Sept. 12, 2006

SAMPLE SURVEY THEORY AND METHODS

A Correspondence Course

Organized by Dr. Phillip S. Kott,
Chief Research Statistician
National Agricultural Statistics Service (USDA)
3251 Old Lee Highway
Fairfax, VA 22030

pkott@nass.usda.gov
phone: 703-877-8000 x 102

PREFACE

Overview

This course provides a solid grounding in modern survey-sampling theory and methods. The lesson plan includes the first 9 chapters of the assigned text (probability sampling, stratification, allocation, multi-stage sampling, ratio and regression estimation, domain estimation, variance estimation in complex surveys, and methods for handling nonresponse) and Chapter 12 (two-phase sampling, small-domain estimation, multiple frames, capture/recapture). This is augmented by the instructor's comments on the text plus supplemental notes on topics like Wilson confidence intervals for small proportions, unequal probability sampling, and the large-sample properties of common estimation strategies. A final lesson covers the multiple-regression estimator.

Required Textbook

Sharon L. Lohr, *Sampling: Design and Analysis*, Duxbury Press, Pacific Grove, CA. 1999. (Go to http://www.duxbury.com/statistics_d/ , and then search for title)

Prerequisites:

At least one previous college-level course in mathematics or statistics and a familiarity with fundamental probabilistic concepts (expectations, variances) is essential. More advanced knowledge will be helpful for some of the supplemental notes.

The ability to download and manipulate (sort, add, take squares, etc.) data on a CD-ROM accompanying the textbook is required to do the homework assignments.

The final lesson will be nearly impossible to follow without previous course-work in matrix algebra. Students without such exposure may omit that lesson.

Course Origin

This course has been developed by the National Agricultural Statistics Service (NASS) of the US Department of Agriculture with the intention of providing training to professional personnel located in various State offices and in Washington, D.C. It replaces a previous course developed by a predecessor USDA agency for the same purpose. Although

tempered by the NASS experience, the course can be of value to anyone needing to learn about the how's and why's of modern survey sampling. There is, for example, material not easily found elsewhere on estimating proportions, which is not a major concern at NASS (although it can be for some clients of NASS services, whether at other USDA agencies or in the states).

Course Organization

The course largely follows Professor Lohr's lucid and thorough book on modern survey sampling. Two chapters from that book on the analysis of survey data are of secondary interest to *producers* of survey statistics and omitted from the course.

Commentary on the chapters and supplemental notes on a number of subjects are contained herein. A good deal of this additional material is not often covered in traditional survey-sampling courses.

Course Objectives

The course will develop the student's ability for choose the proper sampling and estimation strategy when planning a survey. It will familiarize the student with many popular sampling techniques, demonstrate the advantages and disadvantages of alternative strategies (sample plan and estimation method), and discuss the circumstances under which each, or some combination, can be expected to provide statistically and operationally efficiency. The course has broad application to all kinds of sample-survey situations, be they of households, establishments, institutions, animals, vegetation, land masses, or something else. Special emphasis is placed on techniques used by NASS.

More students will begin this course than end it. It is often difficult for working professionals to find the time to master the contents of a course like this one and to do the assignments. Moreover, there is enough material here to fill two full trimesters at an accredited university, one of them at the graduate level.

A student will not have to finish all 14 lessons to profit from the course. Anyone successfully completing the assignments for Lessons 1 through 7, 9, 10, and 12 (the last without reading the accompanying supplemental note) will have finished the equivalent of a full semester's work in the theory and practice of survey sampling.

That being said, the material in Lessons 8, 11, 13, and 14 are of particular interest to NASS. These lessons cover Brewer and maximal-Brewer selection (what has come to be called "multivariate probability-proportional-to-size sampling" or "mpps" at NASS), two-phase sampling, multiple-frame sampling, and calibration weighting. Some students may

want to acquaint themselves with all or part of this material even if they do not complete the accompanying homework assignments.

Course Operation

Each student is held responsible for reading the assigned material, lesson by lesson, and for doing all the assigned homework . It is highly recommended that the reading be done *before* attempting the homework. Many students will have difficulty with some of the supplemental readings, especially those dealing with large-sample properties and variance estimation for complex statistics. Nevertheless, they should be able to satisfactorily complete the course (in all or in part) without mastering that material.

There is no final examination. Instead, a "graduating" student will be encouraged to complete a project utilizing what (s)he has learned, most helpfully on a topic of interest to NASS. Such a student should prepare a journal-length paper summarizing the results of the project and a 15-minute talk advertising the paper of sufficient quality that a typical government statistical agency or contractor would approve of its presentation at a technical session of the Joint Statistical Meetings.

Studying statistics is like studying mathematics in that material nearly incomprehensible when first encountered can seem obvious after the third or fourth exposure. Moreover, "one sound brick must be laid on each earlier brick." Long gaps between lessons make review of earlier lessons necessary and can materially lengthen the total time required for completing the course. Work submitted will be returned to the student after grading. Students need not wait for this work to be returned, but can move ahead to the next lesson. Although this is largely a self-study course, the instructor is available to guide the student through difficulties (s)he may have with the text, commentary, supplemental notes, and homework assignments.

THE LESSONS

(Chapters and Sections refer to *Sampling: Design and Analysis* by Lohr)

Lesson 1 Introduction: Chapter 1. With some additional definitions..... 1

Lesson 2 Simple Probability Samples I: Sections 2.1 to 2.3. 3
Supplemental note on asymptotics in survey sampling*..... 6

Lesson 3 Simple Probability Samples II: Sections 2.4 through 2.9. 15
Supplemental note on Wilson confidence intervals for proportions under simple random sampling..... 19

Lesson 4 Ratio and Regression Estimation: Chapter 3.
Includes additional material on asymptotics and the weighted-residual variance estimator..... 22
Table of large-sample results (for SRS)*..... 28

Lesson 5 Stratified Sampling: Chapter 4. 31

Lesson 6 Cluster Sampling with Equal Probabilities: Chapter 5. 35

Lesson 7 Sampling with Unequal Probabilities I: Sections 6.1 - 6.5. 37
Supplemental note on particular unequal-probability designs including a discussion of the Hartley-Rao variance estimator 40

Lesson 8 Sampling with Unequal Probabilities II: Sections 6.6, 6.7. 48
Supplemental note on Brewer and maximal-Brewer selection..... 51

Lesson 9 Complex Surveys: Chapter 7. 59

Lesson 10 Variance Estimation in Complex Surveys: Chapter 9
(Section 9.1 requires Calculus and can be viewed as optional).
Additional material treats linearization without Calculus and other other subjects..... 62
Supplemental note on the delete-a-group jackknife..... 71

* Asymptotics play a major role in modern survey-sampling theory. Consequently, students should attempt to digest the material presented in these notes. Experience, however, suggests that many will not. As a result, no homework assignment will depend on them.

Lesson 11 Two-Phase Sampling: Section 12.1. 73
Supplemental note on the double and reweighted expansion estimators. 76

Lesson 12 Nonresponse: Chapter 8.
Additional material ties in supplemental note in Lesson 11 to the treatment of nonresponse..... 84

Lesson 13 Other Topics in Sampling: Sections 12.2 - 12.5, excluding Subsection 12.2.4 (Read supplemental note below first.) 91
Supplemental note on multiple-frame estimation 93

Lesson 14* Multiple Regression and Calibration Weighting:
A supplemental note.....108

Additional References 126

* Requires familiarity with matrix algebra

Note: Completing Lessons 1, 2, 3, 4, 5, 6, 7, 9, and 10, coupled with reading Chapter 8 in *Sampling: Design and Analysis* and doing the homework assignment at the end of Lesson 12 (page 90 of these notes) is the equivalent of a rigorous semester-long class is survey sampling. A student may want to end the course at that point and study some or all of the remaining material informally.

Lesson 1: Introduction

Comments on Chapter 1 of *Sampling: Design and Analysis*

One of the problems with survey sampling is that terms do not always have definitive meanings. They can vary from textbook to textbook. In these notes, we will usually conform to the usage Professor Lohr has adopted. Sometimes, very soon in fact, alternative terminology and definitions will be discussed. With the plasticity of meaning in mind, students should take pains to define terms when writing about survey-sampling issues outside of this course.

The Population of Interest

Like many textbooks, *Sampling: Design and Analysis* draws a distinction between the *target* and *sample* populations (p. 3), calling the target population "[t]he complete collection of observations we want to study," and the sample population "[t]he collection of all possible population observation unit that might have been chosen in the sample." For example, a statistician may want to study some behavior within the entire adult population of the US, the target population, but for operational reasons her sample population excludes residents of Alaska, Hawaii, and the territories (Puerto Rico, the Virgin Islands, etc.).

A different set of definitions follows Groves (1989*). In the above example, the entire US adult population is the *population of inference* or *inferential population*, while the adult population in the 48 contiguous states is the target population. The sample population (which is also known as the *frame population*) includes only those members of the 48 states who could have been sampled with the methodology employed. For example, if the survey is conducted in April, the sample population excludes residents of the 48 states who are out of the country during that month.

Another term, the *survey population*, refers to those members of the sample population who would have *responded* to the survey if selected for the sample.

* Most references can be found in *Sampling: Design and Analysis*. Additional references are in the back.

Errors of Observation/Errors of Nonobservation

It is standard to draw a distinction between sampling and nonsampling errors as the text does in Section 1.6. Groves (1989) argues that the more important distinction is between errors of observation and errors of nonobservation. Errors of observation are measurement errors caused by a disconnect between what the analysts wants to measure and what the sample respondent actually provides. They are beyond the scope of this course.

Errors of nonobservation are due to differences between the target population and the respondent sample. They have three potential sources.

1) The sample population is not identical to the target population.
2) Not every element (observation unit) in the sample population is selected for the sample.
3) Not every sampled element responds to the survey.

The primary focus of this course is the second source: sampling error. Lesson 11 is mostly dedicated to the third: survey nonresponse. The first source is referred to as *coverage* or *frame errors*. Sometimes, the sample population is *larger* than the target population; for example, the sample population may be a list of business establishments including dormant entities not in the target population. A simple technique for handling potential overcoverage in this situation is discussed in Lesson 3. Undercoverage, errors of nonobservation due to the sample population missing elements from the target population, is a harder problem to address. One popular method for handling undercoverage is poststratification, which can also be used to adjust for nonresponse. It is discussed in both contexts in Lesson 11.

Unless otherwise noted, the target and sample populations will be assumed to be identical in the remaining lessons of this course.

Homework for Lesson 1

Do Exercises 1, 2, and 8 in Section 1.7 (starting on p. 17) of *Sampling: Design and Analysis.*

Lesson 2: Simple Probability Samples I

Comments on Sections 2.1 to 2.3 of *Sampling: Design and Analysis*

Terminology

Simple random sampling without replacement is sometimes designated *srswtr*. Särndal *et al.* (1992) use *SI*. They likewise use *SIR* for simple random sampling with replacement.

Often the right side of equation (2.10) on page 44 is called the "*estimated* standard error of $\bar{y}$ (not the "standard error"). More often than that, it is denoted $\hat{SE}(\bar{y})$. In my own work, I prefer to use lowercase letters to denote estimators (*se* for estimated standard errors, *v* or *var* for estimated variances, *cv* for estimated coefficients of variation) and uppercase letters for their full-population analogues. Students may find it convenient to adapt this practice in their homework, especially if they find "hats" are difficult to type. Here, we follow the text.

What is often called the "standard error of $\bar{y}$" (but not by Professor Lohr), we will call the *standard deviation* of $\bar{y}$,

$$SD(\bar{y}) = \sqrt{V(\bar{y})},$$

which should not be confused with the *population standard deviation*, S. The text fails to define the coefficient of variation for $\bar{y}$, although it does define the estimated coefficient of variation. The *coefficient of variation* for $\bar{y}$ is

$$CV(\bar{y}) = SD(\bar{y}) / \bar{y}_U.$$

Notice that in estimating $CV(\bar{y})$ (in equation (2.11) on page 33) both $SD(\bar{y})$ and $\bar{y}_U$ are replaced by estimates. The square of $CV(\bar{y})$ is called the *relative variance* of $\bar{y}$:

$$relV(\bar{y}) = V(\bar{y}) / \bar{y}_U^2 = [CV(\bar{y})]^2.$$

The terms "standard error of $\bar{y}$," "coefficient of variation of $\bar{y}$," etc. can be applied to estimators other than the sample mean. One of the homework assignments will be to

define these terms for an estimator $\hat{t}$ of t.

Root Mean Squared Errors and Related Terms

An important concept is the root mean squared error of an estimator. This is because many estimators in survey sampling are not unbiased. The *root mean squared error* of an estimator $\hat{t}$ for t is

$$\mathrm{RSME}(\hat{t}) = \sqrt{\mathrm{MSE}(\hat{t})}.$$

Related concepts are the *estimated root mean squared error*, $\widehat{\mathrm{RMSE}}(\hat{t})$, the *relative root mean squared error*,

$$\mathrm{relRMSE}(\hat{t}) = \mathrm{RMSE}(\hat{t}) / t,$$

and the *estimated relative root mean squared error*, $\widehat{\mathrm{relRMSE}}(\hat{t}) = \widehat{\mathrm{RMSE}}(\hat{t})/\hat{t}$. The squares of the relRMSE and the estimated relRMSE are the relative mean squared error (relMSE) and estimated relative mean squared error, respectively.

When $\hat{t}$ is unbiased, the estimated root mean squared error of $\hat{t}$ is its standard error (as defined in the text), and its relative root mean squared error is its coefficient of variation. The formulae discussed above apply to any estimator, not just estimators of totals.

The Expansion Estimator

The text is ambiguous in Chapter 1 on the meaning of $\hat{t}$. Sometimes it is a general estimator for $t = \sum_U y_i$. At others it is the *expansion estimator*,

$$\hat{t} = N\bar{y} = (N/n)\sum_S y_i,$$

under simple random sampling.

In these notes $\sum_U$ and $\sum_S$ are compact ways of expressing the sum over all elements in the population and sample, respectively.

The Variance for an Estimated Proportion

The text correct states the variance of an estimated proportion based on a simple random sample in equation (2.15) of page 35. When $N \geq 50$, which is usually the case in practical application, $N/(N-1) \approx 1$, and equation (2.15) has the useful approximation:

$$V(\hat{p}) \approx (1 - n/N)\, p(1 - p)/n$$

$$= (1 - f)\, p(1 - p)/n, \qquad \text{(S1)}$$

where $f = n/N$ is called the *sampling fraction* under SRS.

We often will appear more concerned with computing an estimator for variance (see equation (2.16) in this context), rather than with the variance itself. This is because the latter most often cannot be directly measured in practice. In the case of a proportion, however, the variance formula can be used to create Wilson confidence intervals as we will see next lesson.

Useful additional reading – Pages 423 to the middle of 428 in Appendix B of *Sampling: Design and Analysis*, up until the words “close but not equal.”

On Asymptotics in Survey Sampling

Cochran Consistency

Suppose we wanted to estimate the soybean yield (bushels per acres) from three farms by selecting one of those farms at random giving each farm an equal probability of selection. The situation is described in the following table:

Farm	Harvested acres	Bushels	Bushels per acre
1	100	3,000	30
2	200	7,000	35
3	300	12,000	40

The soybean yield is 36.667 bushels per acre; that is (3,000 + 7,000 + 12,000)/(100 + 200 + 300). The expected value of our estimator, however, is 35 bushels per acre, since it can be 30, 35, or 40 with equal probability. Our estimator is biased!

Let $t_y = \sum_U y_i$, $t_x = \sum_U x_i$, and $\bar{x}$ the estimator for the population mean of the x-values. In general, we estimate a ratio like $B = t_y/t_x$ under simple random sampling with $\hat{B} = \bar{y}/\bar{x}$, which need not be unbiased. We will learn more about ratios and ratio estimators in Lesson 4. For now, it is easy to see that had we chosen the entire target population as as the sample, then $\hat{B}$ would equal B.

Cochran (1983) calls an *estimation strategy* – the combination of a sampling plan and an estimator, $\hat{t}$, "consistent" if $\hat{t} = t$ when n, the sample size, equals N, the population size. We will refer to this property as *Cochran consistency*.

The problem with Cochran consistency is that it does not behave as we wish consistency would. An estimator, $\hat{t}$, is thought to be consistent if the absolute difference between $\hat{t}$ and t is small when n is very large. Some government agencies to base their statistics on purposive samples of what they judge to be the n biggest units in the population.

If that judgment were based on y values, both $n\bar{y}_S$ and $N\bar{y}_S$, where $\bar{y}_S$ is the simple average of the y_i in the sample, would be Cochran consistent estimators for $t_y = N\bar{y}_U$. When N is several times the size of n, one of the estimators (and sometimes both) has to be doing a terrible job even when n itself is very large.

Design Consistency

A more useful definition of consistency relates to the accuracy of $\hat{t}$ as both n and N grow *arbitrarily large* (as large as one can think and then maybe larger). Many survey statisticians resisted developing an *asymptotic* (large-sample) theory for two reasons: Target populations in survey sampling are inherently finite, so it seems unreasonable to develop a theory based on their sizes growing arbitrarily large. Moreover, many estimands in survey samples are totals, like t, which themselves would grow arbitrarily large along with the population.

The real issue in asymptotics, however, is the size of the sample, not the population. Sample sizes in survey sampling are usually very large, often in the hundreds and thousands. They are typically much larger than in other branches of statistics. The goal of asymptotic theory is not to study the effects of arbitrarily large samples sizes *per se*, but the impact of a realized large sample sizes using an idealized (i.e., mathematically simplified) framework. Asymptotic theory is actually more relevant in survey sampling than most other branches of statistics.

As for a total, like t, as long as it is positive, as is typically the case in survey sampling, we can concentrate on its *relative error*. The relative error of an estimator $\hat{t}$ for t is $RE(\hat{t}) = (\hat{t} - t)/t$. When the relative error of $\hat{t}$ tends to zero (in probability) as the sample and population grow arbitrarily large, $\hat{t}$ is said to be *design (or randomization) consistent.* Actually, it is not the estimator that is design consistent, but the estimation strategy. That design consistency is not just a property of the estimator, but the estimator and the sampling design taken together, is one reason the prefix “design” is used.

Asymptotic theory is mathematically difficult, so much so that it is usually ignored is survey-sampling courses even though it has a pivotal role in modern theory. We offer here a stripped-down version of the necessary asymptotic theory. For a rigorous treatment of asymptotics in statistics, see Fuller (1976). For a fuller treatment of asymptotics in survey sampling, see Isaki and Fuller (1981). For our purposes, *an estimation strategy for t is design consistent if the relative mean squared error of* $\hat{t}$ *tends to zero as the sample and population grow arbitrarily large.*

We have been vague on how the sample and population grow arbitrarily large. For simple random sampling, we can think as a sequence of populations, $U_1 \subset U_2 \subset U_3 \subset \ldots$ with sample sizes N_1, $N_2 = 2N$, $N_3 = 3N$, ... From each population, U_r, a fresh simple random sample of size $n_r = rn_1$ is selected. The relative mean squared error (and relative variance) of the sample mean, $\bar{y}_r$, as an estimator of the population mean, $\bar{y}_{Ur}$, of U_r is

$$\text{relMSE}(\bar{y}_r) = [(1 - n_r/N_r)S_r^2/n_r] / \bar{y}_{Ur}^2 = [(1 - f)S_r^2/(rn_1)] / \bar{y}_{Ur}^2,$$

where $S_r^2 = [1/(N_r - 1)] \sum^{N_r} (y_i - \bar{y}_{Ur})^2$, $f = n_1/N_1 = n_r/N_r$ for all r. As r grows arbitrarily large, so does the sample and population size. As r grows arbitrarily large, if $\bar{y}_{Ur}$ in the denominator converges to a positive value, say $\bar{y}_\infty$, and S_r^2 converges to something finite, say say S_∞^2, then relMSE($\bar{y}_r$) converges to zero. Thus, $\bar{y}$ is design consistent under simple random sampling. The estimator, $\hat{t}_y = N\bar{y}$, which has the same relMSE as $\bar{y}$, is likewise design consistent under simple random sampling.

Although it is technically r, N_r, and n_r that grow arbitrarily large in asymptotic theory, we conveniently, if sloppily, will speak of n and N getting large in the remainder of this discussion. After all, it is size of the actual sample, the real n, that matters in practice. One useful way of expressing this is to say that n is *sufficiently large* that the results of asymptotic theory have practical relevance.

Another technical point: One needs quantities like $\bar{y}_U$ and S_r^2 to converge to finite values, $\bar{y}_\infty$ and S_∞^2, as r grows arbitrarily large for the asymptotic theory to work. Although a clever statistician can think of sequence of populations where such convergences don't take place, we will always assume they do in these notes.

Some Examples

Let us return briefly to the ratio estimator, $\hat{B} = \bar{y}/\bar{x}$, for $B = \bar{y}_U/\bar{x}_U$. As n grows arbitrarily large, both the mean squared errors of $\bar{y}$ and $\bar{x}$ tend toward zero. Consequently, as long as the limit of $\bar{x}_U$ is positive, it seems reasonable that the mean squared error of $\hat{B}$, which is the average of the square of the difference between $\hat{B}$ and B across all possible samples, also tends toward zero. This is not a formal proof (Lesson 4 hits closer to the mark), but it does give us a feel for what will happen with a sufficiently large sample.

We will see in the next lesson that s^2 (defined in equation (2.8) on page 33) is an unbiased estimator for the finite population variance, S^2 (defined in equation (2.5) on page 29) under simple random sampling. Rarely, however, is s an unbiased estimator for S. To see why the expected value of the square root of a random variable is not necessarily the square root of its expectation, consider the random variable, X, in the following table:

x	P(X = x)	q = ↓x
1	½	1
4	½	2
E(X) = 2.5		E(Q) = 1.5

X takes on the value 1 with probability ½, and the value 4 with probability ½. Its expected value is 2.5 ([1 x ½] + [4 x ½]) . The random variable Q = ↓X takes on the value 1 with probability ½ (when X =1) and 2 with probability ½ (when x = 2). Clearly, E(Q) = 1.5 is not equal to ↓E(X) = ↓2.5 ≈1.58.

Although s is not typically an unbiased estimator for S, it is design consistent under simple random sampling, as we shall soon see.

Asymptotic Orders

Equation (2.7) on page 32 shows how the variance of the estimator $\bar{y}$ relates to the finite population variance, S^2, while equation (2.9) on page 33 does the same for the estimator of $V(\bar{y})$ and s^2. We would like to say that $\hat{V}(\bar{y})$ is a consistent estimator for $V(\bar{y})$, but we have a problem with the relative error of $\hat{V}(\bar{y})$, $RE\{\hat{V}(\bar{y})\} = [\hat{V}(\bar{y}) - V(\bar{y})]/V(\bar{y})$. As the sample size grows arbitrarily large, $V(\bar{y})$ in the denominator of $RE\{\hat{V}(\bar{y})\}$ tends towards zero. How can we tell if the numerator is tending towards zero any faster?

The answer is with asymptotic orders. First, note that, for any n, $V(\bar{y}) = (1 - f)S^2/n$ is a constant, even though $\bar{y}$ is a random variable. We say that as n grows arbitrarily large, a constant, c, is O(1/n) – *of asymptotic order 1/n* – if nc converges to a fixed finite limit, which can be zero. More precisely put, c is a sequence of constants, $\{c_n\}$. The value of c_n can vary as n grows, but nc_n converges to a fixed finite limit.

We can flesh this out a bit with the constant $V(\bar{y})$. Observe that $nV(\bar{y})$ converges to $(1 - f)S_\infty^2$, a finite value. Thus, $V(\bar{y})$ is O(1/n), which is intuitively appealing since it can be expressed as a positive value, $(1 - f)S^2$, times 1/n.

One can replace n in the definition O(1/n) by any power of n. In practice, we usually see constants that are O(1/n), $O(1/\sqrt{n})$, $O(1/n^2)$, $O(1/n^{3/2})$, or $O(1/n^0) = O(1)$. We can also see constants that are O(n), $O(n^2)$, or $O(n^{3/2})$. For example $V(\bar{y}) = (1 - f)S^2/n$ is O(1/n) while $1/V(\bar{y})$ is O(n) as long as S_∞^2 converges to a positive number. This effectively adds the weak assumption that not all the y_i have the same value (so S^2 is not zero). Of course, if all the y_i did have the same value, there wouldn't be much need to draw a sample.

Many random variables are conceptually unbounded. A normally-distributed random variable, for example, has a small probability of being greater than any number you can imagine. For that reason, statisticians have developed the concept of a probability limit.

The formal definition of a probability limit is are beyond the scope of this course. For our purposes, a random variable, X, is $O_P(1/n)$ – *of asymptotic order 1/n in probability* – when $n\{E(X^2)\}^{1/2}$ converges to a fixed finite limit. Notice that X is first squared before taking its expected value. To see why, let $X = \hat{t} - t$. Since $E(X) = 0$, $nE(X) = 0$ for all positive n, yet it makes little intuitive sense to think of the random variable X as $O_p(1/n)$. On the other hand, $\{E(X^2)/n\}^{1/2} = \{E[(\hat{t} - t)^2]/n\}^{1/2} = \{N^2(1 - f)S^2/n^2\}^{1/2} = \{(1 - f)S^2/f^2\}^{1/2}$. So $\{E(X^2)\}^{½}$ is $O(\sqrt{n})$, and X is $O_p(\sqrt{n})$, which makes some sense since $\hat{t}$ and t grow with n, and the difference, $\hat{t} - t$, which averages to zero across *all* samples, tends to get larger in absolute terms for a particular selected sample along with $\{E[(\hat{t} - t)^2]\}^{1/2}$.

If a random variable X is $O_p(1/n)$, then one can show that E(X) is also O(1/n). This does not contradict what was said above. If $nE(X) = 0$ for all n, nE(X) does converge to a fixed finite limit, 0.

The sample mean $\bar{y}$ is easily shown to be $O_P(1)$ (recalling that we have assumed $\bar{y}_\infty$ is finite). Its relative error, $RE(\bar{y}) = (\bar{y} - \bar{y}_U)]/\bar{y}_U$, is $O_P(1/\sqrt{n})$. This is because the relative mean squared error of $\bar{y}$, which is this unbiased random variable's relative variance, $E\{[RE(\bar{y})^2]\} = (1 - f)S^2/(n\bar{y}_U)^2$ is O(1/n).

Most, but not all, of the estimation strategies we will study in this course will have $O_P(1/\sqrt{n})$ relative errors. Under simple random sampling and appropriate mild restrictions on the population values, this includes not only $\bar{y}$ (and $\bar{x}$), but also $\hat{B}$, s^2, and $\hat{V}(\bar{y})$.

We said before that any random variable with a relative mean squared error converging to zero as n grows arbitrarily large is design consistent. It is consequently easy to see that any random variable with an $O_P(1/\sqrt{n})$ relative error (and thus an O(1/n) relative mean squared error) is design consistent.

The following rules will prove helpful in calculations involving asymptotic orders.

If x is $O(n^a)$ and y is $O(n^b)$ with a > b (a and b can be negative),

x + y is $O(n^a)$, which can be written as $x + y \approx x$; x is said to *dominate* y.

xy is $O(n^{a+b})$.

x^c is $O(n^{ac})$.

Morever, if a < 0

$$(1 + x)^c = 1 + cx + [c(c-1)/2]x^2 + q, \text{ where } q \text{ is } O(n^{3a}). \qquad \text{(S2)}$$

$$= 1 + h, \text{ where } h \text{ is } O(n^a)$$

$$\approx 1.$$

This means in the following examples:

$12/n + 50/n^2$ is $O(1/n)$.

$(12/n + 50/n^2)(2/n^{1/2} + 45/n)$ is $O(1/n^{3/2})$.

$(12/n + 50/n^2)^{½}$ is $O(1/n^{1/2})$.

$[1 + 12/n + 50/n^2]^{½} = 1 + 6/n + q$, where q is $O(1/n^2)$

$= 1 + h$, where h is $O(1/n)$.

All the above apply when O(.) is replaced by $O_P(.)$.

Showing that s is Biased, Yet Consistent

Let us assume that s^2 is unbiased under simple random sampling and that its relative error is $O_p(1/\sqrt{n})$. What can we say about s?

We can rewrite s^2 as $s^2 = S^2 + (s^2 - S^2) = S^2\{1 + (s^2 - S^2)/S^2\}$. Consequently, invoking equation (S2),

$$\begin{aligned} s &= S\{1 + (s^2 - S^2)/S^2\}^{1/2} \\ &= S\{1 + \tfrac{1}{2}(s^2 - S^2)/S^2 - \tfrac{1}{8}[(s^2 - S^2)/S^2]^2\} + q, \text{ where } q \text{ is } O_p(1/n^{3/2}) \end{aligned} \qquad \text{(S3)}$$

From which we can conclude

$$\begin{aligned} E(s) &= S\{1 + 0 - \tfrac{1}{8}\,\mathrm{relV}(s^2)\} + E(q), \text{ where } E(q) = O(1/n^{3/2}) \\ &= S + h, \text{ where } h = O(1/n). \end{aligned}$$

That is to say, s has a negative bias as n grows large, but that bias itself tends towards zero as the sample grows. We say that s is *asymptotically unbiased* or *nearly unbiased for sufficiently large n*.

Furthermore, we can deduce from equation (S3),

$(s - S)/S = \tfrac{1}{2}(s^2 - S^2)/S^2 + p$, where p is $O_p(1/n)$, and

$E\{[(s - S)/S]^2\} = \tfrac{1}{4}E\{[(s^2 - S^2)/S^2]^2\} + w$,

where $E\{p(s^2 - S^2)/S^2\}$ is the dominant term in $w = O(1/n^{3/2})$.

The relative mean squared error of s is thus asymptotically ¼ of the relative mean squared of s^2. Put another way, s is design consistent with a $O_P(1/\sqrt{n})$ relative error.

Homework for Lesson 2

How do the variance, relative variance, standard deviation, relative standard deviation, mean squared error, relative mean squared error, root mean squared error, and relative root mean squared error of an estimator $\hat{t}$ for a total and $\hat{\bar{y}}_U = \hat{t}/N$ for a mean compare?

Look at the table in the middle of Page 428 of *Sampling: Design and Analysis*. Treat the four rows of numbers as the four observations in the population.

a. With all the SRS of size 1, compute $\hat{B}$. What is the bias and mean squared error of $\hat{B}$?

b. Do the same with all SRS of size 2 (there are six such samples).

c. Do the same with all SRS of size 3 (there are fours such samples)

d. Do the same with all SRS of size 4.

Lesson 3: Simple Probability Samples II

Comments on Sections 2.4 to 2.8 of *Sampling: Design and Analysis*

Two-Sided Wald Confidence Intervals

The observant student will notice that the book only discusses two-sided confidence intervals. Often one sees one-sided confidence intervals for a population mean based on a simple random sample of the form

$[\bar{y} - z_\alpha SE(\bar{y}), \infty)$ or

$(- \infty, \bar{y} + z_\alpha SE(\bar{y})]$,

where the $-\infty$ can be replaced by 0 when all $y_i \geq 0$ and $\bar{y} > 0$.

For $100(1 - \alpha)\%$ of samples to produce $\bar{y}$ such that $\bar{y}_U$ is within a one-sided confidence interval often requires either a much larger sample than the two-sided interval in the book,

$$[\bar{y} - z_{\alpha/2} SE(\bar{y}), \bar{y} + z_{\alpha/2} SE(\bar{y})], \tag{S4}$$

or a much more symmetric distribution of y_i values than we usually see in survey sampling.

The two-sided interval in equation (S4) is properly called a *two-sided Wald confidence interval*. The large-sample theory behind the Wald interval applies to all unbiased estimation strategies, not just $\bar{y}$ under simple random sampling.

A better interval that the one in equation (S4) would replace $SE(\bar{y})$ by the actually standard deviation of $\bar{y}$. Doing that speeds up the asymptotics considerably; that is to say, makes the confidence interval effective with a much smaller sample. Of course, we do not know the actually standard deviation of $\bar{y}$. Nevertheless, when y_i is a 0/1 variable, so that $\bar{y}_U$ is the proportion, p, we can do something. We can use $Var(\bar{y}) = (1 - f)^2[p(1 - p)]/n$ from equation (S1) to make a better two-side confidence interval than the one equation (S4). A supplemental note on these *Wilson confidence intervals* begins on page 19.

Confidence intervals are constructed *after* a random sample has been drawn. When determining what sample size is needed *before* the sample is drawn, p, which can only be guessed, is the relevant value. Sample-size determinations do not need a Wilson-like adjustment. It is prudent to be conservative when determining sample sizes, since once a sample is drawn, the statistician is stuck with it. Consequently, it makes sense to set S^2, which virtually equals $p(1 - p)$ when $N > 50$, at its maximum possible value, $p = ½$, so that $S^2 \approx p(1 - p) = 1/4$.

Systematic Sampling

Suppose we have a target population with 48 elements, and we want a systematic sample of size 5 to estimate $t = \sum_U y_i$. The alert student will quickly observe that N (48) is not divisible by n (5). A simple solution augments the frame population by two dummy members, elements 49 and 50 – with y_i set to 0 for both. In general, suppose n goes into N k times with remainder $r > 0$; that is, $N = nk + r$, where $0 < r < n$. We then add $n - r$ dummy observations to the frame. Let N_F denote the size of the augmented frame population (in our example, 50). Mathematically, $N_F = N + (n - r) = n(k + 1)$. Our estimator t is $\hat{t}_{SYS} = (N_F/n) \sum_S y_i$. In our example, $N_F/n = 50/5 = 10$.

One can also estimate $\bar{y}_U$ with $\hat{\bar{y}}_U = \hat{t}_{SYS}/N$. Both $\hat{t}_{SYS}$ and $\hat{\bar{y}}_U$ are unbiased under the randomization theory described in Section 2.7, but neither is unbiased under the model-based theory in Section 2.8. If $\hat{t}_{SYS}$ were model unbiased, then $E_M(\hat{t}_{SYS})$ would equal $E_M(t)$. When the sample includes one of the two dummy elements, $E_M(\hat{t}_{SYS})$ is $10(4\mu) = 40\mu$. When the sample does not contain one of the dummies, $E_M(\hat{t}_{SYS})$ is $10(5\mu) = 50\mu$. By contrast, $E_M(t)$ is always 48μ. It is interesting to note that the model expectation of $E_M(\hat{t}_{SYS})$ averaged over all possible samples is indeed 48μ. Model-based analysis, however, keys on the particular sample drawn and views the average across all samples as irrelevant.

Unlike the Sampling: Design and Analysis, we do not use upper case letters to describe random variables under the model.

Although $\hat{t}_{SYS}$ under systematic sampling is unbiased, it may not be design consistent. To see why suppose we have a population containing married couples, and we want to estimate the average body weight of an individual in that population. There are 1,000 couples, and 2,000 individuals in the population. We first sort the couples in alphabetical order by the husband's last name and then list each wife before her husband. If one wanted to select a systematic sample of 100 observations, the sampling interval would be 2,000/100 = 20. If the random integer chosen between 1 and 20 for this purpose were odd, then the first individual selected for the sample would be a woman, as would every subsequently sampled individual. If the random integer were even, the entire sample would be male. The estimator of average individual body weight would not be very good. If the sample contained all men, it would be too high. If the sample contained all women, it would be too low. Moreover, if we let the sample and population double (to 4,000) and then triple *ad infinitum* using the same sampling scheme, the estimator's relative error would not tend toward zero, because the sample would always contain either all males or all females.

We could fix this problem by assuming that there are no patterns of y-values in the list from which the systematic sample is selected. This is tantamount to assuming a model about the data. Many argue that randomization theory alone should be used what drawing inferences from surveys. "The problem with assuming models," they argue, "is that models often fail."

Model-based advocates point out that even the assumption that $\bar{y}_U$ and S^2 converge to finite constants, needed for establishing design consistency under simple random sampling, are model assumptions. That may be true, but those assumptions are very mild ones. By contrast, the reason a statistician uses systematic sampling is because (s)he believes something about the way the list has been ordered, often that elements next to each on the list tend to have similar y-values. Even if neighboring elements don't have similar y-values, the reasoning goes, the expansion estimator under systematic sampling when N is divisible by n should be just as good as (have no more variance than) the expansion estimator under simple random sampling.

The point being made here is that the use of systematic sampling from a ordered list requires an assumption about the y-values on the list, which, however reasonable, renders the estimation strategy design inconsistent. The use of systematic sampling can be made design consistent by randomly ordering the list from which the sample is selected. If this randomization occurs, and N is divisible by n, then systematic sampling is just another way of drawing a simple random sample.

The Random Variable Z_i

The Random variable Z_i, introduced on page 44 of the *Sampling: Design and Analysis*, is the heart of randomization-based sample-survey theory. It is more common to see this sample-inclusion indicator denoted "I_i," but we will follow the text here.

Useful additional reading – The remainder of Appendix B in the Sampling: Design and Analysis.

The Wilson Confidence Intervals for a Proportion

Let

p be a proportion ($p = \sum_U y_i /N$, where y_i is a 0/1 variable), and
$\hat{p}$, its estimator base on a simple random sample ($\hat{p} = \sum_S y_i /n$).

A $100(1 - \alpha)\%$ Wilson confidence interval for $\hat{p}$ contains all p values such that

$$\frac{|\hat{p} - p|}{\{(1 - f)p(1 - p)/n\}^{1/2}} \leq z_{\alpha/2}$$

Consequently,

$(\hat{p} - p)^2 \leq z_{\alpha/2}^2(1 - f)p(1 - p)/n$ or

$\hat{p}^2 - 2\hat{p}\, p + p^2 \leq p\, z_{\alpha/2}^2(1 - f)/n - p^2 z_{\alpha/2}^2(1 - f)/n$, which implies

$[1 + z_{\alpha/2}^2(1 - f)/n]p^2 - [2\hat{p} + z_{\alpha/2}^2(1 - f)/n]p + \hat{p}^2 \leq 0.$

The above is a polynomial in p.
Since $ap^2 + bp + c = 0 \;\Rightarrow\; p = [- b/2 \pm \{(b^2/4) - ac\}^{1/2}] /a$,
where $a = 1 + z_{\alpha/2}^2(1 - f)/n$
$b = - [2\hat{p} + z_{\alpha/2}^2(1 - f) /n]$
$c = \hat{p}^2$,

the extreme points of the inequality are (after some rearranging)

$$p = \frac{\hat{p} + z_{\alpha/2}^2(1 - f)/(2n) \pm z_{\alpha/2}\{ (1 - f)\hat{p}(1 - \hat{p})/n + z_{\alpha/2}^2(1 - f)^2/(4n^2)\}^{1/2}}{1 + z_{\alpha/2}^2(1 - f)/n}$$

Adding and subtracting $\hat{p}$ from the right hand side, we get the *Wilson confidence interval*

$$p = \hat{p} + \frac{(1/2 - \hat{p})\, z_{\alpha/2}^2(1 - f)/n \pm z_{\alpha/2}\{(1 - f)\hat{p}(1 - \hat{p})/n + z_{\alpha/2}^2(1 - f)^2/(4n^2)\}^{1/2}}{1 + z_{\alpha/2}^2(1 - f)/n} \quad . \qquad \text{(S5)}$$

Observe that the center of the Wilson confidence interval for p is $\hat{p}$ only when $\hat{p} = ½$; otherwise, it is somewhere between ½ and $\hat{p}$ when $\hat{p} > ½$ or somewhere between $\hat{p}$ and ½ when $\hat{p} < ½$.

Observe further that

when $\hat{p} = 0$, the confidence interval is $[0, \{z_{\alpha/2}^2(1 - f)/n\}/\{1 + z_{\alpha/2}^2(1 - f)/n\}]$, and
when $\hat{p} = 1$, the confidence interval is $[1/\{1 + z_{\alpha/2}^2(1 - f)/n\}, 1]$.

This contrasts with a two-sided Wald confidence interval,

$$p = \hat{p} \pm z_{\alpha/2}\{(1 - f)\hat{p}(1 - \hat{p})/(n - 1)\}^{½}$$

which collapses to a point when either $\hat{p} = 0$ or $\hat{p} = 1$. That is to say, when the sample has either no y_i equal to 1 or when all the sample y_i values equal 1, a two-sided Wald confidence interval with contain either the single point 0 (when $\hat{p} = 0$) or 1 (when $\hat{p} = 1$).

When $0 < \hat{p} < 1$, the Wilson confidence interval is approximately (dropping $O(1/n^2)$ terms)

$$p \approx \hat{p} + (½ - \hat{p})\, z_{\alpha/2}^2(1 - f)/n \pm z_{\alpha/2}\{(1 - f)\hat{p}(1 - \hat{p})/n\}^{½}.$$

This is within an $O(1/n^2)$ term of the two-sided Wald confidence interval shifted to the right by $(½ - \hat{p})\, z_{\alpha/2}^2(1 - f)/n$ when $p < ½$ and to the left by $(\hat{p} - ½)\, z_{\alpha/2}^2(1 - f)/n$ otherwise. Note that this shift is itself $O(1/n)$, while the Wald confidence interval is $O(1/\sqrt{n})$ when $0 < \hat{p} < 1$.

Empirical studies confirm that Wilson confidence intervals work fairly well in practice even though $\hat{p}$ is far from normally distributed when $p < .3$ or $p > .7$.

An Example

Suppose a city has 23,240 sources of drinking water. The Mayor and City Council think some fraction of the sources contain lead. A perfect test will reveal whether a particular source has lead, but the city only has the means to test 200 sources. A simple random of 200 sources is selected, and 4 turn out to have lead. What is our best guess of the number of drinking-water sources with lead? Can we determine a range within which we are 95%

confident the actual proportion falls?

In the absence of additional information, our best guess of the estimated fraction of drinking-water sources having lead is

$\hat{p} = 4/200 = .02 = 2\%$,

which translates to $\hat{t} = N\hat{p} \approx 465$ sources.

To compute 95% confidence intervals for the fraction of sources with lead, we can make the following simplifications, $f = n/N = 200/23{,}240 \approx 0$, and $z_{.975} \approx 2$ $(1 - \alpha = .95, \alpha = .05, 1 - \alpha/2 = .975)$. A two-sided Wald confidence interval is

$$p = \hat{p} \pm z_{.975}\{\hat{p}(1 - \hat{p})/(n-1)\}^{1/2}$$
$$\approx 0.02 \pm 2\{(0.02)(0.98)/199\}^{1/2} \approx 0.02 \pm 0.02 \quad \text{or} \quad p \in (0, 0.04),$$

More appropriate is the 95% Wilson interval. Approximating $1 + z_{.975}^{2}(1-f)/n \approx 1 + 4/200$ by 1, the Wilson interval is approximately,

$$p \approx \hat{p} + (½ - \hat{p})\, z^2(1 - f)/n \pm z_{.975}\{(1 - f)\hat{p}(1 - \hat{p})/n + z_{.975}^{2}(1 - f)/(4n^2)\}^{½}$$
$$\approx 0.02 + 0.48(4)/200 \pm 2\{0.02(0.98)/200 + 1/200^2\}^{1/2} \approx 0.03 \pm 0.02 \quad \text{or } p \in (0.01, 0.05).$$

Our best estimate for p remain 2%, and for t roughly 465 sources. The 95% Wilson confidence interval, however, is asymmetric. For p, it is between 1% and 5%. For t, it is between roughly 232 (1% of 23,240) and 1,162 (5% of 23,240).

Homework for Lesson 3

Do Exercises 5 (histogram optional), 6, and 18 in section 2.10 (starting on p. 53) of *Sampling: Design and Analysis*. Whenever the text asks for a confidence interval for a proportion, provide both the two-sided Wald and Wilson confidence intervals. Redo Exercise 6 with 3 children not overdue for vaccination.

Lesson 4: Ratio and Regression Estimation

For the remainder of these course notes when the terms "bias," "variance," or "mean squared error" are unadorned by a reference to a model, they refer to design-based properties.

Comments on Chapter 3 of *Sampling: Design and Analysis*

Asymptotic Theory for the Ratio Estimator

The text focuses on estimating $B = t_y/t_x$ based on a simple random sample and then using that estimate to develop ratio estimators for t_y and y_U. The relative errors for all three estimators, $\hat{B} = \hat{t}_y/\hat{t}_x$, $\hat{t}_{yr} = t_x\hat{B}$, and $\hat{\bar{y}}_r = \bar{x}_U\hat{B}$ are the same: $RE(\hat{B}) = (\hat{B} - B)/B$. This is true whether or not the sampling scheme is SRS because t_x and $\bar{x}_U$ are constants.

We will assume here that the estimation strategy and population are such that B is O(1) and greater than zero. Moreover, both $(\hat{t}_y - t_y)/t_y$ and $(\hat{t}_x - t_x)/t_x$ are $O_P(1/\sqrt{n})$. This places very mild restrictions on the population when the sample design is SRS.

Observe that $\hat{B} - B = \hat{t}_y/\hat{t}_x - B = (\hat{t}_y - B\hat{t}_x)/\hat{t}_x = [(\hat{t}_y - t_y) - B(\hat{t}_x - t_x)]/\hat{t}_x$. Using equation (S3) and the equality, $\hat{t}_x = t_x\{1 + (\hat{t}_x - t_x)/t_x\}$,

$$1/\hat{t}_x = (1/t_x)\{1 + (\hat{t}_x - t_x)/t_x\}^{-1}$$

$$= (1/t_x)\{1 - (\hat{t}_x - t_x)/t_x + g\}, \quad \text{where } g = O_P(1/n).$$

Plugging this into $\hat{B} - B$, we have

$$\hat{B} - B = (1/t_x)\,[(\hat{t}_y - t_y) - B(\hat{t}_x - t_x)]\{1 - (\hat{t}_x - t_x)/t_x + g\}$$

$$= (1/t_x)\,[(\hat{t}_y - t_y) - B(\hat{t}_x - t_x)]$$
$$- (1/t_x)\,[(\hat{t}_y - t_y) - B(\hat{t}_x - t_x)]\,(\hat{t}_x - t_x)/t_x \qquad \text{(S6)}$$
$$+ (1/t_x)\,[(\hat{t}_y - t_y) - B(\hat{t}_x - t_x)]\,g.$$

The first term of (S6) is $O_p(1/\sqrt{n})$, the second $O_p(1/n)$, and the third $O_p(1/n^{3/2})$. We assume the sample is sufficiently large that we can ignore the third term in further discussions.

The bias of $\hat{B}$ is approximately the sum of the expectations of the first and second terms terms in equation (S6). The expectation of the first is zero. The expectation of the second is also zero when $B = Cov(\hat{t}_y, \hat{t}_x)/V(\hat{t}_x)$. Unfortunately, we cannot assume that this equality holds, even approximately, without assuming a model for the data. Consequently, we conclude that the bias of $\hat{B}$ is O(1/n) since the second term in equation (S6) is $O_P(1/n)$.

We can divide the left and right hand sides of equation (S6) by B to get

$$(\hat{B} - B)/B \approx [(\hat{t}_y - t_y)/t_y - (\hat{t}_x - t_x)/t_x] - [(\hat{t}_y - t_y)/t_y - (\hat{t}_x - t_x)/t_x]\,(\hat{t}_x - t_x)/t_x.$$

The relative mean squared error of $\hat{B}$ is then

$$E\{[(\hat{B} - B)/B]^2\} \approx E\{[(\hat{t}_y - t_y)/t_y - (\hat{t}_x - t_x)/t_x]^2\}$$
$$- 2E\{[(\hat{t}_y - t_y)/t_y - (\hat{t}_x - t_x)/t_x]^2\,(\hat{t}_x - t_x)/t_x\}$$
$$+ E\{[(\hat{t}_y - t_y)/t_y - (\hat{t}_x - t_x)/t_x]^2\,[(\hat{t}_x - t_x)/t_x]^2\}.$$

The first term on the right-hand size is O(1/n). The other two are $O(1/n^{3/2})$ and $O(1/n^2)$, respectively. The first term dominates, and the relMSE of $\hat{B}$ is O(1/n).

Since B is O(1), the mean squared error of $\hat{B}$ is also O(1/n). Recall that the mean squared error of an estimator is equal to its variance plus the square of its bias (page 28). The bias

of $\hat{B}$, like its relative bias, is $O(1/n)$. The square of its bias is therefore $O(1/n^2)$. The relative contribution of the bias of $\hat{B}$ to its mean squared error is $[E(\hat{B} - B)]^2/MSE(\hat{B})$, which tends to zero as n grows arbitrarily large. That is why the (asymptotic) mean squared error of $\hat{B}$ is sometimes referred to as its variance.

The interested student can verify that the mean squared error of $\hat{B}$ under simple random sampling is

$$E[(\hat{B} - B)^2] = (1/\bar{x}_U^{\,2})V(\bar{d}) + O(1/n^{3/2}),$$

where $d_i = y_i - Bx_i$. Moreover, the bias of $\hat{V}(\hat{B})$ in equation (3.7) on page 58 as an estimator for the mean squared error of $\hat{B}$ is $O(1/n^{3/2})$.

Asymptotic theory for the regression estimator will wait until Lesson 14.

The Model for the Ratio Estimator

Similar to equation (3.16) of the text, suppose the y_i were random variables such that

$$y_i = \beta x_i + \epsilon_i, \tag{S7}$$

where $E_M(\epsilon_i \mid x_i, Z_i) = 0$. It is not always enough to assume, as the text and many other sources do, that $E_M(\epsilon_i) = 0$. We often need the ϵ_i to have mean zero irrespective of, 1, the value of x_i and, 2, whether or not i is in the sample (i.e., $Z_i = 1$).

When the distribution of ϵ_i given x_i is the same for both elements in the sample and not in the sample, we say the sample design is *ignorable* or *noninformative*. Simple random sampling is always ignorable, but $E_M(\epsilon_i \mid x_i)$ is not always zero when the y_i and x_i can be fitted by the model in equation (S7). For example, suppose x_i is a random variable with mean 1, and $y_i = 1 + x_i$ for all i. One can write $y_i = 2x_i + \epsilon_i$, where $\epsilon_i = 1 - x_i$. The unconditional expectation of ϵ_i, $E_M(\epsilon_i)$, is zero, but the conditional expectation, $E_M(\epsilon_i \mid x_i)$, is not.

When $\hat{t}_y = [N/n]\sum_S y_i$, and $\hat{t}_x = [N/n]\sum_S x_i$, regardless of how the sample was drawn,

$$\begin{aligned}
\hat{B} - \beta &= \textstyle\sum_S y_i / \sum_S x_i - \beta \\
&= \textstyle\sum_S (x_i \beta + \epsilon_i) / \sum_S x_i - \beta \\
&= \textstyle\sum_S (x_i \beta) / \sum_S x_i + \sum_S \epsilon_i / \sum_S x_i - \beta \\
&= \textstyle\sum_S \epsilon_i / \sum_S x_i .
\end{aligned}$$

Suppressing the conditioning for notational convenience, as we will often do here, it is easy to see that $E_M(B) = \beta$.

Notice that nothing had to be assumed about the variance of the ϵ_i to establish the model unbiasedness of $\hat{B}$. Determining the model variance of $\hat{B}$ is another matter. For that, we assume that the ϵ_i are uncorrelated conditioned on the Z_i and that $V_M(\epsilon_i \mid x_i, Z_i) = \sigma_i^2$, where σ_i^2 may be a function of x_i but not Z_i. In the text, σ_i^2 is proportional to x_i, but we do not necessarily assume that proportionality here.

Again, suppressing conditioning for notational convenience,

$$\begin{aligned}
V_M(\hat{B}) &= E_M[(\hat{B} - \beta)^2] \\
&= E_M[(\textstyle\sum_S \epsilon_i / \sum_S x_i)^2] \\
&= \textstyle\sum_S E_M(\epsilon_i^2) / (\sum_S x_i)^2 \\
&= \textstyle\sum_S \sigma_i^2 / (\sum_S x_i)^2 \qquad \text{(S8)} \\
&= [1/(n\bar{x}^2)] \textstyle\sum_S \sigma_i^2 / n.
\end{aligned}$$

The problem with our results on the model bias and variance of $\hat{B}$ is that they treat $\hat{B}$ as an estimator for the model parameter, β, rather than the finite-population target, B. Using parallel arguments to those above, we can show that B is also a model-unbiased estimator

for β with model variance equal to $Var_M(B) = [1/(N\bar{x}_U{}^2)\ (\sum_U \sigma_i^2/N)$. When N >> n, and both $\bar{x}_U{}^2$ and $\sum_U \sigma_i^2/N$ converge to positive constants when n grows arbitrarily large, the model variance of B is effectively zero and the distinction between B and β can be ignored.

When B is effectively equal to β, one can estimate the model variance in equation (S8) with

$$\hat{V}_M(\hat{B}) = [1/(n\bar{x}^2)]\ \sum_S (y_i - \hat{B}x_i)^2/(n - 1), \qquad \text{(S9)}$$

Ideally, σ_i^2 would be estimated by $(y_i - \beta x_i)^2$, but since β is unknown, we estimate it with $\hat{B}$, and punish ourselves for that by replacing the n in the denominator by n−1. The interested student can verify that under mild conditions the bias in this estimator for the model variance is $O(1/n^2)$ when N >> n.

Notice that the estimator for the randomization mean squared error of $\hat{B}$ in equation (3.7) on page 68 is very similar. It contains a finite-population-correction term, 1 − n/N, that we have assumed is effectively 1. It also contains $\bar{x}_U{}^2$ in the denominator instead of $\bar{x}^2$. The interested student can again verify that the bias from using $\hat{V}_M(\hat{B})$ to estimate the design design mean squared error of $\hat{B}$ under SRS is $O(1/n^{3/2})$ when N >>n. This is no worse than $\hat{V}(\hat{B})$ in equation (3.7).

It is important to remember that the model variance of $\hat{B}$ depends on the actual sample drawn, and is not the average across all possible samples. The design does not have to be SRS. Nevertheless, since $\hat{V}_M(\hat{B})$ is a good estimator for the model variance of $\hat{B}$ and no worse than $\hat{V}(\hat{B})$ at estimating the design mean squared error of $\hat{B}$ under simple random sampling, it seems the preferable choice in practice for the estimation strategy. Empirical investigations have shown that $\hat{V}_M(\hat{B})$ tends to produce two-sided Wald confidence intervals with better coverage properties than $\hat{V}(\hat{B})$.

The term $e_i = y_i - \hat{B}x_i$ is called the *sample residual* for i under the model in equation (S7). Särndal, Swensson, and Wretman (1989) advises that the sample variance of these sample

residuals, s_e^2, in equations (3.7), (3.8) and (3.9) of the text be replaced by $(\bar{x}_U/\bar{x})^2 s_e^2$. The results they dubbed the *weighted-residual* variance estimators. In particular,

$$\hat{V}_W(\hat{B}) = (1 - f)s_e^2/(n\bar{x}^2), \tag{S10}$$

$$\hat{V}_W(\hat{t}_{yr}) = N^2(1 - f)\,(\bar{x}_U/\bar{x})^2 s_e^2/n, \text{ and} \tag{S11}$$

$$\hat{V}_W(\hat{\bar{y}}_r) = (1 - f)\,(\bar{x}_U/\bar{x})^2 s_e^2/n. \tag{S12}$$

Under the model, the text develops a better finite-population-correction factor than 1 – f, but it requires an assumption about the sizes of the σ_i^2; in particular, that $\sigma_i^2 \propto x_i$. The fpc factor is $1 - f(\bar{x}/\bar{x}_U)$. See equation (3.17) on page 83. In principle, after making an assumption about the relative sizes of the σ_i^2, one can adjust the variance estimators is equations (S10), (S11), and (S12) so that they are exactly model unbiased. This can be done without sacrificing their large-sample design-based properties.

Recall that $\hat{B} = \hat{t}_y/\hat{t}_x$ did not have to be based on a simple random sample to be model unbiased. It is reassuring, however, that, even if the model in equation (S7) fails, $\hat{B}$ is design consistent when the sample is SRS. The term "model-assisted (design-based) survey sampling" is often used to describe the approach that considers both the model and design-based properties of an estimation strategy. Invoking a model allows us to draw inference about the data in the sample actually drawn without pretending it is an average sample. Design consistency provides protection against model failure. The variance estimators in equations (S9), (S10), and (S11) are good when the model is correct and not bad when it fails as long as the sample is sufficiently large.

An exploration into the model-based properties of the regression estimator will wait until Lesson 14. Until then, accept without further explanation that the weighted-residual variance estimator for $\hat{\bar{y}}_{reg}$ is

$$\hat{V}_W(\hat{\bar{y}}_{reg}) = (1 - f)\,[\textstyle\sum_S (g_i e_i)^2 - (\sum_S g_i e_i)^2/n]/[n(n-1)] \tag{S13}$$

where $g_i = \{1 + n(\bar{x}_U - \bar{x})(x_i - \bar{x}) / \sum_S (x_j - \bar{x})^2\}$, and e_i is defined on page 75. When $\bar{x} = \bar{x}_U$, g_i is 1 for all i, and $\hat{V}_W(\hat{\bar{y}}_{reg})$ collapses into $\hat{V}(\hat{\bar{y}}_{reg})$.

Table of Large-Sample Results

(excluding the regression estimator)

Parameters, Estimators, Assumptions

N is the size of U; n is the size of S; $f = n/N > 0$; Sample design is SRS.

$t_y = \sum_U y_i$; $t_x = \sum_U x_i$; $B = t_y / t_x$;

$\bar{y}_U = t_y/N$ converges to $\bar{y}_\infty > 0$ as n and N grow arbitrarily large;

$\bar{x}_U = t_x/N$ converges to $\bar{x}_\infty > 0$ as n and N grow arbitrarily large;

$S_y^2 = \sum_U (y_i - \bar{y}_U)^2/(N-1)$ converges to $S_{y,\infty}^2 > 0$ as n and N grow arbitrarily large;

$S_x^2 = \sum_U (x_i - \bar{x}_U)^2/(N-1)$ converges to $S_{x,\infty}^2 > 0$ as n and N grow arbitrarily large;

$\hat{t}_y = (N/n)\sum_S y_i$; $\hat{t}_x = (N/n)\sum_S x_i$; $\hat{B} = \hat{t}_y / \hat{t}_x$; $\hat{t}_{yr} = t_x\hat{B}$

$\bar{y} = \sum_S y_i/n$; $\bar{x} = \sum_S x_i/n$; $\bar{y}_r = \bar{x}_U\hat{B}$

$s_y^2 = \sum_S (y_i - \bar{y})^2/(n-1)$; $s_x^2 = \sum_S (x_i - \bar{x})^2/(n-1)$;

$s_e^2 = \sum_S (y_i - \hat{B}x_i)^2/(n-1)$;

Asymptotic Orders

t_y and t_x	are $O(n)$	$\hat{t}_y$, $\hat{t}_x$, and $\hat{t}_{yr}$	are $O_P(n)$
$\bar{y}_U$, $\bar{x}_U$ and B	are $O(1)$	$\bar{y}$, $\bar{x}$, $\hat{\bar{y}}_r$, and $\hat{B}$	are $O_p(1)$
$\hat{t}_y - t_y$, $\hat{t}_x - t_x$ and $\hat{t}_{yr} - t_y$	are $O_p(\sqrt{n})$	$E(\hat{t}_y - t_y) = E(\hat{t}_x - t_x) = 0$	
		$E(\hat{t}_{yr} - t_y)$	is $O(1)$
		$V(\hat{t}_y)$, $V(\hat{t}_x)$, $MSE(\hat{t}_{yr})$	are $O(n)$
$(\hat{t}_y - t_y)/t_y$, $(\hat{t}_x - t_x)/t_x$, $\hat{\bar{y}}_r - \bar{y}_U$,		$E(\hat{\bar{y}}_r - \bar{y}_U)$, $E(\hat{B} - B)$,	
$\hat{B} - B$ and $(\hat{B} - B)/B$	are $O_p(1/\sqrt{n})$	and $E[(\hat{B} - B)/B]$	are $O(1/n)$
		$relV(\bar{y})$, $relV(\bar{x})$, $MSE(\hat{\bar{y}}_r)$,	
		$MSE(\hat{B})$, and $relMSE(\hat{B})$	are $O(1/n)$

Variance/Mean-Squared-Error Estimators

$\hat{V}(\hat{\bar{y}}_r) = (1 - f)s_e^2/n$; $\hat{V}(\hat{t}_{yr}) = N^2\hat{V}(\hat{\bar{y}}_r)$; $\hat{V}(\hat{B}) = \hat{V}(\hat{\bar{y}}_r)/\bar{x}_U^2$

$\hat{V}_W(\hat{\bar{y}}_r) = (\bar{x}_U/\bar{x})^2\hat{V}(\hat{\bar{y}}_r)$; $\hat{V}_W(\hat{t}_{yr}) = N^2\hat{V}_W(\hat{\bar{y}}_r)$; $\hat{V}_W(\hat{B}) = \hat{V}_W(\hat{\bar{y}}_r)/\bar{x}_U^2$

Asymptotic Orders

$E\{\hat{V}(\hat{t}_y) - MSE(\hat{t}_y)\}$ and $E\{\hat{V}_W(\hat{t}_y) - MSE(\hat{t}_y)\}$ are $O(\downarrow n)$

$E\{\hat{V}(\hat{B}) - MSE(\hat{B})\}$, $E\{\hat{V}_W(\hat{B}) - MSE(\hat{B})\}$

$E\{\hat{V}(\hat{\bar{y}}_r) - MSE(\hat{\bar{y}}_r)\}$, and $E\{\hat{V}_W(\hat{\bar{y}}_r) - MSE(\hat{\bar{y}}_r)\}$ are $O(1/n^{3/2})$

$E\{\hat{V}(\hat{B}) - MSE(\hat{B})\}/MSE(\hat{B})$ and

$E\{\hat{V}_W(\hat{B}) - MSE(\hat{B})\}/MSE(\hat{B})$ are $O(1/\downarrow n)$

Under the model: $y_i = \beta x_i + \epsilon_i$,

where the ϵ_i are uncorrelated random variables conditioned on the x_i and Z_i,

$E_M(\epsilon_i \mid x_i, Z_i) = 0$, $V_M(\epsilon_i \mid x_i, Z_i) = \sigma_i^2$, and $\sum_U \sigma_i^2/N$ converges to $\bar{\sigma}_\infty^2$ as n and N grow arbitrarily large.

$E_M(\bar{y}_r - \bar{y}_U) = E_M(\hat{t}_{yr} - t_y) = E_M(\hat{B} - B) = 0$

If N >> n, then

$E_M\{\hat{V}_W(\hat{t}_y) - V_M(\hat{t}_y)\}$ is $O(1)$

$E_M\{\hat{V}_W(\hat{B}) - V_M(\hat{B})\}$ and

$E_M\{\hat{V}_W(\hat{\bar{y}}_r) - V_M(\hat{\bar{y}}_r)\}$ are $O(1/n^2)$

$E_M\{\hat{V}_W(\hat{B}) - V_M(\hat{B})\}/V_M(\hat{B})$ is $O(1/n)$

Homework for Lesson 4

Do Exercises 2, 3, and 6 in Section 3.6 (starting on p. 89) of *Sampling: Design and Analysis*. Plotting the data, as requested in a number of the exercises, is helpful but optional.

Lesson 5: Stratified Sampling

Comments on Chapter 4 of *Sampling: Design and Analysis*

The Expansion Estimator

The text uses the subscript hj to denote the jth unit in stratum h. One could just as easily let i denote a unit and write $\hat{t}_{str}$ as

$\hat{t}_{str} = \sum_S y_i/\pi_i = \sum_S w_i y_i$,

where $\pi_i = n_h/N_h$ when $i \in S_h$, and $w_i = 1/\pi_i$. The estimator, $\hat{t}_{str,}$ is a generalization of the expansion estimator under SRS and is also called the expansion estimator.

The notation U_h will be used to denote the set of N_h population units in stratum h.

Asymptotics

There are (at least) two ways to develop large-sample properties for $\hat{t}_{str}$ under stratified simple random sampling (I prefer to add the modifier "simple" to "stratified random sampling") or STSRS. Either the number of strata, H, can stay fixed, while the number of samples per stratum, n_h, grow arbitrarily large along with n and N, or the n_h (once they come into being) can stay fixed while H grows with n and N. Whichever structure is assumed, if $\sum^H (N_h/N)S_h^2$ converges to a positive constant, say $S_{str,\infty}^2$ and $\max\{\pi_i\}/\min\{\pi_i\}$ is bounded, the large-sample properties of $\hat{t}$ under SRS can be extended to $\hat{t}_{STR}$. From now on, however, we will leave the demonstrations of design-based asymptotics to the interested student.

Generalized Wilson Confidence Intervals

The generalized Wilson confidence interval for $\hat{p}_{str}$ defined in equation (4.6) on page 102 is

$$p = \hat{p}_{str} + \frac{(½ - \hat{p}_{str})\, z_{\alpha/2}^2/n^* \pm z_{\alpha/2}\{\hat{p}_{str}(1 - \hat{p}_{str})/n^* + z_{\alpha/2}^2/(4n^{*2})\}^{½}}{1 + z_{\alpha/2}^2/n^*} \qquad \text{(S14)}$$

where $n^* = \hat{p}_{str}(1 - \hat{p}_{str})/\hat{V}(\hat{p}_{str})$ is called the effective size of the STSRS sample. Observe that this is the same equation as (S5) with n* replacing (1-f)n. Note that n* is undefined when $\hat{p}_{str} = 0$ or 1. In that case, model-based theory suggests we let $n^* = (\sum_S w_i)^2 / \sum_S w_i^2$. will be more on the notion of effective sample sizes in Lesson 8.

The Model, Poststratification, and Conditional Selection Probabilities

The model in Section 4.6 (starting on page 113) can be rendered

$$y_i = \mu_h + \epsilon_i, \text{ when } i \in U_h, \qquad \text{(S15)}$$

where the ϵ_i are uncorrelated random variables conditioned on the Z_i, with $E_M(\epsilon_i \mid Z_i) = 0$ and $V_M(\epsilon_i \mid Z_i) = \sigma_h^2$. Under this model and these assumption about ϵ_i, $\hat{t}_{str}$ is an unbiased estimator for t with a model variance that can be estimated by $\hat{V}(\hat{t}_{str})$.

There is thus a model-based justification for the estimator $\hat{t}_{str}$ irrespective of the actual sample design. The model in equation (S15) has been called the *group-mean model*, since each element in group h has the same mean. The term "group" replaces stratum in this context because the model does not require that the groups be used when the sample is is drawn. "Strata," by contrast, refer to design strata in most discourses on survey sampling.

The use of quota sampling (Section 4.8), often relies on a – perhaps implicit – assumption of a group-mean model. One must be aware, however, when using quota sampling to realize that not only must the group-mean model hold, but the sample design must be

ignorable (i.e., $E_M(\epsilon_i \mid Z_i)$ must be zero).

Using a group-mean-model-based approach is particularly appealing with a simple random sample, because SRS is an ignorable design. This is one justification for poststratification as described in Section 4.7. Indeed, without an implicit belief that a group-mean model holds, at least approximately, for a particular set of mutually exclusive groups, one should not use poststratification. For example, given a simple random sample of students, it may make sense to poststratify the sample by sex. It usually make little sense, however, to poststratify based on whether the student's last name has more letters than his (her) first.

Poststratification under SRS can be given a fully design-based justification. Suppose we draw a particular simple random sample and observe $n_h > 1$ units in each group h. Usually randomization theory takes averages across all possible samples, but we can take a conditional approach and only average across simple random producing n_1 selections in group 1, n_2 in group 2, ..., and n_H in group H. For example, if we have a SRS of 20 students, 11 of whom were girls, we can restrict our attention to simple random samples with 11 girls and 9 boys.

Conditional on the H values of the n_h, the probability of selecting a particular unit i in group h is n_h/N_h. In fact, conditional on the realized n_h values, we have a stratified simple random sample. All the results of the early sections of the Chapter apply! The one *caveat* – and it is a big one – is that all the n_h must be greater than 0; in fact, it has to be greater than 1 to estimate the variance of $\hat{t}_{str}$. When deliberately drawing a stratified simple random sample, this isn't an issue because the statistician controls the n_h. With a simple random sample, however, there is no guarantee that at least one unit will be in every group.

The text's requirement that the sample size within groups be reasonably large ($\geq$ 30) is unnecessary. Moreover, the unconditional variance estimator is equation (4.15) is inferior to the simple stratified variance estimator (equation (4.5)) based on the conditional selection (inclusion) probabilities. See Holt and Smith (1994) or Rao (1985). *The warning about data*

snooping being dangerous, by contrast, cannot be overemphasized.

Another practical consideration for poststratification, which may seem obvious, is that all the N_h need to be known. Often, we would like to use poststratification but can't, either because not all the n_h are positive or because not all the N_h known. For example, suppose we have a simple random sample of farms, which we want poststratify by size into those with sales of less than $100,000 in the past year and those with sales of $100,000 or more. From the survey itself, we may know into which of these two domains each of the sampled farms falls, but we do not know the same about every farm in the population. Consequently, we cannot poststratify the sample into the two sales-size groups.

Now suppose we have an educated guess about which farm is in which size category before the survey goes out. We can poststratify on *those* population totals. Realize, however, if a sampled farm was guessed to have small sales, but actually had large sales, it would remain in the small-sales group for poststratification.

Since we do not know the number of farms in the population in each sales category, we cannot estimate the average sales across all farms using actual sales for postratification. We can, however, estimate the mean within each domain (category) using a conditional-probability argument (e.g., the conditional probability of drawing a farm in domain 1 is n_1/N_1). This is an alternative approach to the one described in Section 3.3 (starting on page 77). The same estimates result, the sample mean within each domain, only now each is (conditionally) design unbiased. Unfortunately, since the domain population sizes are unknown, we cannot estimate the variances of the domain sample means in an unbiased fashion. More on this in an advanced homework exercise.

Homework for Lesson 5

Do Exercises 5 and 10, parts a and c in Section 4.9 (starting on p. 118) of *Sampling: Design and Analysis*. Compute a 95% confidence interval for the proportion estimated in 10c.

Lesson 6: Cluster Sampling with Equal Probabilities

Comments on Chapter 5 of *Sampling: Design and Analysis*

The sample statistic s_t^2

The sample statistic s_t^2, defined on page 135, can also be rendered

$$s_t^2 = [1/(n-1)] \sum_S [\hat{t}_i - (\sum_S \hat{t}_j/n)]^2 = [1/(n-1)] \{(\sum_S \hat{t}_i^{\,2}) - (\sum_S \hat{t}_i)^2/n\}.$$

It is not an unbiased estimator for S_t^2 unless each $\hat{t}_i = t_i$. The problem is that although $\hat{t}_i$ is an unbiased estimator for t_i, $\hat{t}_i^{\,2}$ is an unbiased estimator for $t_i^2 + \mathrm{Var}(\hat{t}_i)$, not t_i^2.

The Components-of-Variance Model

The model in equation (5.37) on page 163 can be rendered

$$\begin{aligned} y_{ij} &= \mu + \delta_{ij} \\ &= \mu + \epsilon_{iA} + \epsilon_{ij}, \end{aligned} \tag{S16}$$

where $E_M(\delta_{ij}) = E_M(\epsilon_{iA}) = E_M(\epsilon_{ij}) = 0$, $V_M(\epsilon_{ij}) = \sigma^2$, $V_M(\epsilon_{iA}) = \sigma_A^{\,2}$, regardless of whether i or ij is in the sample. Moreover the variance components, ϵ_{ij} and ϵ_{iA}, are uncorrelated across all i and j, so that $V_M(\delta_{ij}) = \sigma_A^{\,2} + \sigma^2$. The random variables, $\delta_{ij} = \epsilon_{iA} + \epsilon_{ij}$ and $\delta_{ig} = \epsilon_{iA} + \epsilon_{ij}$ $(g \neq j)$ are *correlated*, however, since both ij and ig are in the same cluster. The model-based intracluster correlation coefficient is $\rho_M = \sigma_A^{\,2}/(\sigma_A^{\,2} + \sigma^2)$, paralleling equation (5.38).

The Ratio Estimator

Under this *components-of-variance model*, $\hat{t}_i$ can be written

$$\begin{aligned} \hat{t}_i &= \sum_{j \in S_i} (M_i/m_i)(\mu + \delta_{ij}) \\ &= M_i\mu + (M_i/m_i) \sum_{j \in S_i} \delta_{ij} \\ &= M_i\mu + \delta_{iS}, \end{aligned}$$

where all the summations are over S_i, the subsample of elements in cluster i. Observe that $\delta_{iS} = (M_i/m_i) \sum \delta_{ij}$ is a random variable, uncorrelated across the i with $E_M(\delta_{iS}) = E_M(\delta_{iS}|M_i) = 0$. We can call $V_M(\delta_{iS}|M_i) = \sigma_i^2$, although σ_i^2 clearly is a function of σ_A^2 and σ^2. What function is the advanced homework exercise for this lesson.

The ratio estimator, $\hat{\bar{y}}_r$, in equation (5.28) of page 148 is a variant on the estimator $\hat{B}$. The same can be said about its one-stage "special case," $\hat{\bar{y}}$, in equation (5.16) on page 144. In the general context, $\hat{t}_i$ plays the role of y_i, M_i the role of x_i, and $\bar{y}_U$ the role of the target, B. Under mild conditions, $\bar{y}_r$ is $O_P(1/\sqrt{n})$. Recall that n is the number of psu's, so the asymptotic framework assumes the number of sampled psu's – and *not* the number of element sampled within each psu – grows arbitrarily large.

The model in equation (S7) is implied by the components-of-variance model. Moreover, the expectation of δ_{iS} (which plays the role of ϵ_i in (S7)) is zero conditioned on M_i. Thus, when N >> n, so that $\bar{y}_U \approx \mu$, the model variance of $\hat{\bar{y}}_r$ can be expressed by equation (S9) with the appropriate role conversions. In this spirit of this, the weighted-residual variance estimator for $\hat{\bar{y}}_r$ is equation (5.29) with $\bar{M}$ defined as the sample mean of the M_i, $\bar{M}_S$. The population mean will not do. Similarly, for the special case of a one-stage sample (where $m_i = M_i$ for all i). The weighted-residual approach replaces $\bar{M}_U$ in equation (5.18) on page 145 with $\bar{M}_S$, and the summation in equation (5.19) is pre-multiplied by $(\bar{M}_U/\bar{M}_S)^2$.

Homework for Lesson 6

Do Exercises 3, 12a, and 22 (plotting is optional) in Section 5.9 (starting on p 170) of *Sampling: Design and Analysis*.

Lesson 7: Sampling with Unequal Probabilities I

Comments on Sections 6.1 to 6.5 of *Sampling: Design and Analysis*

The With-Replacement Variance Estimator

An alternative way of looking at a single-stage with-replacement sample is to view it as a sample of n picks or "hits." From each hit, we get an independent estimator of t, which we can denote $\tilde{t}_h = t_h/\psi_h$, where t_h is the value of the unit chosen in the h'th hit, and ψ_h the probability of selecting that unit with the h'th hit. The reason for using a tilde (~) instead of a hat (^) will be made clear shortly.

The estimator $\hat{t}_\psi$ in equation (6.5) on page 188, which is sometimes called the Hansen-Hurwitz estimator, can be rendered

$$\hat{t}_\psi = \sum_{h=1}^{n} \tilde{t}_h/n = \sum t_h/(n\psi_h) = \sum t_h/\pi_h^*, \tag{S17}$$

where π_h^* is the number of times we expect "unit h" (the unit that was selected with the h'th hit) to have been selected overall. It is its pseudo-selection probability, hence the designation, π_h^*. We say "pseudo" because the actual selection probability is $\pi_h = 1 - (1 - \psi_h)^n$, which is slightly less than $\pi_h^* = n\psi_h$.

The variance estimator for $\hat{t}_\psi$ in equation (6.7) can be rendered

$$\begin{aligned}\hat{V}(\hat{t}_\psi) &= (1/n) \sum_{h=1}^{n} ([t_h/\psi_h] - \hat{t}_\psi)^2/(n-1) \\ &= (n/[n-1]) \sum_{h=1}^{n} ([t_h/\pi_h^*] - [\hat{t}_\psi/n])^2 \end{aligned} \tag{S18}$$

The key to establishing the unbiasedness of this variance estimator is that the $\tilde{t}_h = t_h/\psi_h$ are independent, unbiased estimators for t. The situation doesn't change when each t_h in

equations (S17) and (S18) is replaced by $\hat{t}_h$, an unbiased estimator for t_h based on an independently-drawn second (and possibly additional) stage(s) of sampling. Realize that if the same unit is drawn more than once in the first stage of sampling, this argument requires independent subsamples be drawn for each hit.

The with-replacement variance formula in equation (S18) is a very powerful tool. We can use it to estimate the variance of an estimator based on a multi-stage sample, often even when the first stage it drawn *without* replacement, as we shall see. It, or a variant, can also be used to estimate the variance of a estimator based on a stratified sample with one (first-stage) selection in each stratum. Consider the estimator, $\hat{t}_{str} = \sum^H \hat{t}_h$ for $t = \sum^H t_h$. If all the t_h where equal to, say, t_0, then one can replace each $\hat{t}_h/\psi_h$ in equation (S18) by $\hat{t}_h$, $\hat{t}_\psi$ by $\hat{t}_{str}$, and n by H to get

$$\hat{V}_{col}(\hat{t}_{str}) = (H/[H-1]) \sum_{h=1}^{H} (\hat{t}_h - [\hat{t}_{str}/H])^2 \qquad \text{(S19)}$$

The sample arguments for unbiasedness apply.

The variance estimator in equation (S19) is called the *collapsed-stratum* estimator because it essentially collapses all the design strata into one variance stratum. It relies on the assumption that all the t_h are equal. If they aren't, then the variance estimator has an upward bias (it overestimates bias) of $\sum^H (t_h - t/H)^2$, which the interested student can prove in to himself (herself). It is not uncommon to use the collapsed-stratum variance estimator in practice, although often each variance strata consists of only two design strata with thought-to-be-close values for t_h. Design strata must be collapsed into variance strata variance strata before looking at the $\hat{t}_h$.

When C = H/2 is the number of collapsed variance strata,

$$\hat{V}_{col2}(\hat{t}_{str}) = \sum_{c=1}^{C} 2 \sum_{j=1}^{2} (\hat{t}_{cj} - [\hat{t}_{c1} + \hat{t}_{c2}]/2)^2$$

$$= \sum 2 \ \{ \ (\hat{t}_{c1} - [\hat{t}_{c1} + \hat{t}_{c2}]/2)^2 + (\hat{t}_{c2} - [\hat{t}_{c1} + \hat{t}_{c2}]/2)^2 \ \}$$

$$= \sum 2 \ \{ \ ([\hat{t}_{c1} - \hat{t}_{c2}]/2)^2 + ([\hat{t}_{c2} - \hat{t}_{c1}]/2)^2 \ \}$$

$$= \sum^{C} (\hat{t}_{c1} - \hat{t}_{c2})^2, \qquad \text{(S20)}$$

where the subscript cj refers to design-stratum j (= 1 or 2) in collapsed-variance-stratum c.

Sometimes *a variant of the collapsed-stratum method is used with a estimator based on a systematic sample drawn from an ordered list*. Let N/n be an integer for convenience, and let $y_{[1]}, y_{[2]}, ..., y_{[n]}$ denotes the sample in order of selection with n = 2C. One can replace each t_{cj} in equation (S20) with an $(N/n)y_{[k]}$ and compute

$$\hat{V}_{col2}(\hat{t}_{sys}) = (N/n)^2 \{(y_{[2]} - y_{[1]})^2 + (y_{[4]} - y_{[3]})^2 + \ ... + (y_{[n]} - y_{[n-1]})^2 \}.$$

When n is not even, one of the collapsed strata is given three members, and $\hat{V}_{col2}(\hat{t}_{sys})$ is adjusted accordingly.

An alternative way of writing this last variance estimator is as

$$\hat{V}_{col2}(\hat{t}_{sys}) = ([y_{[2]}/\pi_{[2]}] - [y_{[1]}/\pi_{[1]}])^2 + ([y_{[4]}/\pi_{[4]}] - [y_{[3]}/\pi_{[3]}])^2 + \ ... + ([y_{[n]}/\pi_{[n]}] - [y_{[n-1]}/\pi_{[n-1]}])^2 \quad \text{(S21)}$$

where $\pi_{[k]}$ is the selection probability for the k'th systematic selection, which is N/n for all k. Why bother with equation (S21) at all? The variance formula is useful (if conservative) when the sample is drawn using *systematic probability proportional to size (pps) sampling* from an ordered list. This sampling design and other without-replacement pps designs will be described shortly.

The Horvitz-Thompson Estimator

The Horvitz-Thompson estimator in equation (6.12) on page 197, $\hat{t}_{HT} = \sum_S \hat{t}_i/\pi_i$ is another name for what we have been calling "the expansion estimator."

Some Methods of Without-Replacement Probability Sampling

This note discusses four methods of selecting a without-replacement probability sample in a single stage of sampling (which can serve as the first stage of a multi-stage sample). The term "probability proportional (or proportionate) to size" is often used generically whether or not the probability of selection for unit i is related to a measure of unit size. More often than not, π_i *is* related to some measure of size in practice. In that context, if x_i is the unit-i measure of size, and $\psi_i = x_i / \sum_U x_j$ the unit's relative measure of size, then $\pi_i = n\psi_i$.

The table on the next page is based on the data in Table 6.1 on page 186, modified slightly. Each class (the sample unit) was assigned a random number within the interval [0, 1) using the random-number table on page 457. Class 1, for example, was assigned the random number .74970, based on the first number on the table. Class 2 was assigned .74077 based on the second number in the first column, and so forth. These random assignments are called *permanent random numbers* or PRN's. The classes in the following table were reordered based on the sizes of their PRN's. In practice, one usually relies on statistical software to produce a random number from the uniform distribution on [0, 1) rather than a random-number table.

Our goal is to draw a sample of size 3 giving each class a probability of selection proportional to its relative size. The relative sizes are denoted by ψ_i, which is equivalent to the selection probability for each hit in a with-replacement pps sample of size 3. For a without-replacement sample of size 3, the unit selection probabilities are denoted $\pi_i = 3\psi_i$

Poisson, Collocated, and Bernoulli Sampling

A Poisson PRN sample of classes can be selection in the following manner. A class i is selected if and only if $PRN_i < \pi_i$. In the following table, classes 12, 6, 14, and 10 are in the sample. Poisson sample sizes are random. We expected a sample size of 3, but this expectation was not realized.

Class Number	ψ_i	$\pi_i = 3\psi_i$	PRN_i	Collocated PRN_i	Cumulative π_i
12	.03709	.11128	.01991	.06248	.11128
6	.09737	.29212	.06176	.12915	.40340
15	.02318	.06955	.11571	.19582	.47295
3	.04019	.12056	.13557	.26248	.59351
7	.03091	.09274	.17371	.32915	.68624
14	.13456	.40368	.20121	.39582	1.11992
10	.07255	.21765	.20516	.46248	1.30757
8	.06801	.20402	.21223	.52915	1.51159
9	.08346	.25039	.31842	.59582	1.76198
5	.11747	.35240	.36135	.66248	2.11437
13	.07110	.21329	.54377	.72915	2.32767
4	.03400	.10201	.66824	.79582	2.42968
2	.05100	.15301	.74077	.86248	2.58269
1	.06801	.20402	.74970	.92915	2.78671
11	.07110	.21329	.99277	.99582	3.00000

A related sampling technical is called *collocated PRN* sampling. In this design the PRN's are replaced by collocated PRN's. This is done by first ordering the PRN's by size, $PRN_{[1]} < PRN_{[2]} < \ldots PRN_{[N]}$, as we have done, and then replacing the k'th ordered PRN with $PRN_{[k]}^{c} = (k - 1 + r)/N$, where r is a random number selected from the interval [0, 1). Unfortunately, the sample drawn from the information in the above table and r = .93725 (based on the next number in the random-number table) now results in the three classes, 12 6, and 14.

The advantage of collocation is that when the original PRN's bunch within the unit interval, their collocated analogues will not. For example, suppose every class had selection (inclusion) probability $\pi_i = .2$, so that the expected sample size was again 3. Using the PRN's in the table, the sample would consist of five classes, 12, 6, 15, 3, and 7. All have PRN's less than .2. With the collocated PRN's, however, only classes 12, 6, and 15 are selected.

The special case of Poisson sampling where every unit has the same probability of

selection is also called *Bernoulli sampling*. When the expected size of a Bernoulli sample is an integer, collocated sampling always produces a sample of that size.

In practice, Bernoulli samples are often "stratified." That is to say, the population is broken into mutually exclusive groups with a group-specific selection probability, $\pi_{(1)}$ for units in group 1, $\pi_{(2)}$ for units in group 2, etc. If the PRN's are collocated *within each group separately* before drawing the sample, and the expected sample size for each group is an integer, then the result would be equivalent to stratified simple random sampling. Even under non-collocated stratified Bernoulli sampling, one can condition on the realized sample size within each group and treat the resulting sample as if it were a stratified SRS.

What the PRN mechanism adds to stratified simple random sampling is a means of controlling overlap across surveys. One way of doing that is described below. Say we need to draw two stratified simple random samples, A and B, and we want them to be distinct. Let us focus on a particular unit i, which is in group h for sample A and h' for sample B. Denote its probability of selection for sample A as $\pi_{(h)}^{A}$ and for B as $\pi_{(h')}^{B}$, where neither probability exceeds ½. We can avoid unit i being in both samples in the following manner. Choose the unit i sample A if its PRN (or collocated PRN) is less than $\pi_{(h)}^{A}$. Choose the unit for sample B if its PRN (or collocated PRN) is in the range $[½, ½ + \pi_{(h')}^{B})$. It is easy to see that selected probabilities are what we desire them to be, and unit i cannot be in both samples. See Ohlsson (1995) for a broader discussion of how and where PRN techniques are used to control overlap across samples.

Systematic pps Sampling

Systematic pps sampling from a randomly-ordered list works like the cumulative-size method in the text, except that the π_i are first randomly ordered (in the case of the above table, by their PRN's) and then cumulated. Choosing another random value, say r', from [0, 1), the sample consists of those units whose cumulative π-total first exceeds, r', r' + 1, ... r' + n − 1. With r' = .65299, this means that the classes 7, 8, and 1 are selected for the sample. The cumulative π-total for class 7, for example, is .68624, which exceeds .65299, but the cumulative π-total for class 3 does not.

Systematic pps sampling from an ordered list is just like systematic pps sampling from a randomly-ordered list only the list is purposefully ordered. Either way, it is possible to draw

a systematic pps sample when $\pi_i^* = n\psi_i$ exceeds 1 for one of more units. In that case, π_i^* is the expected number of time unit i will be selected for the sample. When π_i^* is an integer, say e, i will be selected exactly e times. Otherwise, the number of times i is selected will either be the smallest integer greater than π_i^* or the largest integer less than π_i^*.

When $n\psi_i$ is greater than 1, a popular alternative is to set $\pi_i = 1$ and select i with certainty. Let c be the number of certainty selections (it may take a few iterations to find them all), and U' the population with the c certainties removed, one can then selects systematic pps sample with probability $(n - c)\psi_i / \sum_{U'} \psi_j$ assigned to non-certainty unit i.

Rao-Sampford Sampling

One of the properties of systematic pps sampling from a randomly-ordered list, which is also called *Goodman-Kish sampling*, is that $\pi_{ij}/(\pi_i\pi_j) \approx (n-1)/n$. This approximation, is better the smaller the maximum value of the π_i. Sometimes, this maximum values is not as small as we would like. If that is the case, a better design in some sense (see Asok and Sukhatme, 1976) is Rao-Sampford sampling. In that design, the first unit is selected with probability ψ_i. The remaining units are selected *with replacement* with probability proportional to $\psi_i/(1 - n\psi_i)$. If any unit is selected more than once, the entire sample is thrown out, and one starts again.

In the following table, the cumulated ψ_i and $\psi_i' = \{\psi_i/(1 - n\psi_i)\} / \sum_U \{\psi_j/(1 - n\psi_j)\}$ have been computed. Using the random draws from [0, 1), .36671, .49870, and .50647, the sample consisting of classes 14, 8, and 8 would be rejected. The next three random numbers from page 457 likewise do not produce a good sample. The numbers, .51796, .41635, and .04382 yield classes 9, 10, and 6.

Class Number	ψ_i	Cumulative ψ_i	ψ_i'	Cumulative ψ_i'
12	.03709	.03709	.03031	.03031
6	.09737	.13447	.09990	.13021
15	.02318	.15765	.01810	.14830
3	.04019	.19784	.03318	.18149
7	.03091	.22875	.02474	.20623
14	.15456	.38331	.20929	.41552
10	.05255	.43586	.04531	.46083
8	.06801	.50386	.06205	.52287
9	.08346	.58733	.08086	.60373
5	.11747	.70479	.13173	.73546
13	.07110	.77589	.06563	.80109
4	.03400	.80989	.02750	.82859
2	.05100	.86090	.04373	.87232
1	.06801	.92890	.06205	.93437
11	.07110	1.00000	.06563	1.00000

The Hartley-Rao Variance Estimator

When the sampling design is single-stage Poisson, the variance of the expansion estimator, $\hat{t}_{HT} = \sum_S y_i/\pi_i = \sum_U y_i Z_i/\pi_i$ is not hard to derive. Since Z_i and Z_j, $i \neq j$, are independent under Poisson sampling, and $V(Z_i) = \pi_i(1-\pi_i)$, the variance of $\hat{t}_{HT}$ is $V(\hat{t}_{HT}) = \sum_U y_i^2(1-\pi_i)/\pi_i$, which is estimated by $\hat{V}(\hat{t}_{HT}) = \sum_S y_i^2(1-\pi_i)/\pi_i^2$.

The Hartley-Rao variance estimator for $\hat{t}_{HT}$ under a systematic pps sampling from a randomly-ordered list or a Rao-Sampford sampling is nearly unbiased when π_i^2 is ignorably small (say less than or equal to .2) for all i. The estimator remains reasonable and convenient to use otherwise. It can be expressed as

$$\hat{V}_{HR}(\hat{t}_{HT}) = [n/(n-1)] \sum_{i\in S}(1 - \pi_i - \sum_{k\in S}\pi_k/n + \sum_{k\in U}\pi_k^2/n)(y_i/\pi_i - [n^{-1}\sum_{j\in S} y_j/\pi_j])^2$$

$$= 1/[n(n-1)] \sum_{i\in S}(1 - n\psi_i - \sum_{k\in S}\psi_k + \sum_{k\in U} n\psi_k^2)(y_i/\psi_i - [n^{-1}\sum_{j\in S} y_j/\psi_j])^2. \quad \text{(S22)}$$

This equation is a bit ugly. What is effectively happening is that the squared difference from the weighted mean for each sampled i, $(y_i/\pi_i - [n^{-1}\sum_S y_j/\pi_j])^2$, has its own *finite population*

correction term: $1 - \pi_i - \sum_S \pi_k/n + \sum_U \pi_k^2/n$. The value of $\sum_U \pi_k^2/n$ does not depend on the sample actually drawn. For the example of the class sample, it is .25349. The value of $\sum_S \pi_k/n$ *does* depend on the sample, but is the same for each sampled unit. If the sample contains classes 7, 8, and 1, then it is .16693. The actual *finite population correction terms* are .907 for class 7 and .796 for classes 8 and 1, which have the same selection probability.

When π_i is ignorably small for all i, which is a much stronger assumption than the π_i^2 being all small, equation (S22) can be approximated with

$$\hat{V}_{WR}(\hat{t}_{HT}) = [n/(n-1)] \sum_{i \in S} (y_i/\pi_i - [n^{-1} \sum_{j \in S} y_j/\pi_j])^2, \qquad \text{(S23)}$$

which is essentially the with-replacement variance estimator in equation (S18). Thus, if we have a multi-stage sample with one of these designs in the first stage and all the first-stage selection probabilities were ignorably small, the with-replacement variance estimator with $\hat{t}_i$ (and $\hat{t}_j$) replacing y_i (and y_j) would be nearly unbiased. Sometimes it is used even when the first-stage selection probabilities are not ignorably small. When the π_i^2 are small, using the with-replacement variance estimator can be shown to be, if anything, conservative (biased upward).

When all the π_i in equation (S22) are equal, the equation collapses into the variance estimator for $\hat{t}$ under simple random sampling in equation (2.14) on page 34. When all the π_i are *not* equal but n is large, the first line on right hand side is often approximately equal to

$$\hat{V}_{WR}(\hat{t}_{HT}) \approx [n/(n-1)] \sum_{i \in S} (1 - \pi_i)(y_i/\pi_i - [n^{-1} \sum_{j \in S} y_j/\pi_j])^2.$$

since $\sum_S \pi_k/n - \sum_U \pi_k^2/n$ is $O_P(1/\sqrt{n})$ under mild conditions. This approximation is dubious when n is not large, which can happen in practice when systematic pps sampling from a randomly-ordered list or Rao-Sampford sampling is used within strata.

Suppose x_i is the measure of size for unit i used in determining the unit selection probabilities; that is, $\pi_i = n\psi_i = nx_i / \sum_U x_j$. Cumberland and Royall (1981), who are responsible for noting that the complex variance estimator proposed by Hartley and Rao is equal to something like the second line on the right hand side of equation (S23), also point out that $\hat{V}_{HR}(\hat{t}_{HT})$ is an unbiased estimator for the variance of $\hat{t}_{HT}$ under the model,

$$y_i = \beta x_i + x_i \epsilon_i,$$

where the ϵ_i are uncorrelated, with $E_\epsilon(\epsilon_i | x_i) = 0$ and $V_\epsilon(\epsilon_i | x_i) = \sigma^2$. *As long as the design is ignorable*, it doesn't matter how the model is drawn for this result to be true, which is why the second line of equation (S23) is rendered without the selection probabilities. It is easy to see that $\hat{t}_{HT} = \sum_S y_i / \pi_i = (1/n)(\sum_S y_i / x_i) \sum_U x_i$ is a model unbiased estimator for $t = \sum_U y_i$. Both have a model expectation of $\beta \sum_U x_i$. It is a bit harder (and left to the interested student) to see that the model expectations of both $(\hat{t}_{HT} - t)^2$ and $\hat{V}_{HR}(\hat{t}_{HT})$ in equation (S23) are $(\sum_U x_i)^2 \{(1/n) - \sum_U x_i^2 / (\sum_U x_i)^2\} \sigma^2$.

Homework for Lesson 7

Using the data in audit.dat from the CD-rom with bookval (i.e., book value) as the measure of size:

a. Draw a pps sample with replacement of size 8 using the cumulative-size method.
b. Draw a systematic pps sample of size 8 using the list as ordered.
c. Draw a systematic pps sample of size 8 ordering the accounts by bookval.
d. Treating the current list as if it were randomly ordered, draw a collocated sample of expected size 8.
e. Draw a random sample of size 10 by first removing certainties and then using systematic pps sampling for the remaining selections.

After each of the above samples are drawn, estimate the total actual value of the population of books. The data set on the CD-ROM only provides the booked value (call it x_i) for each account i in the population. Consequently, we need to provide actual account values (the y_i) here. Let the actual value for sampled account i always be $200 more than its booked value ($y_i = x_i + 200$). Estimate the variance of this total for the samples in parts a and b, assuming the list is randomly ordered for part b.

Lesson 8: Sampling with Unequal Probabilities II

Comments on Sections 6.7 of *Sampling: Design and Analysis*

Recall that the model called M1 in this section can be rewritten as

$$y_{ij} = \mu + \delta_{ij}$$

$$= \mu + \epsilon_{iA} + \epsilon_{ij}, \qquad \text{(S16)}$$

where $E_M(\delta_{ij}) = E_M(\epsilon_{iA}) = E_M(\epsilon_{ij}) = 0$, $V_M(\epsilon_{ij}) = \sigma^2$, $V_M(\epsilon_{iA}) = \sigma_A^2$, and all the ϵ_{ij} and ϵ_{iA} are uncorrelated regardless of whether i or ij is in the sample. Using the notation similar to what have been employing here, the model variance equation on page 211 is can be rendered

$$E_M[(\hat{t}_{HT} - t)^2] = \sigma_A^2\{(K^2/n) - (2K/n)\sum_{i\in S} M_i + \sum_{i=1}^{N} M_i^2\} + \sigma^2\{\sum_{i\in S} K^2/(n^2 m_i) - K\}$$

$$= K^2[\sigma_A^2\{(1/n) - (2/n)\sum_{i\in S} M_i/K + \sum_{i=1}^{N} M_i^2/K^2\} + \sigma^2\{\sum_{i\in S} 1/(n^2 m_i) - 1/K\}]$$

where $t = \sum_{i=1}^{N}\sum_{j=1}^{M_i} y_{ij}$, and $\hat{t}_{HT} = \sum \hat{t}_i/\pi_i$, $\hat{t}_i = (M_i/m_i)\sum_{j\in S_i} y_{ij}$, and $\pi_i = nM_i/K = nM_i/\sum_{k=1}^{N} M_k$.

Although model-based theory never averages over all potential samples, let's do just that. Taking the expectation of the last expression with respect to the design yields:

$$E\{E_M[(\hat{t}_{HT} - t)^2]\} = K^2[\sigma_A^2\{(1/n) - \sum^{N} M_i^2/K^2\} + \sigma^2\{\sum^{N} M_i/(nm_iK) - 1/K\}].$$

This information may be useless from the strictly model-based viewpoint, but consider this. When the design is uninformative, we can reverse the expectation signs to get

$$E_M\{E[(\hat{t}_{HT} - t)^2]\} = K^2[\sigma_A^2\{(1/n) - \sum^{N} M_i^2/K^2\} + \sigma^2\{\sum^{N} M_i/(nm_iK) - 1/K\}]. \qquad \text{(S24)}$$

If our model assumptions are correct the right hand side of (S24) is what we expect the

design mean squared error of $\hat{t}_{HT}$ to be. This value is called the "anticipated variance." The model here is a helpful tool for assessing the properties of a design-unbiased two-stage estimation strategy under idealized conditions.

It is shown in the text that the model variance is minimized for a particular psu sample, S, and a fixed ssu sample size, $m = \sum_S m_i$, when all the m_i are equal. If this is true for the model variance, then it must also be true for the anticipated variance, which is the model variance averaged over all possible samples given the design.

Now let $m_i = m_0$ in every PSU. Equation (S24) becomes

$$E_M\{E[(\hat{t}_{HT} - t)^2]\} = K^2[\sigma_A^2\{(1/n) - \sum^N M_i^2/K^2\} + \sigma^2\{1/m - 1/K\}]$$

$$= K^2[(\sigma_A^2/n)\{1 - n\sum M_i^2/K^2\} + (\sigma^2/m)\{1 - m/K\}],$$

since $nm_0 = m$. Thus, if $\sigma_A^2 > 0$, then the anticipated variance of $\hat{t}_{HT}$ is minimized for a fixed ssu sample size, m, when n, the number of sampled psu's, is maximized; that is, when there is only one ssu sampled per psu – $m_0 = 1$. Cost, however, is not only a function of over-all sample size. There are cost savings to be realized by keeping n down and letting m_0 exceed 1.

The term $1 - n\sum^N M_i^2/K^2$ can be viewed as the model final-population-correction factor for the first stage of sampling, while $1 - m/K$ is that factor for the second stage of sampling.

When both factors are small enough to ignore,

$$E_M\{E[(\hat{t}_{HT} - t)^2]\} \approx K^2(\sigma_A^2/n + \sigma^2/m) = (K^2/m)(m_0\sigma_A^2 + \sigma^2).$$

This simplified formula allows us to see the consequence of an m_0 value greater than 1.

Many surveys involve more than two stages of sampling. For example, a first-stage sample of counties can be followed by subsamples of blocks, then households in those blocks, and finally individuals in the households. It is a messy, but straightforward exercise to extend the component-of variance model to multiple stages.

The anticipated MSE of an estimation strategy is a useful analytical tool before a sample has been drawn. After the sample has been selected, and its members surveyed, the model variance, which conditions on the realized sample, and the design mean squared error, which is free of model assumptions, are more relevant. The anticipated MSE, by contrast, both requires a model and averages across all potential samples.

Brewer and Maximal Brewer Selection

Godambe's Theorem

Suppose the y_i are random variables such that the ratio model:

$$y_i = \beta x_i + \epsilon_i, \tag{S7}$$

holds, where $E_M(\epsilon_i \mid x_i, Z_i) = 0$. In addition, suppose the ϵ_i (conditioned on x_i, Z_i) are uncorrelated, and $E_M(\epsilon_i^2 \mid x_i, Z_i) = x_i^2\sigma^2$. Then, given a fixed sample size of n, the following estimation strategy for $t = \sum_U y_i$ has a number of desirable properties: Set $\pi_i = nx_i/\sum_U x_j$ (assuming $\max\{ nx_i/\sum_U x_j \} \leq 1$), and compute $\hat{t}_{HT} = \sum_S y_i/\pi_i = (\sum_U x_i)(1/n)\sum_S y_i/x_i$. This is called the *mean-of-ratios* strategy, since $(1/n)\sum_S y_i/x_i$ is a mean of ratios.

When all $x_i > 0$ and thus all $\pi_i > 0$, this mean-of-ratios strategy is unbiased under both the design and the model. Moreover, among all design-unbiased strategies with sample size n, it has the smallest possible anticipated variance, which is

$$\begin{aligned}
E_M\{E[(\hat{t}_{HT} - t)^2]\} &= E\{E_M[(\hat{t}_{HT} - t)^2]\} \\
&= E\{E_M[(\textstyle\sum_U x_i \sum_S \epsilon_i/[nx_i] - \sum_U \epsilon_i)^2]\} \\
&= E\{E_M[(\textstyle\sum_U x_i \sum_S \epsilon_i/[nx_i])^2 - 2\, E_M(\sum_U x_i \sum_S \epsilon_i/[nx_i] \sum_U \epsilon_i) + E_M[(\sum_U \epsilon_i)^2]\} \\
&= E\{(\textstyle\sum_U x_i)^2\sigma^2/n - 2\sum_U x_i (\sum_S x_i/n)\sigma^2 + \sum_U x_i^2\sigma^2\} \\
&= (\textstyle\sum_U x_i)^2 \sigma^2\{1/n - \sum_U x_i^2/(\sum_U x_i)^2\},
\end{aligned} \tag{S25}$$

where the $E(\sum_S x_i/n) = \sum_U x_i\pi_i/n = \sum_U x_i^2/(\sum_U x_i)$ is used to derive the last line. Observe the similarity between the right hand sides of equations (S25) and (S24). In (S25), x_i plays the

role of M_i, and there is no second stage of sampling.

The minimum anticipated-variance property of the mean-of-ratios strategy was first established by Godambe (1955). In some practical situations, $nx_i/\sum_U x_j$ will be greater than 1 for one of more population units. To deal with this, first note that for an integer m and a a set Q, $mx_i/\sum_Q x_j > 1$ implies $mx_i > \sum_Q x_j$, $(m-1)x_i > \sum_Q x_j - x_i$, and $(m-1)x_i/(\sum_Q x_j - x_i) > 1$. A more general version of Godambe's result replaces the unit selection probabilities with $\pi_i = \max\{1, (n-c)x_i\sum_{U'} x_j\}$, where U' is U with the c units for which $(n-c)x_i/\sum_{U'} x_j > 1$ removed. The members of U' are usually determined iteratively. The estimator, $\hat{t}_{HT}$, is no longer a mean of ratios. The interested reader can confirm that the last line of equation (S25) becomes $E_M\{E[(\hat{t}_{HT} - t)^2]\} = (\sum_{U'} x_i)^2\sigma^2\{1/(n-c) - \sum_{U'} x_i^2\}$.

Brewer Selection

The problem with Godambe's Theorem is that although we may be willing to accept that the y_i behave like random variables satisfying the ratio model in equation (7), we are seldom willing to accept that the ϵ_i are uncorrelated with variances proportional to x_i^2. *Sampling: Design and Analysis* makes an often more unpalatable, but quite common, assumption that the variances of the ϵ_i are proportional to x_i.

Consider the ratio estimator for $t = \sum_U y_i$: $\hat{t}_r = (\sum_U x_i)\sum_S (y_i/\pi_i)/\sum_S (x_i/\pi_i)$. For many designs and populations, $\hat{t}_r$ is design consistent with a $O_P(1/\sqrt{n})$ relative error. Moreover, it is model unbiased under the model is equation (S7). If, in addition, the ϵ_i are uncorrelated (conditioned on the x_i and Z_i) with $E_M(\epsilon_i^2) = \sigma_i^2$, then

$$
\begin{aligned}
E_M\{E[(\hat{t}_r - t)^2]\} &= E\{E_M[(\hat{t}_r - t)^2]\} \\
&= E\{E_M[(\{\textstyle\sum_U x_i/\sum_S [x_i/\pi_i]\}\sum_S \epsilon_i/\pi_i - \sum_U \epsilon_i)^2]\} \\
&\approx E\{E_M[(\textstyle\sum_S \epsilon_i/\pi_i - \sum_U \epsilon_i)^2]\} \qquad \text{assuming } \textstyle\sum_U x_i/\sum_S [x_i/\pi_i] = 1 + O_P(1/\sqrt{n}) \\
&= E\{E_M[(\textstyle\sum_S \epsilon_i/\pi_i)^2 - 2E_M(\sum_S \epsilon_i/\pi_i \sum_U \epsilon_i) + E_M[(\sum_U \epsilon_i)^2]\}
\end{aligned}
$$

$$= E\{(\textstyle\sum_S \sigma_i^2/\pi_i^2 - 2\sum_S \sigma_i^2/\pi_i + \sum_U \sigma_i^2\}$$

$$= \textstyle\sum_U \sigma_i^2[(1/\pi_i) - 1)]. \qquad \text{(S26)}$$

The *asymptotic anticipated mean squared error* of $\hat{t}_r$ expressed in the last line of equation (S26) does not depend on the joint selection probabilities. In fact, it does not depend on the sample size being fixed.

For a given expected sample size, say n_{exp}, the value for the π_i that minimizes the asymptotic anticipated MSE of $\hat{t}_r$ is $\pi_i = n_{exp}\,\sigma_i / \sum_U \sigma_j$, assuming that $\max\{n_{exp}\,\sigma_i / \sum_U \sigma_j\} \le 1$. The interested reader can confirm this by minimizing $\sum_U \sigma_i^2[(1/\pi_i) - 1)]$ subject to $\sum_U \pi_i = n_{exp}$ using a Lagrangian multiplier. When all $\pi_i = n_{exp}\,\sigma_i / \sum_j \sigma_j < 1$, the value of the asymptotic anticipated MSE is

$$E_M\{E[(\hat{t}_r - t)^2]\} = (\textstyle\sum_U \sigma_i)^2/n_{exp} - \sum_U \sigma_i^2 < (\sum_U \sigma_i)^2/n_{exp}. \qquad \text{(S27)}$$

In practice, we rarely know with any certainty what σ_i^2 is. We can, however, make an educated guess, which we need do only up to a scalar.

It is often assumed that $\sigma_i^2 = kx_i^{2g}$, where g is a value between ½ and 1. There is some evidence that the unit variances for many survey variables increase at a faster rate than x_i but a slower rate than x_i^2. The actual determination for g may depend on the variable. The following example may shed some light. Suppose y_i is the number of high-school seniors in class i planning to go to college, and x_i is the number of seniors in the class. The closer g is to 1 the more the plan-to-college rate is a function of the class. The closer it is to ½, the more the rates a function of the individual student. In general, the more the measurement units of the x-variable behave as a group, the larger the g is.

In *Brewer selection*, we first hypothesize a value for g, choose a value n', and then set $\pi_i = \min\{1, n'x_i^g \sum_U x_j^g\}$. When $n'x_i^g \sum_U x_j^g \le 1$ for every population unit, the expected sample size under Brewer selection is n', which we will call the *initial sample-size target*. Other-

wise, the expected sample size will be less than n'. Nevertheless, as the interested reader can confirm, the asymptotic anticipated MSE of $\hat{t}_r$ is bounded by $(\sum_U \sigma_i)^2/n' = (\sum_U kx_i^g)^2/n'$ under our model assumptions, paralleling equation (S27).

What assumption we make about the size of σ_i^2 is usually on less firm ground that the ratio model itself. Moreover, the ratio model can also fail. Consequently, after the sample is drawn and $\hat{t}_r$ computed, it is prudent to estimate its model variance and design mean squared error with a variant of equation (S11) replacing $(\bar{x}_U / \bar{x})^2$ by $(t/\hat{t}_r)^2$.

Maximal Brewer Selection

In practice, there are usually more than one item of interest on a survey. For example, a farm survey may inquire about the number of currently planted acres dedicated to corn, wheat, soy beans, and a multitude of other crops. Let f = 1, ..., F be a subscript denoting one of the crops in this situation. We are interested in estimating $t_f = \sum_U y_{fi}$ for each f, where y_{fi} is the planted acres of crop f. Suppose for each y_{fi} we have a *control value*, x_{fi}, which may be, say, the harvested acres for crop f reported on the last Census of Agriculture. For each f, we assume the ratio model:

$$y_{fi} = \beta_f x_{fi} + \epsilon_{fi},$$

holds, where $E_M(\epsilon_{fi} \mid x_{fi}, Z_i) = 0$, the ϵ_{fi} are uncorrelated, and $E_M(\epsilon_{fi}^2 \mid x_{fi}, Z_i) = k_f x_i^{g(f)}$.

Although it is possible to use a different ratio estimator, $\hat{t}_{fr} = (\sum_U x_{fi})\sum_S (y_{fi}/\pi_i)/\sum_S (x_{fi}/\pi_i)$, for each f, there is only one sample. What should the selection probabilities be?

Equation (S26) tells us that if we use the ratio estimator for any survey variable, asymptotic anticipated MSE can only go down as each π_i goes up, all other things being equal. This suggests the following selection routine for our F survey variables. For each survey

variable y_{fi}, (e.g., current corn acres for farm i) determine $\pi_i^{(f)}$ using a control variable, x_{fi} (census-year corn acres for i), and a variable (corn) specific n_f'. Under *maximal Brewer selection*, the selection probability for unit i is set at

$$\pi_i = \max_f\{\pi_i^{(f)}\}, \qquad \text{where} \quad \pi_i^{(f)} = \min\{1, n_1' x_{fi}^{g(f)} / \textstyle\sum_U x_{fj}^{g(f)}\}.$$

The asymptotic anticipated MSE for $\hat{t}_{fr}$ under our model assumptions can be no greater than $(\sum_U k_f x_{fj}^{g(f)})^2/n_f'$ for each f.

The expected sample size under this design is, as always, $n_{exp} = \sum_U \pi_i$. When n_{exp} is not an integer, a systematic pps routine using π_i as the measure of size will return a sample with a size equal to one of the two integers closest to n_{exp}. It is not hard to show that when no unit is selected with certainty (i.e., all π_i are less than 1), n_{exp} can be no greater than $\sum^F n_f'$. For many populations, it will be considerably less than that sum.

Some Practical Considerations

There are several impediments to using Brewer and maximal-Brewer selection is practice. The values of k and g – k_f and g(f) in the multiple-survey-variable case – need to be hypothesized. Previous survey information can provide only the roughest of guides. There is some empirical evidence that the design mean squared error of the ratio estimator is not very sensitive to mild misspecification in the choice for g, but k is another matter. In practice, one often has to balance meeting variance targets – using, for example, $(\sum_U k_f x_{fj}^{g(f)})^2/n_f'$ with a liberally chosen k_f as a bound for the variance of survey-variable f – and keeping the sample size small. Moreover, the actual relationship between the n_f' and n_{exp} will often only be revealed by simulation, that is, by drawing potential samples and looking at their properties.

A even more troubling problem arises when for some population unit, x_i (or x_{fi}) is zero. The model with $\sigma_i^2 = kx_i^g$ does not allow y_i to be anything other than zero, since both $E_M(y_i)$ and σ_i^2 are zero. It is precisely this type of model failure that concerns us when we insist that

our estimators be design consistent. In the case of Brewer selection, however, the ratio estimator will not be design consistent because $\pi_i = 0$.

A related, but more subtle problem can afflict the multivariate case. Suppose the population is such that x_{fi} must be greater than zero for at least one f, but can be zero for a particular f. The ratio estimator $\hat{t}_{fr}$ will be design consistent, but may not be very good when $x_{fi} = 0$, $y_{fi} > 0$, and π_i is small.

One practical solution is to never let x_i (or x_{fi}) be zero for a population unit, especially if there is any chance for y_i to be positive. In the example above, basing control values on a single census year was a questionable strategy. The farmer may not have harvested any of the crop that year, but the current year is another matter.

Alternatively, one can set a minimum value for π_i. If that value is π_{min}, say, this can be accomplished by adding an (F +1)'th – or, better yet, a 0'th – control to maximal-Brewer selection with $x_{0i} = 1$ for all i, and setting $n_0' = N\pi_{min}$. This makes sense to do even in the single target variable case where F = 1.

An Example of Brewer and Maximal Brewer Selection

Suppose we want to draw a sample of wholesalers for a monthly survey of fruit and vegetable sales. We have annual sales for each sites based on a previous survey. Let x_{1i} and x_{2i} denote the annual sales for site i of fruit and vegetables respectively. For illustrative purposes, suppose we only have 12 sites in our frame and desire to chose a sample using Brewer or maximal-Brewer selection with g = 3/4 for both fruit and vegetables, $n_1' = 3$ for fruit and $n_2' = 2$ for vegetables.

The following table lays out the frame and the selection probabilities.

Site i	Fruit x_{1i}	Vegetables x_{2i}	$3x_{1i}^{3/4}/\sum x_{1j}^{3/4}$	$2x_{2i}^{3/4}/\sum x_{2j}^{3/4}$	π_i $\max\{\pi_i^{(1)}, \pi_i^{(2)}\}$	PRN_i	Sample Selection
1	110	43	0.15408	0.18801	0.18801	0.18766	2
2	252	30	0.28692	0.14352	0.28692	0.47574	
3	345	0	0.36314	0.00000	0.36314	0.03279	1
4	121	12	0.16550	0.07219	0.16550	0.52571	
5	221	56	0.26002	0.22921	0.26002	0.31079	
6	183	40	0.22571	0.17809	0.22571	0.43955	
7	0	80	0.00000	0.29950	0.29950	0.75728	
8	12	10	0.02925	0.06296	0.06296	0.37724	
9	78	54	0.11906	0.22304	0.22304	0.21690	2
10	190	30	0.23215	0.14352	0.23215	0.43887	
11	39	100	0.07080	0.35407	0.35407	0.03803	1, 2
12	1500	20	1.09339	0.10589	1	0.93453	1

$\pi_i^{(1)} = \min\{1, 3x_{1i}^{3/4} / \sum x_{1j}^{3/4}\} = 1$ for i = 12; $3x_{1i}^{3/4} / \sum x_{1j}^{3/4}$ otherwise

$\pi_i^{(2)} = \text{minax}\{1, 2x_{2i}^{3/4} / \sum x_{2j}^{3/4}\} = 2x_{2i}^{3/4} / \sum x_{2j}^{3/4}$ for all i

The last column contains permanent random numbers for use with Poisson PRN sampling. Using those PRN's a Brewer-selected sample for fruit would contain sites 3, 11, and 12 since for each site $\pi_i^{(1)}$ exceeds PRN_i. Similarly, the Brewer-selected sample for vegetables would contain sites 1, 9, and 11. A maximal-Brewer-selected sample for both fruit and vegetables, with π_i exceeding PRN_i, would contain sites 1, 3, 9, 11, and 12.

The expected sample size for the vegetable sample is 2 since $n_2' = 1$, and no site is chosen with certainty for the vegetable sample. The expected sample size for the fruit sample is

(approximately) 2.91 and for the combined sample is 3.66 ($\sum_U \pi_i^{(1)}$ and $\sum_U \pi_i$ respectively).

The disadvantage of using Poisson sampling for maximal-Brewer selection is that the realized sample size may not be close to what we expect it to be. For the vegetable sample, we expected two selections, but got three. The advantage, as demonstrated in the example, is that we can decompose a maximal-Brewer-selected sample into its components, in this case, the fruit and vegetable samples. This is useful when there are multiple enumerations (measurements) of a single sample but different survey items are of interest on different occasions. For example, if we are are only interested in fruit in odd numbered months and vegetables in even numbered months, we can restrict our contacts to the relevant sample in the appropriate months.

Homework for Lesson 8

Suppose we have a single stage sample where $\hat{t}_i = t_i = y_i$. Rewrite the Horvitz-Thomson and Sen-Yates Grundy variance estimators in equations (6.14) and (6.15) of *Sampling: Design and Analysis* (p. 197), respectively. Show where in Theorem 6.3 the latter requires the sample size to be fixed. (As a consequence, the SYG variance estimator fails for the expansion estimator, $\hat{t} = \sum_S y_i/\pi_i$, under Poisson sampling, but the HT variance estimator does not.)

Using the example of maximal-Brewer selection in the table on the previous page, suppose we increased the selection probability of every site to at least 0.1. What is the change in expected sample size? In actual sample size? Answer the same two questions with the the minimum selection probability raised to 1/3.

Lesson 9: Complex Samples

Comments on Chapter 7 of *Sampling: Design and Analysis*

Estimating S^2

Equation (7.4) can be expressed in a more practical manner as

$$\hat{S}^2 = [N/(N-1)] \sum_S w_i(y_i - \hat{\bar{y}}_U)^2 / \sum_S w_i$$

$$= [N/(N-1)] \{ \sum_S w_i y_i^2 / \sum_S w_i - \hat{\bar{y}}_U^2 \}$$

$$\approx \sum_S w_i y_i^2 / \sum_S w_i - \hat{\bar{y}}_U^2$$

where $\hat{\bar{y}}_U = \sum_S w_i y_i / \sum_S w_i$.

Effective Sample Size

Recall a generalized Wilson confidence interval for a proportion under stratified simple random sampling looks like

$$p = \hat{p}_{str} + \frac{(½ - \hat{p}_{str})\, z_{\alpha/2}^2/n^* \pm z_{\alpha/2}\{\hat{p}_{str}(1 - \hat{p}_{str})/n^* + z_{\alpha/2}^2/(4n^{*2})\}^{½}}{1 + z_{\alpha/2}^2/n^*} \qquad \text{(S14)}$$

where $n^* = \hat{p}_{str}(1 - \hat{p}_{str})/\hat{V}(\hat{p}_{str})$ is the effective size of the STSRS sample. Observe that that $n^* = n\{\hat{p}_{str}(1 - \hat{p}_{str})/n\} / \hat{V}(\hat{p}_{str})\} = n/[\hat{deft}(\hat{p}_{str})]^2$ under STSRS. We can generalized the the concept of an *effective sample size* for any estimation strategy by defining:
$n^* = n/[\text{deft(estimation strategy)}]^2$. It is common to use the same term, "the effective sample size," to refer to its estimate, $n/\hat{deft}^2$.

After a sample has been drawn, it is more productive to estimate design variances directly rather than computing design effects and effective sample sizes. The notion of an (estimated) effective sample size has the most practical relevance in a generalized Wilson confidence interval:

$$p = \hat{p} + \frac{(½ - \hat{p})\, z_{\alpha/2}^2(1 - f)/n^* \pm z_{\alpha/2}\{(1 - f)\hat{p}(1 - \hat{p})/n^* + z_{\alpha/2}^2(1 - f)^2/(4n^{*2})\}^{½}}{1 + z_{\alpha/2}^2(1 - f)/n^*}$$

where S is a potentially stratified and multi-stage sample of elements, $\hat{p} = \sum_S w_i y_i / \sum_S w_i$, y_i is either 0 or 1, and $n^* = (1 - f)\, \hat{p}(1 - \hat{p})/\hat{V}(\hat{p})$.

A further useful generalization is to the estimation of a proportion within a domain; that is, $p_d = \sum_U y_i u_i / \sum_U u_i$, where y_i is again 0 or 1, while $u_i = 1$ when i is in the domain – labeled d – and 0 otherwise. A generalized Wilson confidence interval for p_d is

$$p_d = \hat{p}_d + \frac{(½ - \hat{p}_d)\, z_{\alpha/2}^2(1 - f)/n_d^* \pm z_{\alpha/2}\{(1 - f)\hat{p}_d(1 - \hat{p}_d)/n_d^* + z_{\alpha/2}^2(1 - f)^2/(4n_d^{*2})\}^{½}}{1 + z_{\alpha/2}^2(1 - f)/n_d^*} \qquad \text{(S28)}$$

where $\hat{p}_d = \sum_S w_i u_i y_i / \sum_S w_i u_i$ and $n_d^* = (1 - f)\, \hat{p}_d(1 - \hat{p}_d)/\hat{V}(\hat{p}_d)$.

Even within a generalized Wilson interval, however, the effective domain sample size is no more than an *ad hoc* construct. In actual application, one can replace $(1 - f)/n_d^*$ in the many places it appears in equation (S28) by $\hat{Var}(\hat{p}_d)/[\hat{p}_d(1 - \hat{p}_d)]$.

The generalized Wilson confidence interval for an estimated proportion as described above is undefined when $\hat{p}_d$ (or $\hat{p}$) is either 0 or 1; that is, every sampled unit has the same value for y_i. When that happens, both the numerator and denominator of $\hat{Var}(\hat{p}_d)/[\hat{p}_d(1 - \hat{p}_d)]$ is zero. Design-based theory fails us, and we are forced to assume a model.

The simplest model postulates that the y_i are uncorrelated random variables with mean $\pi_d \approx p_d$ when the sampling faction is very small, which we will assume to be true for convenience. The model further stipulates that the design is ignorable. As a consequence, it is easy to see that the model variance of $\hat{p}_d$ is

$$V_M(\hat{p}_d) = [\sum_S (w_i u_i)^2/(\sum_S w_i u_i)^2]\pi_d(1 - \pi_d) = [\sum_{S_d} w_i^2/(\sum_{S_d} w_i)^2]\pi_d(1 - \pi_d),$$

where S_d is that part of the sample in domain d, while the model variance of the simple domain sample mean is $V_M(\sum_{S(d)} y_i/n_d) = (1/n_d)\pi_d(1 - \pi_d)$. Thus, the model-based effective sample size, which can be plugged into equation (S28), is

$$n_{d\ (model)}^* = n_d\,[V_M(\sum_{S_d} y_i/n_d)]\,/\,V_M(\hat{p}_d) = (\sum_{S_d} w_i)^2 / \sum_{S_d} w_i^2.$$

Homework for Lesson 9

Estimate S^2 based on a stratified simple random sample of y_i values assuming there are H strata, N_h population units per stratum, and n_h sample units per stratum. Is this estimator unbiased? Under what conditions is it design consistent?

Look at the data set nybight.dat provided with *Sampling: Design and Analysis*. Treat the data collected in 1975 as a stratified sample, where data is strata 1 and 2 were selected at a 1 in 100 rate, while data in the other strata were selected at a 1 in 200 rate.

a. Compute the epmf and empirical distribution functions for the number of species caught per trawl.
b. Estimate the mean and median number of species caught per trawl. What is the design effect for the estimated mean?
c. Estimate the fraction of trawls with less than 7 species caught? Provide a 95% confidence interval for this fraction.

Lesson 10: Variance Estimation in Complex Samples

Comments and Extensions on Chapter 9 of *Sampling: Design and Analysis*

Linearization Without Calculus

Section 9.1 shows how differential calculus can be used to linearize complex statistics. In survey sampling the statistics needing linearization often involve the ratio of Horvitz-Thompson estimators, the poststratified expansion estimator, or combinations of ratioing and poststratification. More complicated statistics requiring linearization will be discussed in Lesson 14.

Instead of invoking calculus directly, these notes rely on a variant of equation (S2). When z is $O_P(1/\sqrt{n})$,

$$(1 + z)^{-1} = 1 - z + g \approx 1 - z,$$

where $g = O_P(1/n)$. In particular, when $(\hat{t}_x - t_x)/t_x$ is $O_P(1/\sqrt{n})$ with $\hat{B} = \hat{t}_y / \hat{t}_x$ and $B = t_y / t_x$, the equality can be used (see equation(S6)) to show that, dropping $O_P(1/n)$ terms,

$$\hat{B} - B \approx (1/t_x)\,[(\hat{t}_y - t_y) - B(\hat{t}_x - t_x)]. \qquad \text{(S29)}$$

Notation for a Complex Sample

Before proceeding, we need to establish the notation to be used throughout this discussion. Suppose we have a stratified multi-stage sample of elements, S. Let i denote an element, h = 1, ..., H one of the first-stage strata (H may be 1), S_h the sample of n_h psu's selected from stratum h, and j one of those psu's. Let S_{hj} the set of sample elements selected from psu j of stratum h, hereafter called psu hj for convenience. Analogously, let U_h denote the population of psu's in stratum h, U_{hj} the population of elements in psu hj, and U the union

of the U_{hj} across all the psu's in U_h and all the H strata.

Suppose the psu's are selected from each stratum using Goodman-Kish or Rao-Sampford sampling, both of which are equivalent to simple random sampling when all the psu's in the stratum have identical selection probabilities. The selection of elements from each psu may itself be a stratified multistage sample or, conversely, U_{hj} may contain a single element, which is also the sole member of S_{hj}.

Let π_{hj} be the probability of selecting psu hj, and $\pi_{i|hj}$ the conditional probability of sub-sampling element i within selected psu hj. The selection probability for element $i \in U_{hj}$ is $\pi_i = \pi_{hj}\pi_{i|hj}$. We are interested in this discussion with a Horvitz-Thompson estimator of the form: $\hat{t}_y = \sum_S y_i/\pi_i$, which can also be written: $\hat{t}_y = \sum_S w_i y_i$, where $w_i = 1/\pi_i$ is the sampling weight of element i.

We assume that the population of psu's in each stratum is such that the following variance estimator is reasonable:

$$\begin{aligned} V(\hat{t}_y) &= \sum_{h=1}^{H} [n_h/(n_h-1)] \{ \sum_{j\in S_h} (\sum_{i\in S_{hj}} y_i/\pi_i)^2 - (\sum_{j\in S_h} \sum_{i\in S_{hj}} y_i/\pi_i)^2/n_h \} \\ &= \sum [n_h/(n_h-1)] \{ \sum (\sum w_i y_i)^2 - (\sum \sum w_i y_i)^2/n_h \}. \end{aligned} \tag{S30}$$

There may be an upward bias in equation (S30) from using the with-replacement variance formula (see equation (S23)) for the first stage of sampling. Recall that $\hat{t}_{y(hj)}{}^2 = (\sum_{S_{hj}} y_i/\pi_{i|hj})^2$ is an unbiased estimator of $\hat{t}_{y(hj)}{}^2 = (\sum_{U_{hj}} y_i/\pi_{i|hj})^2 + V(\hat{t}_{y(hj)})$, so that the $(\sum_{S_{hj}} y_i/\pi_i)^2$ terms in equation (S30) carry within them the full contributions to the variance from the later stages of sample selection. Thus, even when the π_{hj} are not ignorably small, equation (S30) will have only a minuscule bias if the contribution to variance from the first stage of sampling is negligible.

The Ratio Estimator

The ratio estimator for $B = t_y / t_x$ under the stratified multi-stage sampling design just described is $\hat{B} = \hat{t}_y / \hat{t}_x = \sum_S (y_i/\pi_i)/\sum_S (x_i/\pi_i)$. Letting $n = \sum^H n_h$, we see from equation (S29) that when $(\hat{t}_x - t_x)/t_x$ is $O_P(1/\sqrt{n})$,

$$\hat{B} - B \approx (1/t_x)\,[(\hat{t}_y - t_y) - B(\hat{t}_x - t_x)].$$

$$= (1/t_x)[(\hat{t}_y - B\hat{t}_x) - (t_y - Bt_x)]$$

$$= (1/t_x)[\textstyle\sum_S (y_i - Bx_i)/\pi_i - \sum_U (y_i - Bx_i)]$$

$$= (1/t_x)[\textstyle\sum_S d_i/\pi_i - \sum_U d_i],$$

where $d_i = y_i - Bx_i$. Consequently, the mean squared error of $\hat{B}$ is approximately $V(\hat{t}_d)/t_x^2$, where $\hat{t}_d = \sum_S d_i/\pi_i$.

Observe that it is the number of sampled psu's, n, not the number of sampled elements that needs to be arbitrarily large for asymptotic theory to be invoked. A common rule of thumb is that n should be at least 20. For a fuller treatment of asymptotics in this context, see Krewski and Rao (1981).

We can estimate the variance of $\hat{B}$ with

$$\hat{V}(\hat{B}) = (1/\hat{t}_x)^2 \sum_{h=1}^{H} [n_h/(n_h - 1)]\,\{\sum_{j\in S_h} (\sum_{i\in S_{hj}} w_i e_i)^2 - (\sum_{j\in S_h} \sum_{i\in S_{hj}} w_i e_i)^2/n_h\}, \qquad \text{(S31)}$$

where $e_i = y_i - \hat{B}x_i \approx d_i$. This estimator has good design-based properties under mild conditions. It also serves as a good estimator of the variance of $\hat{B}$ under the model in equation (S7) when $E(\epsilon_i \epsilon_q)$ is bounded for i and q from the same psu and zero otherwise.

The combined ratio estimator for t_y, $\hat{t}_{yr} = t_x\hat{B}$, requires that t_x be known. When it is, a good estimator for the design mean squared error and model variance of $\hat{t}_{yr}$ is

$$\hat{V}(\hat{t}_{yr}) = (t_x/\hat{t}_x)^2 \sum_{h=1}^{H} [n_h/(n_h-1)] \{ \sum_{j\in S_h} (\sum_{i\in S_{hj}} w_i e_i)^2 - (\sum_{j\in S_h} \sum_{i\in S_{hj}} w_i e_i)^2/n_h \}. \tag{S32}$$

Some textbooks (like *Sampling: Design and Analysis*) advocate estimating the design MSE of combined ratio estimator without the $(t_x/\hat{t}_x)^2$ factor in equation (32). There is also some debate about which version is truly the linearization variance estimator. Both have been given that name since each relies on linearizing $\hat{B}$ as we effectively did in equation (S29).

The Poststratified Estimator

Suppose our goal is to estimate a total like t_y from a stratified multi-stage sample. Sometimes, the population of elements, U, naturally divides into G mutually exclusive groups, U^1, ..., U^G. If the population size, M_g, is known for each group, a postratified estimator for t_y is

$$\begin{aligned} \hat{t}_{yp} &= \sum_{g=1}^{G} M_g \frac{\sum_S y_i u_{gi}/\pi_i}{\sum_S u_{gi}/\pi_i}, \\ &= \sum^G M_g \hat{b}_g \end{aligned} \tag{S33}$$

where $u_{gi} = 1$ when element $i \in U^g$, 0 otherwise, and $\hat{b}_g = \bar{y}_g = \sum_S (y_i u_{gi}/\pi_i) / \sum_S (u_{gi}/\pi_i)$.

For an example of when a poststratified estimator might be used, suppose we wished to estimate the number of registered voters in the US who plan to vote in the next Presidential election. For simplicity, assume that all registered voters reside in the US, and our survey has no nonresponse. We draw a sample of registered voters by first stratifying by region and then drawing a multi-stage sample of individuals, starting with a sample of counties, and then subsamples of blocks within counties, dwelling units within blocks, and finally registered voters within dwelling units. Suppose further that we know the number of registered voters in these four groups: black males, nonblack males, black females, and nonblack females. We can then use $\hat{t}_{yp}$ to estimate the total number of registered voters

who plan to vote. In this example, M_g is the number of registered voters in group g (one of the four groups), u_{ig} is 1 if sampled individual i is in group g, and y_i is 1 if the registered individual plans to vote.

In equation (S33), $\hat{M}_g = \sum_S u_{gi}/\pi_i$ is an unbiased estimator for M_g. Depending of how the groups are defined, if may or may not be reasonable to assume $(\hat{M}_g - M_g)/M_g$ is $O_P(1/\sqrt{n})$ for each g. At the very least, there should be one or more positive u_{gi} value in $O_P(n)$ sampled psu's. Twenty such psu's is a popular rule of thumb, although this rule is often breached in the trenches for practical reasons.

Suppose we *can* assume $(M_g - \hat{M}_g)/M_g$ is $O_P(1/\sqrt{n})$ for all g so that $\hat{t}_{yp}$ is design consistent. Letting $b_g = (\sum_{i\in U} y_i u_{gi} / \sum_U u_{gi})$, we can express t_y as $\sum^G M_g b_g = t_y$. Treating $y_i u_{gi}$ and u_{gi} in each b_g (and $\hat{b}_g$) as y_i and x_i in B, we see that

$$\hat{t}_{yp} - t_y = \sum^G (M_g/\hat{M}_g) \sum_S d_{gi}/\pi_i.$$

where $d_{gi} = y_i u_{gi} - b_g u_{gi}$. Since d_{gi} is only positive when $i \in U^g$ is in group g, we can reexpress the last equation as

$$\hat{t}_{yp} - t_y = \sum_S \tilde{w}_i d_i \approx \sum_S w_i d_i ,$$

where $d_i = y_i - b_g$ and $\tilde{w}_i = (M_g/\hat{M}_g)(1/\pi_i)$ when $i \in S^g$ (that part of U^g in S). Notice that with the *reweighted sampling weights*, $\tilde{w}_i$, $\hat{t}_{yp} = \sum_S \tilde{w}_i y_i$.

Also using the reason we have been employing throughout this discussion, a good estimator for the design mean squared error of $\hat{t}_{yp}$ has the form:

$$\hat{V}(\hat{t}_{yp}) = \sum_{h=1}^{H} [n_h/(n_h - 1)] \{ \sum_{j\in S_h} (\sum_{i\in S_{hj}} w_i \tilde{e}_i)^2 - (\sum_{j\in S_h} \sum_{i\in S_{hj}} w_i \tilde{e}_i)^2/n_h \}, \qquad \text{(S34)}$$

where $\tilde{w}_i$ is defined as above, and $e_i = y_i - (\sum_{q \in S} w_q y_q u_{gq} / \sum_{q \in S} w_q u_{gq})$ for $i \in U^g$.

Some would replace the w_i in equation (S34) with $\tilde{w}_i$. The interested reader can confirm that our version of $\hat{V}(\hat{t}_{yp})$ is the superior estimator for the variance of $\hat{t}_{yp}$ under the group-mean model $y_i = \beta_g + \epsilon_i$ for $i \in U^g$, where the ϵ_i have mean zero, and $E(\epsilon_i \epsilon_q)$ is bounded for i and q from the same psu and zero otherwise. It is a conceptual model like this one that justifies using poststratification in the first place

The model variance/design MSE formula for $\hat{t}_{yr}$ in equation (S32) can be put in the same form as equation (S34) by keeping the same definition for e_i as in (S31) and setting $\tilde{w}_i = (t_x / \hat{t}_x) w_i$. The estimators in equations (S32) and (S34) are closely related to the weighted-residual variance estimator in equation (S11). The "weight" in that name refers to factor adjusting the original sampling weight: $(t_x / \hat{t}_x)$ in equations (S11) and (S32), $(M_g / \hat{M}_g)$ in equations (S34).

Poststratification in an Estimated Ratio

In practice, we often use poststratification to help estimate a ratio like $B = t_y / t_x$. The estimator is that case is $\hat{B}_P = \hat{t}_{yp} / \hat{t}_{xp}$, where $\hat{t}_{xp}$ is defined analogously to $\hat{t}_{yp}$. In the planning-to-vote example, suppose our estimation target is the fraction of registered voters who plan to vote, and rather than knowledge of the number of registered voters in each group, we know only the number of US residents in each group (in fact, we may only know the fraction of US residents in each group). These numbers come from a larger survey, deemed to be virtually accurate for our purposes. Not only are the M_g changed, but the sample is now of all US residents, u_{ig} is 1 if resident i is in group g (0 otherwise), and x_i is 1 if resident i is registered to vote.

Using the tricks we have been employing repeatedly in this discussion, a good mean-squared-error estimator for $\hat{B}_p$ is

$$\hat{V}(\hat{B}_P) = (1/\hat{t}_{xP})^2 \sum_{h=1}^{H} [n_h/(n_h - 1)] \{ \sum_{j \in S_h} (\sum_{i \in S_{hj}} \tilde{w}_i e_i)^2 - (\sum_{j \in S_h} \sum_{i \in S_{hj}} \tilde{w}_i e_i)^2/n_h \}, \qquad \text{(S35)}$$

where $\tilde{w}_i = (M_g/\hat{M}_g)w_i$, and $e_i = (y_i - \hat{B}_P x_i) - \sum_{q \in S} w_q u_{gq}(y_q - \hat{B}_P x_q) / \sum_{q \in S} w_q u_{gq}$ for $i \in S^g$.

The Jackknife

Although the MSE estimator in equation (S35) looks very similar to the one in equation (S31), computing the e_i is considerably more challenging. This is one reason why many would prefer using a jackknife estimator in this case. Armed with the *jackknife replicate weights*, $w_{i(hj)}$, defined on page 305 of *Sampling: Design and Analysis*, one can calculate

$$\hat{B}_{p(hj)} = \frac{\sum^G M_g \sum_S (w_{i(hj)} y_i u_{gi}) / \sum_S (w_{i(hj)} u_{gi})}{\sum^G M_g \sum_S (w_{i(hj)} x_i u_{gi}) / \sum_S (w_{i(hj)} u_{gi})}$$

for every sampled psu. The text's equation (9.8) becomes in this context

$$\hat{V}_{JK}(\hat{B}_p) = \sum^H [(n_h - 1)/n_h] \sum_{j \in S_h} (\hat{B}_{p(hj)} - \hat{B}_p)^2.$$

The jackknife can fail (usually by being biased upward) when the first-stage selection probabilities (the π_{hj}) are not all small, but that weakness is shared by the linearization estimators in equations (S31), (S32), and (S34). Although the jackknife should not be used directly to estimate variances for percentiles, it can be used within Woodruff's method for determining confidence intervals around percentiles.

Variance Strata and Variance Psu's

Suppose we wanted to use BRR to estimate the MSE of $\hat{B}_P$. This can easily be done when $n_h = 2$ for all h, but what is that were not the case? As long as n_h is even for all h, we can randomly assign the design psu's within each S_h to one of two equally-sized *variance psu's*, and then treat the two variance psu's as h1 and h2 in a BRR scheme. This procedure is no more biasing to MSE estimation than pretending the first stage of sampling is with replacement.

In a sample with a large number of design psu's we can use a similar device with the jackknife, thereby reducing the otherwise cumbersome number of replicate weights needed for each element. When n_h is divisible by an integer, say $k < n_h$, we can randomly assign the n_h sampled design psu's into k variance psu's and then treat the variance psu's as the psu's when computing replicate weights and jackknife MSE estimates.

In many design, there is only one psu sampled per stratum. As noted in Lesson 7, a pair of design strata can be treated as a single variance stratum for MSE-estimation purposes. This pairing must be done before looking at the survey results. If anything, using variance strata biases MSE estimation upward.

Another example of a variance stratum is a self-representing psu (see page 242 of *Sampling: Design and Analysis*). These are psu's selected with certainty in the design. For MSE-estimation purposes, they are variance stratum. The second-stage sample units within a self-representing psu (or collections of those ssu's) become variance psu's for MSE-estimation purposes.

Confidence Intervals for Percentiles

The text describes Woodruff's method for determining a confidence interval for a percentile (pages 311 and 312). Let F(.) denote the distribution function of y_i values in the population, q is a target percentile if fractional form, and θ_q the value at the q'th percentile, so that $q = F(\theta_q)$. In this context, an asymmetric confidence interval around the estimated percentile, $\hat{\theta}_q = \hat{F}^{-1}(q)$, is appropriately constructed using a two-sided Wald confidence interval around q rather than a Wilson interval, because q is not an estimate. The picture on page 312 shows how a symmetric confidence interval around q on the y-axis is converted into an asymmetric interval around $\hat{\theta}_q$ on the x-axis.

The Delete-a-Group Jackknife

Many surveys of establishments (businesses, farms, schools) are single-stage element sample where the element is the psu. Consequently, these surveys have hundreds, if not thousands, of design psu's selected from within a smaller, but still often large, number of variance strata. When all the first-stage selection probabilities are small and $n_h > 6$ for all h, Kott (2001) proposes the following delete-a-group (DAG) jackknife routine for MSE estimation purposes.

1. Sort the psu's by stratum, and then systematically assign them to R mutually exclusive groups.

2. Call the complement of each of group r, the r'th jackknife replicate.

3. For element i within a psu in stratum h, define

$$w_{i(r)} = \begin{cases} w_i\, n_h/(n_h - n_{hr}) & \text{when i is in the r'th jackknife replicate} \\ 0 & \text{otherwise (when i is in the r'th group),} \end{cases} \qquad \text{(S36)}$$

   where n_{hr} is the number of psu's in stratum h and group r.

4. When a estimator of a smooth statistic, $\hat{\theta}$, is computed with the w_i, and each of the replicate estimators, $\hat{\theta}_{(r)}$, is computed with its own set of $w_{i(r)}$, the delete-a-group MSE estimator for $\hat{\theta}$ is

$$\hat{V}_{DAGJK}(\hat{\theta}) = [(R-1)/R] \sum^{R} (\hat{\theta} - \hat{\theta}_{(r)})^2. \qquad \text{(S37)}$$

An MSE estimator like that in equation (S37) is often called "a grouped jackknife." Kott showed that when the jackknife replicate weights are computed as in equation (S36), this MSE estimator is close to unbiased under the conditions described above. Kott also proposes an extended DAG jackknife for samples with some smaller n_h values.

The National Agricultural Statistics Service uses the DAG jackknife with 15 replicates (i.e., R = 15). They recommend computing confidence intervals using a Student's t distribution with 14 degrees of freedom. That is good advice for both the DAG jackknife and the random-groups method: confidence intervals should be computed using a t with $R-1$ degrees of freedom.

NASS uses the DAG jackknife with Poisson sampling. Kott (2004a) describes when that approach is appropriate.

Homework for Lesson 10

Do Exercises 9a (treating finalwt as w_i)(starting on p. 317) of Section 9.7 in *Sampling: Design and Analysis*. (When age of the youth is missing, treat it as being over 14.)

Using the data in the first three strata of syc.data, repeat Exercise 9a estimating the variance with a linear variance estimator, with jackknife, and with a DAG jackknife based on 10 groups. Compute a 95% confidence interval for the proportion using the DAG result.

Lesson 11: Two-Phase Sampling

Comments on Section 12.1 of *Sampling: Design and Analysis*

Some may wonder why this course has jumped ahead to a section in Chapter 12 before tackling the important topic of nonresponse, which Professor Lohr treats in Chapter 8. The reason is simple: one of the principal ways of handling nonresponse is to treat the mechanism that results in responses for only a subset of the sample as an additional phase of sampling. Note the word "phase" rather than "stage." Two-stage sampling is an very special case of the much broader concept of two-phase sampling.

Two-Stage Sampling

In two stage sampling, one begins by grouping the elements of interest into clusters and then selecting these clusters, called "primary sampling units" or "psu's," at random. The actually sampling mechanism may be stratified. In addition, different psu's may have different selection probabilities. After this first phase of sampling, a random subsample of elements are selected from each sampled psu. Again, the actual sampling mechanism may be stratified with different elements having different selection probabilities. What makes a two-*stage* sampling procedure special is the second phase of sampling has the following properties:

Independence – The selection of an element in one sampled psu is independent of the selection of an elements from another sampled psu.

Invariance – The probability of selecting an element from a sampled psu does not depend on which of the other psu's were sampled in the first phase.

The independence property of two- (or more) stage sampling is what makes variance estimation so simple when the first stage of selection is conducted with replacement. When j and g are sampled psu's, $\hat{t}_j$ and $\hat{t}_g$ where each is an unbiased estimator based on a sub-sample of elements, are independent.

Invariance is more subtle. This property allows one to easily estimate the overall selection

probability of each sampled element i, π_i, as the product of the selection probability of the psu j containing it, π_{j+}, and the selection probability of i at the second stage of sampling given that psu j was chosen in the first stage, $\pi_{i|j}$; that is, $\pi_i = \pi_{j+}\pi_{i|j}$. Without invariance, $\pi_{i|j}$ not be well defined, because it depends on which psu's are selected in the first phase of sampling.

An Example

One of the simplest examples of a two-phase sample without either the independence or invariance properties is the following. Suppose we had a list of 10,000 potential business addresses. We sampled 1,000 addresses, and mailed each a simple questionnaire. Based on the responses, we determined that 800 addresses could still be valid. The others were returned by the post office as undeliverable or the addressee responded that it was not a business. A simple random subsample of 200 of the 800 addresses was then drawn, with each to be given a lengthy questionnaire by personal interview.

The psu's and ssu's in this example are both elements. Neither are clusters of elements. Effectively, the first-phase sample was divided into two strata. Stratum 1 contained those addresses that could still be associated with businesses. Stratum 2 contained the rest. Since we are interested in estimating totals from stratum 1, we plan to interview the subsample drawn from that stratum. In principal, we treat each element in stratum 2 as if it were selected for the second-phase sample and assigned values of 0 for every survey variable.

Consider two elements selected in the first phase of sampling and assigned to stratum 1. The probability of subsampling each of these distinct first-phase units (psu's) is 200/800. The probability of subsampling both is (200/800)(199/799). They are not independent.

What is the probability of one of those addresses being selected in the two-phase process? It is tempting to compute that value in the following manner. The probability of selecting it in the first phase of selection is 1,000/10,000 or 0.1. The probability of selecting it in the second phase is 200/800 or 0.25. That makes the address's overall selection probability 0.1 x 0.25 or 0.025. Wrong! The actual probability depends on how many of the 10,000 potential business addresses would have been assigned to stratum 1 had they been sampled. Suppose that number, which in practice we are unlikely to know, is 7,500. This

means that of the 10,000 addresses, 7,800 would have been assigned to stratum 1. If we had subsampled 200 addresses no matter how many first-phase addresses were assigned to stratum 1, then the overall selection probability for such an address would be 200/7,500 or approximately 0.0267.

We are ignoring here the small possibility that there may not have been 200 addresses in stratum 1 after the first phase of sampling. Defining exactly what a two-phase sampling process is when there is no invariance property can be a tricky business. In principal, there should be firm rules governing how the second phase sample is to be drawn even though the actual design can vary depending on the first-phase sample. In the above example, we always draw 200 elements from stratum 1 (and, conceptually, all elements from stratum 2), no matter how many first-phase elements fall into that stratum. For completeness, we can add that if fewer than 200 first-phase elements were to be assigned to stratum 1, then we subsample them all.

The Double Expansion Estimator

Without an invariance property, the text correctly conditions D_i on the vector $\mathbf{Z}$ although there is no actual vector algebra in Professor Lohr's treatment. Those uncomfortable with vector notation can replace $\mathbf{Z}$ with $S^{(1)}$, as we do in the following note.

The estimator $\hat{t}_y^{(2)}$ in equation (12.1) of the text is *not* a Horvitz-Thompson estimator when $w_i^{(2)}$ for $i \in S^{(1)}$ varies depending on which other elements are in $S^{(1)}$. In that circumstance $\hat{t}_y^{(2)}$ is a *double expansion estimator.*

The Double and Reweighted Expansion Estimators

The Setup

Suppose $S^{(1)}$ is drawn using a stratified random sample of elements. *Unlike the text, we will use H to denote the number of first-phase strata.* H can be 1. Let F_h denote the set of N_h elements in first-phase stratum h elements before sampling, of which n_h are selected for $S^{(1)}$.

After the first-phase sample has been drawn and surveyed, it is divided into G groups for subsampling, $\mathbb{S}_1, \mathbb{S}_2, ..., \mathbb{S}_G$. As a practical matter, the number of groups (G) cannot be too large, because we will be building an asymptotic framework around the subsample within each group. We let M_g denote the number of first-phase-sample elements in $\mathbb{S}_g$. (This is different from previous lessons, where M_g was the number of population elements in group g.)

We will investigate two subsampling schemes: simple random sampling within groups and analogous Bernoulli sampling. We start by designating a putative sample size per group, m_g, which is large (say, at least 20). When $m_g > M_g$, all of all of $\mathbb{S}_g$ is selected for the subsample under either scheme. Otherwise, under STSRS, m_g elements are subsampled from S_g using simple random sampling. Under the stratified Bernoulli scheme (STBRN), $P(D_i \,|S^{(1)}) = m_g / M_g$ and $P(D_i D_q \,|S^{(1)}) = (m_g / M_g)^2$ for i and q distinct members of $\mathbb{S}_g$. The groups under Bernoulli sampling are not technically quasi-design strata since elements are subsampled independently within them.

Let us assume that the population and first-phase design is such that a reasonable variance estimator for, $\hat{t}_y^{(1)}$, the exists-only-conceptually single-phase expansion estimator is

$$\hat{V}(\hat{t}_y^{(1)}) = \hat{V}(\sum_{i \in S^{(1)}} w_i^{(1)} y_i) = \sum_{h=1}^{H} [n_h/(n_h - 1)]\{ \sum_{i \in F_h} (w_i^{(1)} y_i)^2 - (\sum_{i \in F_h} w_i^{(1)} y_i)^2 / n_h \};$$

that is, the with-replacement variance estimator. Under stratified Bernoulli subsampling, if the variance estimator above is reasonable for $V(t_y^{(1)})$, then the following is reasonable for the variance of the double expansion estimator:

$$\hat{V}_{STBRN}(\hat{t}_y^{(2)}) = \hat{V}(\sum_{i\in S^{(1)}} w_i^{(1)}\hat{y}_i) = \sum_{h=1}^{H} [n_h/(n_h-1)]\{\sum_{i\in F_h}(w_i^{(1)}\hat{y}_i)^2 - (\sum_{i\in F_h} w_i^{(1)}\hat{y}_i)^2/n_h\}, \tag{S38}$$

where $\hat{y}_i = w_2^{(i)}D_i y_i$. This is because the $\hat{y}_i$ are independent under Bernoulli subsampling, just as in two-stage sampling. Invariance has been lost – the definition of $\hat{y}_i$ depends on the size of M_g, but not independence.

As the m_g grow asymptotically large, it turns out that the expected value of the right-hand side of equation (S38) is the same for the STSRS and STBRN subsampling schemes.

That $E[(w_i^{(1)}\hat{y}_i)^2 \mid S^{(1)}]$ is identical under both is not hard to show. Proving the near equality of $(\sum_{F_h} w_i^{(1)}\hat{y}_i)^2 \mid S^{(1)}]$ under the two schemes is a tedious exercise we leave for the hardiest students. Among other things, one must consider different asymptotic assumptions about the n_h.

Although the expectations of $\hat{V}(\hat{t}_y^{(2)})$ as expressed by the right-hand side of equation (S38) are the same under the two schemes, the actual variance will be smaller under STSRS subsampling when all the m_g are large. To see why, observe that the second-phase variance of $\hat{t}_y^{(2)}$ (which treats $\hat{t}_y^{(2)}$ as an estimator for $\hat{t}_y^{(1)}$ given the first-phase sample) is

$$\begin{aligned}
V_{STSRS}(\hat{t}_y^{(2)} \mid S^{(1)}) &= \sum_{g:m_g \le M_g} (1 - m_g/M_g)(M_g^2/m_g)\sum_{i\in S_g}(w_i^{(1)}y_i - \sum_{i\in S_g} w_i^{(1)}y_i/M_g)^2/(M_g-1) \\
&\approx \sum (1 - m_g/M_g)(M_g/m_g)\{\sum_{i\in S_g}(w_i^{(1)}y_i)^2 - (\sum_{i\in S_g} w_i^{(1)}y_i)^2/M_g\} \\
&= \sum \{\sum_{i\in S_g}(w_i^{(2)} - 1)(w_i^{(1)}y_i)^2 - (1 - m_g/M_g)(\sum_{i\in S_g} w_i^{(1)}y_i)^2/m_g\} \\
&\le \sum \{\sum_{i\in S_g}(w_i^{(2)} - 1)(w_i^{(1)}y_i)^2 = V_{STBRN}(\hat{t}_y^{(2)} \mid S^{(1)}),
\end{aligned} \tag{S39}$$

where g: $m_g \le M_g$ restricts the summation to those groups where m_g is no greater than M_g (most, if not all, groups in practice), since the other groups do not contribute to the second-phase variance. The final inequality in (S39) is strict when the mean of the $w_i^{(1)}y_i$ values is positive within at least one S_g with $M_g > m_g$.

Equation (39) suggests the following variance estimator with STSRS subsampling:

$$\hat{V}_{STSRS}(\hat{t}_y^{(2)}) = \hat{V}_{STBRN}(\hat{t}_y^{(2)}) - \sum^G (1 - m_g^{actual}/M_g)(\sum_{i \in s_g} w_i^{(1)} w_i^{(2)} y_i)^2/m_g^{actual}, \tag{S40}$$

where m_g^{actual} is the actual subsample size in group g. The last summation is over the second-phase sample in g, which is denoted s_g. Equation (S40) ignores the impact from the second-phase variance of $(\sum w_i^{(1)} w_i^{(2)} y_i)^2/m_g$, because it is either ignorably small when m_g^{actual} is large or irrelevant when $m_g^{actual} = M_g < m_g$.

Both $\hat{V}_{STBRN}(\hat{t}_y^{(2)})$ in equation (38) and $\hat{V}_{STSRS}(\hat{t}_y^{(2)})$ in equation (S40) can easily extended to include design where the "first phase" of sampling is more complex. NASS draws a stratified, multi-stage cluster sample of farms for its major June survey of crops and stocks. Farms not on NASS's list frame are then restratified based on their June responses and a subsample of them is drawn for enumeration in NASS's December survey. NASS computes a double expansion estimator based on this subsample, but employs an extention of $\hat{V}_{STBRN}(\hat{t}_y^{(2)})$ to estimate its variance. That variance-estimating strategy is biased upward in this context.

The Reweighted Expansion Estimator

A popular alternative to the double expansion estimator is the reweighted expansion estimator:

$$\hat{t}_{yrw}^{(2)} = \sum_{g=1}^{G} \sum_{i \in S_g} w_i^{(1)} \frac{\sum_{i \in s_g} w_i^{(1)} w_i^{(2)} y_i}{\sum_{i \in s_g} w_i^{(1)} w_i^{(2)}}$$

$$= \sum_{g=1}^{G} \sum_{i \in S_g} w_i^{(1)} \frac{\sum_{i \in s_g} w_i^{(1)} y_i}{\sum_{i \in s_g} w_i^{(1)}} \quad . \tag{S41}$$

This estimator need not be unbiased, but it is design consistent under mild conditions when each of the m_g is either large or no less than M_g. We will assume the needed mild conditions always exist in the remainder of this note.

We can write $\hat{t}_y^{(2)} = \sum_s w_i y_i$, where s is the entire second phase sample, and $w_i = w_i^{(1)} w_i^{(2)}$. In a similar vein, $\hat{t}_{yrw}^{(2)} = \sum_s \tilde{w}_i y_i$, where

$$\tilde{w}_i = w_i[\sum_{q \in S_g} w_q^{(1)} / \sum_{q \in s_g} (w_q^{(1)} w_q^{(2)})] = w_i^{(1)} (\sum_{q \in S_g} w_q^{(1)} / \sum_{q \in s_g} w_q^{(1)})$$

for $i \in s_g$, hence the name "reweighted expansion estimator."

When the first-phase sample is self-weighting (all the $w_q^{(1)}$ are equal), the reweighted estimator collapses into the double expansion estimator. It will be convenient later to define $\tilde{w}_i^{(2)} = (\sum_{S_g} w_q^{(1)} / \sum_{s_g} w_q^{(1)})$ for $i \in S_g$, so that $\tilde{w}_i = w_i^{(1)} \tilde{w}_i^{(2)}$.

Let $\bar{y}_g = \sum_{S_g} w_q^{(1)} y_q / \sum_{S_g} w_q^{(1)}$, and $d_i = y_i - \bar{y}_g$ for $i \in S_g$. This lets us rewrite y_i as $\bar{y}_g + d_i$, and and the reweighted expansion estimator as

$$\hat{t}_{yrw}^{(2)} = [\sum_{g=1}^{G} \sum_{i \in s_g} \tilde{w}_i \bar{y}_g] + \sum_s \tilde{w}_i d_i$$

$$= [\sum \sum_{i \in s_g} w_i^{(1)} (\sum_{i \in S_g} w_q^{(1)} / \sum_{i \in s_g} w_q^{(1)}) \bar{y}_g] + \sum_s \tilde{w}_i d_i$$

$$= [\sum \sum_{i \in S_g} w_i^{(1)} \bar{y}_g] + \sum_s \tilde{w}_i d_i$$

$$= [\sum \sum_{i \in S_g} w_i^{(1)} y_i] + \sum_s \tilde{w}_i d_i$$

$$= \hat{t}_{yrw}^{(1)} + \sum_s \tilde{w}_i d_i.$$

From this we see that the second-phase mean squared error of $\hat{t}_{yrw}^{(2)}$ is approximately $V(\sum_s w_i d_i \mid S^{(1)})$ (since $w_i \approx \tilde{w}_i$). Under the STSRS subsampling scheme, that is the same as the variance in equation (S39), only with d_i replacing y_i. Here, however, since $\sum_{S_g} w_i^{(1)} d_i = 0$ for all g, the *same approximate MSE applies for both STSRS and STBRN subsampling.*

To estimate the mean squared of error of $\hat{t}_{yrw}^{(2)}$ under either subsampling scheme, one can use the right-hand side of equation (S38) with

$$\hat{y}_i = \hat{\bar{y}}_g + D_i \tilde{w}_i^{(2)}(y_i - \hat{\bar{y}}_g) \qquad \text{(S42)}$$

$$= y_i + (D_i \tilde{w}_i^{(2)} - 1)(y_i - \hat{\bar{y}}_g) \approx y_i + (D_i[M_g/m_g] - 1)(y_i - \hat{\bar{y}}_g),$$

for $i \in S_g$, where $\hat{\bar{y}}_g = \sum_{s_g} w_i^{(1)} y_i / \sum_{s_g} w_i^{(1)}$, and $e_i = y_i - \hat{\bar{y}}_g$ is used in place of the unknown d_i. The last inequality closely parallels the double expansion case where $\hat{y}_i = D_i[M_g/m_g]y_i = y_i + (D_i[M_g/m_g] - 1)y_i$.

Some would replace $\tilde{w}_i^{(2)}$ with the asymptotically identical $w_i^{(2)} = M_g/m_g$, in $\hat{y}_i$. Our approach has better model-based properties, but we will not show that here (a related point is argued in the following lesson).

An alternative approach to MSE estimation for the reweighted expansion estimator under either subsampling scheme is to use the jackknife:

$$V_{JK}(\hat{t}_{yrw}^{(2)}) = \sum^H [(n_h - 1)/n_h] \sum_{q \in F_h} (\hat{t}_{yrw(q)}^{(2)} - \hat{t}_{yrw}^{(2)})^2,$$

where $\hat{t}_{yrw(q)}^{(2)}$ is computed using equation (S41) with the $w_i^{(1)}$ replaced by $w_{i(q)} = 0$ when $i = q$, and by $[n_h/(n_h - 1)]w_i^{(1)}$ otherwise for $i \in F_h$. See Kott and Stukel (1997).

The MSE estimators described above can remain nearly unbiased under appropriate more-than-two-phase samples. The National Agricultural Statistics Service uses a delete-a-group jackknife with its surveys having multi-phase sampling designs. The theory can also be extended to designs where there is a stratified multi-stage sample of elements the "first"

phase.

The Reweighted and Poststratified Ratio Estimators

If a value x_i is known for every element in the first-phase sample, then a mild extension of the reweighted expansion estimator in equation (S41) is the reweighted ratio estimator:

$$\hat{t}_{yrwr}^{(2)} = \sum_{g=1}^{G} \sum_{i \in S_g} w_i^{(1)} x_i \frac{\sum_{i \in s_g} w_i^{(1)} w_i^{(2)} y_i}{\sum_{i \in s_g} w_i^{(1)} w_i^{(2)} x_i}$$

$$= \sum_{g=1}^{G} \sum_{i \in S_g} w_i^{(1)} x_i \frac{\sum_{i \in s_g} w_i^{(1)} y_i}{\sum_{i \in s_g} w_i^{(1)} x_i} \quad . \tag{S43}$$

The source of these x_i can be a survey of the first-phase sampled units or it can be a combination of information gathered from such a survey and previously held data. For example, a first-phase sample of potential businesses may be surveyed for total annual sales, while a second-phase subsample collects more detailed information, like annual expenditures on electricity. Alternatively, a proxy for annual sales may be available before sampling. The first-phase sample is surveyed merely to determine whether a potential businesses is still in operation, and if so, whether or not it has employees. That information is used to break up the first-phase sample into groups for the second phase of sampling.

Following reasoning analogous to that for the reweighted expansion estimator, the reweighted ratio estimator can be shown to be design consistent under mild conditions. Moreover, its MSE can be estimated with a jackknife or by using a (first-stage) with-replacement variance estimator, like the right-hand side of equation (S38), with

$$\hat{y}_i = x_i \hat{b}_g + D_i \tilde{w}_i^{(2)} (y_i - x_i \hat{b}_g), \tag{S44}$$

$$\approx y_i + (D_i [M_g / m_g] - 1)(y_i - x_i \hat{b}_g),$$

for $i \in S_g$, where $\hat{b}_g = \sum_{s_g} w_i^{(1)} y_i / \sum_{s_g} w_i^{(1)} x_i$. Again, some would replace $w_i^{(2)}$ with M_g / m_g, as in the last near equality.

A useful way of conceptualizing the reweighted ratio estimator is as

$$\hat{t}_{yrwr}^{(2)} = \sum^G \hat{t}_{xg}^{(1)} \hat{b}_g,$$

where $\hat{t}_{xg}^{(1)} = \sum_{S_g} w_i^{(1)} x_i$. When all the $x_i = 1$, we have the reweighted expansion estimator with $\hat{b}_g = x_i \hat{b}_g = \hat{\underline{y}}_g$ for $i \in S_g$, and $\hat{t}_{xg}^{(1)}$ being an estimate of the size of the set of elements in the original population meeting the definition of group g. Call that set U^g, and its size, $M_{(U)g}$ to distinguish it from M_g (the first-phase sample size of S_g).

Now suppose $t_{xg} = \sum_{U_g} x_i$ was known for each g from an outside source. One could then compute a poststratified ratio estimator,

$$\hat{t}_{yprx}^{(2)} = \sum^G t_{xg} \hat{b}_g, \tag{S45}$$

which is design consistent when $\hat{t}_{yrwr}^{(2)}$ is and usually has a smaller MSE because the $\hat{t}_{xg}$ do not have to be estimated using the first-phase sample.

To estimate the MSE of $\hat{t}_{yprx}^{(2)}$ under either STSRS or STBRN subsampling, one can use a jackknife or a (first-stage) with-replacement variance estimator, like the right-hand side of equation (S38), with

$$\hat{y}_i = D_i \tilde{w}_i^{(2)} (y_i - x_i \hat{b}_g) \tag{S46}$$

$$\approx e_i + (D_i [M_g / m_g] - 1) e_i,$$

where $e_i = (y_i - x_i \hat{b}_g)$.

The most common postratified ratio estimator in practice is the special case where all $x_i = 1$. In this context, the poststratified ratio estimator is the poststratifed estimator, which was expressed in equations (S33) with different notation and without a second phase of sampling. Here, it is $\hat{t}_{yp}^{(2)} = \sum^G M_{(U)g} \hat{\underline{y}}_g$.

Homework for Lesson 11

Using the data from the first three strata of syc.dat in *Sampling: Design and Analysis* (p. 445) as the first-phase sample. Draw three simple random subsamples containing 20 whites, 20 blacks, and 20 of all other races (including those not responding to the race question), respectively. From this subsample, using finalwt as $w_i^{(1)}$, estimate the total number of youths in the population who lived with their mother only (treating nonrespondents to that question as if they did NOT live with their mother only) using a double expansion estimator and then a reweighted expansion estimator. Estimate the variances of your answers using a jackknife for the reweighted expansion estimator.

Lesson 12: Nonresponse

Comments on Chapter 8 of *Sampling: Design and Analysis*

Response Homogeneity Groups

Treating the subsample of elements that respond to a survey as if it was selected using a probability sampling mechanism can be called *quasi-randomization*, *quasi-random modeling*, or *response modeling*. This additional phase of sampling is assumed to be Poisson, where an element's probability of responding may be a function of its attributes, but is independent of whether or not other particular elements respond.

It is important to remember that although quasi-random modeling has the same form as randomization-based theory, it depends on a model, which, like all models, can fail. An unbiased estimator under the quasi-random model will be said to be *quasi-design* (or *quasi-randomization)* unbiased and to have a quasi-design mean squared error.

The most common quasi-random model assumes that the population is naturally divided into mutually exclusive *response homogeneity groups* (Särndal *et al.*, 1992), the weighting classes of the text, where each element of a group is assumed equally likely to respond to the survey. The survey respondents, then, are a STBRN subsample of the original sample, where the groups are the "strata." Moreover, given the realized number of respondents in each group, the Bernoulli subsample in each group is *conditionally* a simple random sample.

It is not unusual to treat the design strata themselves as the response homogenity groups, especially with a stratified simple random original sample. Since a (conditional) simple random subsample of a simple random sample is itself a simple random sample, one effectively still has a STSRS sample under the quasi-random model, given the respondent sample size within each stratum,

Given a possibly stratified multi-stage sample with response homogeneity groups that may (or may not) cut across strata, and borrowing notation from the last lesson, the reweighted expansion estimator in equation (S41),

$$\hat{t}_{yrw}^{(2)} = \sum_{g=1}^{G} \sum_{i \in S_g} w_i^{(1)} \frac{\sum_{i \in s_g} w_i^{(1)} w_i^{(2)} y_i}{\sum_{i \in s_g} w_i^{(1)} w_i^{(2)}}$$

$$= \sum_{g=1}^{G} \sum_{i \in S_g} w_i^{(1)} \frac{\sum_{i \in s_g} w_i^{(1)} y_i}{\sum_{i \in s_g} w_i^{(1)}} \quad . \qquad \text{(S41)}$$

$$= \sum_{g=1}^{G} \left(\sum_{i \in S_g} w_i^{(1)} \right) \hat{\bar{y}}_g .$$

applies. Here S_g is the sample of M_g elements in response homogenity group g, $w_i^{(1)}$ is the inverse of the probability of selecting element i for the original sample, and $w_i^{(2)} = 1/P(R_i = 1) = 1/\phi_g$ is a constant for all $i \in S_g$. Observe that we do not need to know what ϕ_g is to compute equation (S41).

Although ϕ_g is the (unconditional) probability an element in group g responds, given a realized subsample size, m_g, $P(R_i = 1|m_g) = m_g/M_g$ is the conditional probability of response for element $i \in S_g$. Either probability (ϕ_g or m_g/M_g) can be used for computing $w_i^{(2)}$ and the result is the reweighted expansion estimator in equation (S41). The quasi-design MSE of $\hat{t}_{yrw}^{(2)}$ can be estimated using methods described in the last lesson for its design-MSE estimation long as each m_g is large (or equal to M_g) and the variance in the absence of nonresponse could reasonably be estimated using the (first-stage) with-replacement variance estimator.

Most survey organization require m_g to be at least 20, although some have accepted group subsample sizes as small as six (see Kott, 1994a; there is anecdotal evidence that this was Australian Bureau of Statistics policy as well). When some m_g are too small, a common practice is collapse initially contemplated response homogeneity groups into a smaller set of groups with larger subsample sizes within each.

Sometimes the population size, $M_{(U)g}$, of the set (U_g) of elements meeting the definition of response homogenity group g is known for all groups, and we can do better than the

reweighted expansion estimator. Consider, for example, a survey of US residents 15 years or older. Suppose it is reasonable to assume the following response homogeneity groups: black men, black women, non-black men, and non-black women. Population numbers for these groups are available from another source. Although those number may themselves be estimates, for the purposes of our much-smaller survey, they can be treated as close enough to truth to be treated as such. The poststratified estimator, $\hat{t}_{ypr}^{(2)} = \sum^{G} M_{(U)g} \hat{\bar{y}}_g$, a special case of the poststratified ratio estimator in equation (S45) (see also (S33)), can be computed and its variance estimated using methods described in the previous lesson.

The post-stratified estimator can also be used to adjust for undercoverage of the population when the sample (or frame) population is smaller than the target population. For example, the target population may be all non-institutionalized residents of the citizens 15 year or older, but the sample population only includes that part of the population residing in households with telephones. By breaking target population into groups with known $M_{U(g)}$ sizes and assuming a quasi-random model in which each individual in a group is equally likely to be in the sample, we can apply our theory of quasi-random response to quasi-random coverage. Implicit in this model, is the assumption that D_i, a random variable equaling 1 when i is in the sample population and zero otherwise, is independent of the sampling mechanism and the survey values. Formally, $(D_i|y_i, Z_i) = \phi_g$ when i is in group g. Since Z_i and D_i are independent, it doesn't matter than undercoverage occurs before sampling and nonresponse after. The same theory applies.

Reweighting Versus Cell-Mean Imputation

In practice, reweighting is an effective way of dealing with unit nonresponse. It is less attractive for handling item nonresponse because the technique requires the creation of a different set of weights for every item with a unique collection of respondents. Imputation is a more popular approach for handling item nonresponse.

Such practical considerations serve to obscure the theoretical differences and similarities between the two approach. To clarify the issues, let us investigate the single-variable (item) framework usually addressed in textbooks on survey sampling. Suppose we have a sample of size M, but only m units respond with y-values. Values for the rest need to be imputed. Suppose further we have exactly the same set up as in our discussion of

reweighting. Now, however, instead of the original sample (and population) being divided into G mutually exclusive response homogeneity groups, it is divided into G mutually exclusive imputation cells. The group-mean prediction model is deemed to be reasonable:

$y_i = \mu_g + \epsilon_i$,

for $i \in U^g$, where $E(\epsilon_i \mid Z_i, R_i) = E(\epsilon_i \epsilon_q \mid Z_i, Z_q, R_i, R_q) = 0$ for i and q distinct, so that both the sampling and response mechanisms are ignorable. It makes some sense to impute a a missing y_q, where $q \in S_g$, by $y_q^* = \bar{y}_g$. It is actually more efficient to impute with $y_q^* = \sum_{s_g} y_i / m_g$, but both are unbiased under the model in the sense that $E_M(y_q - \bar{y}_g \mid Z_i, R_i) = 0$. Imputing each missing y_q in $\hat{t}_y^{(1)} = \sum^G \sum_{s_g} w_i^{(1)} y_i$ with $y_q^* = \hat{\bar{y}}_g$ when $q \in S_g$ yields

$$\hat{t}_{yl}^{(1)} = \sum_{i \in s_g}^{G} \{ \sum w_i^{(1)} y_i + \sum_{i \in S_g - s_g} w_i^{(1)} \hat{\bar{y}}_g \} = \sum^{G} \{ (\sum_{i \in s_g} w_i^{(1)}) \hat{\bar{y}}_g + \sum_{i \in S_g - s_g} w_i^{(1)} \hat{\bar{y}}_g \} = \sum^{G} \sum_{i \in S_g} w_i^{(1)} \hat{\bar{y}}_g = \hat{t}_{yrw}^{(2)}.$$

The estimators are the same, but the model assumptions justifying reweighting and imputation are different. For the former, one assumes that every element in S_g is equally likely to respond. For the later, one assume that every element in the groups has the same expected y-value whether or not it responds. On the one hand, the quasi-random model applies to all possible y-variables, while the prediction model may be true for some variables and not true for others. On the other, the quasi-random denies the real possibility that some sampled elements will refuse to participate in a survey no matter how ordinary – and predictable – their survey values.

Under the prediction model, we are concerned with the combined bias of $\hat{t}_{yl}^{(1)}$ and estimating its combined MSE; that is, $E\{E_M(\hat{t}_{yl}^{(1)} - t_y)\}$, and $E\{E_M[(\hat{t}_{yl}^{(1)} - t_y)^2]\}$, where the first "E" refers to expectation with respect to the original sampling design. As in the previous lesson, let us define $\tilde{w}_i = w_i^{(1)} \sum_{S_g} w_q^{(1)} / \sum_{s_g} w_q^{(1)}$ for $q \in S_g$ so that $\hat{t}_{yl}^{(1)} = \sum_s w_i y_i$ and $\sum_{s_g} \tilde{w}_i = \sum_{S_g} w_q^{(1)}$. Under the group-mean prediction model:

$$\hat{t}_{yl}^{(1)} - t_y = \sum^{G} \sum_{i \in s_g} \tilde{w}_i y_i - \sum^{G} \sum_{i \in U_g} y_i$$

$$= \sum^{G} \sum_{i \in s_g} \tilde{w}_i(\mu_g + \epsilon_i) - \sum^{G} \sum_{i \in U_g} (\mu_g + \epsilon_i)$$

$$= \sum^{G} (\sum_{i \in S_g} w_i^{(1)} - M_{(U)g})\, \mu_g + (\sum_{i \in s} \tilde{w}_i \epsilon_i - \sum_{i \in U} \epsilon_i)$$

$$= \quad A \quad + \quad B.$$

Therefore, $E\{E_M(\hat{t}_{yl}^{(1)} - t_y)\} = E\{E_M(A + B)\} = E\{E_M(A)\} + E\{E_M(B)\} = E(A)$, since A is a constant term from the model-based point of view; that is, given the sample. Define z_i to equal μ_g for $i \in U_g$. Then $\hat{t}_z^{(1)} = \sum_S w_i^{(1)} z_i = \sum^G \sum_{S_g} w_i^{(1)} z_i$ is a design unbiased estimator for $\sum^G \sum_{U_g} z_i = \sum^G M_{(U)g}\mu_g = t_z$, which means that $E(A) = E(\hat{t}_z^{(1)} - t_z) = 0$. Thus, $\hat{t}_{yrw}^{(2)}$ combined unbiased, and so its combined MSE is equal to its combined variance

Similarly, $E\{E_M[(\hat{t}_{yl}^{(1)} - t_y)^2]\} = E\{E_M[(A + B)^2]\} = E(A^2) + E\{E_M(B^2)\}$, again because A is a constant under the model. Letting $V(\epsilon_i) = \sigma_i^2$, and assuming $\sum_s w_i^2\sigma_i^2 >> \sum_U \sigma_i^2$, the (first-stage) with-replacement variance estimator with

$$w_i^{(1)}\hat{y}_i = w_i^{(1)} \hat{\bar{y}}_g + D_i\tilde{w}_i(\hat{y}_i - \bar{y}_g)$$
$$\approx w_i^{(1)} z_i + D_i\tilde{w}_i\, \epsilon_i$$

replacing $w_i^{(1)} y_i$ can be shown to estimate the combined variance of the reweighted expansion estimator in a nearly unbiased fashion. This is precisely what we would use to estimate the design mean squared error under the quasi-random model. We now can see that using $\tilde{w}_i$, as opposed to $w_i^{(1)}(M_g/m_g)$, better captures the model variance component of the combined variance.

To estimate the combined variance of $\hat{t}_{yl}^{(1)}$ with a jackknife, one can use the same jackknife one would compute for the analogous reweighted expansion estimator. This turns out to be identical to computing

$$\hat{t}_{yl(r)}^{(1)} = \sum^{G} \sum_{s_g} w_{q(r)}^{(1)} y_q + \sum^{G} \sum_{S_g - s_g} w_{q(r)}^{(1)} y_{q(r)}^*$$

for the r'th replicate, where $y_{i(r)}^* = \sum_{s_g} w_{q(r)}^{(1)} y_q / \sum_{s_g} w_{q(r)}^{(1)}$.

Ratio Imputation

Usually imputation is used when a survey respondent reports some item values but not others. We consider only one of the simplest of such situations here, so we can get to the to the heart of the matter without changing our notation very much.

Suppose there is a variable, x, that is reported by every element i in the sample, and a related one, y, that is not. To borrow as much as possible from our previous results, we let S denote the full sample, and s the subsample with y-values. Moreover, we assume the sample , and implicitly the population, can be divided into G mutually exclusive groups, such that the following prediction model is sensible:

$$y_i = \beta_g x_i + \epsilon_i,$$

for $i \in S_g$ (the part of S in group g), where $E(\epsilon_i | x_i Z_i, R_i) = E(\epsilon_i \epsilon_q | x_i, x_q, Z_i, Z_q, R_i, R_q) = 0$.

One can easily show using arguments like those used with cell-mean imputation that imputing a missing y_q with $y_q^* = x_q b_g = x_q (\sum_s w_i y_i / \sum_s w_i x_i)$ produces an estimator, $\hat{t}_{ylr}^{(1)}$, identical to the reweighted ratio estimator, $\hat{t}_{yrwr}^{(2)}$, in equation (S43). Moreover, the combined variance of this estimator can be computed in precisely the same manner as the design MSE of the reweighted ratio estimator in the previous lesson. There is also an obvious analogue for the jackknife.

Interestingly, even if the prediction model above fails, $\hat{t}_{ylr}^{(1)}$ can be justified as long as every sampled element in group g is equally likely to provide a y-variable. Because $\hat{t}_{ylr}^{(1)} = \hat{t}_{yrwr}^{(2)}$, estimator is in some sense (nearly) unbiased if either the prediction model holds or the

quasi-random model holds. It does not require both to hold. Of course, both may fail, and $\hat{t}_{ylr}^{(1)} = \hat{t}_{yrwr}^{(2)}$ not be nearly unbiased in any sense.

In practice, we often have both unit and item nonresponse. It is a simple, if mathematically cumbersome, to combine the two. In the above ratio example, the "original" sample may have already included weight adjustment (captured within the $w_i^{(1)}$ for unit nonresponse. The response homogeneity groups used for reweighting may or may have not coincided with the groups formed to impute for missing y-values.

Homework for Lesson 12

Do Exercises 2a, 2d, 2e, 8b, and 10 in section 8.9 (starting on p. 283) of *Sampling: Design and Analysis*. Example 7.4 is on page 235.

Lesson 13: Other Topics in Sampling

The note beginning on the next page, "Multiple-Frame Establishment Surveys" is a shortened and slightly modified version of Chapter 11 of *Business Survey Methods* (1995) by Cox et al. It contains its own set of equations (denoted V1, V2, etc.) and references.

Comments on Sections 12.2 to 12.4 of *Sampling: Design and Analysis*
[Read the note on multiple-frame surveys first.]

The Census of Agriculture

Prior to the 1997 publication, the US Census of Agriculture was conducted by the Census Bureau. The Bureau used a dual-system approach (pp 391, 392) to estimate the undercoverage of its list of farms, treating the list as the first frame. The "sample design" from this frame was ideally a complete census, although in practice the Bureau had to adjust for nonreponse. The National Agriculture Statistics Service's area sample provided the estimates from the second frame.

The official Census numbers were derived from the Bureau's list of potential farms and adjusted for nonresponse. Dual-system estimates capturing farms missing from the list were published in an appendix.

When NASS took over the Census of Agriculture in 1997, it abandoned dual-system estimation and treated the area frame as if it were complete. The agency believed that the "double counting" of area-sample farms with undetected matches on the list frame offset the small number of farms missing from the area frame.

NASS observed a related matching problem with farms not originally on the 1997 Census mailing list but added later. These "late adds" were matched to area-sample tracts, weighted, and then subtracted from the nonoverlap estimate. Unfortunately, the number of farms subtracted from the grand total was only about half of the number of farms added to the list Ideally, the list-based and area-based estimates of the number of late adds should have been identical – give or take sampling error. It is unlikely that the NASS area frame was missing half the 1997 late adds. A more plausible explanation is that there were many late adds that should have been matched with area-frame tracts but were not. A

homework problem describes the method NASS used in 2002 to try to fix this late-add problem.

Composite Estimators

The estimator in equation (V3) of the supplemental note on multiple frames is a composite estimator like that described in a small-domain context on page 399 of *Sampling: Design and Analysis.* When combining two unbiased estimators, t_1 and t_2, of the same quantity, t, any composite estimator, $t_\lambda = \lambda t_1 + (1-\lambda)t_2$, is unbiased. When the two estimator are independent, the optimal value for λ is $V(t_2) / [V(t_1) + V(t_2)]$. See equation (V7).

As pointed out in the text, the synthetic estimator for a domain has a large bias but a small variance. It seems, therefore, a poor candidate for a component of composite estimation. From a model-based point of view, however, the bias of the synthetic estimator can be modeled as a variance component. This bias tends to persist as the sample underpinning the direct estimator gets larger and larger. Hence, a composite estimator combining a direct and synthetic estimator for a domain can be design consistent if the contribution of the direct estimator (α_d in the text) tends to 1 as its sample size grows arbitrarily large. Ghosh and Rao (1994) provides an excellent summary of the methods for estimating small-domains, especially those involving components-of-variance modeling.

Benchmarking

There is another way to combining synthetic domain estimators with a direct estimator. Let us confine our attention to estimators of a total. Suppose we have a direct estimator, $\hat{t}_{(dir)}$ for $t = \sum^D t_d$, the sum of the totals across the domains, with little or no bias and an acceptably small MSE. The MSE's for its components, the $\hat{t}_{d(dir)}$, by contrast, are disturbingly high. We would like to replace each of these domain-level estimated totals with a more reasonable, but biased, estimator, $\hat{t}_{d(biased)}$. These replacements can synthetic, composite, *ad hoc*, or the result of some combination of techniques. A reasonable strategy is to *benchmark* the biased domain estimators to the (virtually) unbiased estimator of the grand total; that is, to compute $\hat{t}_{d(bench)} = \hat{t}_{d(bias)} [\hat{t}_{(dir)} / \sum_{\delta=1}^{D} \hat{t}_{\delta(bias)}]$ for each d. It is easy to see that the sum of these still-biased estimators across the D domains is $\hat{t}_{(dir)}$.

MULTIPLE FRAME ESTABLISHMENT SURVEYS

Frederic A. Vogel
National Agricultural Statistics Service

Phillip S. Kott
National Agricultural Statistics Service

1 INTRODUCTION

Sample surveys of economic establishments are usually designed to provide estimates of characteristics such as total sales, expenditures, number of workers, and inventories for a population of interest. Basic principles from finite population sampling theory apply regardless of the specific sampling design used. A population of elements must be defined (not always a trivial task when business entities have complex internal structures), and a sampling frame must be constructed from which a sample of elements can ultimately be drawn. Every element in the population must have a known probability of selection. Ideally, the frame should also be complete; that is, the selection probability of every element in the population should be positive. In many establishment surveys, but certainly not all (see Subsection 3.1), the population of elements is identical to the population of establishments of interest.

Populations of establishments, whether of farms, retail stores, factories, buildings, schools, or governments often possess common characteristics that impact on the choice of sampling frame and overall sample design. Among these characteristics are skewed distributions, diversity of variables of interest, and changing population membership.

Sample designs for the establishment surveys are often based on list frames incorporating known or projected measures of size for each element in the. Unfortunately, when the variables of interest are diverse and weakly correlated, a single list frame with one measure of size may not suffice. Moreover, the changing nature of the population causes any list frame or combination of list frames to become outdated quickly and therefore incomplete in terms of coverage of the population.

One way to assure the completeness of the frame is to use an area frame covering the entire population of interest so that all the elements in the population and all potential future elements will be located somewhere within the area frame. Whereas a list frame is a list of the elements in the population, an area frame is a collection of geographical units or area segments. In list-frame designs for establishment surveys, the sampling units are usually the establishments themselves, and stratified single stage sample designs are commonly used to select establishments directly from the list. In area-frame designs, area segments are the sampling units, and these units are often selected using stratified multistage designs. Correspondence rules are needed to link the establishment in the population to the area segments in the frame in a unique manner.

Although area frames ensure complete coverage, they do not generally lead to efficient sampling designs because area segments are essentially clusters of elements.

Segment sizes (in terms of numbers of constituent elements) are usually unequal and unknown at the time of the sample design. In fact, there is no guarantee that an area segment contains any establishments at all. Large area sample sizes will be needed to overcome problems with populations containing rare items (variables) and with skewed distributions.

Area frames are most useful for general purpose surveys covering a wide spectrum of items that are fairly evenly distributed geographically or when the sizes (as defined above) of the area segments are available or can be estimated reasonably well. That is why they are used so often in demographic surveys where population censuses provide adequate measures of the size of an area segment. It should also be noted that an area frame has a long life span. It only needs to be updated when geographic features have changed to the point that it becomes difficult to associate population elements with sampled area segments or when the availability of more up-to-date information allows for the development of improved sampling designs. This occurs often in agricultural surveys with the use of recent aerial photographs and/or satellite data.

Area-frame establishment surveys are generally more expensive than list-frame surveys of comparable sample size. This is because area frames are usually much more costly to develop. In addition, sampled establishments from an area frame often have to be personally enumerated, while sampled elements from a list frame can be enumerated by telephone or electronically.

A *multiple-frame* survey uses a combination of frames. The primary reason for using multiple-frame sampling for establishment surveys is to utilize the strengths of one frame to offset the weaknesses of the other. In principle, the theory of multiple-frame sampling can be applied to the use of more than one list frame (see, for example, Bankier 1986). The main focus here, however, will be on combining the use of list and area sampling frames. Area-frame sampling assures completeness, while list-frame sampling can be designed efficiently for large and rare items.

Section 2 outlines the theory of sampling from multiple frames. Section 3 discusses the most common use of multiple-frame sampling in which there is a single list and a single area frame. Section 4 addresses subsampling from an area frame. Section 5 reviews some practical problems with overlapping frames. Section 6 briefly discusses future directions in multiple-frame methodology.

2 FUNDAMENTALS OF MULTIPLE FRAME SAMPLING

This section develops some theory for sampling and estimation from multiple frames. Each frame consists of a set of mutually exclusive primary sampling units, and there exists a one-to-one or many-to-one mapping from the elements in the population of interest to the primary sampling units in a particular frame. The mapping need not be complete for each frame (i.e., some population elements may not be associated with any sampling unit for a particular frame). For simplicity, we will speak of an element belonging to a frame when it maps onto a sampling unit in the frame (sometimes these elements are referred to as the *ultimate* sampling units of the sampling frame). After one or more stages or phases of random sampling, a sample of elements is drawn independently from each frame.

Two important assumptions are made in this section. They are

Completeness - Every element in the population of interest must belong to at least one frame; and

Identifiability - It should be possible to determine whether or not a sampled element from one frame belongs to any other frame and hence could also have been sampled from it.

The completeness assumption is satisfied whenever an area sampling frame covering the entire population of interest is one of the multiple frames. The identifiability assumption, while simple in theory, poses most of the operational difficulties in implementing a multiple-frame survey. This is because it is not always a trivial matter to ascertain whether an element sampled from one frame is also contained in another frame (see Section 5).

The basic theory of multiple-frame sampling was developed by Hartley (1962) and extended by Cochran (1965). Hartley divided the population into mutually exclusive domains defined by the sampling frames and their intersections. For example, if there are two sampling frames, A and B, there are three possible domains:

Domain (a) containing elements belonging only to Frame A;
Domain (b) containing elements belonging only to Frame B; and
Domain (ab) containing elements belonging to both Frames A and B.
With k frames, there will be $2^k - 1$ domains (recall the completeness assumption precludes the existence of a domain without any members from at least one frame).

Let us focus attention on the two-frame example introduced above to clarify some of the issues involved in estimation using multiple frames. Suppose we are interested in estimating a population total $t = \sum_U y_i$. It is possible to decompose t as

$$t = t_a + t_b + t_{ab}\,, \tag{V1}$$

where T_d is the population total in domain d (d = a, b, or ab). Attention in this note will be principally focused on estimating population totals.

When d = a or b, one can estimate t_d with the Horvitz-Thompson or expansion estimator,

$$\hat{t}_d = \sum_{i \in S_d} w_i y_i\,, \tag{V2}$$

where S_d is the set of sampled elements in domain d, and w_i is the sampling weight for element i (with respect to the design under which S_d was selected). A continuum of unbiased *composite estimators* for t_{ab} is given by

$$\hat{t}_{ab(\lambda)} = \lambda \sum_{i \in S_{ab}^{A}} w_i y_i + (1 - \lambda) \sum_{i \in S_{ab}^{B}} w_i y_i \quad = \quad \lambda \hat{t}_{ab}^{A} + (1 - \lambda) \hat{t}_{ab}^{B}, \tag{V3}$$

where $0 \leq \lambda \leq 1$, S_{ab}^{G} is the set of elements in domain ab sampled from frame G (G = A or B), and $\hat{t}_{ab}^{G}$ is the Horvitz-Thompson estimator for t_{ab} based on the frame-G sample. The limits on λ assure that $\hat{t}_{ab(\lambda)}$ will be nonnegative whenever the y_i are nonnegative.

We now have a continuum of unbiased estimators for t:

$$\hat{t}_{(\lambda)} = \hat{t}_a + \hat{t}_b + \hat{t}_{ab(\lambda)}, \tag{V4}$$

where $0 \leq \lambda \leq 1$. The sampling design is independent across frames but not necessarily across domains. Consequently, The variance of $\hat{t}_{(\lambda)}$ is

$$\begin{aligned} V(\hat{t}_{(\lambda)}) \quad = \quad & V(\hat{t}_a) + V(\hat{t}_b) + 2\lambda Cov(\hat{t}_a, \hat{t}_{ab}^{A}) \\ & + 2(1 - \lambda) Cov(\hat{t}_b, \hat{t}_{ab}^{B}) \\ & + \lambda^2 V(\hat{t}_{ab}^{A}) + (1 - \lambda)^2 V(\hat{t}_{ab}^{B}). \end{aligned} \tag{V5}$$

It is reasonable to choose λ so that the variance of $\hat{t}_{(\lambda)}$ is minimized. Some Calculus reveals that the optimal (i.e., variance-minimizing) λ is

$$\lambda_{opt} = \frac{V(\hat{t}_{ab}^{B}) - Cov(\hat{t}_a, \hat{t}_{ab}^{A}) + Cov(\hat{t}_b, \hat{t}_{ab}^{B})}{V(\hat{t}_{ab}^{B}) + V(\hat{t}_{ab}^{A})}. \tag{V6}$$

If the two covariance terms in equation (V6) were zero (or equal to each other), λ_{opt} would take on the form:

$$\lambda_{opt} = \frac{V(\hat{t}_{ab}^{B})}{V(\hat{t}_{ab}^{B}) + V(\hat{t}_{ab}^{A})}, \tag{V7}$$

which always lies between zero and one. As one would expect, the size of λ is directly related to the precision of $\hat{t}_{ab}^{A}$ relative to that of $\hat{t}_{ab}^{B}$. The more relatively precise the Frame A sample is in estimating domain ab (i.e., the smaller the variance), the more weight $\hat{t}_{ab}^{A}$ is given in the estimation of t_{ab}.

The sampling weights, w_i, in the above expressions can be either unconditional or conditional. That is to say, they may reflect either the original probabilities of selection or the recomputed selection probabilities within the domains. For example, if the sample design in Frame A were simple random sampling without replacement, then the unconditional sampling weight for every sampled element from the frame would be N_A/n_A, where N_A and n_A are the respective sizes of the population and sample in Frame A, while the conditional sampling weight for the sampled elements in the intersection of domain d

and Frame A would be $N_{A(d)}/n_{A(d)}$, where $N_{A(d)}$ and $n_{A(d)}$ are the respective sizes of the population and sample in this intersection.

There are theoretical reasons for preferring estimation with conditional rather than unconditional sampling weights (see Rao 1985). In many applications, however, especially those involving an area frame (where population sizes are often unknown), conditional selection probabilities will be either impossible to calculate or impractical. Consequently, unconditional inference will have to suffice.

Although one can, in principle, choose λ so that the variance (conditional or unconditional) is minimized, this is usually impossible to do in practice because the component variance and covariance terms in equation (V5) are unknown. They could be estimated from the sample, but then the choice of λ would not really minimize the actual variance of $\hat{t}_{(\lambda)}$ but the estimated variance of $\hat{t}_{(\lambda)}$. As a consequence, this estimated variance would be biased downward.

Even if the distinction between the variance-minimizing λ and the estimated variance-minimizing λ could be ignored, the following inconvenience remains: the optimal λ can vary from survey item to survey item. This point is demonstrated by Armstrong (1979) with Canadian farm data.

A popular alternative is to eschew issues of optimality or near optimality and fix the value of λ in advance, most commonly at zero. One can then estimate the components on the right hand side of equation (V5) in an unbiased fashion and create an unbiased estimator for the variance of $\hat{t}_{(\lambda)}$. This approach would be slightly biased if λ were estimated from current data instead of being fixed, because the estimated variance of $\hat{t}_{(\lambda)}$ is no longer a linear combination of unbiased estimators.

It is possible to develop a more direct Horvitz-Thompson estimator for t in equation (V1) by treating the samples from the two frames as a single sample and then computing the probability of selection for each sampled element (the sum of its probability of selection from each frame minus the product of these two terms). There is no reason to believe, however, that the resultant estimator will have less variance than $\hat{t}_{(\lambda)}$ when a near optimal λ is used. Moreover, estimating the variance of this alternative will often be quite difficult.

Finally, one can sometimes improve on $\hat{t}_{(\lambda)}$ by using auxiliary information. The interested reader is directed to Fuller and Burnmeister (1972), Bosecker and Ford (1976), Bankier (1986), Skinner (1991), and Rao and Skinner (1993).

3 THE DOMINANT SPECIAL CASE: ONE LIST FRAME, ONE AREA FRAME

The U.S. Census Bureau's 1949 Sample Survey of Retail Stores (Hansen et al. 1953) was one of the first establishment surveys that used a multiple-frame technique. It employed a single list frame and a single area frame. Nevertheless, it is fair to say that the approach currently finds its widest use in surveys of agriculture. See Vogel (1975) and Julien and Maranda (1990) for respective discussions of U.S. and Canadian multiple-frame agricultural surveys.

3.1 Area Sampling

The basic concepts of area-frame sampling are simple. The total area to be surveyed is first stratified by geography or other known characteristics that are related to the variables of interest (population density is one such characteristic for a retail survey, farm density is one for an agriculture survey). A sample of segments -- usually compact blocks of land -- is then selected within each stratum, perhaps using a multistage sampling design.

Gallego and Delincé (1993) discuss sampling designs where points are randomly selected within sampled area segments. In this section, however, we focus on designs for which the last stage of sampling is the selections of area segments. When employing such designs, rules need to be developed to uniquely link (map) the elements in the population with the sampled segments. The area-frame sampling weight for each element is simply the inverse of the selection probability of the segment with which it is linked.

Although the basic concepts are simple, the successful application of area-frame sampling can become very complex, especially if used in a multiple-frame environment. The theory and application of area-frame sampling are well documented and include Jessen (1942), King and Jessen (1945), Houseman (1975), Nealon and Cotter (1987), and Bush and House (1993). Fecso et al. (1986) contains a good list of references covering area-frame sample design issues.

The link between population elements and sampled segments is complicated in those area-frame surveys (e.g., agricultural surveys) where the physical location of establishments (farms) may cross segment boundaries. There are three approaches to redefining the elements of an area frame frequently used in this situation.

In the closed (or tract) approach, an element is defined as the intersection of an establishment and an area segment (i.e., the portion of the establishment that lies within the area segment). Response errors occur when the reporting unit, usually the establishment, is not able to measure and report on that portion of its business that lies within the selected segment's boundaries.

The *closed* approach is statistically robust and relatively efficient for items that are generally distributed evenly over wide areas, such as crop acreages. Outliers can be controlled by the size of the segment for many of these items. For example, if the segment size is 600 acres, the maximum acres for any crop in the segment is 600 acres even though a single establishment located in the segment may account for many times that number of crop acres in its entire operation.

In the *open* approach, an element is defined to be an establishment and is linked only with that segment containing its headquarters. This can lead to difficulties with large establishments with complicated corporate or partnership structures. For example, many medical practices involve several doctors with offices in more than one location and identifying the "headquarters" may not be simple. Complex counting rules must be devised that ensure such establishments do not have a chance of being linked with more than a single area segment.

A sampled segment will likely contain fewer elements with larger item values when using the open approach to defining elements as opposed to a closed approach. As a result, the former is often less statistically efficient than the alternative. A potential advantage of the open approach, however, is that once the establishment has been defined, it may be easier for the respondent to report values for the entire establishment rather than for that portion of the establishment located within an area segment. Segment boundaries will have little intrinsic meaning to many respondents.

In theory, the definition of the element in the *weighted* approach is the same as in the closed approach: the portion of an establishment that lies within an area segment. The difference is in the data values attributed to the element. Data values for an establishment that is the "parent" of several elements (each in a different area segment) are prorated using fixed weights defined so that they sum to unity. For agricultural surveys, it is common to use the fraction of the establishment's (farm's) land within a sampled segment as the weight for the element in that segment.

The weighted approach leads to less sampling variability than the open approach because large operations are spread across many segments. One major problem with this approach, especially when it is applied to agricultural surveys, is that the weights themselves are prone to measurement errors (see Nealon 1984, p. 19). In principle, the weights for all the elements with the same parent establishment should sum to unity. In practice, since only one of these elements is likely to be selected in an area sample, pains must be taken to assure that the weight used to prorate establishment values to sampled elements are not systematically larger or smaller than they should be.

Nealon (1984) provides a theoretical and empirical examination of these three approaches for area-frame and multiple-frame surveys conducted by the U.S's National Agricultural Statistics Service (NASS). The results favor the weighted approach. Houseman (1975) contains a more in depth discussion of the three approaches. Faulkenberry and Garoui (1991) explore additional approaches.

3.2 Estimation

Suppose two frames are to be used, one list and one area frame, and a total is to be estimated. Using the notation of Section 2, let A denote the area frame and B the list frame. Observe that domain b is empty because all elements should be located somewhere on the area frame. Consequently, $t_b = \hat{t}_b = 0$.

In almost all applications of the two frame design, λ in equations (V3) and (V4) is set equal to zero, rendering $\hat{t} = \hat{t}_a + \hat{t}_{ab}{}^B$. This value for λ, besides being convenient for

variance estimation purposes, is often close to optimal (depending on the efficiency of the list-frame stratification) because the variance of the list-frame estimator of domain ab (i.e., $\hat{t}_{ab}^{B}$) is almost always considerably smaller than the variance of the area-frame estimator of this domain ($\hat{t}_{ab}^{A}$) (recall Section 1). Consequently, equation (V7) suggests that pshould be very small. Depending on the content of the list frame, $Cov(\hat{t}_a, \hat{t}_{ab}^{A})$ in equation (V6) may actually be positive because $\hat{t}_a$ and $\hat{t}_{ab}^{A}$ are based on the same sampled segments. This would make the optimal λ even smaller (note: $Cov(\hat{t}_b, \hat{t}_{ab}^{B})$ in (V6) must be zero because $\hat{t}_b$ is always zero).

A little renaming to simplify the notation: let $\hat{t}_L = \hat{t}_{ab(0)} = \hat{t}_{ab}^{B}$, since it is the Horvitz-Thompson estimator from the list frame of domain ab (the overlap domain); and let $\hat{t}_N = \hat{t}_a$, since it is the area-frame estimator for the nonoverlap domain (using either the closed, open, or weighted approach to element definition).

The estimator for the total, $\hat{t} = \hat{t}_L + \hat{t}_N$, is called the *screening multiple-frame* estimator because elements in the area sample are "screened" and only those in the non-overlap domain are enumerated. Its variance has the obvious form:

$$V(\hat{t}) = V(\hat{t}_L) + V(\hat{t}_N). \tag{V8}$$

Estimating $V(\hat{t}_L)$ is straightforward when, as is usually the case, $\hat{t}_L$ is based on a single stage stratified list sample. Estimating $V(\hat{t}_N)$, however, can be a bit more complicated.

Recall that a sample of segments was selected from the area frame. In many cases the design used is roughly equivalent to stratified simple random sampling (see Kott 1989). In practice, sampling fractions are so low within strata that the distinction between with and without replacement sampling can be ignored. Suppose there are H strata and n_h sampled segments within stratum h. Let Y_{hj+} be the sum of the sample-weighted values of all non-overlap elements in segment j of stratum h (i.e., $\sum w_i y_i$, where summation is over all non-overlap elements in segment j). Then

$$\hat{V}(\hat{t}_N) = \sum_{h=1}^{H} n_h/(n_h - 1) \{ \sum_{j=1}^{n_h} Y_{hj+}^{2} - [\sum_{j=1}^{n_h} Y_{hj+}]^2/n_h \} \tag{V9}$$

is an unbiased estimator for $V(\hat{t}_N)$ whenever finite population correction can be ignored.

Although equation (V9) appears straightforward, one needs to be aware that Y_{hj+} will be zero for many sampled segments either because there are no elements linked with the segment or because the elements linked with the segment are also on the list frame.

4 PROBLEMS IN SURVEYS WITH OVERLAPPING FRAMES

As noted in previous sections, an area frame provides complete coverage of the population, but it can lead to inefficient estimation when the population of interest has a skewed distribution. A list frame, by contrast, can provide the means for more efficient estimation, but is often incomplete.

When a list of establishments is used along with an area frame in a multiple-frame environment, the list need not be complete. It should, however, contain those large and rare establishments which are the bane of area-frame surveys.

4.1 List-Frame Issues

All establishments on a list frame must be completely identified by name, address, etc. Operations that are large, have complex management structures, and/or are scattered over different locations must be so identified. This is the identifiability assumption described in Section 2. In a two frame (list and area) sampling design one must be able to determine whether an establishment (or portion of an establishment) sampled from the area frame could also have been selected from the list frame; otherwise, intractable nonsampling errors will result. It must be realized, however, that the identifiability requirement can greatly increase the cost of list development.

When developing a list frame of establishments, more than one source of establishment names are often available. The survey designer must decide whether to use each list as a separate frame in a multiple-frame survey or to combine the lists prior to sampling and create a single list frame for sampling purposes.

Suppose the former approach is chosen with two list frames and a single area frame. The population can then be divided into four mutually exclusive domains: (1) establishment on neither list frame, (2) establishments on the first list frame but not the second, (3) establishments on the second list frame but not the first, and (4) establishments on both list frames.

One can see that the need to identify all domains when there are two or more list frames greatly complicates the survey and estimation process. Many statistical agencies have therefore decided that it is usually more practical to combine all lists and remove duplication prior to sampling. This can still be a significant undertaking requiring the use of record linkage methodologies that may be prone to errors.

4.2 Overlap Detection Issues

The reliance upon matching names on the list frame to sampled elements from the area frame to determine overlap complicates the survey process. Whether an establishment is surveyed via the list or area frame, it is necessary to determine the primary business name, other business names, and the names of individuals who are also associated with the business. Rules of association of individuals' names with establishments must be defined. For example, an area frame unit may contain the residence of Dr. X associated with the medical practice of Drs. X, Y, and Z. Should the

association of the establishment with an area-frame element be based on the residence of each name or on the location of the business itself? This will depend upon the counting rules used to determine the overlap between the frames. This example becomes more complicated if there is a Dr. T involved in the practice, but the list frame only identifies the practice involving Drs. X, Y, and Z. If counting rules are used that would allow Dr. T to report for the entire practice of Drs. X, Y, and Z, then it must be possible to identify the practice as overlap with the list frame. This situation occurs often in agriculture where an individual associated with a partnership can report for the entire partnership, but the name matching process fails to properly link it with a name on the list frame, resulting in an upward bias. A more detailed discussion of these problems appear in Vogel (1975). The survey instrument for both frames must be carefully designed to identify and link names with establishments so that the overlap domain can be properly determined.

Resources need to be available to re-interview "questionable links" so that the domain determination is correct. The domain determination has been shown by several studies to be the single largest source of non-sampling errors in multiple frame surveys (Nealon 1984).

4.3 Estimation Difficulties

A two-frame (list and area) sampling design will generally yield more efficient and robust estimators than an area-frame design by itself. Nevertheless, outliers can still occur when the list frame is not constructed carefully or is not up to date. An establishment missing from the list frame that is sampled in the nonoverlap domain of the area frame can have a much larger sampling weight than it would have had as a list sample (since the costs associated with an area frame are larger, its sampling fractions are usually smaller). Five such establishments in a 1992 U.S. Department of Agriculture survey accounted for six percent of the national estimate.

Although the desire to eliminate potential outliers provides a powerful argument for including as many establishment names as possible on the list frame, there is an equally compelling reason for limiting the list frame to only larger establishments. The smaller the list, the easier it is to check overlap. Moreover, the incremental costs of adding enumerated elements to the nonoverlap domain are relatively small since, a, the fixed cost of developing an area frame has already been expended, and, b, each element linked to a sampled area segment must be contacted to determine its overlap status whether or not it eventually needs to be enumerated.

The use of a multiple-frame design can make it more difficult to measure change over time. Ratio estimators are usually an efficient means of estimating change when a portion of the sample in the reference (denominator) time period remains in the sample in the comparison (numerator) period. Since establishments may move between the overlap and nonoverlap domains as their structures change or as the list frame is updated, the efficiency of ratio estimators of change in multiple-frame surveys is reduced. Extreme caution must also be exercised in multiple-frame designs to ensure that establishments found through the area-frame sample are not added to the list frame during the duration of the design since that can bias estimation (by effectively changing selection probabilities).

Such additions to the list frame should only be allowed when an entirely new sample is selected from both frames. The independence of sample selection between the area and list frames must be maintained for the area frame to properly estimate for the incompleteness of the list frame (see Vogel and Rockwell 1977).

5 THE FUTURE OF MULTIPLE-FRAME SURVEYS

Because of the cost of maintaining an area frame, Canada models the incompleteness of its list frame of retail and wholesale establishments rather than measure list undercoverage with an area sample (Sande 1986). The U.S. Census Bureau may soon follow suit and abandon the area sample component of its Monthly Retail Trade Survey (Konschnik et al. 1991).

This pattern could be reversed in Italy, however, where lists of business establishments are viewed as unreliable. Petrucci and Pratesi (1993) discuss the possibility of introducing multiple-frame business surveys in that country. Moreover, the future use of area frames both by themselves and in a multiple-frame environment is more secure for surveys of agriculture and construction.

REFERENCES

Armstrong, B. (1978), "Test of Multiple Frame Sampling Techniques for Agricultural Surveys: New Brunswick," Survey Methodology, **5**, pp. 178-184.

Bankier, M.D. (1986), "Estimators Based on Several Stratified Samples With Applications To Multiple Frame Surveys," Journal of the American Statistical Association, **81**, pp. 1074-1079.

Bosecker, R.R. and Ford, B.L (1976), "Multiple Frame Estimation With Stratified Overlap Domain," Proceedings of the Social Statistics Section, American Statistical Association, 1976, Part I, pp. 219-224.

Bush, J. and House. C. (1993), "The Area Frame: A Sampling Base for Establishment Surveys," Proceedings of the International Conference on Establishment Surveys," American Statistical Association, pp. 335-354.

Chapman, D.W. (1993), "Cluster Sampling for Personal-Visit Establishment Surveys," Proceedings of the International Conference on Establishment Surveys, American Statistical Association, pp. 645-650.

Chu, A. (1987), "A Proof that the Assignment of Conditional Weights Will Produce Unbiased Estimates" in Weighting Procedures for CBECS III, unpublished report, Rockville, Maryland: Westat.

Cochran J.G. (1977), Sampling Techniques, New York: Wiley.

Cochran, R.S. (1965), Theory and Applications of Multiple Frames Surveys, unpublished Ph.D. dissertation, Ames, Iowa: Iowa State University.

Energy Infromation Administration (1989), Commercial Buildings Characteristics 1989, Washington, DC: Energy Information Administration.

Faulkenberry, G.D. and Garoui, A. (1991), "Estimation a Population Total Using an Area Frame, Journal of the American Statistical Association, **86**, pp. 445-449.

Fecso, R., Tortora R.D., and Vogel, F.A. (1986), "Sampling Frames for Agriculture in the United States, Journal of Official Statistics, **2**, pp. 279-292.

Fuller, W.A. and Burnmeister, L.F. (1972), "Estimates for Samples Selected From Two Overlapping Frames," Proceedings of the Social Statistics Section, American Statistical Association, pp. 245-249.

Gallego, F.J. and Delincé, J. (1993), "Sampling Frames Using Area Samples in Europe,"

Proceedings of the International Conference on Establishment Survey, American Statistical Association, pp. 651-656.

Goldberg, M.L. and Gargiullo, P.M. (1988), "Variance Estimation Using Pseudostrata for a List-Supplemented Area Probability Sample, Proceedings of the Survey Research Section, American Statistical Association, pp. 479-484.

Hansen, M.H., Hurwitz, W.N., and Madow, W.A. (1953), Sample Survey Methods and Theory, Volume 1, pp. 516-556, New York: Wiley.

Hartley, H.O. (1962), "Multiple frame surveys," Proceedings of the Social Statistics Section, American Statistical Association, pp. 203-206.

Hartley, H.O. (1974), "Multiple frame methodology and selected applications," Sankhyá, Ser. C, **36**, pp. 99-118.

Houseman, E (1975), Area Frame Sampling in Agriculture, SRS Report No. 20, Washington, DC: Statistical Reporting Service, U.S.D.A.

Jessen R.J. (1942), Statistical Investigation of a Sample Survey for Obtaining Farm Facts, Research Bulletin 304, Ames, Iowa: Iowa State College of Agriculture and Mechanical Arts.

Julien, C. and Maranda, F. (1990), "Sample Design of the 1988 National Farm Survey," Survey Methodology, **16**, pp. 116-129.

King, A.J. and Jessen , R.J. (1945), "The Master Sample of Agriculture, Journal of the American Statistical Association, **40**, pp. 38-56.

Konschnik, C.A., King, C.S., Dahl, S.A. (1991), "Reassessment of the Use of an Area Sample for the Monthly Retail Trade Survey," Proceedings of the Survey Research Section, American Statistical Association, pp. 208-213.

Kott, P.S. (1989), "The Variance of Direct Expansions from a Common Area Sampling Design," Journal of Official Statistics, **5**, pp. 183-186.

Kott, P.S. (1990), "Variance Estimation when a First Phase Area Sample is Restratified," Survey Methodology, **16**, pp. 99-104.

Lund, R.E. (1968), "Estimators in Multiple Frame Surveys," Proceedings of the Social Statistics Section, American Statistical Association, pp. 282-288.

Nealon, J.P. (1984), Review of the Multiple and Area Frame Estimators, SF&SRB Staff Report 80, Washington, DC: Statistical Reporting Service, U.S.D.A.

Nealon. J.P. and Cotter. J. (1987), Area Frame Design for Agricultural Surveys, Washington, DC: National Agricultural Statistics Service.

Petrucci, A. and Pratesi, M. "Listing Frames and Maps in Area Sampling Surveys on Establishments and Firms (1993)," Proceedings of the International Conference on Establishment Survey, American Statistical Association, pp. 554-559.

Rao, J.N.K. (1985), "Conditional Inference in Survey Sampling," Survey Methodology, Vol. **11**, pp. 15-31.

Rao, J.N.K. and Skinner, C.J. (1993), "Estimation in Dual Frame Surveys With Complex Designs," unpublished paper.

Sande, I. (1986), "Business Survey Redesign Project: Frame Options for the Monthly Wholesale Retail Trade Survey Redesign," unpublished memorandum, Ottawa, Canada: Statistics Canada.

Särndal, C.E. and Swensson, B. (1987), "A General View of Estimation for Two Phases of Selection with Application to Two-phase Sampling and Nonresponse," International Statistical Review, **55**, 1987, pp. 279-294.

Särndal, C.E., Swensson, B., and Wretman, J. (1991), Model Assisted Survey Sampling, New York: Springer-Verlag.

Scott, A.J., Smith, T.M.F. And Jones, R.G. (1977), "The Application of Time Series Methods to the Analysis of Repeated Surveys," International Statistical Review, **45**, 13-28.

Shah, B.V., Barnwell, B.G., Hunt, P.H., and LaVange, L.M. (1991), SUDAAN™ User's Manual, Research Triangle Park, NC: Research Triangle Institute.

Skinner, C.J. (1991), "On the Efficiency of Raking Ratio Estimation for Multiple Frame Surveys," Journal of the American Statistical Association, **86**, pp. 779-784.

Vogel, F.A. (1973), "An Application of a Two-stage Multiple Frame Sampling Design," Proceedings of the Social Statistics Section, American Statistical Association, pp. 617-622.

Vogel, F.A. (1975), "Surveys with Overlapping Frames - Problems in Application," Proceedings of the Social Statistics Section, American Statistical Association, pp. 694-699.

Vogel, F.A., and Rockwell, D. (1977), Fiddling with Area Frame Information in List Development and Maintenance, SF Staff Report 77-01, Washington, DC: Statistical Reporting Service, U.S.D.A.

Homework for Lesson 13

NASS first estimated values for t_L and t_N for the 2002 Census of Agriculture using a list of farms (which produced the estimate for t_L) and an area-frame survey (with which t_N was estimated). When a collection of potential farms were later added to the Census list, this effectively increased t_L to t_L' while decreasing t_N to t_N' by an identical amount. Let $\hat{t}_L$ and $\hat{t}_N$ denote NASS's original estimates and t_A denote the number of late-added farms to t_L. Potential late adds that failed to respond to the census were treated as if they were never contacted in the first place. NASS computed

$t_L' = t_L + t_A$, and

$t_N' = t_N - t_A$.

a. Relate the variances of the NASS estimators, $\hat{t}_L'$ and $\hat{t}_N'$, to the variances of $\hat{t}_L$ and $\hat{t}_N$, assuming the latter two are known quantities, say, $V(\hat{t}_L)$ and $V(\hat{t}_N)$.

b. Explain how $\hat{t}_N'$ can be negative with NASS's methodology.

c. NASS used this approach in each state. Can you think of a modification that would not allow negative estimates for t_N' at the state-level but still (most likely) be unbiased at the US level?

Lesson 14: Multiple Regression and Calibration Weighting

Matrix Preliminaries

This lesson will be nearly impossible to follow without previous exposure to matrices and linear algebra. Let us briefly review the notation we will use and the properties we will need in this lesson.

Let **A** be a F x G matrix, and a_{fg} denote the value in the f'th row and g'th column of **A**. **A** = $\{a_{fg}\}$ is another way of writing **A**.

A 1 x G matrix is called a *row vector*. A H x 1 matric is called a *column vector*. Vectors are usually lower case, while other matrices are upper case. All matrices are written in bold face. Conventionally, most vectors are column vector. We will break that convention later in these notes and allow one key set of vectors be row vectors.

Matrix algebra is conducted with the following rules:

c**A** = $\{ca_{fg}\}$ when c is a scalar (real number).
AB = $\{\sum_{h=1}^{H} a_{fh}b_{hg}\}$, where **A** is F x H and **B** = $\{b_{hg}\}$ is H x G.
$\mathbf{I}_K = \{d_{fg}\}$, where d_{fg} is 1 when f = g and 0 otherwise, is the is the K X K the *identity matrix*.
$\mathbf{CI}_K$ = **C** when **C** is F x K; $\mathbf{I}_K\mathbf{D}$ = **D** when D is K x G.
If **C** and **D** are K x K and **CD** = $\mathbf{I}_K$, then **DC** = $\mathbf{I}_K$, and **C** can be denoted $\mathbf{D}^{-1}$ (**D** *inverse*).
If **A** = $\{a_{fg}\}$ is F x G, then **A'** (**A** *transposed*) = $\{a_{gf}\}$ is G x F.
(**AB**)' = **B'A'**; $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$ if **A** and **B** are invertible (i.e, $\mathbf{A}^{-1}$ and $\mathbf{B}^{-1}$ exist).

The subscript K is often dropped from $\mathbf{I}_K$ for convenience. When working something out yourself, however, it is often helpful only to keep the subscript on $\mathbf{I}_K$, but to write the dimensions below each matrix. When an F x G, **A**, matrix is multiplied by a G x H matrix, **B**, the product, **AB**, is an F x H matrix. Moreover, **AB** only exists because the number of columns in **A** equals the number of rows in **B** (both are G).

We will also need to develop asymptotic properties of matrices. In this discussion when a quantity h is $O(n^{-a})$, we will use the shortcut h = $O(n^{-a})$. Thus, when X is $O(n^{-a})$ and a > 0, we can write X = $O(n^{-a})$, and $(1 + X)^{-2} = 1 - 2X + O(n^{-2a}) = 1 + O(n^{-a})$ (see equation (S2)).

We say the matrix **A** is $\mathbf{O}_{FxG}(n^{-a})$ when $a_{fg} = O(n^{-a})$ for all FG components of **A** (usually the subscript on "**O**" denoting its dimensions is suppressed, but we will not do that here). For probability limits, we write **A** is $\mathbf{O}_{FxG,\,P}(n^{-a})$ when $a_{fg} = O_P(n^{-a})$ for all FG components of **A**.

Suppose **A** and **B** are F x G, and F and G are bounded as n grows arbitrarily large. When $b_{fg} = a_{fg}[1 + O(n^{-a})]$ for all f and g, $\mathbf{B} = [\mathbf{I}_F + \mathbf{O}_{FxF}(n^{-a})]\mathbf{A} = \mathbf{A}\,[\mathbf{I}_G + {}_{OgxG}(n^{-a})]$.
When $\mathbf{D} = \mathbf{O}_{KxK}(n^{-a})$ and K is bounded,

$$(\mathbf{I}_K + \mathbf{D})^{-1} = \mathbf{I}_K - \mathbf{D} + \mathbf{O}_{KxK}(n^{-2a}).$$

If **x** is a random column vector with mean μ, then $\mathbf{V}(\mathbf{x}) = \mathbf{E}(\epsilon\epsilon')$ where $\epsilon = \mathbf{x} - \mu$. Moreover, when $\mathbf{E}(\mathbf{x}|\mathbf{A}) = \mu_A$, $\mathbf{E}(\mathbf{A}\mathbf{x}|\mathbf{A}) = \mathbf{A}\mathbf{E}(\mathbf{x}|\mathbf{A}) = \mathbf{A}\mu_A$, and $\mathbf{V}(\mathbf{A}\mathbf{x}|\mathbf{A}) = \mathbf{A}\mathbf{V}(\mathbf{x}|\mathbf{A})\mathbf{A}' = \mathbf{A}\mathbf{E}(\epsilon_A\,\epsilon_A'|\mathbf{A})$ where $\epsilon_A = \mathbf{x} - \mu_A$.

The Linear Model

We will be interested in the following variant of the linear model:

$$y_i = \mathbf{x}_i\,\beta + \epsilon_i, \tag{S47}$$

where $i \in U$, $\mathbf{x}_i$ is a row vector with G components, β is an unknown column vector also with G components, and U contains N members. Furthermore, ϵ_i is a random variable such that

$$E(\epsilon_i\,|\mathbf{x}_1,\ldots,\mathbf{x}_N;\,Z_1,\ldots,Z_N) = E(\epsilon_i\epsilon_q\,|\mathbf{x}_1,\ldots,\mathbf{x}_N;\,Z_1,\ldots,Z_N) = 0, \quad \text{and}$$
$$E(\epsilon_i^2\,|\mathbf{x}_1,\ldots,\mathbf{x}_N;\,Z_1,\ldots,Z_N) = \sigma_i^2 < \infty \tag{S48}$$

for all i, q, $\in$ U, $i \neq q$. Recall $Z_i = 1$ when is in the sample, 0 otherwise. Observe that σ_i^2 *can* be a function of x_i.

The model in equation (S47) is sometimes called the *multiple regression model*. A degenerate special case is the ratio model in equation (S7) where G=1 and $\mathbf{x}_i$ collapses to a scalar, x_i. It is often helpful to think of this special case even when working in the more

general environment.

Most treatments of the (multivariate) linear model begin with a sample of n elements and express equation (S47) in matrix form as

$\mathbf{y} = \mathbf{X}\beta + \epsilon$,

where $\mathbf{y}$ is an n-component column vector with y_i in the i'th position, ϵ is defined analogously, and $\mathbf{X}$ is a matrix with $\mathbf{x}_i$ in the i'th row. In a survey sampling context, there are practical reasons to prefer the formulation in (S47), but having $\mathbf{x}_i$ expressed as a row vector is a nod to the conventional notation.

A linear (model) unbiased estimator for β based on all the elements in the population is:

$$\mathbf{b_c} = (\textstyle\sum_U c_i\mathbf{x}_i'\mathbf{x}_i)^{-1} \sum_U c_i\mathbf{x}_i'y_i \qquad \text{(S49)}$$

where $\mathbf{c} = \{c_i\}$ is any arbitrary vector of non-negative values such that $E(\epsilon_i | c_1, ..., c_N) = 0$. The c_i and $\mathbf{x}_i$ must be such that the inverse $(\sum_U c_i\mathbf{x}_i'\mathbf{x}_i)^{-1}$ exists. To see that $\mathbf{b_c}$ is unbiased, observe that

$$\begin{aligned}
\mathbf{E}(\mathbf{b_c}|\mathbf{x}_1,..., \mathbf{x}_N; c_1, ..., c_N) &= \mathbf{E}\{(\textstyle\sum_U c_i\mathbf{x}_i'\mathbf{x}_i)^{-1} \sum_U c_i\mathbf{x}_i'[\mathbf{x}_i\,\beta + \epsilon_i]\} \\
&= \mathbf{E}\{(\textstyle\sum_U c_i\mathbf{x}_i'\mathbf{x}_i)^{-1} \sum_U c_i\mathbf{x}_i'\mathbf{x}_i\,\beta + (\sum_U c_i\mathbf{x}_i'\mathbf{x}_i)^{-1} \sum_U c_i\mathbf{x}_i'\epsilon_i\} \\
&= \mathbf{E}(\beta) + (\textstyle\sum_U c_i\mathbf{x}_i'\mathbf{x}_i)^{-1} \sum_U c_i\mathbf{x}_i'E(\epsilon_i) = \beta.
\end{aligned}$$

On the right-hand side of the above derivation we have dropped the list of conditioning variables. This is a common practice. In fact, the $\mathbf{x}_i$, c_i, and later the Z_i, are commonly treated as constants with respect to the expectation operator. We adapt the practice of dropping the conditioning variables for the remainder of the lesson for convenience.

The Gauss-Markov theorem tells us the best linear unbiased estimator for β (best in the sense of having the smallest possible variance) is the version of $\mathbf{b_c}$ with $c_i \propto 1/\sigma_i^2$. When all the ϵ_i are identically distributed, the full-population ordinary-least-squares estimator for β, $\mathbf{b}_{OLS} = (\sum_U \mathbf{x}_i'\mathbf{x}_i)^{-1}(\sum_U \mathbf{x}_i'y_i)$ is the best linear unbiased for β. *Let us assume that is the case*

for now. This presupposes, of course, that $\sum_U \mathbf{x}_i'\mathbf{x}_i$ is invertible. For the remainder of this lesson, we assume for convenience that all the matrices that need to be are invertible.

Suppose we have single-phase random sample, S, of n elements where each element i has selection probability, π_i. Under mild conditions we assume to hold, one can show that

$$\hat{\mathbf{b}}_\pi = (\textstyle\sum_S [1/\pi_i]\mathbf{x}_i'\mathbf{x}_i)^{-1} \sum_S [1/\pi_i]\mathbf{x}_i'y_i = (\sum_U [Z_i/\pi_i]\mathbf{x}_i'\mathbf{x}_i)^{-1} \sum_U [Z_i/\pi_i]\mathbf{x}_i'y_i$$

is a design-consistent estimator for $\mathbf{b}_{OLS}$ with $(\hat{\mathbf{b}}_\pi - \mathbf{b}_{OLS}) = \mathbf{O}_{Gx1,P}(1/\sqrt{n})$. Although $\hat{\mathbf{b}}_\pi$ is model unbiased, another model-unbiased estimator for β based on the sample,

$$\hat{\mathbf{b}}_{OLS} = (\textstyle\sum_S \mathbf{x}_i'\mathbf{x}_i)^{-1} \sum_S \mathbf{x}_i'y_i.$$

has less variance, even if it is not necessarily a design-consistent estimator for $\mathbf{b}_{OLS}$.

The design consistency of $\hat{\mathbf{b}}_\pi$ and smaller model variance of $\hat{\mathbf{b}}_{OLS}$, however, are irrelevant if our goal is to estimate neither $\mathbf{b}_{OLS}$ nor β, but $t_y = \sum_U y_i$. This is a goal we turn toward now.

The Generalized Regression Estimator

Suppose we knew the value $\mathbf{b_c}$ for some set of c_i-values. With that knowledge, we could compute the following generalized difference estimator for t_y :

$$\hat{t}_{ygdiff} = \hat{t}_{yHT} + (\textstyle\sum_U \mathbf{x}_i\mathbf{b_c} - \sum_S (1/\pi_i)\mathbf{x}_i\mathbf{b_c}) = \hat{t}_{yHT} + (\mathbf{t_x} - \hat{\mathbf{t}}_{\mathbf{x}HT})\mathbf{b_c},$$

where $\hat{t}_{yHT} = \sum_S (1/\pi_i)y_i$, $\mathbf{t_x} = \sum_U \mathbf{x}_i$ and $\hat{\mathbf{t}}_{\mathbf{x}HT} = \sum_S (1/\pi_i)\mathbf{x}_i$. Note that $\mathbf{t_x}$ and $\hat{\mathbf{t}}_{xHT}$ are row vectors.

The estimator $\hat{t}_{ygdiff}$ is a generalization of the difference estimator on page 77 of *Sampling: Design and Analysis*. The scalar auxiliary variable x_i in *Sampling* is replaced by the scalar product of two vectors, $\mathbf{x}_i\mathbf{b_c}$. In addition, the sampling design is not necessarily SRS. This generalized difference estimator is design unbiased because $\mathbf{b_c}$ is a constant under randomization inference. The estimator is also model unbiased:

$$E_M(\hat{t}_{ygdiff} - t_y) = E_M[\hat{t}_{yHT} + (\mathbf{t_x} - \hat{\mathbf{t}}_{\mathbf{x}HT})\mathbf{b_c} - t_y]$$

$$= E_M[\textstyle\sum_S (1/\pi_i)(\mathbf{x}_i\,\beta + \epsilon_i) \;+\; (\mathbf{t_x} - \hat{\mathbf{t}}_{\mathbf{x}HT})\mathbf{b_c} - \sum_U (\mathbf{x}_i\,\beta + \epsilon_i)]$$

$$= \hat{\mathbf{t}}_{\mathbf{x}HT}\beta \;+\; (\mathbf{t_x} - \hat{\mathbf{t}}_{\mathbf{x}HT})\beta \;-\; \mathbf{t_x}\beta \;=\; 0.$$

Unfortunately, $\mathbf{b_c}$ is rarely ever known in practice. It can, however, be estimated from the sample. This leads us to the *generalized regression estimator* (GREG):

$$\hat{t}_{yrgreg} = \hat{t}_{yHT} + (\mathbf{t_x} - \hat{\mathbf{t}}_{\mathbf{x}HT})\hat{\mathbf{b}}_{\mathbf{c}}, \tag{S50}$$

where

$$\hat{\mathbf{b}}_{\mathbf{c}} = (\textstyle\sum_S [1/\pi_i]c_i\mathbf{x}_i'\mathbf{x}_i)^{-1}(\sum_S [1/\pi_i]c_i\mathbf{x}_i'y_i).$$

Because $(\sum_S [1/\pi_i]c_i\mathbf{x}_i'\mathbf{x}_i)^{-1}$ is a random variable from the randomization viewpoint, neither $\hat{\mathbf{b}}_{\mathbf{c}}$ nor $\hat{t}_{yr}$ is design unbiased. To get at their asymptotic properties, we assume that the population and sample design are such that $\sum_U y_i/N$ converges to a finite positive constant, $\bar{y}_\infty$, as n and N grow arbitrarily large. Similarly, $N^{-1}\sum_U c_i\mathbf{x}_i'y_i$ and $[N^{-1}(\sum_U c_i\mathbf{x}_i'\mathbf{x}_i)]^{-1}$ converge to a vector and matrix with all finite members. Furthermore,

$$N^{-1}(\textstyle\sum_U c_i\mathbf{x}_i'\mathbf{x}_i - \sum_S [1/\pi_i]c_i\mathbf{x}_i'\mathbf{x}_i) = \mathbf{O}_{GxG,P}(1/\sqrt{n}), \text{ and}$$

$$N^{-1}(\textstyle\sum_U c_i\mathbf{x}_i'y_i - \sum_S [1/\pi_i]c_i\mathbf{x}_i'y_i) = \mathbf{O}_{Gx1,P}(1/\sqrt{n}).$$

As a result, if G is bounded as n (and N) grows arbitrarily large, then

$$N^{-1}\textstyle\sum_S [1/\pi_i]c_i\mathbf{x}_i'\mathbf{x}_i \;=\; [\mathbf{I}_G + \mathbf{O}_{GxG,P}(1/\sqrt{n})]N^{-1}\sum_U c_i\mathbf{x}_i'\mathbf{x}_i, \text{ and}$$

$$N^{-1}\textstyle\sum_S [1/\pi_i]c_i\mathbf{x}_i'y_i \;=\; [\mathbf{I}_G + \mathbf{O}_{GxG,P}(1/\sqrt{n})]N^{-1}\sum_U c_i\mathbf{x}_i'y_i.$$

We can now see that

$$\hat{\mathbf{b}}_{\mathbf{c}} = (N^{-1} \textstyle\sum_S [1/\pi_i] c_i \mathbf{x}_i' \mathbf{x}_i)^{-1} (N^{-1} \textstyle\sum_S [1/\pi_i] c_i \mathbf{x}_i' y_i)$$

$$= ([\mathbf{I}_G + \mathbf{O}_{GxG,P}(1/\sqrt{n})] N^{-1} \textstyle\sum_U c_i \mathbf{x}_i' \mathbf{x}_i)^{-1} ([\mathbf{I}_G + \mathbf{O}_{GxG,P}(1/\sqrt{n})] N^{-1} \textstyle\sum_U c_i \mathbf{x}_i' y_i)$$

$$= (N^{-1} \textstyle\sum_U c_i \mathbf{x}_i' \mathbf{x}_i)^{-1} [\mathbf{I}_G + \mathbf{O}_{GxG,P}(1/\sqrt{n})]^2 N^{-1} \textstyle\sum_U c_i \mathbf{x}_i' y_i$$

$$= (\textstyle\sum_U c_i \mathbf{x}_i' \mathbf{x}_i)^{-1} [\mathbf{I}_G + \mathbf{O}_{GxG,P}(1/\sqrt{n})] \textstyle\sum_U c_i \mathbf{x}_i' y_i$$

$$= [\mathbf{I}_G + \mathbf{O}_{GxG,P}(1/\sqrt{n})]\, \mathbf{b}_{\mathbf{c}} = \mathbf{b}_{\mathbf{c}}[1 + O_P(1/\sqrt{n})].$$

All this heavy lifting allows us to conclude that under our large-sample assumptions, $\hat{\mathbf{b}}_{\mathbf{c}}$ is a design-consistent estimator for $\mathbf{b}_{\mathbf{c}}$. It is also a model-unbiased estimator for β since it has it has the same form as $\mathbf{b}_{\mathbf{c}}$ in equation (S49) with $c_i Z_i / \pi_i$ replacing c_i.

As for the GREG estimator itself,

$$\hat{t}_{ygreg} = \hat{t}_{yHT} + (\mathbf{t}_{\mathbf{x}} - \hat{\mathbf{t}}_{\mathbf{x}HT})\hat{\mathbf{b}}_{\mathbf{c}}$$

$$= \hat{t}_{yHT} + (\mathbf{t}_{\mathbf{x}} - \hat{\mathbf{t}}_{\mathbf{x}HT})\mathbf{b}_{\mathbf{c}} + (\mathbf{t}_{\mathbf{x}} - \hat{\mathbf{t}}_{\mathbf{x}HT})(\hat{\mathbf{b}}_{\mathbf{c}} - \mathbf{b}_{\mathbf{c}})$$

$$= \hat{t}_{ygdiff} + (\mathbf{t}_{\mathbf{x}} - \hat{\mathbf{t}}_{\mathbf{x}HT})(\hat{\mathbf{b}}_{\mathbf{c}} - \mathbf{b}_{\mathbf{c}}). \qquad \text{(S51)}$$

Subtracting t_y from both sides and then dividing by $t_y = N\,\bar{y}_\infty$ yields:

$$(\hat{t}_{ygreg} - t_y)/t_y = (\hat{t}_{ygdiff} - t_y)/t_y + \bar{y}_\infty^{-1} N^{-1} (\mathbf{t}_{\mathbf{x}} - \hat{\mathbf{t}}_{\mathbf{x}HT})(\hat{\mathbf{b}}_{\mathbf{c}} - \mathbf{b}_{\mathbf{c}}).$$

Under our assumptions the generalized difference estimator has a relative error of $O_P(1/\sqrt{n})$.

That dominates $\bar{y}_\infty^{-1} N^{-1} (\mathbf{t}_{\mathbf{x}} - \hat{\mathbf{t}}_{\mathbf{x}HT})(\hat{\mathbf{b}}_{\mathbf{c}} - \mathbf{b}_{\mathbf{c}})$, which is $\mathbf{O}_{1xG,P}(1/\sqrt{n})\mathbf{O}_{Gx1,P}(1/\sqrt{n}) = O_P(1/n)$. From this we can see that the GREG estimator has a relative bias of O(1/n) and a relative error of $O_P(1/\sqrt{n})$.

The GREG estimator in equation (S50) can be rewritten as $\hat{t}_{y_GREG} = \sum_S w_k y_k$, where

$$w_i = 1/\pi_i + (\sum_U \mathbf{x}_q - \sum_S [1/\pi_q]\mathbf{x}_q)(\sum_S [1/\pi_q]c_q\mathbf{x}_q'\mathbf{x}_q)^{-1}[1/\pi_i]c_i\mathbf{x}_i'$$

$$= 1/\pi_i \{1 + (\sum_U \mathbf{x}_q - \sum_S [1/\pi_q]\mathbf{x}_q)(\sum_S [1/\pi_q]c_q\mathbf{x}_q'\mathbf{x}_q)^{-1} c_i\mathbf{x}_i'\}. \quad \text{(S52)}$$

We depart in this lesson from the practice of setting $w_i = 1/\pi_i$. Here w_i is used in place of what was $\tilde{w}_i$ in previous lessons.

Strictly speaking, the w_k in equation (S52) are functions of the realized sample, S, and the c_i, but we suppress that in the notation for convenience. Observe that the w_k satisfy the *calibration equation*: $\sum_S w_k\mathbf{x}_k = \sum_U \mathbf{x}_k$, a term first used in this context by Deville and Särndal (1992). We will return to calibration later.

Some Examples of GREG Estimators

The most common example of the GREG estimator in equation (S50) is the ratio:

$$\hat{t}_{yr} = t_x [\sum_S (y_i/\pi_i) / \sum_S (x_i/\pi_i)],$$

Here $\mathbf{x}_i$ is a scalar, x_i, and $c_i = 1/x_i$, so that

$$\hat{\mathbf{b}}_c = (\sum_S [1/\pi_j]c_j\mathbf{x}_j'\mathbf{x}_j)^{-1}[1/\pi_i]c_i\mathbf{x}_i'y_i = (\sum_S [1/\pi_j]x_j)^{-1}\sum_S [1/\pi_i]y_i = (\hat{t}_{xHT})^{-1}\hat{t}_{yHT} = \hat{b}_\pi, \text{ and}$$

$$\hat{t}_{ygreg} = \hat{t}_{yHT} + (t_x - \hat{t}_{xHT})\hat{b}_\pi = \hat{t}_{yr}.$$

Another example is the poststratified ratio estimator. In that estimator, the population is divided into G mutually-exclusive model groups, and the components of the G-vector $\mathbf{x}_i$ for $i \in U^g$ are all zero, except for the g'th component which equals x_i. One effectively has a different ratio model, $y_i = x_i\beta_g + \epsilon_i$, in each group g.

Again, $c_i = 1/x_i$. As a result,

$\mathbf{t}_x = (t_{x1}, t_{x2}, ..., t_{xG})$, where $t_{xg} = \sum_{i \in U^g} x_i$,

$\sum_S [1/\pi_j] c_j \mathbf{x}_j' \mathbf{x}_j$ is a diagonal matrix with $\sum_{i \in U^g} [Z_i/\pi_j] x_j$ in the g' th diagonal, and

$\sum_S [1/\pi_j] c_j \mathbf{x}_j' y_j$ is a vector with $\sum_{i \in U^g} [Z_i/\pi_j] y_j$ in the g' th position.

So that

$\hat{\mathbf{b}}_\mathbf{c} = (\sum_S [1/\pi_j] c_j \mathbf{x}_j' \mathbf{x}_j)^{-1} [1/\pi_i] c_i \mathbf{x}_i' y_i = (\hat{b}_1, ..., \hat{b}_G)'$, where $\hat{b}_g = \sum_{i \in U^g} [Z_i/\pi_j] y_j / \sum_{i \in U^g} [Z_i/\pi_j] x_j$, and

$\hat{t}_{ygreg} = \hat{t}_{yHT} + (\mathbf{t}_\mathbf{x} - \hat{\mathbf{t}}_{\mathbf{x}HT}) \hat{\mathbf{b}}_\mathbf{c} = \sum^G t_{xg} \hat{b}_g = \hat{t}_{ypr}$.

This is the same as the poststratified ratio estimator in equation (S45) without the second phase of sampling (i.e., when $m_g = M_g$ for all g). When all the x_i are 1, then poststratified ratio estimator collapses into the poststratified estimator.

One can also write $\hat{t}_{ypr} = \mathbf{t}_\mathbf{x} \hat{\mathbf{b}}_\mathbf{c}$. In general, when all $c_i = 1/(\mathbf{x}_i \mathbf{h})$ for some vector **h** (in this case a column of 1's), the GREG can be put in *projection form*: $\hat{t}_{ygreg} = \mathbf{t}_\mathbf{x} \hat{\mathbf{b}}_\mathbf{c}$.

It is possible to have separate multiple regression models in each model group f (f = 1, ...F): $y_k = \mathbf{x}_{(f)k} \beta_f + \epsilon_k$, where $k \in U^f$ and $\mathbf{x}_{(f)k}$ has G_f components. In this context $\mathbf{x}_k$ in the GREG estimator has $G = G_1 + ... + G_F$ components, all of which are zero except for the G_f that have the values of $\mathbf{x}_{(f)k}$.

At the other extreme, a popular x_k in practice is $\mathbf{x}_k = (1, x_k)$. When $c_k = 1$ and $\mathbf{x}_k = (1, x_k)$, we have the *simple regression estimator*. Some use this term only when $\mathbf{x}_k = (1, x_k)$ *and* the sample design is SRS.

The Anticipated MSE of the GREG

Equation (S51) tells us that the mean squared error is asymptotically identical to the variance of the generalized difference estimator, $\hat{t}_{ydiff}$, defined with $\mathbf{b_c}$:

$$MSE(\hat{t}_{ygreg}) \approx V(\hat{t}_{ydiff}) = \sum_{i,q\in U} (d_i/\pi_i)(d_q/\pi_q)(\pi_{iq} - \pi_i\pi_q)$$

$$= \sum_U (d_i^2/\pi_i)(1 - \pi_i) + \sum_{\substack{i,q\in U \\ i\neq q}} (d_i/\pi_i)(d_q/\pi_q)(\pi_{iq} - \pi_i\pi_q),$$

where $d_i = y_i - \mathbf{x}_i\mathbf{b_c}$

$$= (\mathbf{x}_i\beta + \epsilon_i) - \mathbf{x}_i(\sum_U c_q\mathbf{x}_q'\mathbf{x}_q)^{-1}\sum_U c_q\mathbf{x}_q'(\mathbf{x}_q\beta + \epsilon_q)$$

$$= \epsilon_i - \mathbf{x}_i(\sum_U c_q\mathbf{x}_q'\mathbf{x}_q)^{-1}\sum_U c_q\mathbf{x}_q'\epsilon_q.$$

Under mild conditions we assume to hold $E_M(d_id_q) = E_M(\epsilon_i\epsilon_q) + O_P(1/n)$. In this context, i and q can be equal, although they need not be.

As a result of this and the model assumptions in equation (S48), the asymptotic anticipated MSE of $\hat{t}_{ygreg}$ is

$$E_M[MSE(\hat{t}_{ygreg})] \approx \sum_U (\sigma_i^2/\pi_i)(1 - \pi_i) \tag{S53}$$

for many sampling designs (e.g., those for which $\sum_{i,q\in U} (\pi_{iq} - \pi_i\pi_q)/(N^2\pi_i\pi_q)$ is $O(1/n)$ so that the right-hand side of equation (S53) dominates the excluded terms), including most of the designs we have studied in the class, even those with multiple stage and phases. The most common exceptions are sampling designs with systematic sampling from a predetermined list in the first phase of sample selection.

Although the lesson until this point has restricted its attention to single-phase samples of fixed size. It is a trivial matter to expand the analysis to include single-phase designs with random sample sizes, like Poisson sampling. It is also possible (but not as trivial) to extend the asymptotic framework to multi-stage designs, where n refers to the number of sampled psu's rather than elements. A GREG estimator for t_y is not only unbiased under the model in equation (S47), it is design consistent with a relative randomization

mean squared error of $O(1/n_{exp})$ under mild conditions we assume to hold.

Returning to a single-phase elements sample where the model assumptions in equation (S47) and (S48) apply, the right-hand side of equation (S53) does not depend on the choice for the c_i in $\mathbf{b_c}$ in equation (S49). There is no advantage, in terms of reducing the (asymptotic) anticipated MSE of the GREG, in setting $c_i \propto 1/\sigma_i^2$. Many require c_i be proportional to $1/\sigma_i^2$ in their definition of the GREG estimator. In light of what we have just seen, that seems inappropriate. Others have argued that for model-based reasons the choice for $\hat{\mathbf{b}}_{\mathbf{c}}$ in equation (S50) should be $\hat{\mathbf{b}}_{OLS}$ when the σ_i^2 are all identical. There is little justication for that either. If the regression estimator for t_y is forced to be design-consistent to protect against possible model failure, one model-unbiased estimator for β in equation (S50) seems as good as any other – as long as it converges to a finite vector.

Although there is no asymptotically optimal choice for the c_i in the GREG, there is an optimal choice for the selection probabilities given a fixed sample size (or a fixed expected sample size) under the model. It is $\pi_i \propto \sigma_i$. This is Brewer allocation discussed in Lesson 8 for the special case of the ratio model in equation (S7) with uncorrelated errors.

Prediction Theory

Richard Royall (1970, 1976) and others have argued that inference should be conditioned on the realized sample exclusively using a model or family of models. The goal of estimation in their view is to predict the y-values for the elements *not* in the sample. Faced with the model in equation (S47) and error structure in equation (S48), they would use the *prediction-theory estimator* (or, more precisely, the predictor):

$$\hat{t}_{ypred} = \sum_S y_i + \sum_R \mathbf{x}_i \left(\sum_S \mathbf{x}_k'\mathbf{x}_k/\sigma_k^2 \right)^{-1} \sum_S \mathbf{x}_k' y_k/\sigma_k^2, \qquad \text{(S54)}$$

where R is the set of elements in the population, but not in the sample (the remainder set). Observe that $E(\epsilon_i^2) = \sigma_i^2$ only has the be known – or hypothesized – up to a constant. Each

$y_i \in R$ is predicted by $\mathbf{x}_i\mathbf{b}_{BLU}$, where $\mathbf{b}_{BLU} = (\sum_S \mathbf{x}_k'\mathbf{x}_k/\sigma_k^2)^{-1} \sum_S \mathbf{x}_k'y_k/\sigma_k^2$ is the best linear unbiased estimator for β under the model in equation (S47).

When $\mathbf{x}_i$ is a scalar, x_i, and $\sigma_i^2 \propto x_i$, $\hat{t}_{ypred}$ collapses into what looks like the ratio estimator under a self-weighted sampling plan, $\hat{t}_{yr} = t_x (\sum_S y_i / \sum_S x_i)$. Although it looks like a GREG estimator, it is need not be. In fact, it is not hard to show that the most efficient sampling plan for a fixed sample size given the model in equation (S7) purposefully chooses the elements with the largest x-values.

Prediction theory recognizes the possibly that the model used to determine the estimator may be wrong. In particular, the specification of the σ_i^2 may be erroneous. Fortunately, getting the σ_i^2 wrong does not affect the model unbiasedness of $\hat{t}_{ypred}$, however. Consequently, it is prudent to estimate the model variance of $\hat{t}_{ypred}$ without specifying the σ_i^2.

A more troublesome case of specification error occurs when auxiliary variables are missing from the model. To deal with such a possibility, suppose a *working model* like that in equation (S7) has been used to determine the estimator, in this case the simple ratio. A more general model can then be employed to choose the sample. Royall and Herson (1973) showed that by choosing the sample so that $\sum_S x_i/n = \sum_U x_i/N$, which they called being *balanced on x*, the simple ratio estimator becomes unbiased under the more general model: $y_i = \alpha + \beta x_i + \epsilon_i$. This can be seen by noticing that balancing the sample on x renders the ratio estimator equal to the expansion estimator: $\hat{t}_y = (N/n) \sum_S y_i$. Royall and Herson point out that a statistician can choose a sample that is balanced on additional variables suggested by a more general model. They focus particularly on integer powers of x.

A practical way to find samples that are balanced or nearly balanced on particular variables is to use simple random sampling and examine the results. For simple random sampling $\sum_S z_i/n - \sum_U z_i/N = O_P(1/\sqrt{n})$ under mild conditions for any z-variable. Thus, this sampling plan returns asymptotically balanced samples for almost every conceivable auxiliary

variable. It may nonetheless be ineffective for particular variables and particular samples. Kott (1986) showed that the asymptotics can often be speeded up for all powers of x by drawing a systematic sample from a list ordered by the x-values. Recall, however, that such a sampling plan does not allow design-consistent estimation.

A Purely Randomization-Based Approach

Let us now take the opposite tack and concentrate fully on randomization mse. Consider Consider an estimator for t_y with the same form as the generalized-difference estimator $\hat{t}_{ygdiff} = \hat{t}_{yHT} + (t_{\mathbf{x}} - \hat{t}_{\mathbf{x}HT})\mathbf{b}$, but where the vector $\mathbf{b}$ is chosen to minimize the randomization variance of the estimator. It is not hard to show that this variance-minimizing value is

$\mathbf{b} = [\mathbf{V}(\hat{t}_{\mathbf{x}HT})]^{-1}\,\mathbf{Cov}(\hat{t}_{\mathbf{x}HT}, \hat{t}_{yHT})$, where

$\mathbf{V}(\hat{t}_{\mathbf{x}HT}) = \sum_{i,q\in U} \mathbf{x}_i'\mathbf{x}_q\,[(\pi_{iq} - \pi_i\pi_q)/(\pi_i\pi_q)]$, and $\mathrm{Cov}(\hat{t}_{\mathbf{x}HT}, \hat{t}_{yHT}) = \sum_{i,q\in U} \mathbf{x}_i'y_q\,[(\pi_{iq} - \pi_i\pi_q)/(\pi_i\pi_q)]$.

Sometimes it will be possible to calculate $\mathbf{V}(\hat{t}_{\mathbf{x}HT})$, but calculating $\mathbf{Cov}(\hat{t}_{\mathbf{x}HT}, \hat{t}_{yHT})$ usually requires knowledge of all the y-values in the population. This matrix can be estimated in a design-unbiased fashion by $\hat{\mathbf{Cov}}(\hat{t}_{\mathbf{x}HT}, \hat{t}_{yHT}) = \sum_{i,q\in S} \mathbf{x}_i'y_q\,[(\pi_{iq} - \pi_i\pi_q)/(\pi_i\pi_q\pi_{iq})]$, which is the Horvitz-Thompson formula applied to $\mathbf{Cov}(\hat{t}_{\mathbf{x}HT}, \hat{t}_{yHT})$. Computing $\hat{\mathbf{Cov}}(\hat{t}_{\mathbf{x}HT}, \hat{t}_{yHT})$ requires that all the π_{iq} be known and positive.

The following asymptotically optimal randomization-based estimator can be calculated in practice:

$$\hat{t}_{yrand} = \hat{t}_{yHT} + (t_{\mathbf{x}} - \hat{t}_{\mathbf{x}HT})\,[\hat{\mathbf{V}}(\hat{t}_{\mathbf{x}HT})]^{-1}\,\hat{\mathbf{Cov}}(\hat{t}_{\mathbf{x}HT}, \hat{t}_{yHT}), \tag{S55}$$

where $\hat{\mathbf{V}}(\hat{t}_{\mathbf{x}HT})$ is defined in the obvious way. This estimator is also model-unbiased. Under Poisson sampling, things simplify considerably because $\pi_{iq} = \pi_i\pi_q$ for $i \neq q$. One can easily see that $\hat{t}_{yrand}$ is the GREG estimator in equation (S50) with $c_i = (1 - \pi_i)/\pi_i$.

Calibration Estimators

Deville and Särndal (1992) introduced the idea of a calibration estimator for t_y. Such an estimator is of the form, $\hat{t}_{ycal} = \sum_S w_i y_i$, with a set of *calibration weights*, $\{w_k\}$, that minimize some loss function yet satisfy the calibration equation, $\sum_S w_i \mathbf{x}_i = \sum_U \mathbf{x}_i$. The loss function is chosen to assure that the calibration weights are close to the original sampling weights, the $1/\pi_i$.

Any estimator with the form of $\hat{t}_{ycal}$ having weights that satisfy the calibration equation is unbiased under the model in equation (S47). The prediction estimator in equation (S53) can be put in calibration form with $w_i = 1 + \sum_R \mathbf{x}_k (\sum_S \mathbf{x}_q'\mathbf{x}_q /\sigma_q^2)^{-1} \mathbf{x}_i'/\sigma_q^2$, yet these weights have no reason to be close to the original sampling weights and surely are not selected to minimize any loss function relating the w_i to the $1/\pi_i$. Chambers (1996) called the prediction-estimator weights, "case-based weights."

The definition of a calibration estimator we will use here expands on the form but not the intent of Deville and Särndal. We will say that $\hat{t}_{ycal}$ is a *calibration estimator* if the w_i as a group satisfy the calibration equation, and each $w_i = (1/\pi_i)[1 + O(1/\sqrt{n_{exp}})]$ under mild conditions which are assumed to hold.

We have seen that the GREG estimator in equation (S50) can be put in the form of a calibration estimator. Brewer (1994) showed that by choosing the c_i for each $i \in S$ so that $c_i = (1 - \pi_i)/(\mathbf{x}_i\mathbf{h})$ for some vector $\mathbf{h}$, the GREG estimator can be put in prediction form:

$$\hat{t}_{ygreg} = \sum_S y_i + \sum_R \mathbf{x}_i\mathbf{b_c},$$

where $\mathbf{b_c}$ retains its definition in equation (S50). For the ratio model,

$$\hat{t}_{yBrew} = \sum_S y_i + \sum_R x_k \{\sum_S (y_i [1-\pi_i]/\pi_i) / \sum_S (x_i [1-\pi_i]/\pi_i)\}$$

$$= \sum_S \{1 + (\sum_R x_k / \sum_S (x_k [1-\pi_k]/\pi_k)\} ([1-\pi_i]/\pi_i)y_i$$

Brewer called the process of computing the w_i by using his c_i *cosmetic calibration*. He argued GREG estimators that could be put in prediction form were less likely to have any

weights less than unity than GREG estimators in projection form. Brewer noted that when every calibration weight is at least 1, each element can be thought to at least represents itself in $\hat{t}_{ycal}$. In the case of the $\hat{t}_{yBrew}$ above, all $w_i = 1 + (\sum_R x_k / \sum_S (x_k [1-\pi_k] / \pi_k)$ are at least 1 as long as all the x_i are nonnegative and one x_k with $\pi_k < 1$ is positive.

Bankier (2002) made a similar claim about the purely randomization-based estimator, $\hat{t}_{yrand}$, in equation (S54). One major reason both of these approaches reduce the potential for a calibration weights to be less than unity is that when $\pi_i = 1$, $c_i = 0$ under each. As a result, w_i is forced to equal unity, rather than being allowed to fall beneath that value.

Deville and Särndal (1992) focus on calibration weights of the form:

$$w_i = (1/\pi_i) f(c_i \mathbf{x}_i \lambda), \tag{S56}$$

where f is a monotonic, twice-differentiable function with $f(0) = f'(0) = 1$, and λ is chosen so that the calibration equation holds. When $f(z) = 1 + z$,

$$\lambda' = (\sum_U \mathbf{x}_j - \sum_S [1/\pi_j] \mathbf{x}_j)(\sum_S [1/\pi_j] c_j \mathbf{x}_j' \mathbf{x}_j)^{-1},$$

and Deville and Särndal's calibration estimator is the GREG (see equation (52) and observe that $\lambda' \mathbf{x}_i' = \mathbf{x}_i \lambda$). For nonlinear f, Deville and Särnal propose a iterative search method for determining λ. See also, Singh and Mohl (1996) for more details on computation.

Brewer (1994) and others have proposed iterative routines to recompute calibration weights originally in linear form when any $w_i = (1/\pi_i)(1 + \mathbf{x}_i \lambda)$ is less than unity. An appealing alternative builds constraints, often of the form $L_c \leq w_i \pi_i \leq B_c$, into the f(.) function. Both Deville and Sarndal (1992) and Folsom and Singh (2000) discuss this approach.

Folsom and Singh and Kott (2004b) discuss using a non-linear calibration in equation (S56) to handle sample nonresponse and population undercoverage. Briefly, $1/f(\mathbf{x}_i \lambda)$ serves as an estimate of the quasi-random probability of response (or of coverage). In this context, f(0) and f'(0) need not be zero.

Raking was discussed in *Sampling: Design and Analysis* (pp. 169-171) as a means of adjusting for nonresponse. It turns out that raking is identical to finding a set of calibration weights (and a λ) such that $w_i = (1/\pi_i)\exp(\mathbf{x}_i\lambda)$ for an appropriately defined $\mathbf{x}_i$. The value for each component of $\mathbf{x}_i$ in raking is limited to be either 0 or a 1. This limitation may be true for raking, but it is not the case for nonlinear calibration in general, even when $f(\mathbf{x}_i\lambda) = \exp(\mathbf{x}_i\lambda)$.

Variance Estimation

Särndal et al. (1989) proposed this model-variance/design-mean-squared-error estimator for the GREG in equation (S50):

$$v_{SSW} = \sum_{i\in S}\sum_{q\in S}[(\pi_{iq} - \pi_i\pi_q)/\pi_{iq}](w_ir_i)(w_qr_q),$$

where $r_i = y_i - \mathbf{x}_i(\sum_S w_qc_q\mathbf{x}_q'\mathbf{x}_q)^{-1}\sum_S w_qc_q\mathbf{x}_q'y_q$. This estimator have desirable properties for many element samples designs under certain conditions. Deville and Sarndal argued that the same variance/MSE estimator can be used for any calibration estimator, based on weights satisfying equation (S56) (with $f(0) = f'(0) = 1$). This turns out not to be true when calibration is used to adjust for nonresponse or undercoverage (because neither $f(0)$ or $f'(0)$ need be zero), but that is beyond the scope of this course. Demanti and Rao (2004) and the discussion by Kott that follows have more on this.

Under a multi-stage sample, it may no longer be reasonable for elements from the same psu to have uncorrelated errors. Assuming elements from *different* psu's have uncorrelated errors and the first-stage selection probabilites are ignorably small, the following generalized of equation (S32) should have both good model-based and randomization-based properties:

$$\hat{V}_{WR}(\hat{t}_{ycal}) = \sum_{h=1}^{H}(n_h/[n_h - 1])\{\sum_{j\in S_h}(\sum_{i\in S_{hj}} w_ir_i)^2 - (\sum_{j\in S_h}\sum_{i\in S_{hj}} w_ir_i)^2/n_h\}, \qquad \text{(S57)}$$

where h denotes a first-stage stratum of psu's, n_h the number of sampled psu in stratum h, S_h the set of sampled psu's in h, and S_{hj} the set of subsampled elements from psu j of stratum h (there can be many stages of sampling involved). Recall that in equation (S32),

$w_i = 1/\pi_i$. The w_i in equation (S57) is equivalent to $(t_x / t_x)w_i$ in (S32).

From a model-based point of view, when $\sum_U \epsilon_i \approx 0$, the model variance of a calibration estimator is approximately

$$E_M[(\sum_S w_i y_i - \sum_U y_i)^2] = E_M[(\sum_S w_i \epsilon_i - \sum_U \epsilon_i)^2]$$

$$\approx E_M[(\sum_S w_i \epsilon_i)^2]$$

$$= \sum^H \sum_{j \in S_h} E_M(\sum_{i \in S_{hj}} w_i \epsilon_i)^2. \qquad \text{(S58)}$$

It is easy to see that $\hat{V}_{WR}(\hat{t}_{ycal})$ in equation (S57) is a model unbiased estimator for the right-hand side of equation (S58).

Means and Ratios

This lesson has focused on estimating the finite population total, $t_y = \sum_U y_i$. Often, we are interested in other finite-population statistics. In this section, we consider the estimation of a population ratio, $R = t_y / t_z$, where $t_z = \sum_U z_i$. A special case occurs when $z_k = 1$ for all $i \in U$, so that R is the mean of the y_k in the population U. For simplicity's sake, subscripts denoting the variables involved have not been placed on R. The same simplification will be made with the estimators for R.

It is not hard to show that under mild conditions, $\hat{R}_{HT} = \hat{t}_{yHT} / \hat{t}_{zHT}$ is a design-consistent estimator for R. Its randomization bias and mean squared error are both of asymptotic order $1/n_{exp}$, so that randomization bias is an asymptotically insignificant fraction of mean squared error.

These properties are shared by the calibrated estimator for R, $\hat{R}_{cal} = \sum_S w_k y_k / \sum_S w_k z_k$, where

the w_k belong to a set of calibration weights of the form $w_k = (1/\pi_k)[1 + O_P(1/\sqrt{n_{exp}})]$ that satisfy the calibration equation, $\sum_S w_i \mathbf{x}_i = \sum_U \mathbf{x}_i$. Rao *et al.* (2002) have noted in a somewhat different context that $\hat{t}_{ycal}$ and $\hat{t}_{zcal}$ can have smaller MSE's than their Horvitz-Thompson analogues without $\hat{R}_{cal}$ having a smaller MSE than $\hat{R}_{HT}$. It is possible that the y- and z-values tightly correlated and that the calibration of both the numerator of denominator in $\hat{R}_{cal}$ adds mostly unwanted noise to the ratio-estimation process.

Other Extensions

Extending jackknife and BRR variance/MSE estimation routines to GREG estimators follows closely the treatment of the ratio estimators in Lesson 11. As Deville and Sarndal argued, what applies to the GREG estimator applies to the analogous nonlinear calibration estimation.

Extensions to GREG and calibration principles to two- (and multi-) phase sampling can be very messy, since the set of auxiliaries can be different at different phases. Estevao and Särndal (2002) have a clear treatment of this complex topic.

There has been much recent work on applying local-polynomial-regression and spline-fitting techniques to sample-survey data. Two papers in this area are Breidt and Opsomer (2000) and Zheng and Little (2003).

One open question in using multiple-regression techniques within a randomization-based survey-sampling context is when the sample is large enough to justify the asymptotic theory. Kott (1994b) and Bell and McCaffrey (2002) describe using models to uncover and correct small-sample problems in some of the methods described in this lesson and throughout this course.

Homework for Lesson 14

Show that with a self-weighting sample, $\mathbf{x}_i$ equal to a scalar, x_i, and $c_i = 1/x_i$, Brewer's cosmetic calibration estimator for t_y collapses the usual ratio estimator.

Look at the regression estimator under SRS on page 74 of *Sampling Design and Analysis.* It is a GREG estimator. What is c_i? What is $\mathbf{x}_i$? What does this estimator look like in projection form? Why did we claim that the model-variance/design-mean squared-error estimator for this estimator is equation (S13) on page 27?

.

Additional References (not found in *Sampling: Design and Analysis* or for "Multiple Frame Establishment Surveys," which has its own set of references)

Asok, C. and Sukhatme, B.V. (1976). On Sampford's Procedure of Unequal Probability Sampling Without Replacement. *Journal of the American Statistical Association*, **71**, 912-918.

Bankier, M. (2002). 2001 Canadian census weighting: Switch from projections GREG to pseudo-optimal regression estimation. Presented at the International Conference on Recent Advances in Survey Sampling, Ottawa.

Basu, D. An essay on the logical foundations of survey sampling, part 1 (with discussion). In *Foundations of Statistical Inference*, V.P. Godambe and D.A. Sprott, Eds., Holt, Rinehart and Winston, Toronto, 203-242.

Bell, R.M. and McCaffrey, D.F. (2002). Bias reduction in standard errors for linear regression with multi-stage samples. Survey Methodology, **28**, 169-181.

Breidt, F.J. and Opsomer, J. D. (2000). Local polynomial regression in survey sampling. *Annals of Statistics*, **28**, 1026-1053.

Brewer, K.R.W. (1994). Survey sampling inference: some past perspectives and present prospects. *Pakistan Journal of Statistics*, **10(1)A** 213-233.

Chambers, R.L. (1996). Robust case weighting for multipurpose establishment surveys. *Journal of Official Statistics*, **12**, 3-52.

Cox, B.G., Binder, D.A., Chinnappa, B. N., Christianson, A., Colledge, M.J., and Kott, P.S., eds. (1995). *Business Survey Methods.* New York: Wiley.

Cumberland, W.G. and Royall, R.M. (1981). Prediction models and unequal probability sampling. *Journal of the Royal Statistical Society B*, 43, 353-367.

Demanti, A. and Rao, J.N.K. (2004). Linearization variance estimators for survey data (with discussion). Survey Methodology, 30, 17-34.

Deville, J-C. and Särndal, C-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, **87**, 376-382.

Estevao, V.M., and Särndal, C-E. (2000). A functional form approach to calibration. *Journal of Official Statistics*, **16**, 379-399.

Folsom., R.E. and Singh, A.C. (2000). The generalized exponential model for sampling weight calibration for extreme values, nonresponse, and poststratification. ASA *Proceedings of the Section on Survey Research Methods*, 598-603.

Fuller, W.A. (1976). *Introduction to Statistical Time Series*. John Wiley & Sons, New York.

Kim, J.K., Navarro, A., and Fuller, W.A. 2000. Variance estimation for 2000 census coverage estimates. *ASA Proceedings of the Section of Survey Research Methods*, 515-520 (forthcoming in *JASA*).

Holt, D. and Smith, T.M.F. (1994). Post-stratification. *Journal of the Royal Statistical Society A*, **142**, 37-44.

Godambe, V.P. (1955). A unified theory of sampling from finite populations. *Journal of the Royal Statistical Society B*, **17**, 269-278.

Godambe, V.P. and Joshi, V.M. (1965). Admissibility and Bayes estimation is sampling finite populations, 1. *Annals of Mathematical Statistics*, **36**, 1707-1722.

Goodman, R. and Kish, L. (1950). Controlled selection – a technique in probability sampling, *Journal of the American Statistical. Association*, **45**, 350-372.

Hartley, H.O. and Rao, J.N.K. (1962): Sampling with unequal probabilities and without replacement. *Annals of Mathematical Statistics*, **33**, 350-374.

Hájek J. (1971). Discussion of "An essay on the logical foundations of survey sampling, part 1." In *Foundations of Statistical Inference*, V.P. Godambe and D.A. Sprott, Eds., Holt, Rinehart and Winston, Toronto, 236.

Isaki, and Fuller, W.A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical. Association*, **77**, 89-96.

Kott, P.S. (1986). Some asymptotic results for the systematic and stratified sampling of a finite population. *Biometrika*, **73** , 485-91.

Kott, P.S. (1994a). Reweighting and variance estimation for the Characteristics of Business Owners Survey. *Journal of Official Statistics*, **10**, 407-418.

Kott, P.S. (1994b). Hypothesis testing of linear regression coefficients with survey data. *Survey Methodology*, **20**, 159-64.

Kott, P.S., (2001). The Delete-a-group jackknife, *Journal of Official Statistics*, **17**, 521-6.

Kott, P.S., (2004a). Delete-a-group variance estimation for the general regression estimator under Poisson sampling.

Kott, P.S., (2004b). Using calibration weighting to adjust for nonresponse and coverage errors. Presented at the Mini Meeting on Current Trends in Survey Sampling and Official Statistics, Calcutta, India. http://www.nass.usda.gov/research/reports/calcutta-for-web.pdf

Kott, P.S. (2005). Randomization-assisted model-based survey sampling. *Journal of Statistical Planning and Inference*, **129**, 236-277.

Kott, P.S. and Stukel, D.M. (1997). Can the jackknife be used with a two-phase sample? *Survey Methodology*, **23**, 81-89

Ohlsson, E. (1995). Coordination of samples using permanent random numbers. In *Business Survey Methods* by Cox, B. *et al.*

Rao, J.N.K. (1985). Conditional inference in survey sampling. *Survey Methodology*, **11**, 15-31.

Rao, J.N.K. (1994). Estimating totals and distribution functions using auxiliary information at the estimations stage. *Journal of Official Statistics*, **10**, 1-21.

Rao, J.N.K., Yung, W. and Hidiroglou, M.A. (2002). Estimating equations for the analysis of survey data using poststratification information. *Sankhyā*, **A64**, 1-15.

Royall, R.M. (1976). Current advances in sampling theory: Implications for human observational studies. *American Journal of Epidemiology*, **104**, 463-473.

Samford, M.R. (1967). On sampling without replacement with unequal probabilities of selection. *Biometrika*, **54**, 499-513.

Särndal, C-E, Swensson, B., and Wretman, J. (1989). The weighted residual technique for estimating the variance of the general regression estimator of a finite population total. *Biometrika*, **76**, 527-537.

Singh, A.C. and Mohl, C.A. (1996). Understanding calibration estimators in survey sampling. *Survey Methodology*, 22, 107-115.

Zheng, H. and Little, R.J.A. (2003). Penalized spline model-based estimation of the finite population total form probability-proportional-to-size samples. *Journal of Official Statistics*, **19**, 99-117.