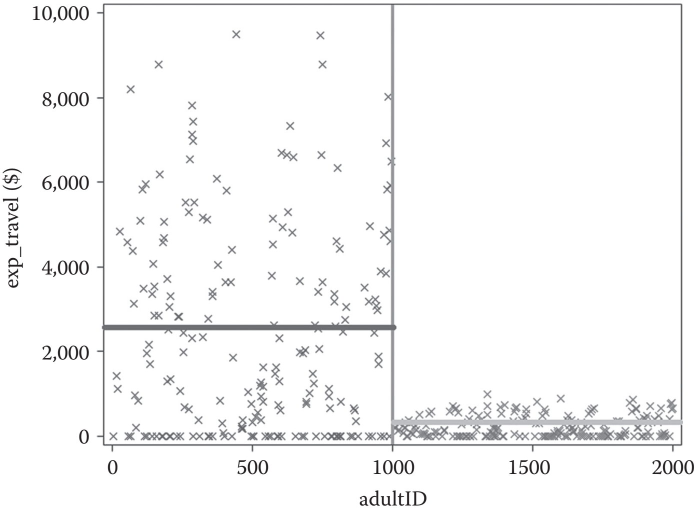


FIGURE 1.4
Visual depiction of the effect of stratification on variance computations.

Figure 1.4 is provided to help visualize what is happening in Equation 1.1 when the SAMPLE_STR_SRSWOR data set is analyzed by PROC SURVEYMEANS with CITYSIDE specified in the STRATA statement. The vertical reference line represents the boundary between the two strata. For the west side ($h=1$), the variance is a function of the sum of the squared distances between the points plotted and the horizontal reference line around $2500, the stratum-specific sample mean. For the east side ($h=2$), the same can be said for the horizontal reference line around $250. When the stratification is ignored, the vertical boundary disappears and a single horizontal reference line would replace the two stratum-specific lines around $1400, the mean expenditure for data pooled from both strata. The point is that the sum of the squared distances to this new reference line would be much greater, overall, which helps explains why measures of variability are larger in the second PROC SURVEYMEANS run when the stratification is ignored.

1.4.4 Clustering

The third feature of complex survey data is *clustering*. This occurs whenever there is not a one-to-one correspondence between sampling units and population elements; instead, the sampling unit is actually a cluster of two or more population elements. A few examples include sampling households in a survey measuring attitudes of individuals, sampling doctor's offices in a survey measuring characteristics of patient visits to doctors' offices, or sampling classrooms in an education survey measuring the scholastic aptitude of students. Clustering is rarely ideal as it generally decreases precision, but it is often a logistical necessity or used to control data collection costs. For instance, most nationally representative, face-to-face surveys in the United States sample geographically clustered units to limit interviewer travel expenses.

Whereas homogeneity within strata leads to precision gains, homogeneity within clusters has the opposite effect. Let us illustrate this phenomenon by considering the following alternative sample design for the expenditure survey example. Suppose the 2000 residents are evenly distributed across the town's $C = 100$ blocks—that is, exactly $N_c = 20$ adults reside on each unique block—and that the market research firm decides data collection would be easier to orchestrate if the sampling units were blocks themselves. Perhaps, they use a town map to enumerate all blocks and select an SRS of $c = 20$ of them, collecting expenditure data on all adults living therein. Note that this alternative design still maintains a sample size of 400. Suppose the survey is administered and the results are stored in the data set called SAMPLE_CLUS. To isolate the effect of clustering, we will assume there was no stratification and, for simplicity, we will ignore the FPC.

Whenever the underlying sample design of the complex survey data set involves clustering, we should place the cluster identifier variable(s) in the CLUSTER statement of the given SURVEY procedure. In the present example, this identifier is the variable BLOCKID. Program 1.3 is comprised of two PROC SURVEYMEANS runs, one assuming simple random sampling and another properly accounting for the clustering. As before, we are requesting the sample mean and its estimated variance, only this time for both expenditure variables, EXP_OTCMEDS and EXP_TRAVEL.

Program 1.3: Illustration of the Effect of Clustering on Measures of Variability

```
title1 'Cluster Sampling';
title2 'Estimating a Sample Mean and its Variance Ignoring the
Clustering';
proc surveymeans data=sample_clus mean var;
  var exp_OTCmeds exp_travel;
run;

title2 'Estimating a Sample Mean and its Variance Accounting
for the Clustering';
proc surveymeans data=sample_clus mean var;
  cluster blockID;
  var exp_OTCmeds exp_travel;
run;
```

Cluster Sampling

Estimating a Sample Mean and Its Variance Ignoring the Clustering

SURVEYMEANS Procedure

Data Summary	
Number of observations	400

Statistics			
Variable	**Mean**	**Std Error of Mean**	**Var of Mean**
exp_OTCmeds	18.430203	0.709202	0.502968
exp_travel	1271.310549	101.074843	10,216

Cluster Sampling

Estimating a Sample Mean and Its Variance Accounting for the Clustering

SURVEYMEANS Procedure

Data Summary	
Number of clusters	20
Number of observations	400

Statistics			
Variable	**Mean**	**Std Error of Mean**	**Var of Mean**
exp_OTCmeds	18.430203	0.814593	0.663563
exp_travel	1271.310549	320.315188	102,602

This is yet another instance where ignoring a feature of the complex survey data set does not affect the point estimate since the sample means are identical in both PROC SURVEYMEANS runs, but the clustering does impact measures of variability. Failing to account for clustering is especially risky because clustering can prompt a significant increase in the estimated variances. One might notice the increase is far more dramatic for EXP_TRAVEL than EXP_OTCMEDS. The explanation has to do with the degree of homogeneity or how correlated adults' responses are within clusters with respect to the given outcome variable.

The reader might find a visualization of homogeneity useful prior to the presentation of an equation commonly used to quantify it. To this end, Figure 1.5 plots the distribution of the two expense variables within the sampled clusters. The cluster-specific means are represented by a dot and flanked by 95% confidence interval end points (not accounting for the any design features, only to illustrate the within-cluster variability). The idea is that the further away the dots are from one another or the more dissimilar the confidence intervals appear, the larger the expected increase in variance when factoring in the clustering in the sample design. Contrasting the right panel to the left helps explain why the variance increase for travel expenditures trumps that for OTC medications.

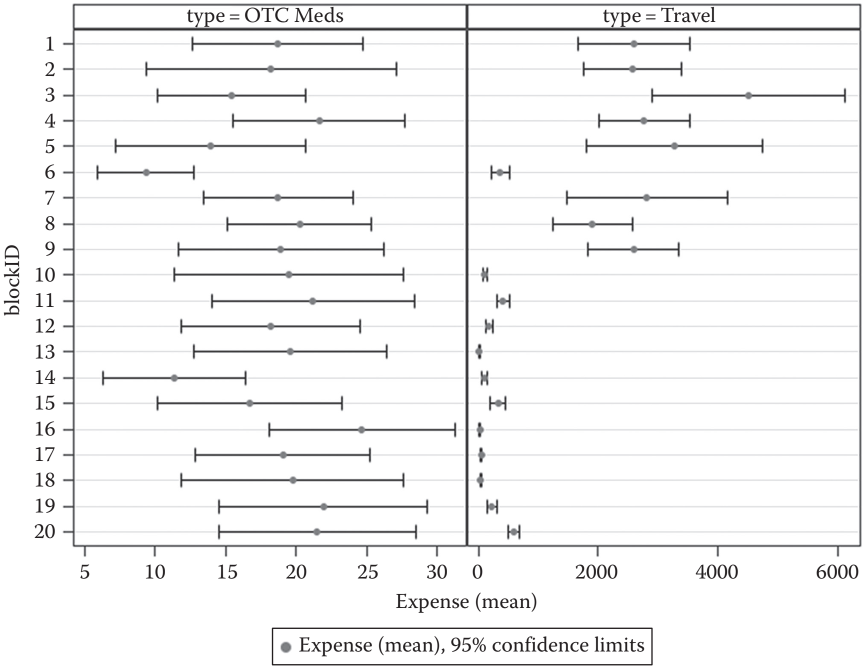


FIGURE 1.5
Expenditure distributions within blocks selected as part of a clustered sample design for the hypothetical expenditure survey.

For the special case of equally sized clusters, an alternative method to calculate the variance of a sample mean further elucidates this concept. Specifically, one can provide a summarized data set containing only the cluster-specific means to PROC SURVEYMEANS and treat as if it were an SRS. Program 1.4 demonstrates this method. It begins with a PROC MEANS step storing the cluster means of expenditure variables in a data set named CLUSTER_MEANS. The resulting data set of 20 observations is then analyzed by PROC SURVEYMEANS without any statements identifying complex survey features.

Program 1.4: Illustration of the Reduced Effective Sample Size Attributable to Clustering

```
proc means data=sample_clus noprint nway;
  class blockID;
  var exp_OTCmeds exp_travel;
output out=cluster_means mean=;
run;

proc surveymeans data=cluster_means mean var;
  var exp_OTCmeds exp_travel;
run;
```

SURVEYMEANS Procedure

Data Summary	
Number of observations	20

Statistics			
Variable	**Mean**	**Std Error of Mean**	**Var of Mean**
exp_OTCmeds	18.430203	0.814593	0.665363
exp_travel	1271.310549	320.315188	102,602

Indeed, we can confirm the measures of variability output match those from the second run of Program 1.3 in which we provided a data set of the full sample (400 observations) to PROC SURVEYMEANS but specified BLOCKID in the CLUSTER statement. The point of this exercise is to illustrate how clustering reduces the effective sample size. Kish (1965) coined the phrase *design effect* to describe this phenomenon.

The design effect of an estimate $\hat{\theta}$ is defined as the ratio of the variance accounting for the complex survey features to the variance of an SRS of the same size or

$$Deff = \frac{\text{Var}_{complex}(\hat{\theta})}{\text{Var}_{SRS}(\hat{\theta})} \tag{1.2}$$

A *Deff* of 2 implies the complex survey variance is twice as large as that of a comparable SRS. Equivalently, this is to say that the effective sample size is one-half of the actual sample size. While it is possible for certain complex survey designs to render design effects less than 1, meaning designs that are more efficient than an SRS, clustering typically causes this ratio to be greater than 1.

In the special case of an SRS of equally sized clusters, an alternative expression for Equation 1.2 is

$$Deff = 1 + (N_c - 1)\rho \tag{1.3}$$

where

N_c is the number of population elements in each cluster

ρ is the *intraclass correlation coefficient* (ICC), a measure of the clusters' degree of homogeneity

The ICC is bounded by $-1/(N_c - 1)$ and 1. The extreme value on the lower end corresponds to all clusters sharing a common mean. At the other extreme, a value of 1 implies perfect homogeneity within clusters or all elements therein sharing a common value. In practice, negative values of ρ are rare. Most common are slightly positive values. Despite a seemingly small