Chapter 11: Sample Design Cross-Cutting Protocol

The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures

Created as part of subcontract with period of performance September 2011 – September 2016

This document was republished in September 2017 after a thorough review; no substantive changes were made. This supersedes the version originally published in April 2013.

M. Sami Khawaja, Josh Rushton, and Josh Keeling
The Cadmus Group
Portland, Oregon

NREL Technical Monitor: Charles Kurnik

NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Subcontract Report
NREL/SR-7A40-68567
September 2017

Contract No. DE-AC36-08GO28308

NREL
NATIONAL RENEWABLE ENERGY LABORATORY

Chapter 11: Sample Design Cross-Cutting Protocol

The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures

Created as part of subcontract with period of performance September 2011 – September 2016

This document was republished in September 2017 after a thorough review; no substantive changes were made. This supersedes the version originally published in April 2013.

M. Sami Khawaja, Josh Rushton, and Josh Keeling
The Cadmus Group
Portland, Oregon

NREL Technical Monitor: Charles Kurnik

Prepared under Subcontract No. LGJ-1-11965-01

NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

National Renewable Energy Laboratory
15013 Denver West Parkway
Golden, CO 80401
303-275-3000 • www.nrel.gov

Subcontract Report
NREL/SR-7A40-68567
September 2017

Contract No. DE-AC36-08GO28308

Disclaimer

These methods, processes, or best practices ("Practices") are provided by the National Renewable Energy Laboratory ("NREL"), which is operated by the Alliance for Sustainable Energy LLC ("Alliance") for the U.S. Department of Energy (the "DOE").

It is recognized that disclosure of these Practices is provided under the following conditions and warnings: (1) these Practices have been prepared for reference purposes only; (2) these Practices consist of or are based on estimates or assumptions made on a best-efforts basis, based upon present expectations; and (3) these Practices were prepared with existing information and are subject to change without notice.

The user understands that DOE/NREL/ALLIANCE are not obligated to provide the user with any support, consulting, training or assistance of any kind with regard to the use of the Practices or to provide the user with any updates, revisions or new versions thereof. DOE, NREL, and ALLIANCE do not guarantee or endorse any results generated by use of the Practices, and user is entirely responsible for the results and any reliance on the results or the Practices in general.

USER AGREES TO INDEMNIFY DOE/NREL/ALLIANCE AND ITS SUBSIDIARIES, AFFILIATES, OFFICERS, AGENTS, AND EMPLOYEES AGAINST ANY CLAIM OR DEMAND, INCLUDING REASONABLE ATTORNEYS' FEES, RELATED TO USER'S USE OF THE PRACTICES. THE PRACTICES ARE PROVIDED BY DOE/NREL/ALLIANCE "AS IS," AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING BUT NOT LIMITED TO THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL DOE/NREL/ALLIANCE BE LIABLE FOR ANY SPECIAL, INDIRECT OR CONSEQUENTIAL DAMAGES OR ANY DAMAGES WHATSOEVER, INCLUDING BUT NOT LIMITED TO CLAIMS ASSOCIATED WITH THE LOSS OF PROFITS, THAT MAY RESULT FROM AN ACTION IN CONTRACT, NEGLIGENCE OR OTHER TORTIOUS CLAIM THAT ARISES OUT OF OR IN CONNECTION WITH THE ACCESS, USE OR PERFORMANCE OF THE PRACTICES.

Preface

This document was developed for the U.S. Department of Energy Uniform Methods Project (UMP). The UMP provides model protocols for determining energy and demand savings that result from specific energy-efficiency measures implemented through state and utility programs. In most cases, the measure protocols are based on a particular option identified by the International Performance Verification and Measurement Protocol; however, this work provides a more detailed approach to implementing that option. Each chapter is written by technical experts in collaboration with their peers, reviewed by industry experts, and subject to public review and comment. The protocols are updated on an as-needed basis.

The UMP protocols can be used by utilities, program administrators, public utility commissions, evaluators, and other stakeholders for both program planning and evaluation.

To learn more about the UMP, visit the website, https://energy.gov/eere/about-us/ump-home, or download the UMP introduction document at http://www.nrel.gov/docs/fy17osti/68557.pdf.

Suggested Citation

Khawaja, M.S.; Rushton, J.; and Keeling, J. (2017). *Chapter 11: Sample Design Cross-Cutting Protocol, The Uniform Methods Project: Methods for Determining Energy-Efficiency Savings for Specific Measures*. Golden, CO; National Renewable Energy Laboratory. NREL/ SR-7A40-68567. http://www.nrel.gov/docs/fy17osti/68567.pdf

Acronyms

CI	confidence interval
CV	coefficient of variation
DOE	U.S. Department of Energy
ER	error ratio
FPC	finite population correction
GWh	gigawatt hour
HVAC	heating, ventilating, and air conditioning
kWh	kilowatt hour
M&V	measurement and verification
mMcf	million thousand cubic feet
NTG	net-to-gross
SAS	Statistical Analytics Software
SE	standard error
SRS	simple random sampling
UMP	Uniform Methods Project

Table of Contents

List of Figures

List of Tables

1 Introduction

Evaluating an energy efficiency program requires assessing the total energy and demand saved through all of the energy efficiency measures provided by the program. For large programs, the direct assessment of savings for each participant would be cost-prohibitive. Even if a program is small enough that a full census could be managed, such an undertaking would almost always be an inefficient use of evaluation resources.

A cost-effective alternative is to directly assess energy savings for a sample of the program population. However, when a study is based on a random sample rather than a full census, the outcomes of the study are influenced by the particular sample selected for direct evaluation. This random influence is called sampling error. Sampling error introduces an element of uncertainty to every sample-based estimate.

Determining reasonable estimates for quantities of interest is usually a straightforward arithmetic exercise, but quantifying the uncertainty behind such estimates is far more challenging. This document describes the broad principles that apply to all sample-based studies, and it provides specific guidance for applying the procedures most commonly needed in energy efficiency evaluations.

A significant challenge in energy efficiency evaluation is the lack of direct measurement. We can measure energy *consumption*, but energy *savings* is the difference between actual consumption and what consumption *would have been* had energy efficiency measures not been installed. Savings calculations combine consumption measurements with various adjustments to account for technical and behavioral baseline conditions.

Uncertainty can be introduced at every stage of the evaluation, including the sampling, measurement, and adjustment. It is often difficult or impossible to quantify the effect of every potential source of error. Evaluation reports often limit uncertainty discussions to random error (especially sampling error and regression error), because there are well-understood methods for quantifying uncertainty due to random errors. However, a high-quality evaluation should include strategies for mitigating all major sources of uncertainty, and a high-quality report should discuss unquantifiable aspects of uncertainty so research consumers can fully assess the research rigor.

The bulk of this chapter describes methods for minimizing and quantifying sampling error. Measurement error and regression error are discussed in various contexts in other chapters. A broader view of uncertainty is presented in Chapter 12: *Survey Design* and in this chapter's Appendix A.[1]

[1] As discussed in the section "Considering Resource Constraints" of the Introduction chapter to this report, small utilities (as defined under U.S. Small Business Administration regulations) may face additional constraints in undertaking this protocol. Therefore, alternative methodologies should be considered for such utilities.

1.1 Chapter Organization

The main body of this chapter provides a high-level discussion of the sample design and analysis principles that arise most often in evaluation work. Generally non-technical, this discussion is intended for a wide audience. A more technical, detailed account of important statistical concepts and methods is provided in the appendices.

- Section 2 reviews the statistical terms and concepts routinely encountered in evaluation work.
- Section 3 describes how complex evaluations are broken into components and how component-level research tasks are prioritized.
- Section 4 illustrates the evaluation process through several examples.
- Section 5 discusses validity threats and cost considerations.
- The appendices provide detailed descriptions of the statistical principles and methods that are referenced throughout this document.
    - Section 6: Appendix A discusses general sources and types of errors.
    - Section 7: Appendix B presents fundamental estimates and uncertainty calculations.
    - Section 8: Appendix C presents important sample designs and weighted estimates.

2 Overview

This section presents basic sampling concepts and terminology.

2.1 Sampling and Sample Design

The target group to be studied is called the **population**, and each member of the population is associated with one or more **variables**. The population could be any group of interest, such as program participants, installed measures, or retrofitted sites. A variable can either be a descriptive attribute (such as building type or climate zone) or a numerical quantity (such as square footage, *ex ante* (claimed) savings, *ex post* (evaluated) savings, or air-conditioning tonnage). The primary research objective in a sample-based study is to estimate the population average or total of one or more variables (for example, the total energy and demand savings for all program participants).

Some variables are known through the program database (for example, *claimed* savings) for every member of the population. Other variables (especially *evaluated* savings) can only be obtained through primary data collection and direct estimation. Variables whose values are known for all members of the program population are called **auxiliary**.[2]

A **sample** is a subset of a population selected for direct assessment of one or more variables of interest. The **sample design** describes the exact method by which population members are selected for inclusion in the sample. Sample designs are often informed by auxiliary data such as *claimed* savings estimates or building square footage. **Sample analysis** is the process of estimating population averages or totals and then quantifying the uncertainty in these estimates. The sample analysis may use both sample data and population-level auxiliary data.

Every sample design specifies some element of randomness in the sample selection procedure, but the nature of this randomness varies from one design to the next. Randomization in the sample design forms the basis for calculations that quantify uncertainty in the final estimates, so uncertainty calculations directly depend on the sample design. To yield valid results, the sample analysis must account for the sample design. For example:

- In **simple random sampling** (SRS), each member of the population has probability n/N of being selected,[3] and each individual's inclusion in the sample is unaffected by the particular identities of other members in the sample. If a sample is selected via SRS, then the usual sample mean and standard error formula will yield valid results.
- In **stratified sampling**, auxiliary data are used to partition the population into distinct groups, or strata, and then SRS is performed within each group. In this case, stratum weights are needed to obtain valid analytical results.

[2] In the case of two-phase sampling (Section 8.7), auxiliary data are collected for a large sample through a phone survey or other low-cost interaction. A smaller sample is then selected from the large sample and subjected to intensive measurement and verification. In this case, auxiliary data are known only for the larger sample, but not the entire population.

[3] Here, n is the sample size and N is the population size.

2.2 Uncertainty and Efficiency

Sample design is typically approached with one of two goals:

1. ***To minimize estimator uncertainty, given a fixed amount of study resources.*** In this case, time and budget are the primary constraints. For these projects, the goal is to design a sample that generates the most precise estimate within those constraints.
2. ***To minimize the resources needed to reduce uncertainty to some stated level.*** Often, the evaluation is required to meet a specified confidence-and-precision requirement (typically stipulated by a regulating body or forward-capacity market). In this case, the goal is to minimize time and cost subject to the constraint of meeting this target.

A design is **efficient** if it leads to minimal uncertainty for a fixed research budget. There are many strategies available for designing an efficient study. Energy efficiency program evaluations commonly use one or more of these (in various combinations):

- SRS
- Stratified sampling
- Cluster/multi-stage sampling.

The final design should always be selected to minimize estimation error in light of all available information—including both what is learned through sampling and what is known in advance through auxiliary data. For example, when participant-level *claimed* saving estimates are available, the sample design and analysis plan should use this information to increase efficiency (typically through stratification and/or ratio estimation).

An **estimator** is the particular function (mathematical expression or equation) through which sample data are used to estimate a population quantity. In general, an estimate will not precisely equal its target (for example, the sample mean is unlikely to equal the population mean exactly). The difference between the two—the **sampling error**—can be statistically estimated and, to some degree, controlled through sample design.

Descriptive estimators—such as the mean and standard deviation—can be calculated for any data set. The **mean** is the arithmetic average of the values, while the **standard deviation** is a measure of the variability among observations in the data. In normally distributed data, about 68% of observations are within one standard deviation of the mean, and 95% are within two standard deviations. (Note that a large standard deviation indicates greater dispersion of individual observations about the mean.)

As previously mentioned, the exact value of an estimate depends on the particular sample drawn. Thus, if an entire evaluation were repeated multiple times with a different sample drawn each time, a different estimated value would result for each evaluation.

An estimator is **unbiased** if it tends to be centered at its target quantity. This means that if the entire evaluation (selecting a sample and calculating the estimate based on the sample) were repeated many times, the average of the resulting values would be very near the target population value. The **standard error** (SE) of an estimator quantifies the dispersion that would be observed

among these values.[4] The distinction between the standard deviation and the standard error is important. The standard deviation describes variability of the data, while the standard error describes variability of the estimator (for instance, the variability of the sample means obtained from repeated sampling).

For example, in measuring the capacity of a sample of 100 heating, ventilating, and air-conditioning (HVAC) units, the standard deviation for this sample was found to be 25% of the value of the mean capacity. Assuming a normal distribution, approximately 95% of HVAC units in the population should have a capacity within ±50% of the sample mean. However, the standard error is 2.5% of the sample mean ($25\%/\sqrt{100}$). Thus, if we drew repeated samples of 100 HVAC units, the sample means would be within 2.5% of the population mean approximately 95% of the time.

2.3 Confidence and Precision

When data are collected via SRS, the standard error of the sample mean equals the standard deviation of the data, divided by the square root of the sample size.[5] In general, the standard error increases as the standard deviation of the underlying data increases or the sample size decreases.

Statistical methods are available for calculating standard errors for a wide range of estimators. Once an estimator's standard error is known, it is a simple matter to express the estimator's uncertainty through, for example, a **confidence interval** (CI). A CI is a range of values that is believed—with some stated level of confidence—to contain the true population quantity. The **confidence level** is the probability that the interval actually contains the target quantity.

Precision provides convenient shorthand for expressing the interval believed to contain the estimator (for example, if the estimate is 530 kilowatt-hours [kWh], and the relative precision level is 10%, then the interval is 530 ±53 kWh).[6] In reporting estimates from a sample, it is essential to provide both the precision and its corresponding confidence level (typically 90% for energy efficiency evaluations).

For a given data set, an estimate's uncertainty can be expressed in precision terms at any level of confidence. To have higher confidence, it is necessary to take a wider interval, which results in less precision. In other words, when all else is held constant, there is a tradeoff between precision and confidence.[7] As a result, any statement of precision without a corresponding confidence level is incomplete and impossible to interpret. For example, assume the average savings among participants in an ENERGY STAR appliance program is estimated as 1,000 kWh per year, and

[4] This can be thought of as the standard deviation of the estimator itself, and it may account for multiple sources of random error, including sampling error.

5 This formulation ignores the finite population correction (FPC) (see "Sample Means with FPC" in Appendix C).

6 Note the counterintuitive implication of this standard definition. Low-precision values correspond to narrow intervals and, hence, describe tight estimates. This can lead to confusion when estimates are described as having "low precision."

7 Although there is a close relationship between confidence and precision, these terms are not direct complements of each other. If the confidence level is 90%, there is no reason that the precision needs to be 10%. It is just as logical to talk about 90/05 confidence and precision as 90/10.

the analyst determines this estimate to have 16% relative precision at the 90% confidence level. The same data set and the same formulas may be used to estimate 10% relative precision at the 70% confidence level. If the confidence level is not reported, the second formulation would appear to have less uncertainty when, in reality, the two are identical.

The estimators commonly used in energy efficiency evaluations generally have sampling errors that are approximately normal in distribution.[8] To calculate the bounds for such an estimator, first multiply the estimator's standard error by a *z*-value.[9] Then add this product to the estimate itself to obtain the CI upper bound, and subtract the product from the estimate to obtain the lower bound.

Note that the *z*-value depends only on the confidence level chosen for reporting results. That is, for a given estimate $\hat{x}$, the confidence interval is:[10]

$$\hat{x} - z \cdot \widehat{\mathrm{SE}}(\hat{x}) \leq x \leq \hat{x} + z \cdot \widehat{\mathrm{SE}}(\hat{x})$$

In this equation, a *z*-value of 1.645 is used for the 90% confidence level and a value of 1.960 is used for the 95% confidence level. (These values are tabulated in most statistics textbooks and can be calculated with a spreadsheet.) The absolute and relative precision at the selected confidence level is estimated as:

$$\text{Absolute Precision } (\hat{x}) \;=\; z \cdot \widehat{\mathrm{SE}}(\hat{x})$$

$$\text{Relative Precision } (\hat{x}) \;=\; \frac{z \cdot \widehat{\mathrm{SE}}(\hat{x})}{\hat{x}}$$

The standard error always has the same physical units as the estimator, so absolute precision always has the same physical units as the estimation target. Relative precision, however, is always unit-free and expressed as a percentage.[11]

---

[8] This means that if the entire evaluation (drawing a sample and calculating the estimator from the sample) were repeated many times, the resulting estimator values would roughly follow a normal distribution.

[9] If the sample size, *n*, is small, a *t*-value with *n*-1 degrees of freedom is more appropriate than a *z*-value, as *z*-values will lead to an overstatement of achieved precision. At the 90% confidence level, the choice of *t*- versus *z*-value makes little difference for sample sizes greater than 30. The TINV() function in Microsoft Excel can be used to calculate *t*-values.

[10] We have added a "hat" to the SE in this expression. This is to emphasize that any real-life CI would have to rely on a sample-based estimate of the standard error, because the true standard deviation of an estimator cannot be known without perfect knowledge of the population. Inferential statistics in practice substitutes the standard deviation of the sample for the standard deviation of the population. The uncertainty associated with this substitution is treated as negligible. This treatment is usually appropriate, but at very small sample sizes the uncertainties associated with this substitution may become more significant.
Also, strict notational correctness would require a lower case "se" in this equation instead of the "$\widehat{\mathrm{SE}}$." We appreciate the distinction, but do not believe that the failure to distinguish between a function and its generic instance will lead to any errors in practice.

[11] Absolute precision is most frequently applied when estimating quantities such as population proportions, which are themselves percentages. In such cases, the expression "… has 5% precision" is ambiguous. It is better to say

Example 1-1

If a program's average savings are estimated as 10.31 kWh and the standard error is calculated as 1.70 kWh, then we have 90% confidence that the true population mean lies within the interval:

$$\begin{aligned} 10.31\ \text{kWh} - 1.645 \cdot 1.70\ \text{kWh} &\leq \quad \text{average savings} \\ &\leq \ 10.31\ \text{kWh} + 1.645 \cdot 1.70\ \text{kWh} \end{aligned}$$

And the precision formulas are

$$\text{Absolute Precision } (\hat{x}) \ = \ 1.645 \cdot 1.70\ \text{kWh} \ = \ 2.80\ \text{kWh}$$

$$\text{Relative Precision } (\hat{x}) \ = \ \frac{2.80\ \text{kWh}}{10.31\ \text{kWh}} \ = \ 27.2\%$$

In other words, based on the selected sample, the best estimate of the true (unobserved) population mean is the sample mean (10.31 kWh). We are 90% confident that the true value is within 2.80 kWh or 27.2% of this estimate.

[End of Example]

If the estimated outcome is large relative to its standard error, the estimator will tend to have a small relative precision value at a given confidence level. (Small precision values are desirable.) However, if the amount of variability is large relative to the estimated outcome, the precision will be poor. For example, if the observed average savings is 1,000 kWh and the associated relative precision (at, say, 90% confidence) is 150%, then we are 90% confident that the true average savings is somewhere between negative 500 kWh (which means that the measure actually caused consumption to increase) and 2,500 kWh.

---

either "…has 5% absolute precision" or "… is precise to within five percentage points." (See *Estimating Population Proportions* in Appendix B.)

3 Complex Evaluations: Designing for Multiple Objectives

This section describes sample design and analysis procedures for the research tasks most commonly encountered in energy efficiency evaluations. Evaluations vary in size and complexity. The scope of a given study can be:

- A single program, encompassing several distinct measure groups
- A full portfolio, spanning multiple programs and sectors
- Some collection of measure groups of particular interest to a client.

In the material that follows, the term *study* refers to any of these possibilities. Also, this material mentions—but does not thoroughly discuss—several important statistical concepts; however, these are discussed in detail in *Appendix B. Fundamental Estimates and Uncertainty Calculations* and *Appendix C. Sample Design and Weighted Estimates*.

Most energy efficiency portfolios support a wide range of measures and serve multiple sectors. Complex portfolio evaluations generally include multiple precision requirements at different levels of aggregation. For example, a single evaluation may need to satisfy each of the following:

- Estimate savings to within 10% at the 90% confidence level for each sector (residential, commercial, government/nonprofit, industrial)
- Estimate savings to within 10% at the 90% confidence level for all nonresidential lighting projects combined
- Estimate savings to within 20% at the 90% confidence level for each program in the portfolio.

It would not be difficult to design an efficient study that meets any one of these requirements, but it is much more challenging to design an efficient study that meets all of the requirements simultaneously.

To design an efficient study, the researcher usually engages in some back-and-forth between high-level evaluation requirements and component-level study design details. In all cases, the study design must:

- Lead to valid and essentially unbiased estimates of the object(s) of study
- Meet prescribed confidence and precision targets through valid means
- Be cost-efficient.

The following general steps describe a simplified approach to sample design that relies—to some degree—on trial and error. This approach will lead to an effective and efficient research design for most evaluations. Section 4: *Worked Examples* provides examples illustrating the essential steps, and Appendices A and B give further examples and detailed technical guidance.

1. ***Describe the portfolio structure and the requirements for confidence and precision.*** A complex study may span multiple programs that cover different sectors and

technology groups (for example, custom versus prescriptive). Also, evaluators may be required to provide savings estimates at the study, sector, program, and measure levels.

Often the confidence and precision requirements are imposed through a regulatory process or forward capacity market standard. These values are most commonly set at 90% confidence and 10% precision at the portfolio or sector level, but requirements vary. The evaluator needs to understand which confidence and precision requirements apply to which levels. (That is, at what level—measure, program, sector, portfolio—are savings to be estimated with the stated confidence and precision?) In addition to regulatory precision requirements, clients often require disaggregated results at other levels of precision. A population segment for which an estimate must be reported is called a **reporting domain**.

2. ***Identify the basic sampling and analysis domains.*** At the highest level, the sampling groups usually reflect the structure of the reporting domains. For example, if sector-level savings need to be reported, then residential sampling and analysis will normally be independent of industrial and commercial evaluation activities.[12]

   The basic groups for sampling and analysis are called **domains of study**. There can be multiple evaluation tasks within a study domain. For example, HVAC and lighting savings both need to be evaluated within the commercial sector, but because these measures interact, their evaluation tasks may not be independent. However, each domain's analysis is essentially self-contained and independent of other domains. In the remaining steps, we assume the reporting domains are the same as the domains of study.[13]

3. ***Determine the appropriate stratification.*** The sample sizes and associated data collection costs are directly related to the amount of variability (usually measured with a coefficient of variation or error ratio) in the population. If unit-level savings vary greatly between domain subgroups (for example, measure groups or building types), divide the domain into more homogeneous subgroups (strata). This is called **stratification**. Stratification reduces the sample size needed to obtain a given domain-level precision. (It also allows the evaluator to ensure representation among various subgroups.)

   For example, if domains correspond to sectors, the commercial domain may include the following strata:

Small Retail Lighting	Medium Retail Lighting	Large Retail Lighting
Office Lighting	Office HVAC	Office Plug Load
Small Retail HVAC	Large Retail HVAC	Grocery Refrigeration
Grocery Lighting		

[12] There are exceptions. In some cases, the basic sampling/analysis groups cut across reporting domains, as when sampling and analysis are performed independently within sector-pooled technology groups.

[13] The general principles provided in the appendices remain valid for alternative approaches, but we do not provide step-by-step guidance for all possible approaches.

4. ***Determine the data requirements and estimation strategies within each domain.*** For each group (for example, prescriptive commercial program) or subgroup (for example, offices), use the program database to identify important measure categories (for example, lighting). Then, for each measure category, determine estimation procedures and data needs based on the prevailing measurement and verification (M&V) protocol.

5. ***Record* claimed *contribution,* ex post *uncertainty, and M&V costs for each stratum.*** For each stratum within a domain, determine total *claimed* savings. Based on the M&V protocols (Step 4), note the approximate evaluation cost-per-sample-unit within each measure category. When possible, also include an estimate of the uncertainty parameter (CV or error ratio [ER]) within each category.[14] Measures contributing significantly to total savings and exhibiting significant variability will receive highest levels of evaluation resources.[15] This will reduce the standard error and improve confidence intervals.

6. ***Estimate sample sizes within each domain.*** In the most straightforward cases, the previous step will yield reliable cost, uncertainty, and claimed total estimates. In such a case, implement the cost-weighted Neyman formula (*Appendix C. Sample Design and Weighted Estimates*) to obtain the domain's optimal sample allocation as a function of total sample size *n*. Adjust *n* to obtain an efficient domain-level sample allocation, which should meet the precision requirement.

   Sometimes there may be insufficient basis for estimating variation or the reporting requirements may be too complicated to permit a straightforward Neyman allocation. In such cases, the planning process may be simplified by prioritizing measure categories with high *claimed* totals and high uncertainty. The evaluator can then assign initial planning targets of, say, of 10% precision with 90% confidence for each high-priority category. For categories that are not high priority, choose more liberal targets (for instance, 90/20). (These targets may be revised in Step 7.) Sample sizes are then calculated using the formulas provided in *Appendix C. Sample Design and Weighted Estimates*.

7. ***Aggregate Precision to Reporting Requirement Level***. For each reporting level (such as the sector- and study-levels), calculate the expected precision based on the sample allocations obtained in Step 6. If the expected precision at some level falls short of its target, increase the sample sizes in lower-level groups until all precision expectations meet their targets.

   This step is difficult to optimize through a simple formula, but if the calculations in the previous step have been automated, then a gradient-descent algorithm may be used to identify categories that yield the greatest impact on higher-level precision per

[14] This may be based on previous studies' estimates of coefficient of variation. Otherwise, variability may assessed qualitatively (for example, low, medium, or high), based on the evaluator's judgment.
[15] There are, of course, other considerations. See Section 5, *Additional Considerations*, for further discussion.

evaluation dollar and to increase evaluation resources for these categories until higher-level precision estimates meet the evaluation targets.

In cases where a domain's sample allocation is based on evaluator-prioritized precision targets, these targets should be adjusted directly if higher-level precision estimates are significantly higher or lower than the evaluation targets.

8. ***Document the Assumptions and Sampling Plan.*** Document the sampling plan obtained through these steps. Include assumptions about data variability (CVs and ERs) and calculations showing that all precision targets will be met if the observed variability is no greater than what is assumed. At this point, the client and evaluator should agree on the measures to be taken, if any, to adjust sample sizes should early data collection provide evidence that variability assumptions are in error.

Appendix C. Sample Design and Weighted Estimates provides technical guidance about optimizing sample design components. However, the hands-on approach—in which the evaluator prioritizes measure categories and then assigns (and adjusts) precision requirements for each category—is very flexible and sufficient for many applications.

4 Worked Examples

Section 3 described the general procedure for planning a portfolio evaluation at a high level. This section illustrates the basic components of this procedure. The general approach is to begin with lower-level evaluation tasks and then show how these build to a portfolio-level evaluation plan. The discussion makes frequent use of the formulas described in appendices B and C.

4.1 Measure- and Site-Level Evaluation Planning

In most energy efficiency evaluations, populations are segmented by sector: residential, commercial, and industrial.[16] Residential populations tend to be large in number and homogeneous, while the commercial and industrial segments are often smaller and more heterogeneous. Two major considerations drive the sample planning for any measure-level evaluation task:

- The heterogeneity of the relevant population segment (especially with respect to equipment usage patterns)
- The segment's size (in terms of both the number of units in the population and the average savings per unit).

Evaluations in the residential sector often use many different estimators and a variety of data sources. For example, proportions may be estimated from telephone survey data, ratios may be estimated from site visit data, and means may be estimated from end-use metering data. Because residential populations tend to be relatively homogeneous, SRS is the most common sample design in this sector.

Commercial and industrial populations are composed of multiple subsectors (for example, retail, office, grocery, manufacturing, and food processing). Nonresidential portfolios generally offer both prescriptive and custom measures for these sectors. Because the population members vary greatly in size, the expected savings for each measure installation varies from site to site. For example, a convenience store may convert 20 T12 florescent lamps to T8s, but a large office may convert 500 lamps. A well-maintained program database, which would include site-level *claimed* savings estimates, is critical to the efficient evaluation of nonresidential savings. Stratified ratio estimation is a central evaluation tool for these sectors.

4.1.1 Telephone Surveys

Telephone surveys are one of the most common methods of primary data collection in residential evaluations. These surveys are rich sources of data from which a number of population characteristics may be estimated, such as attitudes and opinions, purchasing behaviors, and demographics. Most of the data collected are categorical and are used to estimate proportions (such as the proportion of customers satisfied with the program, or the proportion of customers who actually installed a measure recorded in the program database).

[16] This list is not exhaustive. Other possible segments include: low-income, agricultural, public/institutional, and transportation.

For attitudinal, demographic, and other questions used to inform process evaluation, the uncertainty of a proportion estimate is usually described in terms of absolute precision (see *Appendix B. Fundamental Estimates and Uncertainty Calculations*). Write $e_{\text{abs.}}$ for the absolute precision level. Then the sample size needed to achieve this degree of precision is calculated as:

$$n = \left(\frac{z}{e_{\text{abs.}}}\right)^2 \cdot p(1-p)$$

Here, z is the z-value for the corresponding level of confidence, and p is the true population proportion. The expression $p(1-p)$ obtains its maximum when $p = 0.5$, so an n computed with this value will obtain the desired precision in all cases.

Example 4-1

For part of a process evaluation of a residential energy-education program, a participant survey is used to estimate the proportion of participants who changed their thermostat setting due to the program. The utility wants the survey-based estimate to be within five percentage points (absolute) of the true population proportion, with 90% confidence. If we have no *a priori* knowledge of the true proportion, we use the value with $p = 0.5$ to plan our survey. Then the sample size is:

$$n = \left(\frac{1.645 \cdot 0.5}{0.05}\right)^2 \approx 270.6$$

Thus, a survey sample of 271 participants is needed to ensure the desired level of confidence and precision.

[End of Example]

Note that the finite population correction (FPC) is not used in this formula. The FPC is typically negligible in the residential sector, as program populations tend to be quite large compared to evaluation survey samples.

Telephone surveys may also be used for impact evaluation, but this application should be limited to measures for which:

- No special training is needed to specify the measure and determine that it is installed correctly (For example, energy-efficient showerheads and compact fluorescent lamps satisfy this requirement, but attic insulation does not, because a homeowner may not know the effective R-value of insulation and may not be able to assess installation quality.)
- Average measure savings is well known through other resources.

When these conditions are satisfied, the only information needed to estimate total measure savings is the number of measures installed, and this quantity can be estimated with phone survey data.

When survey-level results are being reported for an impact evaluation, the uncertainty of a proportion estimate is often reported in terms of relative precision. Write $e_{\text{rel.}}$ for the target relative precision level. Then the sample size needed to achieve this degree of precision is calculated as:

$$n = \left(\frac{z}{e_{\text{rel.}}}\right)^2 \cdot \frac{1-p}{p}$$

The expression $(1-p)/p$ does not have any maximum; it increases without bound as p decreases to zero. Thus, some *a priori* lower bound on plausible values for p is needed to calculate the necessary sample size.

If savings at the measure level are not directly reported, but are instead rolled into estimated savings at a higher level for reporting, then measure-level savings is treated as a stratum within the higher level for sample planning.

Example 4-2

Continuing the energy-education example, assume that (1) the results of the participant survey will be used to inform an impact evaluation and (2) average savings among individuals who adjust their thermostats is known through a previous study. Then to estimate program savings, estimate the proportion of participants who adjusted their thermostats.

Consider two possible circumstances:

a. The utility wants the survey-based estimate to be within 20% (relative) of the true population proportion, with 90% confidence. Based on an informal internal evaluation, the utility is confident that at least 40% of the participants have adjusted their thermostats.

   Using

   $$\frac{(\mathbf{1}-\boldsymbol{p})}{\boldsymbol{p}} \geq \frac{(\mathbf{1}-\mathbf{0.4})}{\mathbf{0.4}} = \mathbf{1.5},$$

   the sample size is calculated as:

   $$n = \left(\frac{1.645}{0.2}\right)^2 \cdot 1.5 \approx 101.5$$

   Thus, a survey sample of 102 participants is needed to ensure the desired level of confidence and precision.

b. The utility does not want results reported at the program level. Instead, estimated program savings are to be rolled into residential sector-level savings for reporting.

   Then this program will be treated as a stratum within the residential domain. Its sample size will be determined through a cost-weighted Neyman allocation applied to the residential sector.

   For this, we will need to record the number of program participants (*N*), the marginal cost of surveying a single participant (*c*), the average savings among participants who adjust

their thermostats (X), and an *a priori* estimate of the proportion of participants who adjust their thermostats (p_0).

The unit-level standard deviation used in the Neyman allocation is this:

$$s = X \cdot \sqrt{p_0 \cdot (1 - p_0)}$$

This stratum's share of the residential sample will be proportional to $N \cdot s/\sqrt{c}$.

[End of Example]

4.1.2 Verification Site Visits

Verification site visits can be conducted for parameters that are not easily measured by telephone surveys. Common examples are:

- Installation rates (for example, proportion of program-provided CFLs installed)
- Measure Coverage (for example, percent of insulation installed)
- End-use parameters (for example, efficiency rating or thermostat set point).

4.1.2.1 Installation Rates

If there is only one measure per household—as is often the case with water heat, HVAC, and certain appliance measures—then the estimate is a sample proportion, which is analyzed as illustrated in examples 4-1 and 4-2. Note, however, that the marginal cost of a site visit is higher than that of a phone survey, so all else being equal, measures requiring on-site verification will receive smaller shares of the domain-level sample than those requiring only phone surveys. Savings for measures that can have multiple installations at each household or that have measures that vary greatly between sites should be estimated using a mean- or ratio-based method.

Example 4-3

For the evaluation of a direct-mail program that sent three CFLs to each residence within a utility's service territory, assume that the average hours of use and average wattage of replaced lamps are reliably known through a previous study. Write X for the product of the average hours of use and the average difference between replaced lamps and program lamps.

Then the research focus is on estimating the number of program bulbs that have been installed. Each residence may have installed 0, 1, 2, or 3 program bulbs (or more if some customers give unwanted CFLs to friends or neighbors). A visited site's savings is estimated as X times the number of program bulbs installed at the site. Estimate the average number of installed program bulbs as a simple mean.

To plan this evaluation task, information is used from an earlier evaluation that found the number of program lamps installed at a site was 2.1 on average, with a standard deviation of 1.3.

Consider two possible circumstances:

a. The utility wants the total program savings to be estimated to within 20% (relative precision), with 90% confidence.

Using CV $= \mathbf{1.3/2.1 = 0.62}$, the sample size is calculated as:

$$n = \left(\frac{1.645}{0.2}\right)^2 \cdot (0.62)^2 \approx 25.9$$

Thus, a survey sample of 26 participants is needed to meet the precision target at the stated confidence level.

b. The utility does not want results reported at the program level. Instead, estimated program savings are to be rolled into residential sector-level savings for reporting.

Thus, the program will be treated as a stratum within the residential domain, and its sample size will be determined through a cost-weighted Neyman allocation applied to the residential sector.

For this, record the number of program participants (N), the marginal cost of visiting a single participant (c), the average savings per installed CFL (X), and the *a priori* estimate of the standard deviation of the number of installed lamps per residence (from the previous report, this is 1.3).

The unit-level standard deviation used in the Neyman allocation is $\boldsymbol{s = X \cdot 1.3}$, and the stratum's share of the residential sample should be proportional to $\boldsymbol{N \cdot s/\sqrt{c}}$.

[End of Example]

4.1.2.2 Measure Coverage

Some site visits are made to estimate the proportion of reported savings measures that were actually installed—for example, the proportion of rebated CFLs installed in a home, or the quality and quantity of installed attic insulation. In these cases, the estimation strategy is based on a ratio estimator rather than a proportion- or mean-based estimator (see *Appendix B. Fundamental Estimates and Uncertainty Calculations*).

When measure-level savings must be estimated with a prescribed level of precision and confidence, the sample size formula for the ratio estimator is:

$$n = \left(\frac{z}{e_{\text{rel.}}}\right)^2 \left(\frac{s^{(\text{ratio})}}{\bar{y}}\right)^2$$

Here, $e_{\text{rel.}}$ refers to relative precision and $s^{(\text{ratio})}$ is similar to the standard deviation, but it only captures deviations between *ex post* savings (y_i) and realization-rate-adjusted claimed savings (see *Appendix B. Fundamental Estimates and Uncertainty Calculations*).

When there is no measure-level precision target, the measure is treated as a stratum within sector-level savings. In this case, the measure's share of the sector-level sample should be proportional to

$$N \cdot s^{(\text{ratio})}/\sqrt{c}$$

Where N is the number of participants in the stratum, c is the marginal cost of collecting data for a single participant, and $s^{(\text{ratio})}$ is as above.

Example 4-4

A weatherization program rebates material costs for attic insulation. The program database records the R-value and quantity of rebated insulation for each participant and calculates participant-level *claimed* savings estimates from these data.

To evaluate the program, technicians will visit a sample of participating sites and record the effective R-value (taking into account both the nominal R-value and the installation quality) and the installed quantity. Based on the data collected, *ex post* savings will be estimated for each site, and program savings will be estimated using a ratio-based realization rate. Write x_i for the *claimed* savings of the i^{th} visited site and write y_i for the *ex post* savings. Then

$$\text{Realization Rate} = \frac{\sum_{\text{sample}} y_i}{\sum_{\text{sample}} x_i}$$

The total savings estimate is the realization rate multiplied by the population total of the *claimed* savings values.

In this example, the evaluator is planning the current study using results from the previous year's evaluation. The previous evaluation estimated a realization rate of 75% from a sample of 100 participants. This estimate achieved a relative precision of ±8% with 90% confidence.

Calculate the error ratio, $\text{ER} = s^{(\text{ratio})}/\bar{y}$, based on the values given in last year's report:

$$\frac{s^{(\text{ratio})}}{\bar{y}} = \frac{\sqrt{n} \cdot e_{\text{rel.}}}{z} = \frac{\sqrt{100} \cdot 8\%}{1.645} \approx 0.49$$

Consider two possible circumstances:

a. Program-level results are to be estimated to within 20% (relative precision), with 90% confidence. The sample size is then:

$$n = \left(\frac{z}{e_{\text{rel.}}}\right)^2 (\text{ER})^2 = \left(\frac{1.645}{0.20}\right)^2 (0.49)^2 \approx 16.2$$

Therefore, the evaluator should plan to visit 17 participants to meet the 90/20 target for the realization rate. Because total savings is estimated as the realization rate multiplied by the claimed total, the total savings has the same relative precision as the realization rate.

b. The utility does not want results reported at the program level. Instead, estimated program savings are to be rolled into the sector-level saving estimates for reporting.

Because the program will be treated as a stratum within the residential domain, its sample size will be determined through a cost-weighted Neyman allocation. For

this, record the number of program participants (N), the marginal cost of visiting a single participant (c), and the *a priori* estimate of the standard deviation of the quantity $s^{(\text{ratio})}$.

The stratum's share of the sector sample will be proportional to $\boldsymbol{N \cdot s^{(ratio)}/\sqrt{c}}$.

[End of Example]

4.1.2.3 End-Use Parameters

In some cases, the purpose of a site visit is to estimate the value of some end-use parameter, such as the number of linear feet of pipe wrap installed or the technical specifications of an HVAC system. If the program database contains participant-level ex ante information, then total measure savings should be estimated using a ratio estimator. Otherwise, the estimates must be based on the sample mean. In both cases, sample planning for the measure-level evaluation task proceeds as illustrated in the previous examples.

Example 4-5

A site visit is required to estimate the heating capacity of ductless mini-split installed air conditioners (AC) for which customers will receive (or have received) rebates from a residential HVAC program. Unlike the previous residential examples, this program is relatively small, having only 200 participants.

As this is the first evaluation of this program, there is no prior information on the target population. However, the regional technical resource manual refers to a metering study that determined the cooling capacity had a standard deviation of 5.4 kBtu/h. The program implementer assumed that the average mini-split installed AC had a capacity of 18 kBtu/h. Thus, the best estimate of the CV is this:

$$\text{CV} = \frac{s}{\bar{x}} = \frac{5.4}{18} = 0.3$$

To achieve measure-level results having 90% confidence and ±10% relative precision, calculate the initial and, subsequently, the final sample sizes (with finite population correction) as:

$$n_0 = \left(\frac{1.645}{0.10}\right)^2 \cdot (0.3)^2 \approx 24.4$$

$$n = \frac{24.4 \cdot 200}{24.4 + 200} \approx 21.7$$

Thus, visit 22 households to achieve the desired level of precision.

[End of Example]

4.1.3 End-Use Metering

In most cases, end-use metering data are used to estimate some site-specific parameter, such as the average daily hours of use or the average kilowatt (kW) draw. Meter-based estimates are then used to evaluate *evaluated* savings for each metered measure installation. Sampling for end-use

metering proceeds as outlined above, with ratio-based estimates used when there is meaningful ex ante information, and mean-based estimates used when no such information is available.

4.2 Domain-Level Evaluation Planning

Sample plans for various levels of reporting domains can be developed after measure-level evaluation tasks have been analyzed and documented, as above. These plans may be based purely on optimization calculations, or they may involve a more hands-on approach (see Step 7 in Section 3).

Example 4-6

For a commercial and industrial (C&I) custom program evaluation, the distribution of participants is shown in Table 1.

Table 1. Example C&I Program Details

Subsector	Participants	End Uses	Percent of Ex Ante Savings
Retail	80	Lighting	25%
Office	65	Lighting, HVAC, Appliances	21%
Restaurant	30	Lighting, Appliances	9%
School	13	Lighting, HVAC	12%
Light Manufacturing	11	Lighting, Motors	33%
Total	**199**	**Lighting, HVAC, Appliances, Motors**	**100%**

To estimate satisfaction with a lighting measure, the evaluator chose to draw a stratified sample. This sample needed to provide a program-level estimate with 10% absolute precision, at the 90% confidence level. Thus, the first step is to determine the overall sample size needed (which is done in the same way as an SRS is determined for a proportion).

$$n_0 = \left(\frac{1.645 \cdot 0.5}{0.10}\right)^2 \approx 67.7$$

$$n = \frac{67.7 \cdot 199}{67.7 + 199} \approx 50.5$$

The results show that calling a total of 51 businesses will achieve the desired level of precision.

To determine how to distribute the sample, use the Neyman allocation, assuming that the variation is proportional to savings. The subsector sample sizes are then calculated as:

$$n_{\text{retail}} = 50.5 \cdot \left(\frac{25\%}{25\% + 21\% + 9\% + 12\% + 33\%}\right) \approx 12.6$$

$$n_{\text{office}} = 50.5 \cdot \left(\frac{21\%}{25\% + 21\% + 9\% + 12\% + 33\%}\right) \approx 10.6$$

$$n_{\text{rest.}} = 50.5 \cdot \left(\frac{9\%}{25\% + 21\% + 9\% + 12\% + 33\%}\right) \approx 4.5$$

$$n_{\text{school}} = 50.5 \cdot \left(\frac{12\%}{25\% + 21\% + 9\% + 12\% + 33\%}\right) \approx 6.1$$

$$n_{\text{light mfg.}} = 50.5 \cdot \left(\frac{33\%}{25\% + 21\% + 9\% + 12\% + 33\%}\right) \approx 16.7$$

After rounding the values up to the nearest integer and accounting for the fact that there are only 11 sites in the light manufacturing sector, the final subsector sample sizes are 13, 11, 5, 7, and 11, for a total 47, which is slightly lower than the original 51.

[End of Example]

Example 4-7

To evaluate total savings for the C&I program described by Table 1, regulatory requirements stipulate that results must be within 10% relative precision at the 90% confidence level. Previous experience has shown that, typically, the overall realization rate is approximately 90%, with an ER of approximately 0.4, so the total sample size for the program is:

$$n_0 = \left(\frac{1.645}{0.1}\right)^2 (0.4)^2 = 43.3$$

$$n = \frac{43.3 \cdot 199}{43.3 + 199} \approx 35.6$$

Thus, the initial plan is to visit 36 sites. As before, distribute the sample using the Neyman allocation. There are no data on subsector-specific ERs or CVs, so assume variation within each sector is proportional to ex ante savings.[17] Then for sector h, the share of the sample will be proportional to:

$$\begin{aligned} N_h \cdot \frac{s_h}{\sqrt{c_h}} &\propto N_h \cdot \frac{[ex\ ante \text{ total for stratum } h]/N_h}{\sqrt{c_h}} \\ &= \frac{[ex\ ante \text{ total for stratum } h]}{\sqrt{c_h}} \\ &\propto \frac{[\text{stratum } h\text{'s percent of the } ex\ ante \text{ total}]}{\sqrt{c_h}} \end{aligned}$$

Also, evaluation costs differ among subsectors; engineers estimate the following hours are required to evaluate a site for each subsector:

[17] To be precise, assume that within each stratum, the standard deviation of savings is proportional to the stratum's claimed savings average. (If necessary, stratify by size in addition to building type.) For this reasoning, *standard deviation* can either have the usual definition, s, or the ratio version, $s^{(\text{ratio})}$. (See *Appendix C*.)

Table 2. Evaluation Times and *Claimed* Savings by Subsector

Subsector	Hours	Proportion of *Claimed* Savings	
Retail	2	25%	
Office	4	21%	
Restaurant	2	9%	
School	4	12%	
Light Manufacturing	8	33%	

Using these estimates as a proxy for cost, allocate sample sizes to each subsector using the cost-weighted Neyman allocation as follows:

$$n_{\text{retail}} = 35.6 \cdot \left(\frac{25\%/\sqrt{2}}{25\%/\sqrt{2} + 21\%/\sqrt{4} + 9\%/\sqrt{2} + 12\%/\sqrt{4} + 33\%/\sqrt{8}/} \right) \approx 5.2$$

$$n_{\text{office}} = 35.6 \cdot \left(\frac{21\%/\sqrt{4}}{25\%/\sqrt{2} + 21\%/\sqrt{4} + 9\%/\sqrt{2} + 12\%/\sqrt{4} + 33\%/\sqrt{8}} \right) \approx 7.4$$

$$n_{\text{rest.}} = 35.6 \cdot \left(\frac{9\%/\sqrt{2}}{25\%/\sqrt{2} + 21\%/\sqrt{4} + 9\%/\sqrt{2} + 12\%/\sqrt{4} + 33\%/\sqrt{8}} \right) \approx 5.2$$

$$n_{\text{school}} = 35.6 \cdot \left(\frac{12\%/\sqrt{4}}{25\%/\sqrt{2} + 21\%/\sqrt{4} + 9\%/\sqrt{2} + 12\%/\sqrt{4} + 33\%/\sqrt{8}} \right) \approx 7.4$$

$$n_{\text{light mfg.}} = 35.6 \cdot \left(\frac{33\%/\sqrt{8}}{25\%/\sqrt{2} + 21\%/\sqrt{4} + 9\%/\sqrt{2} + 12\%/\sqrt{4} + 33\%/\sqrt{8}} \right) \approx 10.4$$

After rounding the values up to the nearest integer, the final subsector sample sizes are 6, 8, 6, 8, and 11, for a total 39. This represents the allocation that optimizes the balance between precision and cost.

4.3 Portfolio-Level Evaluation Planning

This section illustrates the planning process outlined in Section 1.3 through an extended example of an energy efficiency portfolio evaluation. The utility promotes efficiency measures in the residential, institutional (government and nonprofit), commercial, and industrial sectors. Table 3 shows program sizes.

Table 3. Claimed Savings by Sector

Sector	*Claimed* kWh Total
Residential	2,900,000
Institutional	2,200,000
Commercial	3,300,000
Industrial	3,000,000
Total	**11,400,000**

This evaluation entails estimating total savings to within 10% for each sector and to within 5% for the entire portfolio (all precision values assume 90% confidence). Sampling and analysis are to be performed separately within each sector (thus, data collected in the commercial sector has no bearing on estimates related to the industrial sector).

- Steps 1 and 2 are immediate: Report the savings for each of the four sectors, and the sectors are the domains of study.
- For Step 3, stratify each domain by measure group and size.
- For Step 4, examine the program database to determine the specific measures and measure groups that contribute to savings within each sector.

Table 4 shows savings by measure category for the residential program.

Table 4. Residential Program Data

Measure Group	*Claimed* kWh
Lighting	1,800,000
HVAC	600,000
ENERGY STAR Appliances	500,000
Total	**2,900,000**

This utility recently completed a study of ENERGY STAR appliances, so deemed values are considered acceptable for that program, so long as installation rates are directly evaluated. Then telephone surveys will provide acceptable data, and a proportion estimator will be appropriate for estimating savings. Stratification may also be appropriate if there are distinct participant groups for which installation rates may vary.

After reviewing the M&V protocols, the evaluator determines that (1) usage loggers are needed for evaluating savings from lighting measures and (2) interval metering is needed for evaluating HVAC savings. The final verified savings for both measure types will be determined through engineering calculations. After calculating savings for measures in the sample, ratio estimators will be used to evaluate total program savings for both measure groups.

For Step 5, consider the data to be used in the savings calculations to (1) determine average M&V costs for sampled units within each measure category and (2) anticipate variability within each group. (This process was illustrated in Section 4.1: *Measure- and Site-Level Evaluation Planning*.) Then use the cost-optimized allocation formula to determine the sample fraction for each group (Step 6). The results are summarized in Table 5.

Table 5. Cost, Variability, and Sample Fractions for Residential Sector

Measure Group	Evaluation Cost per Unit	Anticipated Variability	Average Claimed kWh	*Claimed* Standard Deviation	Sample Fraction
Lighting	$2,000	0.4 (ER)	200	80	48.3%
HVAC	$2,500	0.6 (ER)	2,400	1,440	21.6%
ES Appliances	$100	0.2 (CV)	250	50	30.0%

In Table 5, variability entries are based on experience with similar evaluation tasks. Average *claimed* values are based on program data and the standard deviations are the products of average savings and the error ratios or coefficients of variation. The sample fractions are calculated using the formula from *Planning and Optimizing Stratified Designs* (Appendix C).

Continuing Step 6, use the standard error formulas to determine the standard error for estimated total savings as a function of sample size. After some experimentation, the evaluator determines a residential sample allocation that should yield the 90/10 target for the sector. In Table 6, measure-level standard errors are based on estimator-specific standard error formulas. The total standard error is the square root of the sum of squared measure-level standard errors.

Table 6. Preliminary Sample Allocation for Residential Sector

Measure Group	Claimed kWh Total	*Claimed* Standard Deviation	Sample Size	Standard Error (Evaluated Total)	Relative Precision
Lighting	1,800,000	80	30	131,453	12.0%
HVAC	600,000	1,440	13	99,846	27.4%
ES Appliances	500,000	50	19	22,942	7.5%
Total	2,900,000	NA	62	166,660	9.5%

Repeat this process for the institutional, commercial, and industrial sectors. This is the more hands-on approach to Step 6, which begins with stipulated group-level precision targets, and usually leads to more back-and-forth iterations. Note that the more technical approach is also valid.

For Step 7, collect sector-level *claimed* savings totals and standard errors and use the formula for the standard error of a sum of independent estimates to estimate the standard error and precision at the portfolio level.

Table 7. High-Level Standard Errors

Sector	*Claimed* kWh Total	Precision	Standard Error
Residential	2,900,000	9.5%	166,660
Institutional	2,200,000	10%	133,739
Commercial	3,300,000	10%	200,608
Industrial	3,000,000	10%	182,371
Total	**11,400,000**	**4.6%**	**318,243**

The implied portfolio-level precision is $1.645 \cdot 318{,}243/11{,}400{,}000 = 4.6\%$, so this sample allocation will meet all precision targets if our CV and ER assumptions hold.

If the estimated precision value had been higher than the target, the evaluator would increase the sample sizes incrementally for the influential sector(s) with the lowest marginal sampling costs until the overall precision was achieved.

[End of Example]

5 Additional Considerations

The following sections discuss important considerations when choosing both a sample size and design.

5.1 Threats to Validity

The fundamental assumption in a design-based sample analysis is that population members have been sampled according to the rules specified in the sampling plan. When factors external to the sample plan affect the final sample, the study's validity may be compromised. In particular, specific external factors may lead to biased estimators and incomplete pictures of uncertainty.

The following are validity threats that commonly arise in impact evaluations.[18]

1. **Non-Coverage.** Validity is threatened when significant population segments are not included in the sample frame. The result is that values calculated from the sample cannot then be said to be representative of the entire population.
2. **Non-Response.** This type of threat occurs in every sample-based study for which population members have the option of refusing to be included. If certain types of households are more likely to refuse to participate or to respond to certain questions, the values calculated from the sample will understate the contribution of this portion of the population.
3. **Self-Selection**. In evaluation activities where participation is voluntary, some groups of people may be more likely to participate than others. This may be associated with demographics, education level, personal attitudes, or any number of unobservable factors. If this is the case, the estimate from these samples may not be completely representative.
4. **Measurement Error.** At times, data collection done either through metering or survey instruments may not be completely accurate.[19] Metering results can be biased by equipment failure, incorrect placement, or poor calibration. Survey instruments are vulnerable to a variety of threats that can be thought of as types of measurement error, such as: construct error, ambiguous wording of questions, and respondent social bias.

5.2 Cost Considerations

There is always a tradeoff between cost and precision. Although some gains in precision can be made through a thoughtful sample design, increasing the sample size always leads to better precision. However, the cost of doing so can be prohibitive.

---

[18] Threats to validity and strategies for mitigating their effects are explored in greater detail in *Appendix A*. For issues specific to survey instruments, see also the "Survey Design and Implementation for Estimating Gross Savings" chapter of this document.

[19] In most metering applications, this measurement error is ignored, particularly when data sources are utility-grade electricity or natural gas meters. However, other types of measurements—such as flow rates in water or air distribution systems—can have significant errors. The magnitude of such errors is often not large enough to warrant concern in a program evaluation and is largely provided by manufacturer's specifications.

The general precision equation can be written in this form:

$$\text{Precision} = \text{confidence level} \sqrt{\frac{\text{variance}}{\text{sample size}}}$$

Precision is a function of three factors: the confidence level (z), variance (s^2), and the sample size (n). The confidence level is fixed for a given study (typically at 90% for energy efficiency evaluations). The population variance does not change with sample size either, so the only factor under the evaluator's control in this equation is the sample size. However, precision is not improved at rate proportional to the sample size, but by the square root of the sample size. This is an important consideration in evaluation planning, as the cost-sample-unit is often linear, while improvements in precision are not.

Example 5-1

In conducting a metering study of commercial lighting to determine average hours of operation, the evaluator first performs a literature review. The effort reveals past studies showing that commercial lighting hours of operation typically vary with a CV of 0.5. When considering costs, the evaluator estimates each site will cost $1,000 for travel, data collection, and analysis. Figure 1 compares cost to precision.

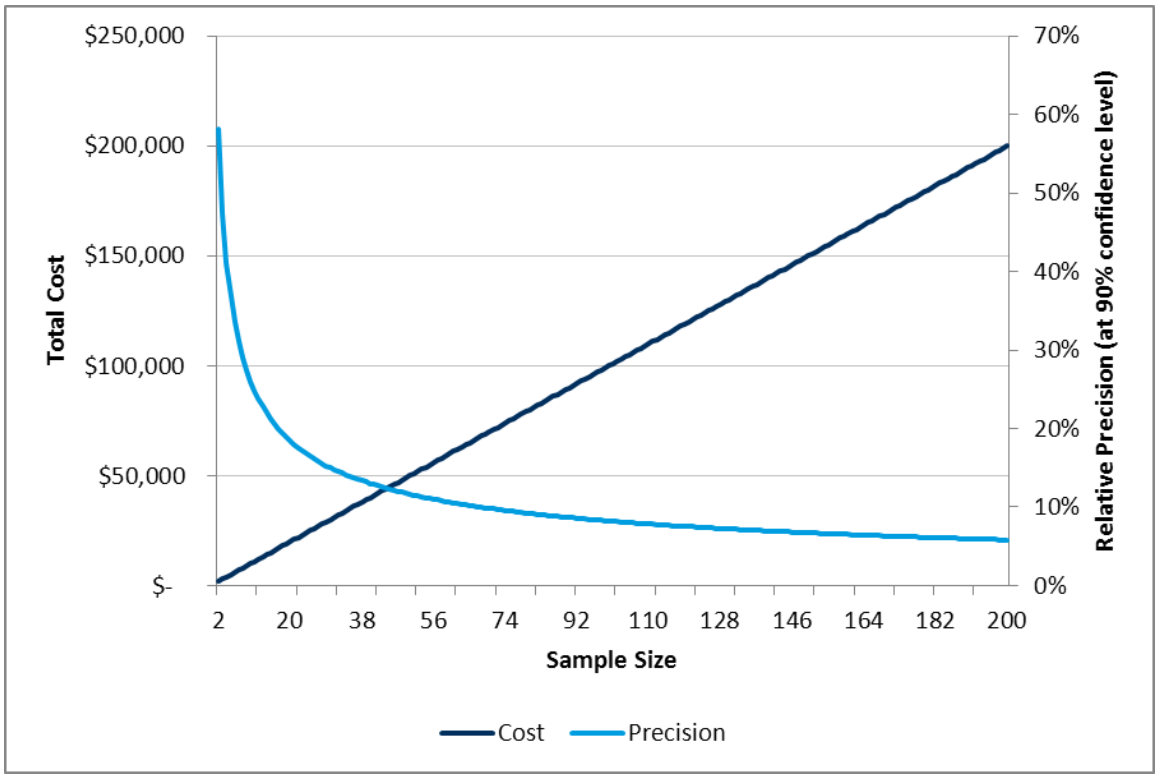

Figure 1. Example: cost vs. precision

So, visiting 70 sites to achieve ±10% relative precision (at the 90% confidence level) will cost $70,000. However, visiting only two sites (the minimum to calculate precision) would result in

relative precision of ±58% at a cost of $2,000. Thus, given repeated experiments, a 1% improvement in precision can be expected to cost an average of approximately $1,417.

If the evaluator chose to sample an additional 70 sites, the results would have a relative precision of ±7% at a total cost of $140,000. While the costs doubled, the precision only improved by approximately one third. Thus, average cost for a 1% increase in precision has now ballooned to approximately $23,333.

[End of Example]

5.3 Varying Uncertainty

In some cases, variation in the estimates of interest may differ in magnitude. If these measures are being combined, then the overall uncertainty of the final outcome is a function of those measures with large and small variation. As precision increases with variability (shown in the general equation repeated here), the overall sample will be more efficient when those measures with higher savings variation are allotted larger samples.

$$\text{Precision} = \text{confidence level} \sqrt{\frac{\text{variance}}{\text{sample size}}}$$

It is common practice in energy efficiency evaluations to estimate different parameters of an algorithm by different methods. One parameter may come from a phone survey, another from site visits, and a third may come from a secondary source. It is critical in these evaluations to identify the parameters having the greatest potential impact on overall uncertainty and then target them accordingly.

For example, in an evaluation conducted to estimate the savings of a residential energy-efficient showerhead program, the main inputs are hours of use, flow rate, and the installation rate. While installation rate and hours of use can be measured by phone survey, the flow rate must be measured on site. In this study, the evaluator knows that the CV of hours of use is much higher than the CV of flow rate. Thus, applying a sampling strategy that allots more of the sample to phone surveys and less to site visits could be more efficient than an equal allotment.

5.4 Outcome of Interest

As shown in the preceding example, it is critical to determine the true value of increased precision. Making this determination entails not only cost considerations, but knowing the value to the overall measure of interest. In an energy efficiency evaluation, this is most often total portfolio gross and/or net energy savings. If precision targets are set at the portfolio level, then the relative precision of a portfolio of programs is calculated as follows:

$$\text{Relative Precision of Portfolio} = 1.645 \cdot \left(\frac{1}{\sum_{i=1}^{m} \widehat{\text{savings}}_i} \right) \cdot \sqrt{\sum_{i=1}^{m} \left(SE[\widehat{\text{savings}}_i] \right)^2}$$

This formula follows from results presented in Appendix B. Fundamental Estimates and Uncertainty Calculations.

In Example 4-1, a 3% improvement in precision may justify an additional $70,000 in costs if the savings in this stratum represents a large proportion of total savings. If, however, a given measure makes up only 10% of total program savings, then a 1% improvement in precision at the measure level only contributes approximately 0.1% to the precision at the program level. Thus, both cost and value should be considered when choosing how to allocate resources effectively.

6 Appendix A. Sources and Types of Error

This appendix provides an introduction to how uncertainty is classified in evaluation applications, and it discusses systematic error and random error unrelated to sampling.

6.1 Sources of Uncertainty

As a measure of the "goodness" of an estimate, *uncertainty* refers to the amount or range of doubt surrounding a measured or calculated value. Any report of gross or net program savings, for example, has a halo of uncertainty surrounding the reported relative value to the true values (which are not known). As defined this way, uncertainty is an overall indicator of how well a calculated or measured value represents a true value. Without some measurement of uncertainty, it is impossible to judge an estimate's value as a basis for decision-making.

Program evaluation seeks to estimate energy and demand savings with reasonable accuracy. This objective may be affected by:

- **Systematic error** (that is, not occurring by chance), such as non-coverage, non-response, self-selection, and some types of measurement errors
- **Random error** (that is, occurring by chance), attributable to using a population sample rather than a census to develop the calculated or measured value. This error type can also be the result of some types of measurement error.[20]

The distinction between systematic and random sources of error is important because different procedures are required to identify and mitigate each. Although the amount of random error can typically be estimated using statistical tools, other means are required to estimate the level of systematic error. Because additional investment in the estimation process can lead to reductions in both types of error, tradeoffs between evaluation costs and reductions in uncertainty are inevitably required.

6.2 Sources of Systematic Error

Systematic errors typically occur from the way data are measured, collected, and/or described:

1. **Measured.** At times, equipment used to measure consumption may not be completely accurate. Human errors (for example, errors in recording data) may also cause this type of error. Metering results can be biased by equipment failure, incorrect placement, or poor calibration.[21] Survey instruments are vulnerable to a variety of threats that can be thought of as types of measurement error, such as construct error, ambiguous wording of questions, and respondent social bias.

[20] Note that measurement error may be systematic or random. For example, a meter that is not properly calibrated and consistently under- or overestimates a measurement exhibits systematic error. A meter that is only accurate within a given interval is said to have random error within that interval.

[21] Such errors will bias measurements within a site. However, because the magnitude and direction of the bias may differ from one site to the next, these errors may be viewed as random (not systematic) from the point of view of the broader evaluation, provided the errors are not similar across sites.

Measurement error is reduced by investing in more accurate measurement technology, establishing clear data collection protocols, and reviewing data to confirm they were accurately recorded. In most applications, this error source is ignored, particularly when data sources are utility-grade electricity or natural gas metering equipment. However, other types of measurements can have significant errors.

2. **Collected.** *Non-coverage errors* can occur when some parts of a population are not included in the sample. This can be a problem because the value calculated from the sample will not accurately represent the entire population of interest. Non-coverage error is reduced by investing in a sampling plan that addresses known coverage issues. For example, a survey implemented through several modes (such as phone, Internet, and mail) can sometimes address known coverage issues, assuming that non-coverage is related to the means of communication. However, in some cases there is little to do beyond clearly stating that some hard-to-reach segment of the population was excluded from the study.

   Non-response errors occur when some portion or portions of the population having certain attitudes or behaviors are less likely to provide data than are other population portions. In a load research or metering study, if certain types of households are more likely to refuse to participate—or if researchers are less likely to be able to obtain required data from them—the values calculated from the sample will understate the contribution of this portion of the population and over-represent the contribution of sample portions more likely to respond. In situations where the underrepresented portion of the population has different consumption patterns, non-response error is introduced into the value calculated from the sample. Non-response error is addressed through investments that increase the response rate, such as incentives and multiple contact attempts.

   The converse of non-response errors are *self-selection errors*. In evaluation activities where participation is voluntary, some groups of people may be more likely to participate than others. This may be associated with demographics, education level, personal attitudes, or any number of unobservable factors. If this is the case, the estimate from these samples may not be completely representative. Self-selection bias is best addressed by conducting studies in which participation is mandatory, although this is typically infeasible. Establishing representative quotas by demographics believed to be associated with self-selection may also mitigate these effects.

   Researchers often use "weights" in deriving their final estimates. These weights are means of adjusting the representativeness of the sample to reflect the actual population of interest. For example, if the proportion of single-family respondents is 70% in the sample but is 90% in the population, a weight of 90/70 can be used to increase the representativeness of single-family responses.

3. **Described (modeled).** Estimates are created through statistical models. Some are fairly simple and straightforward (for example, estimating the mean), and others are fairly complicated (for example, estimating response to temperature through regression models). Regardless, modeling errors may occur due to using the wrong model, assuming inappropriate functional forms, including irrelevant information, or

excluding relevant information (for example, in modeling energy use of air conditioners, the evaluator used cooling degree days only). In another example, home square footage or home type may not be available, so the statistical model will attribute all the observed differences in energy use to temperature, although clearly a portion of the use is attributable to the home size. This model will introduce systematic error.

Bias in regression estimates resulting from the omission of a relevant variable is also a well-known phenomenon. While evaluators use experience, economic theory, and engineering principles to prevent this type of bias, there is no statistical procedure to testing for this bias.

Reference manual assumptions are another potential source of modeled error. Technical reference manuals describe estimation procedures that are designed to balance evaluation rigor with practical concerns. Engineering assumptions and stipulated or deemed parameter values can introduce bias.

However, if a deemed value is obtained from a study that reports the value's standard error, then this standard error can be incorporated into a later evaluation, provided the study's target population is similar to the population being evaluated. In this case, the unknown bias can be accounted for within the evaluation's standard error calculations.

6.3 Sources of Random Error

Most random errors are due to sampling, measurement, or regression/extrapolation.

1. **Sampling.** Whenever a sample is selected to represent the population—whether the sample is of appliances, meters, accounts, individuals, households, premises, or organizations—there will be some amount of random sampling error. Any selected sample is only one of a large number of possible samples of the same size and design that could have been drawn from that population. Sampling error and strategies for mitigating it are discussed in detail in the rest of this document.

   The primary topic of this chapter is the mitigation and quantification of sampling error.

2. **Measurement.** In a survey, random measurement error may be introduced by factors such as respondents' incorrectly recalling dates, expenses, or by differences in a respondents' mood or circumstances, which affect how they answer a question. Technical measurements can also be a source of measurement error. (See item 1 and footnote 20 in the systematic error list.)

   These types of random measurement error are generally assumed to even out, so that they do not introduce systematic bias, but only increase the variability. For this reason, researchers often do not attempt to quantify the potential for bias due to random measurement error. However, measurement error can still be a source of variability, and

researchers are encouraged to include this source of uncertainty in standard error calculations when it presents a significant threat to validity.[22]

3. **Regression.** Regression error may arise at either the measure/site level, or at the population/stratum level.

   Site-level regression error arises when site-level savings estimates are obtained through regression (where a separate model is fitted to each site's data, and each site's savings is estimated through some function of the fitted parameters). For most site-level regression procedures, standard regression theory will provide a way to estimate the standard error of each site's savings estimate. These standard errors can then be accounted for in an evaluation's uncertainty calculations using methods similar to those applied in two-stage sampling. (See Section 8.7: *Two-Phase (Nested) Sampling*, of Appendix C. Also, ASHRAE Guideline 14 provides further details.)

   Population-level regression error arises when a single regression model is fit to data from multiple sites—possibly the entire population of sites that installed some program measure of interest. For example, a billing analysis may estimate program-wide natural gas savings due to high-efficiency residential furnaces by fitting a regression to billing data from all program participants and a control group of nonparticipants. The standard error of such regression-based estimates can be calculated with standard regression-related methods. Because the standard error applies to the estimate of total savings due to a measure—rather than site-level savings—this standard error is rolled up into sector- or portfolio-level savings uncertainty using the root-sum-of-squared-error formula. (In other words, it is treated in precisely the same manner as stratum-level sampling error.)

6.4 Mitigating Systematic Error

Determining the steps needed to mitigate systematic error is a more complex problem than mitigating random error, because various sources of systematic error are often specific to individual studies and procedures. To mitigate systematic error, evaluators typically need to invest in additional procedures (such as meter calibration, a pretest of measurement or survey protocols, a validation study, or a follow-up study) to obtain additional data to assess differences between participants who provided data and those who did not.

To determine how rigorously and effectively an evaluator has attempted to mitigate sources of systematic error, the following may be examined:

1. Were measurement procedures (such as the use of observational forms or surveys) pretested to determine if sources of measurement error could be corrected before the full-scale fielding?
2. Were validation measures (such as repeated measurements, inter-rater reliability, or additional subsample metering) used to validate measurements?
3. Was the sample frame carefully evaluated to determine what portions of the population, if any, were excluded in the sample? If so, what steps were taken to

[22] ASHRAE Guideline 14-2002 and Guideline 2002R offer extensive guidance on accounting for measurement error. Also, see Section 8.6, *Two-Stage Sampling for Large Projects* in this document for a related discussion.

estimate the impact of excluding this portion of the population from the final results?

4. Were steps taken to minimize the effect of non-response or self-selection in surveys or other data collection efforts? If non-response appears to be an issue, what steps were taken to evaluate the magnitude and direction of potential non-response bias?
5. Has the selection of formulas, models, and adjustments been conceptually justified? Has the evaluator tested the sensitivity of estimates to key assumptions required by the models?
6. Did trained, experienced professionals conduct the work? Was the work checked and verified by a professional other than the one conducting the initial work?

Many evaluation reports do not discuss any forms of uncertainty other than sampling error, which is quantified through confidence intervals for energy or demand savings. This is misleading because it suggests that (1) the confidence interval describes the total of all uncertainty sources (which is incorrect) or (2) the other sources of uncertainty are not important relative to sampling error. Sometimes, however, uncertainty due to other sources of error can be significant. A quality report should discuss all potentially significant sources of uncertainty so that research consumers can fully assess the evaluation's rigor.

6.4.1 Measurement Error

Measurement error can result from inaccurate mechanical devices (such as meters or recorders), inaccurate recording of observations by researchers, or inaccurate responses to questions by study participants. Basic human error occurs in taking physical measurements or conducting analyses, surveys, or documentation activities.

For mechanical devices—such as meters or recorders—it is theoretically possible to perform tests with multiple meters or recorders of the same make and model to assess the variability in measuring the same value. However, for meters and most devices regularly used in energy efficiency evaluations, it is more practical to use manufacturer or industry study information on the likely amount of error for any single piece of equipment.

Assessing the level of measurement error for data obtained from researchers' observations or respondents' reports is usually a subjective exercise, based on a qualitative analysis. This is because it is often impossible to make objective quantitative measures of these processes. The design of recording forms or questionnaires, the training and assessment of observers and interviewers, and the process of collecting data from study participants are all difficult to quantify.

Special studies of a subsample can be used to provide an assessment of the uncertainty potential in evaluation study results. For example:

- It is possible to have more than one researcher rate the same set of objects to evaluate the level of agreement between ratings.
- By conducting short-term metering of specific appliances for a subsample, an evaluator can verify information about appliance use.

- Participants can be re-interviewed to test their answers to the same question at different times.
- Pretests or debriefing interviews can be conducted with participants to determine how they interpreted specific questions and constructed their responses.

6.4.2 Non-Coverage and Non-Response

Another challenge is estimating the effect of excluding a portion of the population from a sample (sample non-coverage) or of the failure to obtain data from a certain portion of the sample (non-response). The data needed to assess these error sources are typically the same as those needed to resolve the errors; but such data are usually unavailable.

However, for both non-coverage and non-response, it is sometimes possible to design special studies to estimate the uncertainty level introduced.

- If a particular portion of the population was not included in the original sample design, it is possible to conduct a small-scale study on a sample of the excluded group. For example, conducting a special study of respondents who are in a particular geographical area or who are living in a certain type of housing can help determine the magnitude and direction of differences in calculated values for this portion of the population.
- In some situations—such as a survey—it is also possible to conduct a follow-up study of a sample of members from whom data were not obtained. This follow-up would also provide data to determine if non-respondents were different from respondents, as well as an estimate of the magnitude and direction of the difference.

7 Appendix B. Fundamental Estimates and Uncertainty Calculations

This section describes basic estimators commonly used in energy efficiency evaluations. Standard errors and other important formulas are also provided. These are fundamental to quantifying uncertainty, and they provide the foundation for basic sample design. For all formulas and examples in this section assume the data are collected through a simple random sample of size n from a very large population.[23]

Many research questions can be phrased in terms of:

- A population *average*, such as average savings among program participants or proportion of participants with gas heat
- A population *total*, such as total savings among all program participants or total number of customers with gas heat.

For consistency, this section's results are generally expressed in terms of averages. To estimate a population total, simply multiply the estimated average by the population size. The resulting estimate's standard error is the population size times the standard error of the average estimate. Because both the estimator and its standard error are multiplied by the population size, the relative precision is unaffected when translating between estimates of population averages and estimates of population totals.

7.1 Estimating a Population Proportion

Many energy efficiency evaluation tasks use survey data, which are typically used to estimate proportions. To estimate the proportion of the population having characteristic x (such as the proportion of utility customers who are aware of a given program), we use this formula:

$$\hat{p} = \frac{n_x}{n}$$

Where:

n_x = the number of sample points with characteristic x

n = the sample size.

To quantify the uncertainty surrounding this estimate, calculate the standard error and then calculate the precision.

[23] When the population is not very large, a non-negligible finite population correction will apply to standard errors. Simple random samples with finite population corrections are discussed in detail in Section 8.1, *Simple Random Sampling* in Appendix C.

The standard error of a proportion is most often[24] calculated as:

$$\widehat{\mathrm{SE}}(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

The absolute precision is then calculated as:

$$\text{Absolute Precision}(\hat{p}) \;=\; z \cdot \widehat{\mathrm{SE}}(\hat{p})$$

Note that the absolute precision equation does not involve dividing by the original estimate. This is different from energy savings estimates, where uncertainty is generally expressed in terms of relative precision. However, in process-related contexts, relative precision for a proportion can be a confusing measure, as the next example shows.

Example B-1

In a survey of 400 participants regarding their experience with a rebate program, we estimate the proportion of program participants satisfied with their rebate amount as $\hat{p} = 92\%$. We can then calculate the absolute precision at the 90% confidence level:

$$\text{Absolute Precision}(\hat{p}) \;=\; 1.645\sqrt{\frac{0.92(1-0.92)}{400}} \;=\; 2.2\%$$

Thus, we are 90% confident that the proportion of participants satisfied with the rebate is between 89.8% and 94.2%.

The relative precision, however, is calculated as:

$$\text{Relative Precision}(\hat{p}) \;=\; \frac{1.645\sqrt{\frac{0.92(1-0.92)}{400}}}{0.92} \;=\; 2.4\%$$

The relative and absolute formulations are both describing the same range of values, but the relative version expresses the confidence interval (CI) width as a proportion of a proportion. It says the CI has a width of 2.4% of 92%.

Not only is this confusing, it also leads to precision values that depend on how study results are communicated. The same study results could be communicated in terms of the proportion of participants who are *not* satisfied with the rebate amount. In this case, we have:

[24] When $\hat{p}$ is very close to one or zero, confidence intervals should be calculated through alternative means, such as the exact binomial method (see Example B-2). An oft-cited rule is that the exact method should be used if either n_x or $n - n_x$ is less than five.

$$\text{Absolute Precision}(1-\hat{p}) = 1.645\sqrt{\frac{0.08(1-0.08)}{400}} = 2.2\%$$

$$\text{Relative Precision}(1-\hat{p}) = \frac{1.645\sqrt{\frac{0.08(1-0.08)}{400}}}{0.08} = 27.8\%$$

While the absolute precision is the same as before, the relative precision is more than 10 times larger than previously calculated. As a result, someone reading the results might think the "unsatisfied" estimate is less precise than the "satisfied" estimate, despite the fact they convey identical information.

[End of Example]

In general, we recommend that precision for population proportions be expressed in absolute terms, especially when the research question is attitudinal or demographic. However, when the research target is a direct indicator of savings (such as the proportion of program-provided measures that are actually installed), relative precision may be preferred.

In Example B-1, the population proportion was estimated as $\hat{p} = 92\%$. Because the sample was of size $n = 400$, the data must have comprised $n_x = 368$ positive survey responses and $n - n_x = 32$ negative responses. Neither of these is less than five, so we were justified in using methods that assume $\hat{p}$ has an approximately normal sampling error. The next example illustrates the **exact binomial method**, which does not require the normality assumption.[25]

Example B-2

To verify the installation of measures that are recorded in a program database, we survey 50 participants, of whom 48 indicate they have installed the measure noted in the database. Thus, we estimate the percentage of participants who have installed the measure as $\hat{p} = 96\%$. However, with only two negative survey responses, we cannot say that the sampling error of $\hat{p}$ is approximately normal. Therefore, we need a method for obtaining a confidence interval that does not appeal to normality through a *z*-value. One option is the exact binomial method.

In a survey of $n = 50$ randomly selected people, the number of positive responses, n_x, follows a binomial distribution with 50 trials and an unknown "success" probability *p* for each trial. To construct a 90% CI for *p*, we calculate the upper and lower CI bounds separately.

[25] The exact binomial never understates uncertainty, but it often overstates it. This conservatism may be appropriate for some applications, and inappropriate for others. See Agresti (2003) or Brown, Cai, and DasGupta (2001) for details and alternative methods. In spite of the apparent simplicity of estimating a population proportion, there is no full consensus on the most desirable confidence interval for this problem among practicing statisticians. Alan Agresti, Brent Coull, George Casella, and others have attached insightful comments to the Brown, Cai, and DasGupta paper.

For the CI lower bound, we must answer the question, "*What is the smallest p for which the probability of obtaining 48 or more 'successes' is less than 5%?"* In Excel, this question can be answered using

$$=\texttt{Binom.inv}(50, p, 0.95)$$

For a given value of p, this function returns the smallest integer m for which the probability that $n_x \leq m$ is at least as large as 0.95.

If we choose a value p for which the function returns $m = 48 - 1$, then we know that the probability of 48 or more successes is no greater than 5% for the chosen p.

After finding a p for which the function returns a value of 47, we adjust p upward until the function returns a value of 48. Write $\hat{p}_{\text{lower}}$ for the largest p for which the function returns a value of 47. Then we are 95% confident that $p \geq \hat{p}_{\text{lower}}$.

In this example, the exact binomial method yields $\hat{p}_{\text{lower}} = 87.9\%$. A similar process yields the CI upper bound, $\hat{p}_{\text{upper}} = 99.3\%$. Thus, our estimate is $\hat{p} = 96\%$, and the exact binomial 90% confidence interval for p is

$$87.9\% \leq p \leq 99.3\%$$

For comparison, the normal-based confidence interval is

$$91.4\% \leq p \leq 100\%$$

The normal-based confidence interval understates uncertainty relative to the exact binomial confidence interval.

[End of Example]

In an extreme case, all survey responses may be affirmative. Then with no variability in the data, there is no basis for constructing a normal-based CI. However, it would not be credible to report 100% confidence that 100% of the population is in the affirmative category. The exact binomial method will yield a credible CI in such cases.

7.2 Using a Sample Mean to Estimate a Population Mean

Evaluations often need to estimate the average energy consumption for particular equipment types, such as residential refrigeration. When no useful auxiliary information is available,[26] the population average is estimated by the sample mean,

$$\bar{x} = \frac{\sum x_i}{n}$$

[26] Auxiliary information is discussed in the next section.

To quantify the uncertainty surrounding this estimate, calculate the standard error and then the precision. The sample mean's standard error is:

$$\widehat{\mathrm{SE}}(\bar{x}) = \frac{s}{\sqrt{n}}$$

Here, the sample standard deviation, *s*, is calculated as:

$$s = \sqrt{\frac{\sum(\bar{x} - x_i)^2}{n-1}}$$

The absolute and relative precision are then calculated as:

$$\text{Absolute Precision}(\bar{x}) \;=\; z \cdot \widehat{\mathrm{SE}}(\bar{x}) \;=\; z \cdot s/\sqrt{n}$$

$$\text{Relative Precision}(\bar{x}) \;=\; z \cdot \frac{\widehat{\mathrm{SE}}(\bar{x})}{\bar{x}} \;=\; z \cdot \frac{s/\sqrt{n}}{\bar{x}}$$

Example B-3

A metering study of 70 CFLs finds the hours of use to average 2.0 per day, with a standard deviation of 0.82 hours. Precision can then be estimated as:

$$\text{Absolute Precision}(\bar{x}) \;=\; 1.645 \cdot \frac{0.82 \text{ hrs/day}}{\sqrt{70}} \;=\; 0.16 \text{ hrs/day}$$

$$\text{Relative Precision}(\bar{x}) \;=\; 1.645 \cdot \left(\frac{0.16 \text{ hrs/day}}{2.15 \text{ hrs/day}}\right) \;=\; 7.5\%$$

Thus, we are 90% confident that average CFL usage is between 1.84 and 2.16 hours per day. Alternately, we can say that the mean hours of use is 2 hours per day, with ±9.8% precision at the 90% confidence level.

[End of Example]

7.3 Using a Ratio Estimator to Estimate a Population Mean

When estimating the population mean of some variable *y* that is closely correlated with some other variable *x*—which is known for every member of the population—a ratio estimator should be used to take advantage of the correlation. The known variable *x* is called an **auxiliary variable**. In energy efficiency evaluations, this is most often seen in realization rates, where the goal is to estimate the *evaluated* savings total, and the program database includes *claimed* savings estimates for each member of the population.

For commercial and industrial projects, *claimed* savings values often incorporate site-specific information, such as square footage of conditioned space and hours of operation. In these cases,

claimed values vary from project to project and the values can reasonably be expected to correlate with *evaluated* savings values.

The primary interest is in estimating the population mean of some variable y (denoted μ_y), where the variable x_i is known for every member of the population. (Thus, μ_x, the population mean of the x_i, is also known.) Then the ratio-based estimate of μ_y is[27]

$$\hat{\mu}_y = \frac{\sum y_i}{\sum x_i} \cdot \mu_x$$

The ratio estimator is technically biased, but its (unquantifiable) bias will generally be negligible compared to its standard error, provided the sample is not too small (ideally, the sample size should be at least 30). This can be a problem when separate ratio estimators are used for small strata; to avoid this issue savings from small strata should be estimated using a combined stratified ratio estimator, as described in *Appendix C. Sample Design and Weighted Estimates*.

The ratio estimator is similar to the estimator obtained by fitting the regression model $y = bx$. However, software that is not survey-oriented generally does not treat uncertainty correctly for (design-based) ratio estimators.[28] This deficiency is especially pronounced with weighted estimators, because design-based weights describe selection probabilities (see *Appendix C. Sample Design and Weighted Estimates*), whereas ordinary regression weights quantify observation-level standard errors.

The only source of uncertainty in this estimate is the uncertainty in the estimated realization rate,

$$\hat{r} = \frac{\sum y_i}{\sum x_i}$$

Estimator uncertainty is quantified through the standard error. The realization rate's standard error is:[29]

$$\text{Standard error of realization rate} = \widehat{\text{SE}}(\hat{r}) = \frac{1}{\sqrt{n}} \sqrt{\sum \frac{(y_i - \hat{r} \cdot x_i)^2}{\bar{x}^2 \cdot (n-1)}}$$

[27] All summations in this section are taken over the sample, not the population. This point can sometimes lead to confusion when working with ratio estimators.

[28] Sample-based inference, which is based on the selection probabilities inherited from the sample design, is often called *design-based*. By default, regression software usually applies *model-based* inference.

[29] The denominator in this expression uses the sample mean $\bar{x}$, rather than the population mean μ_x. This is consistent with Särndal 1992 (page 181, eq. 5.6.12) and the California Evaluation Protocol, but Lohr 1999 (page 68, eq. 3.7) uses the population mean instead. None of these references explicitly compares the two choices. Both possibilities are mentioned in Cochran 1977 (page 155, eqns. 6.12 and 6.13) and in Thompson 2002 (page 69, eqns. 5 and 7), but neither reference states a clear preference. One reason for our preference is that the standard error could be "gamed" by choosing small-scale projects if the population mean were used.

Thus, the standard error of the ratio-based estimate of μ_y is:

$$\widehat{\mathrm{SE}}\left(\hat{\mu}_y\right) \;=\; \widehat{\mathrm{SE}}(\hat{r}\cdot\mu_x) \;=\; \widehat{\mathrm{SE}}(\hat{r})\cdot\mu_x \;=\; \frac{1}{\sqrt{n}}\cdot\frac{\mu_x}{\bar{x}}\cdot\sqrt{\sum\frac{(y_i-\hat{r}\cdot x_i)^2}{n-1}}$$

To express these standard errors more succinctly, write:

$$s^{(\text{ratio})} = \sqrt{\sum\frac{(y_i-\hat{r}\cdot x_i)^2}{n-1}}$$

Then the expressions become:

$$\widehat{\mathrm{SE}}(\hat{r}) \;=\; \frac{s^{(\text{ratio})}}{\sqrt{n}}\cdot\frac{1}{\bar{x}}$$

$$\widehat{\mathrm{SE}}\left(\hat{\mu}_y\right) \;=\; \widehat{\mathrm{SE}}(\hat{r}\cdot\mu_x) \;=\; \frac{s^{(\text{ratio})}}{\sqrt{n}}\cdot\frac{\mu_x}{\bar{x}}$$

To see how ratio-based estimates leverage auxiliary data to increase study efficiency, compare this formula with the standard error of the sample mean in the previous section. The ratio-based standard error only has to account for the portion of variability in the y_i that is not explained by the realization-rate-adjusted x_i.

In cases where the realization rate itself is of primary interest, precision may be best described in absolute terms. However, when a population average (or total) is the estimation target, relative precision is usually needed. Depending on context, the precision is calculated with one of the following expressions.

$$\text{Absolute Precision}(\hat{r}) \;=\; z\cdot\widehat{\mathrm{SE}}(\hat{r})$$

$$\text{Relative Precision}(\hat{r}) \;=\; z\cdot\frac{\widehat{\mathrm{SE}}(\hat{r})}{\hat{r}}$$

$$\text{Relative Precision}(\hat{\mu}_y) \;=\; z\cdot\frac{\widehat{\mathrm{SE}}\left(\hat{\mu}_y\right)}{\hat{\mu}_y} \;=\; z\cdot\frac{\widehat{\mathrm{SE}}(\hat{r}\cdot\mu_x)}{\hat{r}\cdot\mu_x} \;=\; z\cdot\frac{\widehat{\mathrm{SE}}(\hat{r})}{\hat{r}}$$

Note that the relative precision of the estimated *evaluated* mean, $\hat{\mu}_y = \hat{r}\cdot\mu_x$, is exactly the same as the relative precision of the realization rate, $\hat{r}$. This is because $\widehat{\mathrm{SE}}(\hat{r}\cdot\mu_x) = \widehat{\mathrm{SE}}(\hat{r})\cdot\mu_x$, so the *evaluated* total's relative precision expression has cancelling factors of μ_x in its numerator and denominator.

Example B-4

In an impact evaluation for a commercial efficiency program, $n = 20$ projects are randomly selected for on-site verification. For each site, we have both *claimed* and *evaluated* savings

estimates.[30] The *claimed* total for the sampled sites is 607,415 kWh and the *evaluated* total for the sampled sites is 745,104 kWh, so the estimated realization rate is 1.227.

The data and the line $y = 1.227 \cdot x$ are plotted in Figure 2.

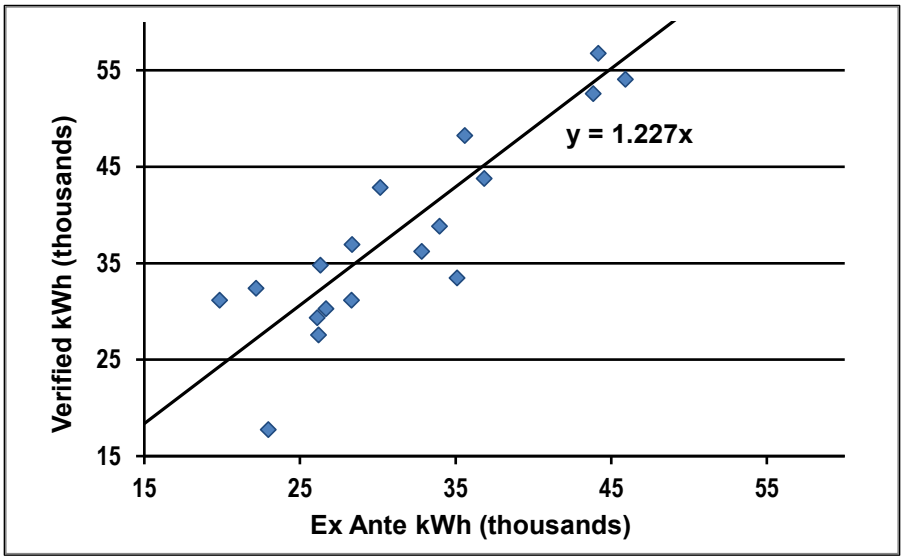

Figure 2. Verified versus claimed savings values

For these data, $s^{(\text{ratio})} = 6{,}176$ kWh and $\bar{y} = 39{,}216$ kWh. Thus, at the 90% confidence level, the relative precision is:

$$\text{Relative Precision}(\hat{r} \cdot \mu_x) \;=\; 1.645 \cdot \frac{6{,}176\ /\sqrt{20}}{39{,}216} \;=\; 5.8\%$$

If we ignored the auxiliary (*claimed*) data and used the sample mean estimator, $N \cdot \bar{y}$, instead of the ratio estimator, we would need to replace $s^{(\text{ratio})}$ with the standard deviation of the sample's verified savings numbers (in this case, $s = 12{,}132$ kWh). We would then obtain this:

$$\text{Relative Precision}(N \cdot \bar{y}) \;=\; 1.645 \cdot \frac{12{,}132\ /\sqrt{20}}{39{,}216} \;=\; 11.4\%$$

Here, the ratio estimator's precision is roughly one-half of the mean-based estimator's precision. This is because the ratio estimator's *s*-factor only needs to account for deviations between verified savings values and realization rate-adjusted *claimed* values ($y_i - \hat{r} \cdot x_i$). However, the mean-based *s*-factor (the usual sample standard deviation) must account for deviation between each verified savings value and the mean of the verified savings values ($y_i - \bar{y}$).

Figure 3 shows the spread of the two types of deviations for this example.

[30] Claimed values are the values in the program database, and evaluated values are engineering estimates based on data collected on-site during the evaluation.

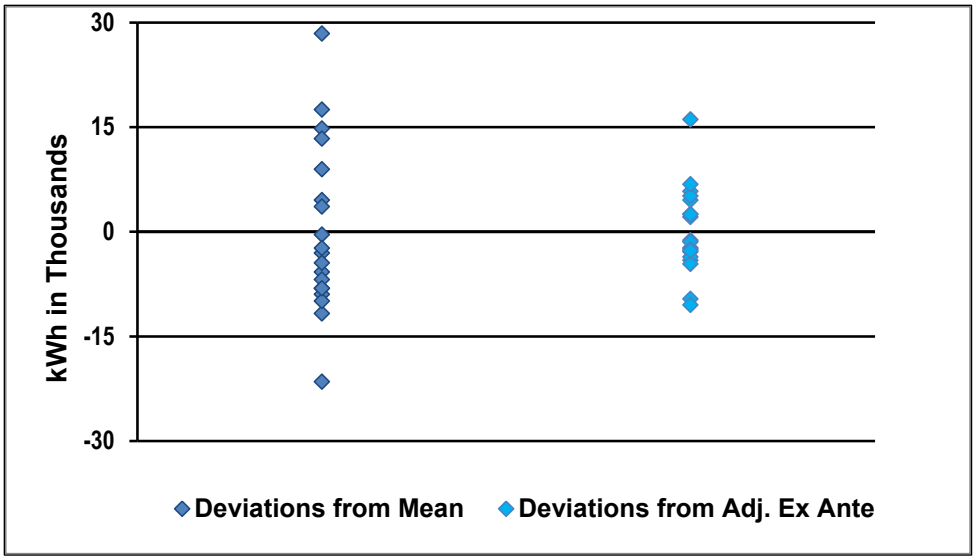

Figure 3: Comparison of verified savings deviations

[End of Example]

To develop intuition, it is helpful to think of the sizes of $s^{(\text{ratio})}$ and s relative to $\bar{y}$, rather than in absolute terms. Example B-4 had $s^{(\text{ratio})}/\,\bar{y} = 15.7\%$ and $s/\,\bar{y} = 30.9\%$. The expression $s^{(\text{ratio})}/\,\bar{y}$ is called the **error ratio** (ER), and $s/\,\bar{y}$ is the **coefficient of variation** (CV). These quantities describe the typical deviation size as a percentage of the typical project size.

In general, the deviations captured by $s^{(\text{ratio})}$ and s may reflect a number of unpredictable factors. For $s^{(\text{ratio})}$, the deviations between verified savings and adjusted *claimed* savings may result from factors such as poor data handling at the time of implementation, changes in site conditions since implementation, or changes in the number of shifts operating at the site. The standard deviation s may be influenced any of these factors, plus general variability among project sizes. As a result, the ER and CV do not obey any firm rules, except that the ER will generally be smaller than the CV whenever verified savings is roughly proportional to *claimed* savings.[31] (Also, most evaluators would agree that an ER of 15.7% and a CV of 30.9% are quite small for a commercial program.)

Example B-5

The program database for a commercial gas efficiency program indicates 9.42 million Mcf [thousand cubic feet] of claimed (*claimed*) savings program-wide, so we will conduct 40 site visits to verify the claimed savings. The 40 sampled sites account for a total of 2.00 mMcf in claimed savings, and our site visits verify a total of 1.70 mMcf in savings. Then we have:

[31] In general, the ratio estimator will be more efficient than the mean-based estimator if the correlation between x and y is greater than $0.5 \cdot \text{CV}(x) \cdot \text{CV}(y)$ (Cochran, 1977, page 157).

$$\hat{r} = 1.7 \text{ mMcf}/2.0 \text{ Mcf} = 85.0\%$$
$$\bar{y} = 1.7 \text{ mMcf}/40 = 0.0425 \text{ mMcf}$$
$$\bar{x} = 2.0 \text{ mMcf}/40 = 0.0500 \text{ mMcf}$$

Our data yields $s^{(\text{ratio})} = 0.0233$ mMcf, so the error ratio is:

$$\text{ER} = 0.0233 \text{ mMcf}/0.0425 \text{ mMcf} = 54.8\%$$

At the 90% confidence level, the realization rate's absolute precision is:

$$\text{Absolute Precision}(\hat{r}) = 1.645 \cdot \frac{s^{(\text{ratio})}}{\sqrt{n}} \cdot \frac{1}{\bar{x}} = 1.645 \cdot \frac{0.0233}{\sqrt{40}} \cdot \frac{1}{0.05} = 0.121$$

In other words, we have 90% confidence that the population realization rate is within 12.1 *percentage points* of 85%.

We estimate the program-wide total savings as $0.85 \cdot 9.42$ mMcf $= 8.01$ mMcf.

To calculate the relative precision of this estimate, we use:[32]

$$\text{Relative Precision}(\hat{r} \cdot \mu_x) = 1.645 \cdot \frac{s^{(\text{ratio})}/\sqrt{n}}{\bar{y}} = 1.645 \cdot \frac{0.0233/\sqrt{40}}{0.0425} = 14.3\%$$

So, we are 90% confident that the actual program savings is within 14.3% percent of 8.02 mMcf.

If we ignored the auxiliary (*claimed*) data and used the sample mean estimator, $N \cdot \bar{y}$, instead of the ratio estimator, we would have to replace the error ratio, $s^{(\text{ratio})}/\bar{y} = 54.8\%$, with the coefficient of variation, $s/\bar{y}$.

As noted earlier, the CV will be greater than the ER when *evaluated* and *claimed* values are strongly correlated. For example, if the CV in this example is 93.1%, then the mean-based estimator would be much less precise:

$$\text{Relative Precision}(N \cdot \bar{y}) = 1.645 \cdot 0.931 \cdot \frac{1}{\sqrt{40}} = 24.2\%$$

[End of Example]

[32] Recall that the relative precision of the population total estimate is the same as the relative precision of the population mean estimate, because both of the estimates and their standard errors differ by a factor of N from one setting to the other.

7.4 Estimating a Difference or Sum

- Sums and differences of estimated quantities arise frequently in evaluation work. Two prominent examples are:
- ***Combining savings across domains or strata.*** Large studies are often composed of multiple distinct research tasks for which the savings from the various research domains are to be summed to estimate the composite savings.
- ***Calculating savings as a difference.*** Savings is the difference between consumption in an inefficient scenario and consumption in an efficient one. Because energy efficiency evaluations seek to estimate these savings, evaluators often need to estimate a difference rather than a mean or proportion.

Assume independent, unbiased estimates, $\hat{x}$ and $\hat{y}$, of target quantities x and y. The difference or sum of the two estimates is an unbiased estimate of the difference or sum of the targets:

$$\widehat{x \pm y} = \hat{x} \pm \hat{y}$$

The standard error of the estimated difference or sum is then a function of both estimators. In general, this is:

$$\text{SE}(\hat{x} \pm \hat{y}) = \sqrt{\text{SE}(\hat{x})^2 + \text{SE}(\hat{y})^2 + 2 \cdot \text{Cov}(\hat{x}, \hat{y})}$$

Here, $\text{Cov}(\hat{x}, \hat{y})$ is the covariance of the two estimators. When the two estimators are based on separate, independently drawn samples, their sampling errors will be independent and their covariance will equal zero. In such cases, the formula reduces to:

$$\text{SE}(\hat{x} \pm \hat{y}) = \sqrt{\text{SE}(\hat{x})^2 + \text{SE}(\hat{y})^2}$$

When the sampling errors are not independent, the evaluator will either need to estimate the covariance[33] or employ an alternate method, such as the bootstrap.

The absolute and relative precision are then estimated as:

$$\text{Absolute Precision}(\hat{x} \pm \hat{y}) = z \cdot \widehat{\text{SE}}(\hat{x} \pm \hat{y})$$

$$\text{Relative Precision}(\hat{x} \pm \hat{y}) = z \cdot \left(\frac{\widehat{\text{SE}}(\hat{x} \pm \hat{y})}{\hat{x} \pm \hat{y}} \right)$$

Example B-6

A utility ran a CFL program and a refrigerator-recycling program, so the evaluator randomly sampled 30 projects from the CFL program and independently sampled 35 projects from the recycling program. The CFL sample led to an estimated program savings of 20 GWh, and the

[33] The procedure for evaluating the covariance will depend on the particular estimators and their relationship to one another.

refrigerator-recycling program had an estimated savings of 5 GWh. The total portfolio savings was then estimated as 25 GWh.

Assume both program-level estimators had 10% relative precision at the 90% confidence level. To evaluate the uncertainty of total savings, we first calculate the standard error for each program:

$$\widehat{\mathrm{SE}}(\text{CFL Savings}) = \frac{10\% \cdot 20\ \text{GWh}}{1.645} = 1.22\ \text{GWh}$$

$$\widehat{\mathrm{SE}}(\text{Refrigerator Savings}) = \frac{10\% \cdot 5\ \text{GWh}}{1.645} = 0.30\ \text{GWh}$$

The total program relative precision is then:

$$\text{Relative Precision}(\text{Portfolio Savings}) = \frac{1.645 \cdot \sqrt{(1.22)^2 + (0.30)^2}}{20 + 5} = 8.2\%$$

[End of Example]

7.5 Estimating a Product

In some instances, the product of two estimates is required. A common example of this is in using installation rates, where the proportion of measures installed is multiplied by an estimated per-unit savings to arrive at final verified savings.

In general, the exact standard error of a product is quite complicated,[34] but when the two estimators' sampling errors are independent, the standard error is:

$$\mathrm{SE}(\hat{x} \cdot \hat{y}) = \sqrt{(\hat{x} \cdot \mathrm{SE}(\hat{y}))^2 + (\hat{y} \cdot \mathrm{SE}(\hat{x}))^2 + (\mathrm{SE}(\hat{x}) \cdot \mathrm{SE}(\hat{y}))^2}$$

[34] The delta method yields a reasonably simple approximation that includes a covariance term. However, in evaluation work, there are few circumstances in which a product of two non-independent estimators is needed. In these rare cases, one should either apply the bootstrap method or, if the covariance can be estimated, the delta method.

Example B-7

For an evaluation of an HVAC program, the estimated gross annual unit energy savings is 200 kWh, with a standard error of 12.2 kWh/year. (This corresponds to 10% relative precision.)

The client and regulator have agreed that net savings will be calculated using the net-to-gross (NTG) ratio from a previous year's evaluation. The earlier evaluation reported an NTG estimate of 80% with a SE of 3.2% (absolute precision) at the 90% confidence level. Net unit savings is then estimated as $200 \text{ kWh} \cdot 0.8 = 160 \text{ kWh}$ per year.

Because the NTG estimate is independent of the gross estimate, the relative precision of net per-unit savings is:

$$\frac{1.645\sqrt{(80\% \cdot 12.2)^2 + (200 \cdot 3.2\%)^2 + (12.2 \cdot 3.2\%)^2}}{160} = 12.0\%$$

Note that the net savings estimate is less precise than the gross savings estimate (12% versus 10% relative precision, respectively). This is due to the additional uncertainty introduced through the NTG factor.

[End of Example]

7.6 Summary of Analytical Techniques

Table 8 summarizes the basic formulas used for analysis of simple random samples.

Table 8. Sample Analysis Formulas for Large Populations

Estimator and Target Quantity	Expression	Standard Error	Data Type
Sample proportion ($\hat{p}$); Population proportion (p)	$\frac{n_x}{n}$	$\frac{1}{\sqrt{n}} \cdot \sqrt{\hat{p}(1-\hat{p})} = \frac{s^{(p)}}{\sqrt{n}}$	Binomial
Sample mean ($\bar{x}$); Population mean (μ_x)	$\frac{\sum x_i}{n}$	$\frac{1}{\sqrt{n}} \cdot \sqrt{\frac{\sum(\bar{x} - x_i)^2}{n-1}} = \frac{s}{\sqrt{n}}$	Quantitative
Ratio estimator ($\hat{r} \cdot \mu_x$); Population mean (μ_y)	$\frac{\sum y_i}{\sum x_i} \cdot \mu_x$	$\frac{1}{\sqrt{n}} \cdot \sqrt{\frac{\sum(y_i - \hat{r}x_i)^2}{n-1}} \cdot \frac{\mu_x}{\bar{x}} = \frac{s^{(\text{ratio})}}{\sqrt{n}} \cdot \frac{\mu_x}{\bar{x}}$	Quantitative
Sum (or difference)*	$\hat{x} \pm \hat{y}$	$\sqrt{\text{SE}(\hat{x})^2 + \text{SE}(\hat{y})^2}$	Either
Product*	$\hat{x} \cdot \hat{y}$	$\sqrt{(\hat{x} \cdot \text{SE}(\hat{y}))^2 + (\hat{y} \cdot \text{SE}(\hat{x}))^2 + (\text{SE}(\hat{x}) \cdot \text{SE}(\hat{y}))^2}$	Either

*The indicated standard error formula is only valid if estimators are statistically independent (see the previous two subsections).

8 Appendix C. Sample Design and Weighted Estimates

For the estimators in Appendix B, it was assumed the sample was drawn through simple random sampling from a large population. This section discusses estimation with more general sample designs. Much of the discussion focuses on stratified designs and related topics, such as weighted estimators and sample optimization. We also discuss sampling with probability proportional to size and two-stage sampling for assessing savings for large projects.

8.1 Simple Random Sampling

In many ways, simple random sampling (SRS) is the most natural and intuitive sample design. In fact, more complicated designs can often be thought of as modifications or combinations of SRS. As the name suggests, SRS without replacement is the simplest random sampling approach, equivalent to "drawing *n* names from a hat."[35] The defining feature is that the final sample could be any set of *n* distinct names, and all such sets are equally likely. Thus, for an SRS of size *n* from a population of size *N*, each individual unit has selection probability n/N.

8.1.1 Sample Means with FPC

The only difference between this section and the sample mean discussion in Appendix B is that a very large population is no longer assumed.

Example C-1

For estimating the average number of incandescent bulbs still operating in residences within some utility's territory, the estimation target is the population mean,

$$\mu_x = \frac{x_1 + x_2 + \cdots + x_N}{N}$$

Here,

N = utility's total number of residential customers (the population size)

x_i = the number of incandescent bulbs operating at the i^{th} residence.

To estimate μ_x, we directly verify the number of incandescent bulbs in each of *n* homes, where the homes are selected via SRS. Based on these data, the most natural estimate of μ_x is the sample mean:

$$\bar{x} = \frac{1}{n} \sum_{\text{sampled } i} x_i$$

[35] The names are drawn without replacement, which means once a name is drawn, it is excluded from subsequent selection rounds. Thus, no name can be drawn more than once.

The standard error of the sample mean of an SRS is:

$$\widehat{\mathrm{SE}}(\bar{x}) \;=\; \sqrt{1-\frac{n}{N}}\cdot\frac{1}{\sqrt{n}}\cdot\sqrt{\sum_{\text{sample}}\frac{(x_i-\bar{x})^2}{(n-1)}} \;=\; \sqrt{1-\frac{n}{N}}\cdot\frac{s}{\sqrt{n}}$$

[End of Example]

Readers who are familiar with the statistical properties of sample means but not familiar with finite population inference may be surprised by the factor of $\sqrt{1-n/N}$ in the standard error expression.

This is called the **finite population correction (FPC)**, and it is a direct result of the SRS sample design. The FPC can be thought of as accounting for the fact that when the sample represents a significant fraction of the population, the uncertainty about the population mean is reduced. Note that when the population size is very large compared to the sample size, the ratio n/N will be close to zero, so the FPC will be close to one. In other words, the FPC is negligible for large populations.[36] In contrast, when the sample size is large so that n/N is close to one, the FPC (and hence the standard error) will be close to zero. A very large sample size means that most of the population has been measured directly, leaving little uncertainty about the population mean.

Determining an appropriate sample size is a critical step in planning a study. This determination is generally based on an agreed-upon precision target and some fixed confidence level. The general procedure uses the relevant precision formula and the target precision and confidence levels to express the necessary sample size in terms of important population quantities.

For the sample mean under SRS, the relative precision formula is typically used:

$$\text{Relative Precision}(\bar{x}) = \mathrm{z}\cdot\frac{\widehat{\mathrm{SE}}(\bar{x})}{\bar{x}}$$

The simplest way to calculate the sample size proceeds in two steps:

1. Calculate an initial sample size, $\boldsymbol{n_0}$, using the large-population standard error formula (that is, the formula without the FPC).
2. Adjust the initial sample size to account for the FPC in the true standard error.

The next example illustrates Step 1 and is followed by a brief discussion of the parameters that drive sample sizes. Step 2 is discussed at the end of this section.

Example C-2

To estimate the population mean to within 10% of its true value with 90% confidence, Step 1 ignores the FPC to obtain the initial sample size, n_0. This is the smallest integer that yields

[36] The proportion, sample mean, and ratio estimator sections of *Appendix B* provided standard error formulas that are valid under the assumption that the FPC is negligible.

$$0.10 \geq 1.645 \cdot \frac{s/\sqrt{n_0}}{\bar{x}}$$

Equivalently, n_0 is the smallest integer that satisfies this equation:

$$n_0 \geq \left(\frac{1.645}{0.10}\right)^2 \cdot \left(\frac{s}{\bar{x}}\right)^2$$

The quantity $s/\bar{x}$ is called the sample **coefficient of variation** (CV). This factor will not be known until after the data are collected. Past experience is the best guide for determining plausible values for the CV.

If the sample-based CV is greater than was expected when the sampling plan was developed, the study will fail to meet the agreed-upon confidence/precision target. For large studies, it may be advisable to (1) conduct a pilot study to estimate the CV in advance of the primary data collection effort or (2) plan for staged data collection so that sample sizes for later stages can be adjusted to reflect the CV observed through earlier stages. In all cases, the evaluator and the client should agree in advance on the measures to be taken to ensure an adequate sample size.

[End of Example]

As shown in the calculation in Example C-2, the large-population sample size formula is:

$$n_0 = \left(\frac{z \cdot \mathrm{CV}}{e_{\mathrm{rel.}}}\right)^2$$

Where:

CV is the coefficient of variation, the standard deviation divided by the mean
$e_{\mathrm{rel.}}$ is the desired level of relative precision
z is the critical value of the standard normal distribution value for the desired confidence level

For example, for 90% confidence, 10% precision, and a CV of 0.5, the initial sample size is:

$$n_0 = \left(\frac{1.645 \cdot 0.5}{0.10}\right)^2 = 67.7$$

Therefore, a sample of size 68 should be used here if the FPC is negligible. (Researchers often assume a CV of 0.5 when determining sample sizes, and because 90/10 confidence/precision is a common target, samples of size 68 are very common.)

One reason CVs of 0.5 are often reasonable in evaluation work is that the savings values are typically positive for all (or nearly all) projects. If 95% of a program's projects have savings between zero and 200% of the mean savings, and if the savings values are approximately normally-distributed, then a CV of 0.5 will apply.[37] This value, however, should not be applied

[37] Recall that for a normal distribution, approximately 95% of the population will fall within two standard deviations (SD) of the mean. If the CV equals 0.5, then the SD is one half of the mean. Thus, the 95% interval, mean ± 2

without due consideration of the expected nature of program savings. The justification noted here does not apply if project savings are heavily skewed towards large savers (in this case, the normality assumption fails). A stratified design (described later in this appendix) can often resolve this sort of skew and yield an effective CV that is closer to 0.5. In general, comparable previous studies and evaluation experience are the best guides for assessing likely CV values.

Because the FPC reduces standard error, it also reduces sample size required for any fixed levels of precision and confidence and fixed CV. The finite population adjustment reduces the necessary sample size as follows:

$$n = \frac{n_0 \cdot N}{n_0 + N}$$

In Example C-2, if the target population is of size $N = 200$, then the population is only three times the size of the sample. In this case, the finite population adjustment reduces the required sample size from 68 to 50:

$$n = \frac{68 \cdot 200}{68 + 200} \approx 50$$

8.1.2 Population Proportions and Ratio Estimators With FPC

Proportion estimates and ratio estimates can both be interpreted as versions of sample means. Thus, under SRS, these estimators' standard errors and sample sizes undergo finite population adjustments that are identical to their sample mean analogues.

The estimators themselves are unchanged from the large population case:

$$\begin{aligned} \hat{p} &= \frac{n_x}{n} \\ \hat{r} \cdot \mu_x &= \frac{\sum y_i}{\sum x_i} \cdot \mu_x \end{aligned}$$

Their standard errors, however, are multiplied by a finite population correction, just as in the sample mean case:

$$\begin{aligned} \widehat{\mathrm{SE}}(\hat{p}) &= \sqrt{1 - \frac{n}{N}} \cdot \frac{\sqrt{\hat{p} \cdot (1 - \hat{p})}}{\sqrt{n}} \\ \widehat{\mathrm{SE}}(\hat{r} \cdot \mu_x) &= \sqrt{1 - \frac{n}{N}} \cdot \frac{1}{\sqrt{n}} \cdot \sqrt{\sum \frac{(y_i - \hat{r} \cdot x_i)^2}{n - 1}} = \sqrt{1 - \frac{n}{N}} \cdot \frac{s^{(\text{ratio})}}{\sqrt{n}} \cdot \frac{\mu_x}{\bar{x}} \end{aligned}$$

SD, is the same as mean ± mean (the mean, plus or minus itself). In other words, if the CV is 0.5 and the data are normal, the 95% CI will range from 0 to 200% of the mean. Again, if one is willing to assert that the data will be normal and that most of the members of the population will fall between 0 and 200% of the mean, then a CV of 0.5 is appropriate.

Sample size calculations for both population proportions and ratio estimators are similar to the sample mean calculations. Calculate an initial sample size, n_0, using the large-population standard error formula and then apply a finite population adjustment.

For population proportions the large-population precision formula is:

$$e_{\text{abs.}} = z \cdot \widehat{\text{SE}}(\hat{p}) = z \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n_0}}$$

So the initial sample size formula is:

$$n_0 = \left(\frac{z}{e_{\text{abs.}}}\right)^2 \cdot p(1-p)$$

In this formula, z is as before and $e_{\text{abs.}}$ is the absolute precision target. If there is no basis for making *a priori* assumptions about p, then use $p = 0.5$, because $p(1-p)$ obtains its maximum with this value.

For both population proportions and ratio estimators, the FPC reduces the necessary sample size as before. In both cases, the final sample size is:

$$n = \frac{n_0 \cdot N}{n_0 + N}$$

Example C-3

For a large population, the requirement for estimating a population proportion to within 5 percentage points, with 90% confidence, is this:

$$n_0 \geq \left(\frac{1.645}{0.05}\right)^2 \cdot p(1-p)$$

The quantity $p(1-p)$ can never be greater than $0.5(1-0.5) = 0.25$, so the precision target is guaranteed to be met if:

$$n_0 \geq \left(\frac{1.645}{0.05}\right)^2 \cdot (0.5)^2 = 270.6$$

Thus, if the population is very large and there is no *a priori* knowledge of p, then to meet the 90/5 standard, plan for the study to achieve at least 271 complete responses.

Now assume there are only $N = 550$ individuals in the target population. Then the FPC reduces the required sample size to:

$$n = \frac{270.6 \cdot 550}{270.6 + 550} = 181.4$$

In this case, plan for 182 complete survey responses.

[End of Example]

When the ratio estimator $\hat{r} \cdot \mu_x$ is used to estimate the population mean μ_y, the large-population precision formula is:

$$e_{\text{rel.}} \;=\; z \cdot \frac{\widehat{\text{SE}}(\hat{r} \cdot \mu_x)}{\hat{r} \cdot \mu_x} \;=\; z \cdot \frac{s^{(\text{ratio})}/\sqrt{n_0}}{\bar{y}}$$

Therefore, the initial sample size formula is:

$$n_0(\hat{r} \cdot \mu_x) \;=\; \left(\frac{z}{e_{\text{rel.}}}\right)^2 \left(\frac{s^{(\text{ratio})}}{\bar{y}}\right)^2$$

This formula is identical to the one obtained for the sample mean, except that the standard deviation, *s*, has been replaced with $s^{(\text{ratio})}$, which quantifies only that portion of variability not explained through the auxiliary information.

The quantity $s^{(\text{ratio})}/\bar{y}$ is called the **error ratio (ER)**.[38] When the *x* and *y* variables are correlated, the error ratio will tend to be smaller than the CV, so the ratio-based estimator will be more efficient than the sample mean.

As indicated above, the FPC reduces the necessary sample size precisely as before. In both cases, the final sample size is:

$$n = \frac{n_0 \cdot N}{n_0 + N}$$

[38] The California Evaluation Framework prescribes a model-assisted approach, based on evidence that deviations between evaluated values y_i and adjusted claimed values $\hat{r}x_i$ tend to scale in proportion to x_i^{γ} for some $\gamma \approx 0.8$. This approach leads to a different procedure for estimating the error ratio. When greater efficiency may be gained through this well-studied model-based approach, researchers are encouraged to apply it.

Summary of SRS Estimators

The important equations for SRS are listed in Table 9.

Table 9. Results for Simple Random Samples

Estimator	Expression	Standard Error	Initial Sample Size	Sample Size With FPC
Sample mean	$\frac{\sum x_i}{n}$	$\sqrt{1-\frac{n}{N}}\cdot\frac{s}{\sqrt{n}}$	$n_0 = \left(\frac{z}{e_{\text{rel.}}}\right)^2 \cdot (\text{CV})^2$	$\frac{n_0 \cdot N}{n_0 + N}$
Sample proportion	$\frac{n_x}{n}$	$\sqrt{1-\frac{n}{N}}\cdot\frac{\sqrt{p(1-p)}}{\sqrt{n}}$	$n_0 = \left(\frac{z}{e_{\text{abs.}}}\right)^2 \cdot p(1-p)$	$\frac{n_0 \cdot N}{n_0 + N}$
Ratio estimator	$\frac{\sum y_i}{\sum x_i}\cdot\mu_x$	$\sqrt{1-\frac{n}{N}}\cdot\frac{s^{(\text{ratio})}}{\sqrt{n}}\cdot\frac{\mu_x}{\bar{x}}$	$n_0 = \left(\frac{z}{e_{\text{rel.}}}\right)^2 \cdot (\text{ER})^2$	$\frac{n_0 \cdot N}{n_0 + N}$

8.2 Stratified Random Sampling

Stratified sampling entails partitioning the population into distinct groups (called *strata*) and drawing samples independently from each stratum. In some cases, the groupings reflect qualitative population characteristics. For example, participants in a commercial HVAC program may be stratified by business type, or participants in a comprehensive nonresidential program may be separated by custom versus prescriptive projects. Strata may also be created to group the population into size categories according to *claimed* savings values in the program database.

The main reason for using stratified sampling is to reduce the variance in a population-wide estimator by separating the population into homogeneous groups. Population-level uncertainty is then driven exclusively by within-stratum variation. As a result, when homogeneous groupings are available, stratified random sampling is almost always more efficient than simple random sampling. In addition, in cases of study domains with particularly small populations, stratification ensures that every relevant stratum is represented in the sample. (This may not be case in simple random sampling.)

Stratification is a very flexible tool in its application. For instance, the population of program participants may first be divided into sector and fuel type groupings and then stratified by size. The particular choice of stratification variable(s) will depend on context.

For this section, assume that (1) the population has been partitioned into H non-overlapping strata and (2) the stratum population sizes are given by $N_1, N_2, \ldots, N_H$. Also assume that each stratum's sample is selected via simple random sampling within the stratum.[39] For example, within stratum h, an SRS of size n_h is been drawn from a group of N_h individuals, so each

[39] Stratification can also be employed with more general probability sampling within each stratum. (This is described in most sample design textbooks.) When an alternative scheme is used, the researcher should clearly describe the sampling scheme and the estimator with references (or direct calculations) explaining why standard error calculations are valid indicators of uncertainty.

sampled unit represents N_h/n_h members of the population. Thus, the weight of a unit sampled from stratum h is $w_h = N_h/n_h$.

Stratified designs bring new notational requirements. For most objects, a subscripted *h* will indicate stratum number, and a subscripted *all* will indicate that an object spans all strata. Most stratified approaches are more easily understood when research tasks are expressed in terms of population totals (and their estimators) rather than population means, so the notation also makes this distinction.

The general conventions for this section are as follows.

Population Quantities

X_{all} and Y_{all} are the x_i and y_i population totals
N_{all} is the total number of population members, $N_{\text{all}} = N_1 + N_2 + \cdots + N_H$
μ_{all} is the population mean of the x_i, $\mu_{\text{all}} = X_{\text{all}}/N_{\text{all}}$
X_h and Y_h are stratum-h population totals of the x_i and y_i
$\mu_{x,h}$ is the stratum-h population mean of the x_i, $\mu_{x,h} = X_h/N_h$

Sample Quantities and Estimators

n_{all} is the total sample size, $n_{\text{all}} = n_1 + n_2 + \cdots + n_H$
$\bar{x}_h$ and $\bar{y}_h$ are the stratum-h sample means of the x_i and y_i
$w_h = N_h/n_h$ is the weight that applies to stratum-h sample members
$\bar{x}_{\text{all}}^{(\text{w})}$ and $\bar{y}_{\text{all}}^{(\text{w})}$ are the weighted sample means of the x_i and y_i
$h(i)$ is the stratum containing unit *i*

As before, the procedures for determining appropriate sample sizes will be demonstrated after the basic properties of the estimators are established. Stratified versions of sample means, proportions, and ratio estimators are described in this section.

8.2.1 Stratified Means

The basic idea behind the independent-estimators approach is illustrated in the following example.

Example C-4

For this evaluation, the object is to estimate the total air-conditioning tonnage among all commercial retailers in a particular service territory. A sample mean applied to a simple random sample would be very inefficient, because a small number of commercial retailers are orders of magnitude larger than most of the population. (This skew would translate to a very large CV.)

If retailer size categories are known through auxiliary data, these size categories may be used as strata for the study. Within each stratum, skew would be limited, so stratum-level CVs should be moderate.

Assume three retailer size categories: stratum one covers small retailers, stratum two covers medium retailers, and stratum three covers large retailers. Write s_1 for the stratum-one sample standard deviation, and likewise for $s_2, \ldots, s_H$. Then the estimated stratum one total is $\hat{X}_1 = N_1 \cdot$

$\bar{x}_1$, and its standard error is:

$$\mathrm{SE}\left(\hat{X}_1\right) = \mathrm{SE}(N_1 \cdot \bar{x}_1) \; = N_1 \cdot \sqrt{1 - \frac{n_1}{N_1}} \cdot \frac{s_1}{\sqrt{n_1}}$$

Calculate $\hat{X}_2$ and $\hat{X}_3$ the same way, and estimate the population total as:

$$\hat{X}_{\text{all}}^{(\text{w})} = \; \hat{X}_1 + \hat{X}_2 + \hat{X}_3 \; = \; N_1 \cdot \bar{x}_1 + \; N_2 \cdot \bar{x}_2 + N_3 \cdot \bar{x}_3$$

The superscripted "w" emphasizes that this is a weighted estimator. Its standard error is:

$$\mathrm{SE}\left(\hat{X}_{\text{all}}^{(\text{w})}\right) \; = \sqrt{\mathrm{SE}\left(\hat{X}_1\right)^2 + \mathrm{SE}\left(\hat{X}_2\right)^2 + \mathrm{SE}\left(\hat{X}_3\right)^2}$$

To estimate the population-wide mean, use:

$$\hat{X}_{\text{all}}^{(\text{w})}/(N_1 + \; N_2 + N_3).$$

This estimate's standard error is:

$$\mathrm{SE}(\hat{X}_{\text{all}}^{(\text{w})})/(N_1 + \; N_2 + N_3).$$

[End of Example]

The general formula for the stratified-means estimator of the population total is:

$$\hat{X}_{\text{all}}^{(\text{w})} \; = \sum_{h=1}^{H} \hat{X}_h = \sum_{h=1}^{H} N_h \cdot \bar{x}_h$$

This estimator can also be written as a weighted sum,

$$\hat{X}_{\text{all}}^{(\text{w})} \; = \sum_{\text{sampled } i} \frac{N_{h(i)}}{n_{h(i)}} \cdot x_i \; = \sum_{\text{sampled } i} w_{h(i)} \cdot x_i$$

The weighted sum's standard error is calculated as follows. (Notice that only the within-stratum standard deviations, s_h, affect the standard error.)

$$\mathrm{SE}\left(\hat{X}_{\text{all}}^{(\text{w})}\right) \; = \sqrt{\sum \mathrm{SE}\left(\hat{X}_h\right)^2} \; = \sqrt{\sum N_h^2 \cdot \mathrm{SE}(\bar{x}_h)^2} \; = \sqrt{\sum \frac{N_h^2}{n_h} \cdot \left(1 - \frac{n_h}{N_h}\right) \cdot s_h^2}$$

To estimate the population *mean*, divide the estimated total by the population size:

$$\bar{x}_{\text{all}}^{(\text{w})} = \frac{\hat{X}_{\text{all}}^{(\text{w})}}{N_{\text{all}}}$$

This estimator is called the weighted mean.

8.3 Stratified Proportions

The reasoning in the previous section also applies to population proportions. To estimate the fraction of the population having some particular characteristic, first estimate the total number of individuals with the characteristic and then divide by the population size.

To express these results, we must expand on the notation of Appendix B:

N_{all}^{x} is the total number of individuals in the population who have characteristic *x*.
N_h^x is the total number of individuals from stratum *h* who have characteristic *x*.
p_{all} is the population proportion, $p_{\text{all}} = N_{\text{all}}^{x}/(N_1^x + N_2^x + \cdots + N_H^x)$
n_h^x is the number of *sampled* individuals from stratum *h* who have characteristic *x*.
$\hat{p}_h$ is the proportion of the stratum *h* sample with the characteristic, $\hat{p}_h = n_h^x/n_h$.
$\hat{p}_{\text{all}}^{(\text{w})}$ and $\widehat{N}_{\text{all}}^{x}$ are our estimates of p_{all} and N_{all}^{x}.

The weighted estimators related to population proportions are:

$$\widehat{N}_{\text{all}}^{x} = \sum_{h=1}^{H} N_h \cdot \hat{p}_h$$

$$\widehat{\text{SE}}\left(\widehat{N}_{\text{all}}^{x}\right) = \sqrt{\sum N_h^2 \cdot \widehat{\text{SE}}(\hat{p}_h)^2} = \sqrt{\sum \frac{N_h^2}{n_h} \cdot \left(1 - \frac{n_h}{N_h}\right) \cdot \hat{p}_h(1-\hat{p}_h)}$$

$$\hat{p}_{\text{all}}^{(\text{w})} = \frac{\widehat{N}_{\text{all}}^{x}}{N_1 + N_2 + \cdots + N_H} = \frac{\sum N_h \cdot \hat{p}_h}{N_1 + N_2 + \cdots + N_H}$$

$$\widehat{\text{SE}}\left(\hat{p}_{\text{all}}^{(\text{w})}\right) = \frac{\widehat{\text{SE}}\left(\widehat{N}_{\text{all}}^{x}\right)}{N_1 + N_2 + \cdots + N_H}$$

8.3.1 Stratified Ratio Estimators

The stratified ratio estimator is based on the ratio of the weighted sum of the sampled y_i to the weighted sum of the sampled x_i. Rather than applying a different realization rate within each stratum, we apply this single weighted realization rate to all strata. In the preceding section on stratified means, $\hat{X}_{\text{all}}$ represented the weighted total of the x_i, and the weighted mean was $\bar{x}_{\text{all}}^{(\text{w})} = \hat{X}_{\text{all}}/N_{\text{all}}$.

The weighted realization rate can be thought of either as the ratio of estimated totals or as the ratio of estimated means:

$$\hat{r}_{\text{all}}^{(\text{w})} = \frac{\sum_{\text{sample}} w_{h(i)} \cdot y_i}{\sum_{\text{sample}} w_{h(i)} \cdot x_i} = \frac{\hat{Y}_{\text{all}}^{(\text{w})}}{\hat{X}_{\text{all}}^{(\text{w})}} = \frac{\bar{y}_{\text{all}}^{(\text{w})}}{\bar{x}_{\text{all}}^{(\text{w})}}$$

The ratio-based estimate of the population total of the y_i is:

$$\hat{Y}_{\text{all}}^{(\text{w})} = \hat{r}_{\text{all}}^{(\text{w})} \cdot X_{\text{all}} = \frac{\bar{y}_{\text{all}}^{(\text{w})}}{\bar{x}_{\text{all}}^{(\text{w})}} \cdot X_{\text{all}}$$

The standard error is:[40]

$$\begin{aligned}
\text{SE}\left(\hat{Y}_{\text{all}}^{(\text{w})}\right) &= \left(\frac{\mu_{\text{all}}}{\bar{x}_{\text{all}}^{(\text{w})}}\right) \cdot \sqrt{\sum_{h=1}^{H} \frac{N_h^2}{n_h}\left(1 - \frac{n_h}{N_h}\right) \sum_{\substack{\text{stratum } h \\ \text{sample}}} \frac{\left(y_i - \hat{r}_{\text{all}}^{(\text{w})} \cdot x_i\right)^2}{n_h - 1}} \\
&\approx \left(\frac{\mu_{\text{all}}}{\bar{x}_{\text{all}}^{(\text{w})}}\right) \cdot \sqrt{\sum_{h=1}^{H} \left(\frac{N_h}{n_h}\right)^2 \left(1 - \frac{n_h}{N_h}\right) \sum_{\substack{\text{stratum } h \\ \text{sample}}} \left(y_i - \hat{r}_{\text{all}}^{(\text{w})} \cdot x_i\right)^2} \\
&= \left(\frac{\mu_{\text{all}}}{\bar{x}_{\text{all}}^{(\text{w})}}\right) \cdot \sqrt{\sum_{\text{sample}} w_{h(i)}\left(w_{h(i)} - 1\right)\left(y_i - \hat{r}_{\text{all}}^{(\text{w})} \cdot x_i\right)^2}
\end{aligned}$$

Typically, $\mu_{\text{all}}/\bar{x}_{\text{all}}^{(\text{w})}$ will be close to one, because it is the ratio of the actual mean to the estimated mean. So to see the basic features of the standard error formula, we can ignore this factor. What remains in the first equation in the chain above is very similar to the standard error of the weighted sum, $\hat{X}_{\text{all}}^{(\text{w})}$. The only difference is that the s_h^2 of the weighted sum's standard error is now replaced by:

$$\left(s_h^{(r,\,\text{w})}\right)^2 = \sum_{\substack{\text{stratum } h \\ \text{sample}}} \frac{\left(y_i - \hat{r}_{\text{all}}^{(\text{w})} \cdot x_i\right)^2}{n_h - 1}$$

The last formula in the chain is identical to the formula provided in the *California Evaluation Framework.* Although the FPC is obscured in the *Framework*'s weight-based presentation, the middle expression clearly shows that the formulation does account for the FPC.

8.3.2 Summary of Estimators for Stratified Samples

The next two tables summarize results for the estimators developed in this section. Table 10 gives the estimators themselves and their standard errors.

[40] See Särndal 1992, page 181.

Table 10. Formulas for Stratified Estimators

Estimator	Expression	Standard Error
Weighted sum	$\hat{X}_{\text{all}}^{(w)} = \sum N_h \cdot \bar{x}_h = N \cdot \bar{x}_{\text{all}}^{(\text{w})}$	$\sqrt{\sum \frac{N_h^2}{n_h}\left(1-\frac{n_h}{N_h}\right)s_h^2}$
Weighted proportion	$\hat{p}_{\text{all}}^{(w)} = \frac{\sum N_h \cdot \hat{p}_h}{\sum N_h}$	$\sqrt{\sum \frac{N_h^2}{n_h}\left(1-\frac{n_h}{N_h}\right)\hat{p}_h(1-\hat{p}_h)}$
Weighted Ratio Estimator	$\hat{Y}_{\text{all}}^{(r,\,\text{w})} = \hat{r}_{\text{all}}^{(r,\,\text{w})} \cdot X_{\text{all}}$	$\sqrt{\sum \frac{N_h^2}{n_h}\left(1-\frac{n_h}{N_h}\right)\left(s_h^{(r,\,\text{w})}\right)^2}\left(\frac{\mu_{\text{all}}}{\bar{x}_{\text{all}}^{(\text{w})}}\right)$

Table 11 provides supplementary formulas.

Table 11. Additional Formulas

Estimator	Unit-level Standard Deviation Estimates	Other Expressions
Weighted sum	$s_h^2 = \sum_{\text{sample } h} \frac{(x_i - \bar{x}_h)^2}{(n_h - 1)}$	NA
Weighted proportion	$s_{p,h}^2 = \frac{n_h^x}{n_h} \cdot \left(1 - \frac{n_h^x}{n_h}\right)$	$\hat{p}_h = \frac{n_h^x}{n_h}$
Weighted Ratio Estimator	$\left(s_h^{(r,\,\text{w})}\right)^2 = \sum_{\text{sample } h} \frac{\left(y_i - \hat{r}_{\text{all}}^{(\text{w})} \cdot x_i\right)^2}{n_h - 1}$	$\hat{r}_{\text{all}}^{(\text{w})} = \frac{\bar{y}_{\text{all}}^{(\text{w})}}{\bar{x}_{\text{all}}^{(\text{w})}}$

8.4 Planning and Optimizing Stratified Designs

The basic result in the optimization of stratified designs is called the **Neyman allocation**. Among all possible allocations of the n sample units to the H strata, the lowest overall variance will be achieved if:

$$n_h = n \cdot \left(\frac{N_h \cdot s_h}{N_1 \cdot s_1 + \cdots + N_H \cdot s_H}\right)$$

This formula has one major shortcoming that may render it unacceptable for planning large scale studies—it does not consider cost-efficiency. If units from Stratum 1 are much more expensive to survey than units from Stratum 2, then the cost-optimal sample design should allocate fewer units to the more expensive stratum.

The **cost-weighted Neyman allocation** addresses this concern. Use c_h for the marginal cost of sampling a single unit from stratum *h*. Assume a fixed budget for data collection. Then among all possible resource allocations, the lowest overall variance will be achieved if, for some *n*,

$$n_h = n \cdot \left(\frac{N_h s_h \ / \sqrt{c_h}}{N_1 s_1 \ / \sqrt{c_h} + \cdots + N_H s_H \ / \sqrt{c_h}} \right)$$

Both the Neyman allocation and the cost-weighted Neyman allocation work the same with other estimators. Simply replace the stratum-level standard deviation s_h with the appropriate selection from Table 11.

Table 12. Sample Allocation Formulas

Step	Formula
Estimate maximum acceptable overall variance	$\text{Var}(\hat{X}_{\text{all}}) = (X_{\text{all}})^2 \cdot \left(\frac{e_{\text{rel.}}}{z}\right)^2$
Allocate sample among strata.	$n_h = n \cdot \left(\frac{N_h s_h \ / \sqrt{c_h}}{N_1 s_1 \ / \sqrt{c_h} + \cdots + N_H s_H \ / \sqrt{c_h}} \right)$

At the planning stage, of course, data-driven estimates of stratum-level standard deviations are not available. Planning estimates may come from other studies, general past experience, or agreed-upon values based on known database quality standards.[41]

8.5 General Probability Samples and PPS

In simple random sampling without replacement, it was demonstrated that with a sample of size *n* from a population of size *N*, each individual unit has selection probability of:

$$\pi_i = \frac{n}{N}$$

More general sample designs are available, however, such as **probably proportional to size** (PPS). The idea behind PPS is to sample *n* units from the population, each with probability proportional to its size. Because such a scheme necessarily requires auxiliary information for determining the π_i, the typical auxiliary information notation is used for this section.

> x_i is the auxiliary information for site *i*. (In evaluation work, this is usually the claimed savings estimate from the program database.)
>
> y_i is the variable of primary interest for site *i*.

[41] This is especially relevant for ratio estimators, because large deviations between evaluated and claimed values often reflect problems in the program database, rather than variation in consumer behavior.

The goal is to estimate the population total, $Y = y_1 + \cdots + y_N$.

In practice, auxiliary data (the x_i) are used as a proxy for the true savings sizes (the y_i) in calculating the π_i. Insofar as the x_i are consistently proportional to the y_i, PPS estimation will result in very low standard errors.[42]

Strict PPS can be difficult to implement in a manner that both (1) yields no repeat entries in the sample and (2) produces a sample of fixed size, *n*.[43] However, there are several available variants that are easy to implement, but loosen one or both of the requirements noted.

The variant called Poisson sampling (illustrated in Example C-5) produces samples with no repeat entries, but with variable sample sizes. This variant does not require size stratification, because project sizes are appropriately accounted for through probability weighting.

Example C-5
Determine the sample size target, *n,* and use the auxiliary data to set selection probabilities.

$$\pi_i \;=\; n \cdot \frac{x_i}{x_1 + x_2 + \cdots + x_N}$$

In a spreadsheet, generate a random number (distributed uniformly between 0 and 1) for each project and then designate each project as sampled if its random number is less than its π_i value.

Then standard estimator of the population total is:

$$\hat{Y} \;=\; \sum_{\text{sampled } i} \frac{y_i}{\pi_i}$$

This estimator's standard error is estimated as:

$$\widehat{\mathrm{SE}}\left(\hat{Y}\right) \;=\; \sqrt{\sum_{\text{sampled } i} (1 - \pi_i)\left(\frac{y_i}{\pi_i}\right)^2}$$

[End of Example]

Other PPS variants are available (see Särndal, *et al.*, pp. 85-99).

[42] The same statement holds for ratio estimators, so PPS does not have any general efficiency advantage over ratio methods. It is only an alternative approach that avoids the need for size stratification and, thus, may be simpler to employ in some contexts (especially for within-site subsampling, which is described in the next section).
[43] See Särndal, *et al.*, pp. 90-7. A principle difficulty is that the second-order inclusion probabilities can be difficult to evaluate for any given scheme that produces the desired first-order probabilities. Advanced statistical software packages (such as STATA and SAS) can draw samples and analyze data for most PPS variants, so these difficulties are not fatal. However, as the algorithms would not be easy to implement in a spreadsheet, these methods may not be practical for field work.

8.6 Two-Stage Sampling for Large Projects

Nonresidential programs often include a small number of very large projects. In many cases, direct evaluation of every measure within a large project would impose an unacceptable burden on the customer. As a result, evaluators must rely on a subsample of measures within each large project in the set of sampled projects. This is called two-stage sampling.[44]

The principles described in the preceding sections apply both to the overall sample and to each subsample. This section explains how to integrate subsample results with the broader program evaluation. Our guidance is similar to that given in ASHRAE Guideline 14.

Example C-6

An industrial energy efficiency program is being evaluated using a stratified design that includes a single stratum for very large projects (designated as stratum *H*). For this example, assume the following: (1) a weighted-sum estimator will be used to combine stratum-level results and (2) all measures at any sampled site that is not a member of the large projects stratum will be directly evaluated.

For each stratum other than stratum *H,* the estimated total savings is:

$$\hat{X}_h = N_h \cdot \bar{x}_h \quad \text{and} \quad \text{SE}(\hat{X}_h) = \sqrt{\frac{N_h^2}{n_h}\left(1 - \frac{n_h}{N_h}\right)s_h^2}$$

For a sampled site *i* within stratum *H*, we do not directly evaluate the savings x_i. Instead, we estimate x_i using verified values $x_{i,1}, x_{i,2}, \ldots, x_{i,m}$ for some sample of measures within site *i.* The particular method for estimating x_i based on the sampled $x_{i,j}$ depends on the site-level sample design and evaluation plan. However, in all cases it is possible to calculate the estimate, $\hat{x}_i$, and its standard error, $\text{SE}(\hat{x}_i)$. The total savings estimate for stratum *H* is then:

$$\hat{X}_H = N_H \cdot \frac{\hat{x}_1 + \hat{x}_2 + \cdots + \hat{x}_{n_H}}{n_H} = N_H \cdot \bar{\hat{x}}_H$$

The standard error of this estimate includes both the usual sampling error (as with the other $\hat{X}_h$) and within-site sampling errors:

$$\text{SE}(\hat{X}_H) = \sqrt{\frac{N_H^2}{n_H}\left(1 - \frac{n_H}{N_H}\right)s_H^2 + \sum_{\text{sample } H} \text{SE}(\hat{x}_i)^2}$$

[44] The distinguishing feature of two-stage sampling is that a sample of secondary units (for example, measures) is selected within each sampled primary unit (for example, project). *One-stage sampling* refers to the case where all secondary units are selected from each sampled primary unit. *Cluster sampling* is usually synonymous with two-stage sampling, but some textbooks reserve this term for one-stage sampling.
Also, two-*stage* sampling is not the same as two-*phase* sampling, in which a large initial sample is observed through low-cost interactions (for example, phone surveys), and the initial sample data are used to increase efficiency for a small sample involving more expensive interactions (for example, site visits). (Two-phase sampling is discussed in Section 8.7, *Two-Phase [Nested] Sampling*.)

It is not uncommon to conduct a full census of very large sites. In such cases, $n_H = N_H$, so the first term in the standard error is zero. Therefore, the terms $\mathrm{SE}(\hat{x}_i)^2$ are the sole contributors to the estimator's standard error for any census stratum.

As always, the total program savings is estimated as:

$$\hat{X}_{\text{all}}^{(\text{w})} = \sum_{h=1}^{H} \hat{X}_h \quad \text{and} \quad \mathrm{SE}\left(\hat{X}_{\text{all}}^{(\text{w})}\right) = \sqrt{\sum \mathrm{SE}\left(\hat{X}_h\right)^2}$$

[End of Example]

Example C-6 illustrates an important feature of two-stage sampling—each finite population correction applies only to the level at which the relevant sampling occurs. Thus, the FPC due to first-stage sampling applies to program-level estimates, while within-site sampling may lead to FPCs which apply within the $\mathrm{SE}(\hat{x}_i)$.

ASHRAE Guideline 14 presents this same approach, but with a slightly different perspective on the origin of random deviations between the $\hat{x}_i$ and x_i. In Guideline 14, the standard errors of the $\hat{x}_i$ are assumed to account for measurement, modeling, and similar sources of random error.

This section's guidance is compatible with Guideline 14. In general, dominant error sources should always be accounted in the $\mathrm{SE}(\hat{x}_i)$, and the dominant errors may be due to modeling error in one context and sampling error in another, depending on site-level evaluation strategies.

The following example illustrates an important point regarding the proper handling of auxiliary data when site-level sub-sampling is used.

Example C-7

For an industrial energy efficiency program, the evaluator is using a stratified design and has created a single stratum containing the program's largest projects (designated as stratum *H*). The evaluator plans to evaluate savings directly for every measure at sampled sites that are not members of stratum *H*. For this example, assume the evaluator plans to use a weighted ratio estimator to estimate the total program savings.

For a sampled site *i* in stratum *H*, the evaluator uses whatever means are available to estimate y_i efficiently—that is, to minimize $\mathrm{SE}(\hat{y}_i)$.[45] For some sites, this may include within-site ratio estimation or a PPS estimator. In such cases, the evaluator may review *claimed* savings assumptions on site and adjust *claimed* values to reflect actual hours of use and similar inputs, provided that the adjustments are (1) applied to sampled and non-sampled measures alike and (2) based on information that is equally available for sampled and non-sampled measures.

For example, if the *claimed* values in the program database assume a 16-hour daily schedule for every measure at a given site, but the site actually operates for 24 hour per day, the measure-

[45] Recall that for ratio estimators, y_i represents verified savings and x_i represents claimed savings estimates.

level *claimed* values may be adjusted accordingly. The main requirement is that such adjustments be made without giving the site's sampled measures any special consideration.[46]

Also, because *claimed* values cannot be adjusted for every site in the population, this sort of *a priori* adjustment applies only to measures within a sampled site and only to the calculation of $\hat{y}_i$ and $\mathrm{SE}(\hat{y}_i)$. The original *claimed* values must still be used in calculating the program-level standard error.

In this case, the estimated the realization rate is determined as:

$$\hat{r}_{\text{all}}^{(\text{w})} = \frac{N_1 \cdot \bar{y}_1 + N_2 \cdot \bar{y}_2 + \cdots + N_{H-1} \cdot \bar{y}_{\text{H-1}} + N_H \cdot \bar{\hat{y}}_H}{N_1 \cdot \bar{x}_1 + N_2 \cdot \bar{x}_2 + \cdots + N_{H-1} \cdot \bar{x}_{\text{H-1}} + N_H \cdot \bar{x}_{\text{H}}}$$

The only difference between this expression and the weighted-sum ratio given in the preceding section on stratified ratio estimators is that this expression uses estimated (rather than directly observed) $\hat{y}$ values for the stratum-H sample. With this minor adjustment, estimate the population total Y_{all} as:

$$\hat{Y}_{\text{all}}^{(\text{w})} = \hat{r}_{\text{all}}^{(\text{w})} \cdot X_{\text{all}}$$

In these equations, the x_i refer to the *claimed* savings values from the program database (unadjusted) and the X_{all} is the *claimed* total (unadjusted) for the entire population. The standard error is estimated as:

$$\widehat{\mathrm{SE}}\left(\hat{Y}_{\text{all}}^{(\text{w})}\right) = \left(\frac{\mu_{\text{all}}}{\bar{x}_{\text{all}}^{(\text{w})}}\right) \sqrt{\sum_{h=1}^{H} \frac{N_h^2}{n_h}\left(1 - \frac{n_h}{N_h}\right)\left(s_h^{(r,\,\text{w})}\right)^2 + \left(\frac{N_H}{n_H}\right)^2 \sum_{\text{sample } H} \widehat{\mathrm{SE}}(\hat{y}_i)^2}$$

Here, the standard errors of the $\hat{y}_i$ may reflect adjustments to measure-level *claimed* values, as discussed above.

8.7 Two-Phase (Nested) Sampling

When an M&V protocol requires on-site metering or other labor-intensive procedures at sampled sites, a *two-phase (nested)* design can often reduce study costs without compromising rigor. A two-phase study is conducted as follows:

1. Select a large sample of projects/sites/measures (the Phase 1 sample). Conduct low-cost evaluation research for sites in the Phase 1 sample (for example, phone surveys may be used to verify installation and size or quantity). Use the information obtained to update *claimed* savings values for all sites in the Phase 1 sample.

[46] These claimed adjustments need not be highly detailed, because the final estimate $\hat{y}_i$ will be adjusted to reflect empirical data and rigorous measure-level analysis. The goal is only to reduce $\mathrm{SE}(\hat{y}_i)$ by taking advance measures to diminish the deviations between measure-level verified and claimed savings values.

2. Select a subsample of Phase 1 projects for intensive M&V (this is the Phase 2 sample). Use the M&V data to evaluate verified savings for each of the Phase 2 projects.
3. Analyze the Phase 2 data using a ratio estimator with Phase 1 *claimed* updates as auxiliary data.

In a two-phase study, the total savings is estimated as:

$$\hat{Y} = \hat{r} \cdot \hat{X} = \left(\frac{\sum_{\text{Sample 2}} y_i}{\sum_{\text{Sample 2}} x_i}\right) \cdot \left(N \cdot \frac{\sum_{\text{Sample 1}} x_i}{n_1}\right)$$

Because the *claimed* values have been updated to reflect basic verification data, a large source of variation between *claimed* and *evaluated* has been eliminated. This can result in drastic reductions in the effective error ratio. However, the standard error formula needs to be adjusted to reflect the fact that the auxiliary data are only available for a sample and not the whole population. With the adjustment, the standard error is:

$$\widehat{\text{SE}}(\hat{Y}) = N \cdot \sqrt{\left(1 - \frac{n_1}{N}\right)\frac{s_y^2}{n_1} + \left(1 - \frac{n_2}{N}\right)\frac{s_{\text{ratio}}^2}{n_2}}$$

Here, s_{ratio} calculated from the deviations between the updated *claimed* values (Phase 1) and the final evaluated savings values (Phase 2).

This approach reconciles two important aspects of evaluation rigor:

- ***Program-level sampling rigor.*** This refers to minimizing sampling error, which is a function of sample size, population size, and variability between reported and verified savings values. (This variability is captured by the error ratio.)
- ***Site-level estimation rigor.*** This refers to minimizing the errors in site-level savings estimates. In other words, minimizing the deviations between a site's verified savings value and its actual savings.

Two-phase sampling may be used to increase sampling efficiency (equivalently, to increase sampling rigor for a given study cost) without reducing site-level evaluation rigor.