BONNEVILLE POWER ADMINISTRATION

Sampling for M&V: Reference Guide

May 2024

Sampling for M&V: Reference Guide

Version 3.0

May 2024

Prepared for

Bonneville Power Administration

Prepared by

Facility Energy Solutions

Stillwater Energy

SBW Consulting

Contract Number BPA-2-C-92283

Table of Contents

1. Introduction

1.1. Purpose

Sampling for M&V: Reference Guide (Sampling Guide) is a complement to the Measurement & Verification (M&V) protocols used by the Bonneville Power Administration (BPA). The *Sampling Guide* assists the engineer in selecting how many of what types of points to monitor or meter when it would not be cost-effective to perform these actions at all points.

Originally developed in 2012, this *Sampling Guide* is one of ten documents produced by BPA to direct M&V activities; an overview of the ten documents is given in the *Measurement and Verification (M&V) Protocol Selection Guide and Example M&V Plan (M&V Selection Guide).*

Chapter 6 of this guide provides full citations (and web locations, where applicable) of documents referenced and an appendix provides a glossary specific to this guide.

1.2. Protocols Version 3.0

BPA revised the M&V protocols described in this guide in 2024. BPA published the original documents in 2012 as Version 1.0, which were updated to Version 2.0 in 2018. The current documents are Version 3.0.

1.3. How is M&V Defined?

BPA's *Implementation Manual* (the IM) defines measurement and verification as "the process for quantifying savings delivered by an energy conservation measure (ECM) to demonstrate how much energy use was avoided. It enables the savings to be isolated and fairly evaluated."[1] The IM describes how M&V fits into the various activities it undertakes to "ensure the reliability of its energy savings achievements." The IM also states:

> *The Power Act specifically calls on BPA to pursue cost-effective energy efficiency that is "reliable and available at the time it is needed."[2] "[…] Reliability varies by savings type: For UES measures, and calculators.[3,4], measure specification and savings estimates must*

---

1 2017-2019 Implementation Manual, BPA, October 1, 2017. *https://www.bpa.gov/EE/Policy/IManual/Documents/IM_2017_10-11-17.pdf*

2 Power Act language summarized by BPA.

3 UES stands for Unit Energy Savings and is discussed subsequently. In brief, it is a stipulated savings value that region's program administrators have agreed to use for measures whose savings do not vary by site (for sites within a defined population). More specifically UES are specified by either the Regional Technical Forum – RTF (referred to as "RTF approved") or unilaterally by BPA (referred to as BPA-Qualified). Similarly, Savings Calculators are RTF approved or BPA-Qualified.

4 Calculators estimate savings that are a simple function of a single parameter, such as operating hours or run time.

be RTF approved or BPA-Qualified.[5] *Custom projects require site-specific measurement and verification (M&V) to support reliable estimates of savings. BPA M&V Protocols direct M&V activities and are the reference documents for reliable M&V."*

The *M&V Selection Guide* includes a flow chart providing a decision tree for selecting the M&V protocol appropriate to a given custom project and addressing prescriptive projects using UES estimates and Savings Calculators.

M&V is site-specific and required for stand-alone custom projects. BPA's customers submit bundled custom projects (projects of similar measures conducted at multiple facilities) as either an M&V Custom Program or as an Evaluation Custom Program; the latter requires evaluation rather than the site-specific M&V that these protocols address.

1.4. Background

BPA contracted with a team led by Facility Energy Solutions to assist the organization in revising the M&V protocols used to assure reliable energy savings for the custom projects it accepts from its utility customers. The team conducted a detailed review of the 2018 M&V Protocols and developed the revised version 3.0 under Contract Number BPA-2-C-92283.

The Facility Energy Solutions team is comprised of:

- Facility Energy Solutions, led by Lia Webster, PE, CCP, CMVP
- Stillwater Energy, led by Anne Joiner, CMVP
- SBW Consulting, led by Santiago Rodríguez-Anderson, PE

BPA's Todd Amundson, PE, PMVE was project manager for the M&V protocol update work. The work included gathering feedback from BPA and regional stakeholders, and the team's own review to revise and update this 2024 *Sampling Reference Guide*.[6]

[5] https://www.bpa.gov/-/media/Aep/energy-efficiency/document-library/24-25-im-april24-update.pdf, page 1.

2. Overview of Sampling

2.1. Description

Proper measurement and verification (M&V) requires actual measurements of the affected systems. Except in rare instances, it is not cost-effective to measure the performance of every piece of equipment. Typically, we can measure the performance of a representative *sample* of the population and extrapolate the results to the whole. The process of identifying which and how many items to measure and evaluating the results to assess reliability is known as *sampling*.

2.2. Applicability

Sampling is often used where the population of affected systems is too large to measure each point cost-effectively. The primary requirement for sampling to be effective is that the samples selected must be representative of the entire population.

Any end-use technology or building category can be evaluated using sampling techniques: lighting upgrades, controls upgrades, rooftop unit replacements, residential upgrades, industrial equipment retrofits, etc.

Several BPA protocols specify that sampling can be used, or the protocol methodology is compatible with sampling techniques (see Table 2-1).

Table 2-1: Protocols for Which Sampling May be Appropriate

Protocol Name	Application	IPMVP* Option	Sampling Applicable
Verification by Meter-Based Energy Modeling	Projects in large buildings with an existing condition baseline or equipment that is sub-metered	B/C	
Verification by End-Use Metering	A measure that changes both load and operating hours	B	YES
Existing Building Commissioning	Multiple measures improving performance of building systems	C	
Engineering Calculations with Verification	Estimation of savings using equipment characteristics and accepted assumptions about operating parameters	N/A	YES

* International Performance Measurement and Verification Protocol

2.3. Advantages of Sampling

The primary advantage of sampling is cost-effectiveness. By measuring the performance of a sample of affected items rather than the entire population, the level of effort, and associated cost, is significantly reduced.

2.4. Disadvantages of Sampling

The fundamental disadvantage of sampling is the introduction of uncertainty, because not every item is measured. With careful sample plan development, selection, and evaluation, uncertainty can be kept to minimal levels and be quantified.

Also, a sampled measurement process can be compromised if the selected samples are not representative of the population, if data collection problems reduce the sample size to less than acceptable numbers, or if there proves to be much greater variability in the population than originally assumed. These limitations may not be discovered until the data has been collected and evaluated, making corrective steps difficult.

2.5. Types of Sampling

For most measurement and verification purposes, there are two types of sampling: *simple random sampling* and *stratified sampling*.

- **Simple random sampling draws a sample from a population without respect to the underlying characteristics of the population**. This approach works well with homogeneous populations. One example is measuring the average power draw of several light fixtures containing two 4-foot T-8 lamps to establish the typical power draw where hundreds of identical fixtures exist. There is no guarantee that a random sample will be representative of the population, but it will be unbiased because each member of the population has equal likelihood of being selected. The likelihood of a representative sample increases with sample size.

- **Stratified sampling partitions the population by some characteristics and selects separate samples from each sub-group.** Stratified sampling is used where the population is not homogeneous and needs to be segregated by some defining characteristic. While the population could be divided into discrete groups and simple random sampling applied to each group, this would result in a greater overall sample size and level of effort. Stratified sampling considers the contribution to the overall uncertainty from each group (strata) and allocates the samples to minimize the overall uncertainty and sample size. An example might be residential energy use, where the population could be divided into three strata of single-family homes, multifamily homes, and condominiums.

2.6. Definitions

Sampling is most efficient when samples must be drawn from a homogeneous population. Homogeneous in this context means that all members of the population have similar characteristics. A group of single-family, three-bedroom homes of similar vintage, with four occupants, all in the same city, would be a homogenous population.

If a group contains members of differing characteristics, the population is considered heterogeneous. A group of small commercial buildings that includes offices, banks, convenience stores, and self-storage buildings is heterogeneous because the schedule, function, and internal

load differ across the building types. In this example, grouping the buildings by function will make several small homogeneous populations.

For most projects, it is reasonable to assume that a homogeneous population can be defined by a single average value and that most of the members will have values very close to that average. Only a few members will vary significantly from the average, with some having values higher and some lower than the average value. The farther from the average value, the fewer members will be found. This type of population is considered a *normal* distribution. An example might be the operating hours of lighting fixtures in an office. Most will operate about 2,200 hours per year, but a few may operate for 1,100, and a few for 3,300 hours. If the population distribution were plotted as a function operating hours, it would form a bell-shaped curve centered at 2,200 hours.

However, there are projects where we need to establish the presence or absence of a condition, rather than determining a value for population members. Examples include situations that test for pass/fail, installed/not installed, operating/broken, or free-rider/not free-rider. These are referred to as *binomial distributions*, as each element can occupy one of two states. The sampling techniques described here can be applied to binomial distribution cases with some adjustments to the mathematics.

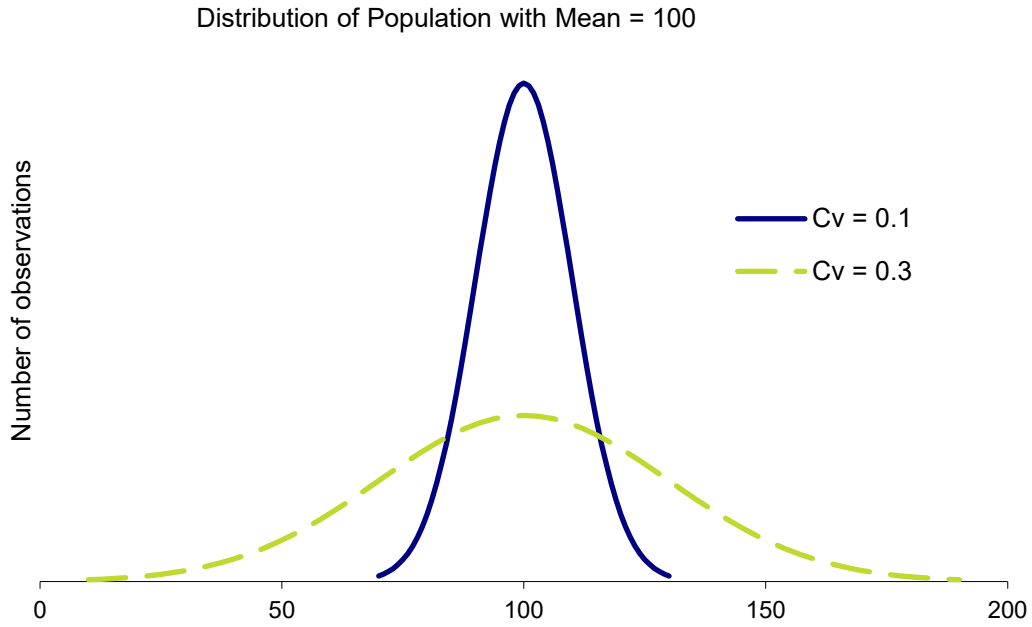
There are also projects where the goal is to determine specific values rather than proportions in a population. In this case, a power analysis must be conducted. Power analysis has different reporting requirements and is beyond the scope of this guide.

The *coefficient of variation* (CV) is a mathematical expression of the dispersion of a data set from its mean. It can be visualized as how wide or narrow the population distribution (or bell curve) is when data is plotted on a two-dimensional graph. The CV is defined as the *standard deviation* of the population divided by the average value (or *mean*) of that population, but it is more important to understand its significance.

Populations with a broad range of values have a large CV; tightly clustered populations have a small CV. The previous example of lighting operating hours has a moderate distribution ranging between 1,100 and 3,300, centered at 2,200. An example of a population with a tight distribution would be the speed of cars on a crowded freeway – they all need to travel at close to the same speed for traffic to move.

For many energy efficiency projects, the CV will range in value from 0.1 to 1.0, depending on what is being evaluated. Figure 2-1, below, illustrates the distributions of two different populations, each with a mean value of 100, but with different coefficients of variation.

Figure 2-1: Illustration of Two Distributions with Different Coefficients of Variation



We expect a certain amount of confidence in our measured results. Mathematically, *confidence* (or confidence level) describes how repeatable the measurement process is. If the desired confidence level is 90%, we would expect that our sampled measurement process will provide the same or similar result nine times out of ten. Increasing the desired confidence level requires expending more effort and often greater expense, the trade-off for increased reliability.

The related *confidence interval* is the range around a measurement that conveys how accurate the measurement is. If the confidence level is 90%, then there is a 90% probability that the true value of the population lies within the stated confidence interval.

The purpose of a sampled measurement process is to quantify a value that represents the entire population. While the average value of the measured samples is known exactly, sampling introduces some uncertainty as to the true average value of the population. The acceptable difference between the sampled measurement and the population average values is expressed as the *precision*. If we accept a 10% precision in our measurement, we are saying that the average value of the population is within ±10% of our measured value.

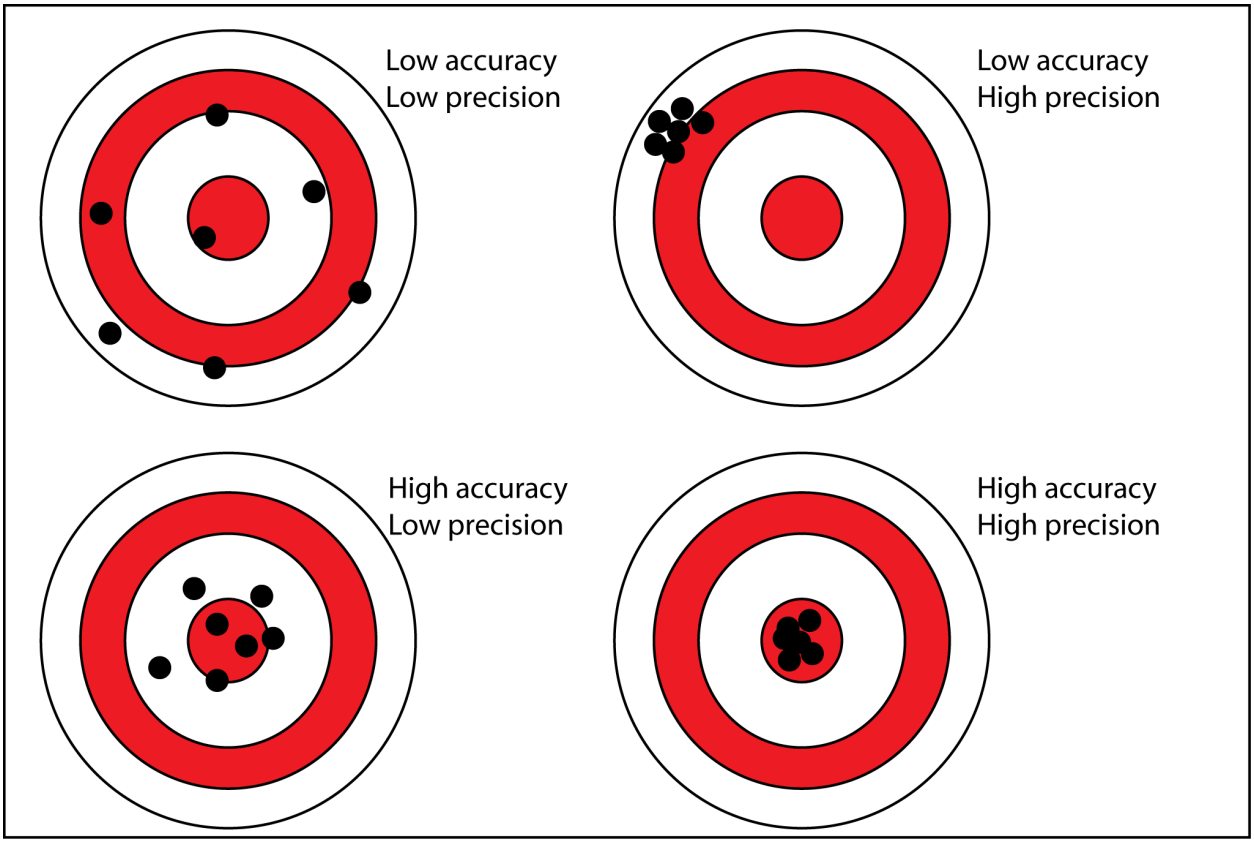
It is important to understand several points:

- **Confidence and precision must be reported together** – a specification of either one alone is meaningless. When reporting results, a statement of ±10% precision at 90% confidence means that: 1) the true value of the population is within ±10% of the measured value; and 2) we are 90% certain of this result. To be 99% certain of the result would require accepting a precision of ±16%. We could also state a greater precision, but with a much-reduced confidence.

➡ **Do not confuse precision and accuracy – they are not the same thing**. *Precision* refers to the repeatability of the measurement by equipment or process; *accuracy* refers to how close to the true value the measurement is. We often assume that a precise value is an accurate value. However, a biased measurement device or process can yield a very precise, but inaccurate, result. Figure 2-2 illustrates the difference between accuracy and precision.

➡ **Checking for bias requires an independent measurement device or process** – one cannot measure and calibrate with the same equipment! However, checking for systematic bias is beyond the guidance that can be provided here. For our current purpose, we will continue to assume that a precise value is an accurate value.

Figure 2-2: Accuracy vs. Precision



3. The Sampling Process

3.1. Process Steps

The following are the steps to be taken in a sampled measurement process:

1. **Define the population.** The first step in a sampled measurement process is to identify the population of interest. This could be all of the equipment affected by a retrofit or all of the buildings participating in a program. Usually an equipment inventory or participant database is available that defines the population.

2. **Decide if the population is homogenous or heterogeneous.** Consider the population and decide if the members are homogeneous or heterogeneous. Do they all have identical or nearly-identical characteristics, or are there characteristics that identify members of the population as unique? Are all fan motors the same size, or do they span 2- to 50-hp? Do all 20-hp motors serve the same purpose, or are some fire pumps and others driving conveyer belts? If the population can be considered homogenous, then use simple random sampling. If the population is heterogeneous, then either divide the population into homogeneous groups and use simple random sampling on each or apply stratified sampling to the entire population. Where the entire population is comprised of several distinct categories (strata), the population can be segregated into these categories or strata. Each stratum is then sampled as an independent sub-population, out of which individual elements can be randomly selected. Strata are selected by considering which categories might have the greatest influence on the primary outcome variable one wishes to measure. Stratified sampling yields a smaller total sample size (and cost), but requires more calculation and evaluation effort.

3. **Define the desired confidence and precision.** BPA recommends sampled measurement to ±10% precision at 90% confidence but allows some flexibility in the target precision in certain cases. Where individual measurements are expensive, or the populations have large variances (high CV), reducing the acceptable precision results in smaller sample sizes. Because precision is proportional to the square of the sample size, relaxing the precision from 10% to 20% results in a four-fold decrease in sample size. Increasing the precision from 10% to 5% requires a four-fold increase in sample size.

4. **Assume an initial coefficient of variation.** The sample size to achieve a specific precision and confidence depends on the population coefficient of variation CV. But the CV is often not known until after the measurements are taken, so the sample size cannot be determined in advance. To get around this paradox, we must assume a CV value to develop a sampling plan. If measurements have been performed previously and the CV is known for the characteristic of interest (e.g., residential square footage), one can initially apply the CV determined by earlier research. When estimating the true CV, a statistician should use a default value for the CV of not less than 0.5 for homogeneous samples and 1.0 for heterogeneous samples, until such a time that the population CV can be estimated from the project sample population (PJM 2016, 36). However, the actual CV must be calculated

afterwards and compared to the assumed value. (In some cases, researchers can continually monitor the CV of the population as the study progresses and additional data is acquired, allowing researchers to add additional samples if necessary.)

5. **Calculate the sample size.** With the population(s) defined and the precision and confidence targets set, calculate the required number of samples for each population or group. Where populations are small, or the desired precision is very high, the calculated sample size may be close to or even exceed the population size. The sample size can be adjusted using the *Finite Population Correction* equation, given below.

6. **Select random samples.** From the equipment or participant list, select samples at random until the desired number have been identified. A quasi-random selection process can be used, such as selecting every k^{th} element from a randomized list or selecting an element randomly and selecting every k^{th} element. Alternatively, a random number generator can be used to select samples.

7. **Implement the measurement process.** Using the randomly selected samples, take measurements or conduct surveys. Occasionally, the identified items may not be able to be measured (people are not home, etc.). Rather than giving up a measurement and reducing the actual sample size, have a plan to provide substitutions in the field (find another room, go to another house, etc.). Researchers and evaluators will typically select an initial pool of randomly drawn elements that is larger than the sample size required for the study. While there is always a concern that substitutions can introduce bias, it is usually better to accept a substitute member than reduce sample size. If the populations are truly homogeneous, substitutions will not introduce bias.

8. **Evaluate results.** For each population or measurement group, calculate: the *average value*, the *standard deviation*, the *coefficient of variation*, and the *size* of the final sample. Compare the CV of each group – if the actual CV is less than the assumed value, then the desired precision and confidence targets have been met. For each group, calculate the actual precision achieved. If time and budget permit, additional measurements can be taken to increase the sample size and improve the precision. In all cases, report the actual precision achieved.

9. **Learn from experience.** Because the sample size is a function of the assumed CV, use the calculated CV value for next year's sampling plan or for another, similar project researching the same primary outcome variable.

3.2. Random Samples

A sampling process is used to maximize the useful information from a minimum amount of data. With the information available on each system or building, investigate the defining characteristics to determine if the population is homogeneous or heterogeneous. Consider both the performance and the usage aspects. Table 3-1 illustrates performance and usage characteristics for typical equipment, systems, or building upgrades. This list is only intended to provide guidance; it is not exhaustive.

Note that parameters like *efficiency*, *kW/ton*, and *R or U value* are not directly (or easily) measurable quantities. Each requires multiple (and sometimes difficult) measurements, adding to the complexity, cost, and uncertainty of the measurement process. It is better to focus on directly measurable quantities like *power* (kW) rather than trying to rigorously quantify equipment performance.

If the population is reasonably homogeneous, random samples can be selected from the population. If either the usage or performance characteristics vary widely, the population should be divided into groups containing items of similar characteristics. However, too many groups with small numbers in each will be difficult to evaluate as well. The grouping should be a balance between resolution and number of groups. In general, no more than a dozen groups should be necessary.

Table 3-1: Performance and Usage Characteristics for Typical Technologies and Applications

Technology or Application	Performance Characteristics	Usage Characteristics
Lighting Fixtures	Power (watts)	Space type, operating hours
Motors	hp, RPM, rated efficiency	Constant speed and load, variable speed and load, operating hours
Rooftop Units	Capacity (tons), EER / SEER	Operating hours, local climate
Office Equipment	Function, power (Watts)	Operating hours, cycles
Water Heating	Power (Watts / kBTUh), gallons	Inlet and outlet temperatures, gallons/day
Envelope Improvements	R value, U value, SHGC	Indoor temperature, outdoor temperature (climate), window orientation
Homes	Size, type (SF, MF, condo), number of bedrooms, overall R and U values	Number of occupants, age of occupants / hours home, local climate, vintage, type of heating equipment
Small Commercial Buildings	Size, function, internal equipment	Schedule, number of occupants, local climate

If the population has been divided into groups, consider whether to use simple random sampling for each group or stratified sampling. Simple random sampling is easy to implement and the results from one group have no influence on another group. Stratified random sampling allocates resources more effectively, providing lower overall measurement costs. However, the calculation process is more complex and results from one group may affect the overall uncertainty of the entire process.

The acceptable level of confidence and precision needs to be established before the research has begun. BPA recommends measured sampling with 10% precision at 90% confidence level but allows some flexibility in the target precision in certain cases. Where individual measurements are expensive, or the populations have large variances (high CV), reducing the acceptable precision results in smaller sample sizes. (The evaluator should avoid revising downward the anticipated CV to reduce the required sample size.) Consider the savings expected from the project or program, the cost of each measurement, and the overall evaluation budget. Choosing an acceptable precision level may be an iterative process as the estimated measurement and evaluation cost is compared to the available budget or project/program savings. Absent a defined budget, initial M&V costs should be no more than 10% of the project budget; and annual M&V costs should be no more than

10% of the annual cost savings. The rationale is to provide enough verification to reliably assess project performance, while preventing the M&V effort from eroding its cost-effectiveness.

3.2.1. Simple Random Samples

Most projects use a simple random sampling process, even when multiple groups are identified. For each group, the parameter to be measured needs to be identified, the population size known, and the relative variance of the parameter estimated.

The initial sample size for a population is given by:

- Sample Size Equation: $n = \frac{Z^2 C_v^2}{P^2}$

  where: n = initial sample size

  Z = The z-statistic[7] for the desired level of confidence. Equal to 1.645 for a confidence level of 90%

  CV = assumed coefficient of variation of the item being measured

  P = desired precision, typically 10%

The value of n must be an integer. The calculated value should be rounded up to the nearest integer value (e.g., 67.2 becomes 68).

Because the coefficient of variation is not known in advance, it must be estimated first and calculated afterwards to see if the estimate resulted in the correct sample size. Table 3-2: lists some typical measured values and reasonable estimates of initial CV values that can be used to develop sampling plans. It is not intended to be a prescriptive, comprehensive, or definitive list; experience with previous similar projects (if available) will provide better information on likely CV values.

Table 3-2: Technologies Categorized by Typical Amount of Variance

Items With Low Variance (Assume CV = 0.25)	Items With Medium Variance (Assume CV = 0.5)	Items With High Variance (Assume CV = 1.0)
• Lighting fixture power • HVAC operating hours under energy management & control system or building automation system control	• Lighting fixture hours under manual control • Lighting fixture energy use under daylight harvesting control	• Lighting fixture hours under motion sensor control • Variable-speed or variable-load motor load factors

[7] Use of a z-statistic implies normality. The Central Limit Theorem shows that the means of sufficiently large random samples drawn from a population will follow a normal distribution, even if the population that is the source of the sample is not normally distributed. However, for sample sizes smaller than 30, the Central Limit Theorem begins to break down and the normality assumption no longer is valid. A t-distribution is the appropriate distribution for M&V practitioners to consider when drawing samples of fewer than 30 units. The t-statistic replaces the z-statistic in the sample size equation and is calculated using the degrees of freedom (sample size minus the number of estimates).

- Identical rooftop unit full load kW
- Constant-speed motor load factors (motors of similar hp)
- Office equipment hours/day or cycles/day

Where populations are small (< 500) and/or variances large (CV > 0.5), the calculated initial sample may be significantly greater than 10% of the population. If the initial sample size is more than 10% of the population, the sample size may be adjusted downward to account for the small population size. (This is obviously necessary if the sample size exceeds the population size!) However, the correction may still yield a sample size greater than 10% of the population. The finite population correction is:

- Finite Population Correction Equation: $n^* = \frac{Nn}{N+n}$

where: n^* = adjusted sample size

n = initial sample size

N = Population size

Table 3-3: illustrates how sample size depends on the assumed CV, all calculated at 10% precision and 90% confidence (z=1.645). Note that the assumed CV strongly influences the necessary sample size. Populations with low variances (CV ~0.25) can be measured using less than 20 samples, but populations with large variances (CV ~1.00) may need nearly 300 samples to achieve the same precision.

Table 3-3: Illustration of Dependence of Sample Size on Assumed CV

Sample Characteristics		Illustrative Value	
Precision		10%	
Confidence		90%	
Z-Statistic		1.645	
Assumed CV	0.25	0.5	1.0
Population Size, N		**Sample Size, n***	
3	3	3	3
5	4	5	5
10	7	9	10
25	11	19	23
50	13	29	43
100	15	41	74
250	16	54	131
500	17	60	176
Infinite	17	68	271

Once the sample size has been determined, the last step is to select the samples from the population. This can take the form of quasi-random sampling where each k^{th} element is selected from the population until the desired number is achieved (assuming the list is random with regard to the primary variable being measured). This method is acceptable so long as k is not so small that all of the samples are concentrated at the beginning of the list. The selected samples should be distributed fairly evenly across the population list.

For true randomness, a random number generator (such as the *RAND()* function in *Microsoft Excel*) can be used to assign a random value between 0 and 1. If 10% of the population is to be sampled, then selecting all items with random values between 0 and 0.1 (or any other range spanning 0.1) will yield a random sample of the appropriate sample size. It is helpful to add a few additional observations to the sample size so that the quota will still be met if some observations are compromised or invalidated by data collection or non-response shortfalls.

Also, a process should be developed in advance for selecting alternate samples in the field. Rather than ignore a location or piece of equipment because access is restricted (or for any other valid reason), allow the field inspectors to choose an alternative but similar sample, document why an alternate was selected, and what the actual sample was. Occasional sample substitution will yield far better results than rigid adherence to the selected random samples that cannot be measured.

3.2.2. Stratified Sampling

There are several reasons to use stratified sampling instead of simple random sampling:

- To cost-effectively allocate resources
- To reduce the overall uncertainty (increase the precision)
- To provide information about each stratum

Once the overall population has been identified, the characteristics of the members need to be investigated to see if the population is homogeneous and, if not, what groups within the population might be homogeneous. For example, a population of small commercial buildings that includes offices, banks, convenience stores, and self-storage buildings should be *stratified* based on building function. However, too many groups with small numbers in each will be difficult to evaluate as well. The grouping should be a balance between resolution and number of groups. In general, no more than a dozen groups should be necessary.

Next, the acceptable level of confidence and precision needs to be established. BPA recommends sampled measurement to 10% precision at 90% confidence but allows some flexibility in the target precision in certain cases. Where individual measurements are expensive, or the populations have large variances (high CV), reducing the acceptable precision (such as accepting 20% precision instead of 10%) results in smaller sample sizes.

The advantage of stratified sampling is that evaluating the precision at the project level instead of the group level reduces the overall sample size and cost relative to group level sampling. Stratified sampling achieves this cost reduction by allocating resources according to each stratum's contribution to the overall variance. Strata that contribute little to the overall variance – either

because they are small or because they have a low coefficient of variation – have a smaller sample size than larger or higher variance groups.

As in simple random sampling, an initial coefficient of variation for each stratum needs to be estimated. If all strata have the same CV, the sample size will be allocated proportional to the projected energy use or savings from each group.

For each project, the information shown in Table 3-4: is necessary:

Table 3-4: Recommended Format for a Stratified Sampling Plan

Stratum Name	Population	Estimated Savings	Assumed Coefficient of Variation
Group 1	N_1	kWh_1	CV_1
Group 2	N_2	kWh_2	CV_2
Group *i*	N_i	kWh_i	CV_i
Total	N	kWh_T	

If savings for each group cannot be estimated with a high degree of reliability, use the expected energy use in place of the estimated savings. After all, we measure energy use, not savings.

With this information, the total sample size *n* can be calculated as:

- Sample Size Equation: $n = \dfrac{[\sum_i (kWh_i \cdot CV_i)]^2}{\left[\dfrac{P \cdot kWh_T}{Z}\right]^2 + \sum_i \dfrac{(kWh_i \cdot CV_i)^2}{N_i}}$

  where: n = total sample size required

  kWh_i = estimated savings from group i

  CV_i = assumed coefficient of variation for group i

  P = desired precision, typically 10%

  kWh_T = estimated total savings

  Z = 1.645 for a confidence level of 90%

  N_i = population size of group i

While this equation looks intimidating, it is relatively straightforward to implement using a spreadsheet calculation tool. The total sample size must be an integer, so round up the result to the next highest integer. Note that for small populations (<500), it is possible that the total sample size may still exceed 10% of the population. However, the total sample size will still be less than that for group-level sampling.

Next, the total number of samples needs to be allocated to each group according to the expected contribution to the variance. Group sample sizes are calculated as:

- Equation For Group Sample Sizes: $n_i = n\left[\frac{kWh_i \cdot CV_i}{\sum_i kWh_i \cdot CV_i}\right]$

where: n_i = sample size for group i

n = total sample size

kWh_i = estimated savings from group i

CV_i = assumed coefficient of variation for group i

Again, the group sample size needs to be rounded up to the next highest integer. In rare cases where a single group has a very small population and a high CV, it is possible to calculate a stratum sample size greater than the population. In this event, simply cap the sample size at the population (measure all items). Even if the group has a high CV, it will make no contribution to the overall uncertainty if all items are measured.

Once the group sample sizes have been determined, selecting appropriate samples is the same as for simple random sampling.

3.2.3. Binomial Distributions

Binomial distributions are a special case sometimes encountered in energy efficiency projects and programs. Average values of the probability of encountering a specific state (e.g., pass/fail) can be calculated by assigning a value of zero to the fail (or absent) condition and one to the pass (or present) condition. However, the calculation of the standard deviation and uncertainty need to be treated differently than for a normally distributed population.

The definition of a binomial distribution is one where there are complementary probabilities of occurrence of a state. Let p be one condition and q the other. Then, by definition:

- $p + q = 1$

The standard deviation (s) and standard error (SE) of this type of population can be shown to be:

- Standard Deviation and Error: $s = \sqrt{pq}$ $SE = \sqrt{\frac{pq}{n}}$

where: n = sample size

The precision (P) of sampled measurement of the value of p is then:

- Precision of Sampled Measurement: $P = Z \cdot SE = Z\sqrt{\frac{pq}{n}}$

where: P = achieved precision

p = probability of state p (equal to 1-q)

q = probability of state q (equal to 1-p)

n = sample size

Z = 1.645 for a confidence level of 90%

Rearranging to determine sample size yields:

- Sample Size: $n = pq\frac{Z^2}{P^2} = p(1-p)\frac{Z^2}{P^2}$

  where: n = initial sample size

  p = probability of state p

  q = probability of state q

  Z = 1.645 for a confidence level of 90%

  P = desired precision

The value of n must be an integer, so the calculated value should be rounded up to the nearest integer value (e.g., 67.2 becomes 68). If p and q cannot be estimated, the safest assumption to make is $p = q = 0.5$, which will maximize the sample size. Once sample size has been determined, sample selection is like that for simple random sampling. In practice, binomial populations tend to be very large. Therefore, it is rare that the finite population correction would need to be applied, so it is not discussed here.

When selecting a desired precision, a rule-of-thumb is to target less than ½ of the expected smaller value of p or q. For example, evaluating a situation where the expected failure rate is 10% requires a target precision of ±5% or less. If the target precision were greater than the quantity being evaluated, it would be challenging to make inferences from the results.

A quick check (when working at the 90% confidence level)[8] to determine if the sample size is adequate is to see that the conditions $np \geq 5$ and $nq \geq 5$ are met. For example, if p is estimated to be 0.1, then the minimum sample size needs to be 50 to ensure that (50)*(0.1) ≥ 5 is true. Failure to meet this condition will result in under-sampling and unreliable results. If p is expected to be very small, the corresponding sample will be very large (e.g., $p = 0.01$ results in n=500).

3.3. Sample Size and Results Calculator

An *Excel* file entitled *BPA Sample Size and Results Calculator.xls*[9] accompanies this protocol and is available from BPA upon request.

[8] When working at the 95% confidence level a quick check of $np \geq 10$ and $nq \leq 10$ is a more appropriate threshold.

[9] Mark Stetz, the primary author of Version 1.0 of the *Sampling Guide*, developed the calculator.

4. Applications and Examples

4.1. Application 1: Simple Random Sample

An industrial energy project replaced 50 standard-efficiency conveyer-belt motors with premium-efficiency motors. All motors are 10-hp in size, serve similar loads, and have similar operating schedules. The measurement and verification plan calls for measuring the actual energy use of a motor over a 24-hour period in the baseline case and retrofit case.

Because all baseline motors are similar in size, function, operating hours, and nameplate efficiency, they can be treated as a single homogeneous group. The same applies to the retrofit motors. Simple random sampling will be used, and the selected samples will have true-power, data loggers connected to them for short intervals. If the motors varied in size or served different loads, then the population would need to be divided into distinct groups.

Because the incremental demand reduction from each motor is small, the energy use must be measured to a high degree of precision to reliably assess the savings. The proposed measurement process is to be carried out with 5% precision at 90% confidence to minimize uncertainty in the savings estimate (instead of the usual 10% precision).

Although conveyer-belt motors are subject to large instantaneous load changes, the energy use over the course of a day is expected to remain relatively constant. The expected variance on daily energy use among all motors is therefore small, as each motor serves a portion of the same conveyer line and is subject to the same variations. The coefficient of variation is assumed to be 0.25 for purposes of selecting the sample size.

Calculating, the initial sample size from the equations for $P = 5\%, C = 90\%, CV = 0.25,$ *and* $N=50$ is 29. This sample size needs to be repeated twice – once in the baseline case and once in the post-retrofit case. Given the magnitude of the savings and the M&V budget, installing 29 data loggers twice is considered too expensive. It is decided to accept a reduced measurement precision, recognizing that the resulting uncertainty in the savings calculation would be significantly greater. The target precision is reduced to a more typical value of 10%. Referring to the sample-size table, the proposed sample size for $P = 10\%, C = 90\%, CV = 0.25,$ *and* $N=50$ is 13. Two additional samples are added to ensure sufficient coverage, bringing the total sample size to 15. Repeating this process twice (30 measurements) falls within the proposed budget.

From the motor inventory, every third motor is selected, so that fifteen motors are identified. During the data logger installation, two of the selected motors are found to be in an overhead location that would require a scissor jack to reach. Instead, two nearby motors at ground level are substituted and noted.

After removing the data loggers, the energy use over a 24-hour period is determined for each. One data logger was misconfigured, resulting in unreliable data; the data was discarded. Table 4-1 presents the results.

Table 4-1: Example Simple Random Sample Plan

Sample Characteristics	Illustrative Value
Average Energy Use, kWh	89
Standard Deviation, kWh	31
Coefficient of Variation, Standard Deviation / Average	0.35
Actual Sample Size	14

The measured coefficient of variation is 0.35 – greater than the assumed value of 0.25 – indicating that the target precision level was not met. The actual precision is determined by "unadjusting" the sample size and then calculating the resulting precision.

With a population of $N=50$ and an adjusted sample size n^* *of 14*, the unadjusted sample size n is:

- $n = Nn^*/(N-n^*) = (50)(14)/(50 - 14) = 19.4 \approx 20$

The precision for $n=20$ and $CV = 0.35$ at 90% confidence is:

- $p = ZCV / n^{1/2} = (1.645)(0.35) / (14)^{1/2} = 15\%$

To achieve the desired target precision would require

- $n = Z^2CV^2 / p^2 = (1.645)^2(0.35)^2 / (0.10)^2 = 33.1 \approx 34$
- $n^* = Nn/(N+n) = (50)(34) / (50 + 34) = 20.2 \approx$ **21 samples** (adjusted for population size)

The cost of seven additional samples needs to be checked against the M&V budget and schedule to ensure that additional measurements can be taken. The measured CV for the baseline motor measurements will be applied to the new motor measurements as well, bringing the entire process to a total of 42 measurements (21 baseline, 21 post-retrofit), budget permitting. The greater variance than originally assumed increased the required sample size. The M&V plan calls for measuring the same motors after replacement rather than a new random sample. The justification for this decision is beyond the guidance of this document.[10]

4.2. Application 2: Stratified Random Sample

A utility provided incentives for an HVAC upgrade program that offset the costs of high-efficiency rooftop units. Incentives are based on the level of improvement relative to ASHRAE 90.1-2016 specifications. To assess the actual program savings, a representative sample of replaced rooftop units is to be measured to verify the actual energy use and estimated savings from each unit and for the program.

[10] Measuring the same motors allows the use of a paired *Student's t-test* to verify the statistical validity of the savings reduction. Using a different random sample would allow only an unpaired *Student's t-test*, which may not be as sensitive as a paired test.

Energy saved will be calculated based on measured energy use and the rated EER relative to ASHRAE 90.1-2016 baseline as follows:

$$kWh_{saved} = kWh_{measured}\left(\frac{EER_{rated}}{EER_{90.1}} - 1\right)$$

where: kWh_{saved} = actual energy savings from each unit during the monitoring period

$kWh_{measured}$ = measured energy use of the new unit during the monitoring period

EER_{rated} = nameplate EER rating at ARI conditions

$EER_{90.1}$ = ASHRAE 90.1 EER rating at ARI conditions

The challenge is to measure the energy use on a sufficient number of rooftop units to achieve an overall precision in the savings estimate of 10% at 90% confidence. Over 400 rooftop units of various sizes have been deployed at a number of different building types; significant variation is expected in their loads and operating hours, depending on the building function and schedule. Since only total program savings need to be known with a high degree of precision, the population of rooftop units can be stratified by building type and function. This is possible because all the rooftop units are similar in size for each building type. If rooftop units were of radically varying sizes within each group, further strata would be required (e.g., Mall remote terminal unit (RTU) 10 tons, Mall RTU 50 tons, etc.)

Initial savings can be estimated by making assumptions about operating hours for each space type and function. Additionally, the initial coefficient of variation for each group will have to be assumed. Based on estimated and assumed values shown in Table 4-2, values and calculated fields can be constructed based on the equations presented previously.

Table 4-2: Example Stratified Random Sample Plan

Usage Group	Estimated Energy Savings kWh_i	Assumed CV_i	Population Size N_i	$(kWh_i * cv_i)$	$(kWh_i * cv_i)^2 / N_i$	Samples Required
Office	200,000	0.25	100	50,000	25,000,000	5
Schools	120,000	0.50	50	60,000	72,000,000	6
Malls	500,000	0.25	200	125,000	78,125,000	11
Grocery Stores	150,000	0.50	75	75,000	75,000,000	7
Worship	10,000	1.00	10	10,000	10,000,000	1
Total	980,000		435	320,000	260,125,000	30
		Total Samples			27	

Note that although *Malls* has a low CV, it has the largest sample size because it has the greatest contribution to the overall variance. The *Worship* group has the smallest sample size because it has the lowest contribution to the overall variance despite its high CV. The total initial sample size of 27 is allocated according to each group's variance; additional rounding upwards increases the

total sample size to 30. Note that simple random sampling on each group would require a total of 106 samples to achieve 10% precision at 90% confidence. Stratified sampling has reduced the level of effort by more than a factor of three!

Based on the sample size table, data loggers are deployed at locations selected at random from the project inventory. After one month, the data is retrieved and evaluated. The energy use of each rooftop unit is used to estimate the energy savings based on its nameplate rating relative to the ASHRAE 90.1 baseline. The average energy savings and standard deviation for each group is calculated and the group CV determined. Data logger collection errors reduced the sample size from 11 to 9 in the *Malls* group. Based on the previously defined equations, Table 4-3 can be constructed:

Table 4-3: Example Stratified Random Sample Results

Usage Group	Population Size N_i	Samples Taken	Measured Energy Savings, kWh	Measured Standard Deviation	Actual CV	Contribution to Variance, Overall Precision
Office	100	5	220,000	66,000	0.30	8.28E+08
Schools	50	6	110,000	70,000	0.64	7.19E+08
Malls	200	9	400,000	90,000	0.23	8.60E+08
Grocery Stores	75	7	170,000	80,000	0.47	8.29E+08
Worship	10	1	12,000	0	0.00	0.00E+00
Total	435	28	912,000			10.3%

The *Worship* group has no standard deviation because only one measurement was taken. It therefore does not contribute to the total variance despite the originally assumed CV of 1.0. The resulting overall precision of 10.3% is close enough to the target level of 10% that the precision goal has been achieved.

The measured CVs could be used to develop next year's sampling plan if one is necessary.

4.3. Application 3: Binomial Distribution

A utility company offers standardized incentives for lighting upgrades but does not want to provide incentives for repairing failed fixtures. The utility requires a sampled survey of lighting fixtures to document that the facility contains 90% or more operating, functional fixtures. For facilities where the number of operating fixtures is less than 90%, the utility will adjust the number of fixtures eligible for incentives, explaining that they do not offer incentives to subsidize maintenance.

Because a fixture is either *operating* or *failed*, the population distribution is binomial. (To remove ambiguity, the utility defines any fixture with a single failed lamp as *failed*, regardless of the number of lamps in the fixture.) At a large facility where the population of eligible fixtures exceeds 10,000, an inspector needs to determine how many fixtures are *operating* and how many are *failed* to determine whether the failure rate exceeds 10% and needs to have 90% confidence in the results. Because the target precision needs to be smaller than the failure rate, a target precision of 2% was

selected, meaning that a measured error rate of 10% could reflect a true value between 8% and 12%.

The population of fixtures does not need to be grouped by type or schedule, so long as it is of a type eligible for an upgrade incentive. Therefore, separation into discrete groups or strata is not required.

Because the population N is very large, and the estimated values of p and q are known, the initial sample size is calculated as:

- $$n = pq\frac{Z^2}{P^2} = p(1-p)\frac{Z^2}{P^2}$$

where: n = initial sample size

p = probability of state p (0.10)

q = probability of state q (0.90)

Z = 1.645 for a confidence level of 90%

P = desired precision (0.02)

The calculation results in a sample size of 609, which also meets the requirements that $np > 5$ and $nq > 5$. From the facility description, 300 spaces are randomly selected, where each space has at least two fixtures. The spaces are inspected, noting at each how many fixtures are present and how many are considered *operating* and how many *failed*. Because many spaces have more than two fixtures, the actual sample size will be much greater than 609.

When the survey is completed, it is found that 800 total fixtures were surveyed, 88 fixtures failed, and 712 were functional. This yields the following results:

n = actual sample size, 800

p = probability of state p (failed), 88/800 = 0.11

q = probability of state q (operating), 712/800 = 0.89

Although the calculated failure rate of 11% exceeds the threshold of 10%, the precision needs to be calculated to see if the measured results are different from 10% in a statistical sense. The error of the estimate is calculated as:

- $$P = Z\sqrt{\frac{pq}{n}}$$

where: P = Precision

Z = 1.645 for 90% confidence

p = probability of state p, 0.11

q = probability of state q, 0.89

n = sample size, 800

The precision is 0.0182 (1.82%). This indicates that the true value of the failure rate is 0.110 ±0.0182, meaning that we are 90% certain that the true value of the measurement lies between 0.0918 and 0.1282 (9.18% to 12.82%). Because the lower bound of this range is less than 10%, we cannot state that the measured result is statistically different from the 10% threshold, so we accept the results and allow incentives for all fixtures. With 800 samples, 96 failures (12% failure rate) would be required to exceed the uncertainty bounds and adjust the number of fixtures eligible for incentives.

5. Minimum Reporting Requirements

The user of this document should follow the minimum reporting requirements of the M&V protocol for which the sampling is undertaken. In addition, the sampling activities need to be supported in the M&V plan and verification report by a thorough statement of the assumptions made about the population and the nature of the sample. (The verification report need not duplicate sampling information reported in the M&V plan if the verification report appends the M&V plan.)

- **Population Characteristics:** When testing for heterogeneity and determining if the sample should be stratified, the engineer must consider the defining aspects of the systems and facilities under study. Most commonly, the defining characteristics include performance and usage. These and any other defining characteristics should be described and any needs for stratification should be discussed.
- **Distribution:** The M&V plan and verification report should state the assumed probability distribution of the population characteristics. If the engineer has reasons to assume a distribution other than the normal distribution or the binomial distribution, these reasons should be stated in support of the reported distribution.
- **Population Size (N):** The population size of the group or groups needs to be defined.
- **Sample Type:** The documents should specify whether a simple random sample or a stratified sample is required. If stratification is necessary, the M&V plan and verification report should state whether each group should be (or was) treated separately or if the strata should be (or were) used in a combined model.
- Variability:
  - **For simple and stratified sampling**, state the assumed coefficient of variation (CV) for each group and how it will be calculated after sampling is complete.
  - **For binomial sampling**, state the assumed presence / absence rate, or success vs. failure rate.
- **Precision (P):** State the target precision level and how it will be calculated after sampling is complete.
- **Confidence Level (Z):** The desired confidence level (such as 90%).
- **Sample Size (n):** The initial sample size, per strata if multiple strata, and after applying the finite population correction in the case of smaller populations.
- **Finite Population Correction:** State whether the finite population correction was used.
- **Power Analysis:** If a power analysis is warranted (see Section 2.6), its parameters should be described; however, power analysis and its reporting requirements are beyond the scope of this guide.

In addition to these elements, the M&V plan needs to discuss the planned course of action if the target precision is not met, and the verification report needs to address the approach followed. There are three options:

- Report the poorer precision level in the verification report;
- Increase the sample size (take more measurements during field work); or
- Convert a stratified sampling plan to a simple sampling plan and report the overall precision; note that one is then no longer able to describe strata findings with any statistical confidence.

6. References and Resources

ASHRAE. 2014. *ASHRAE Guideline 14-2014 – Measurement of Energy, Demand, and Water Savings*. Atlanta, Ga.: American Society of Heating, Refrigerating and Air-Conditioning Engineers.

Purchase at: https://www.techstreet.com/standards/guideline-14-2014-measurement-of-energy-demand-and-water-savings?product_id=1888937

ASHRAE. 2016. *ASHRAE Standard 90.1-2016 – Energy Standard for Buildings Except Low-Rise Residential Buildings*. Atlanta, Ga.: American Society of Heating, Refrigerating and Air-Conditioning Engineers.

Purchase at: *https://www.techstreet.com/standards/ashrae-90-1-2016-i-p?product_id=1931793*

Brase, Charles Henry, and Corrinne Pellillo Brase. 2009. *Understandable Statistics: Concepts and Methods* (Ninth Ed.). New York, N.Y.: Houghton Mifflin Company.

Cochran, William Gemmell. 1977. *Sampling Techniques* (Third ed.). Hoboken, N.J.: John Wiley & Sons, Inc.

De Leeuw, Edith D., Joop Hox, and Don A. Dillman (editors). 2008. *International Handbook of Survey Methodology.* Mahwah, N.J.: Lawrence Erlbaum Associates.

IPMVP. 2012. International Performance Measurement and Verification Protocol Volume 1: Concepts and Options for Determining Energy and Water Savings. EVO 10000 – 1:2012. Washington, D.C.: Efficiency Valuation Organization.

Available at: *https://evo-world.org/en/products-services-mainmenu-en/protocols/ipmvp.*

Kish, Leslie F. 1965. *Survey Sampling.* New York, N.Y.: John Wiley & Sons, Inc.

PJM Forward Market Operations. 2010. *PJM Manual 18B: Energy Efficiency Measurement and Verification.* Revision: 03, Effective Date: November 17, 2016. Norristown, Penn.: PJM Interconnection.

Available at: http://www.pjm.com/-/media/documents/manuals/m18b.ashx

Schaeffer, Richard L., William Mendenhall III, and R. Lyman Ott. 2006. *Elementary Survey Sampling* (Sixth Ed.). Pacific Grove, Calif.: Duxbury Press.

TecMarket Works Team. 2006. California Energy Efficiency Evaluation Protocols: Technical, Methodological, and Reporting Requirements for Evaluation Professionals. April 2006. San Francisco, Calif.: California Public Utilities Commission.

Available at: www.cpuc.ca.gov/PUC/energy/Energy+Efficiency/EM+and+V.

Thompson, Steven K. 2002. *Sampling.* (Second Ed.) New York, N.Y.: John Wiley & Sons, Inc.

7. Appendix: Glossary of Statistical Terms

This Glossary provides definitions for the statistical terms used in this *Sampling Reference Guide*. Additional M&V terms are defined in the companion document *Glossary for M&V: Reference Guide.*

Accuracy: An indication of how close the measured value is to the true value of the quantity in question. Accuracy is different from precision.

Binomial Distributions: A population consisting of items that can only occupy one of two states (i.e., present/absent, on/off, pass/fail).

Coefficient of Variation (CV): An indication of how much variability or randomness there is with any given data set. It quantifies variation within the population relative to the average and is dimensionless. The larger it is, the more variation there is in the population relative to the average. It is calculated as the ratio of the standard deviation to the average:

$$CV = \sigma / \bar{x}$$

Confidence Interval: A range of uncertainty expected to contain the true value within a specified probability. The probability is referred to as the *confidence level*.

Confidence Level: A population parameter used to indicate the reliability of a statistical estimate. The confidence interval expresses the assurance (probability) that given correct model selection, the true value of interest resides within the proportion expressed by the confidence interval.

Error: Deviation of a measurement from the true value.

Finite Population Correction Equation: When calculating sample size, the finite population correction equation reduces the number of samples required (n) when the population size is small (N) relative to the number of samples (n) assuming a very large population. It is calculated as:

$$n^* = \frac{Nn}{N+n}$$

Heterogeneous Population: A population in which members' characteristics vary and thus there are subpopulations of distinct types.

Homogeneous Population: A population in which all members have similar characteristics and can be considered of the same type.

Independent Variable: Also termed an *explanatory* or *exogenous variable*; a factor that is expected to have a measurable impact on the dependent, or outcome variable (e.g., energy use of a system or facility).

Mean: The most widely used measure of the central tendency of a series of observations. The Mean ($\overline{Y}$) is determined by summing the individual observations (Y_i) and dividing by the total number of observations (n), as follows:

$$\overline{Y} = \frac{1}{n}\sum Y_i$$

Normal Distribution: A continuous and symmetric population distribution in which the frequency of occurrence decreases exponentially as values deviate from the mean (or central) value. In a regression equation, the distribution of errors (residuals) at a given value of x is a normal distribution and the mean of residuals is zero. It is also referred to as a *Gaussian* or *bell curve*.

***p*-value:** The probability that a coefficient or dependent variable is not related to the independent variable. Small *p*-values, then, indicate that the independent variable or coefficient is a significant (important) predictor of the dependent variable in a regression model. The *p*-value is an alternate way of evaluating the *t*-statistic for the significance of a regression coefficient and is expressed as a probability.

Precision: The indication of the closeness of agreement among repeated measurements; a measure of the repeatability of a process. Any precision statement about a measured value must include a confidence level. A precision of 10% at 90% confidence means that we are 90% certain the measured values are drawn from samples that represent the population and that the "true" value is within ±10% of the measured value. Because precision does not account for bias or instrumentation error, it is an indicator of predicted accuracy only given the proper design of a study or experiment.

Quasi-Random Sampling: A sampling technique in which each k^{th} element is selected from the population until the desired number is achieved.

Random Number Generator: A computer algorithm that generates quasi-random numbers that can be used for sample selection or for other purposes. Note that computers cannot generate truly random numbers and that algorithms may generate results with either uniform or quasi-normal population distributions.

Reliability: When used in energy evaluation, refers to the likelihood that the observations can be replicated.

Sampling: The process of identifying which and how many items to measure and evaluating the results to assess reliability.

Simple Random Sampling: A sampling technique that draws representative samples from a single population where any sample is expected to be representative of the entire population.

Standard Deviation (s): The square root of the variance, which brings the variability measure back to the units of the data. (With variance units in kWh^2, the standard deviation units are kWh.) The sample standard deviation (s) is defined as:

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

Standard Error (SE): An estimate of the standard deviation of the coefficient. For simple linear regression, it is calculated separately for the slope and intercept: there is a *standard error of the intercept* and *standard error of the slope*. SE is defined as:

$$SE_{\bar{x}} = \frac{s}{\sqrt{n}}$$

Stratified Sampling: A sampling technique used when the population is not homogeneous and needs to be segregated by some defining characteristic. Sample size for each stratum is determined by the contribution of that stratum to the overall variance such that the total sample size is minimized.

***t*-statistic:** A measure of the probability that the value (or difference between two values) is statistically valid. The calculated *t*-statistic can be compared to critical *t*-values from a *t*-table. The *t*-statistic is inversely related to the *p*-value; a high *t*-statistic (t>2) indicates a low probability that random chance has introduced an erroneous result. Within regression, the *t*-statistic is a measure of the significance for each coefficient (and, therefore, of each independent variable) in the model. The larger the *t*-statistic, the more significant the coefficient is in the estimation of the dependent variable. The t-statistic is calculated as:

$$t_{\hat{\beta}} = \frac{\hat{\beta} - \beta_0}{s.e.(\hat{\beta})}$$

Uncertainty: The range or interval of doubt surrounding a measured or calculated value within which the true value is expected to fall within some stated degree of confidence. Uncertainty in regression analysis can come from multiple sources, including *measurement uncertainty* and *regression uncertainty*.

Variance (S^2): A measure of the average distance between each of a set of data points and their mean value, and it is equal to the sum of the squares of the deviation from the mean value, or the square of the standard deviation. Variance is computed as follows:

$$S^2 = \frac{\sum(Y_i - \bar{Y})^2}{n-1}$$

z-statistic: (Also known as the *Standard Score*.) The z-statistic indicates how many standard deviations an observation or datum is above or below the population mean. It is a dimensionless quantity derived by subtracting the population mean from an individual raw score and then dividing the difference by the population standard deviation.

$$z = \frac{x - \mu}{\sigma}$$

Bonneville Power Administration

DOE\BP-4353 • July 2018