Sample Size Selection in Energy Efficiency Research and Evaluation – The Use and Abuse of the Coefficient of Variation

Andie Baker, Research Into Action, Portland, Oregon[1]

ABSTRACT

Energy efficiency research and evaluation have become more mathematical and statistical in nature as computational methods have augmented traditional experimental approaches. As such, it is no longer sufficient to base sample sizes on available funding alone, nor on energy efficiency research/evaluation routine practices. Deciding the number of residences, instruments, or appliances to be audited, metered, or tested in some recent Northwestern United States studies has seemed daunting, sometimes compounded by an incomplete understanding of the trade-offs between population variability and study confidence. Cochran's formula (Cochran 1977), used for the proportional sampling of large populations, is frequently applied in energy efficiency research and evaluation for determining statistically valid sample sizes.

Using this formula, researchers calculate sample sizes based on the degree of heterogeneity in the test population and the acceptable experimental risk. The coefficient of variation (CV) in the formula expresses the test population's diversity on the characteristic of interest (e.g., residential square footage), and CV is defined mathematically as the standard deviation of the characteristic's values divided by their mean. The CV for any population variable often is often unknown and must be estimated. This is where problems in sampling may arise. We have observed that, when funders require a rigorous 95/5 study[2] and have fixed budget parameters, population variances sometimes are underestimated. This can create studies that are both under-sampled and underpowered. The inappropriate use of low CVs in calculating sample sizes for energy efficiency studies raises the methodological and interpretive problems outlined in this paper. This paper further cites creative strategies applied in current research that attempt to ameliorate the effects of underestimating the CV in sampling calculations, and points to the need for funders, consultants, and oversight committees to adopt more sophisticated statistical approaches.

Competing Perspective in Energy Efficiency Research and Evaluation

Sound energy research and evaluation require a balance of competing considerations, particularly funding. Whether we conduct impact or process evaluations of efficiency programs or design research to determine savings potential or market characterization, we seek to know more about how well energy efficiency measures work. In determining the level of rigor required for a study, we first consider how the study might be used. The degree of confidence we require in the results of a study or evaluation depends on the applications planned; a higher degree of risk (usually financial or safety) necessitates a greater degree of confidence in the study outcome. For example, studies underlying the feasibility, safety, or cost-effectiveness of a new technology may require greater rigor due to increased risks of misinterpretation or misapplication of the data. Informational or initial investigative studies, such as appliance saturation studies, may require a lower degree of rigor, since there is less immediate risk due to any misapplication of study results. The costs

[1] Research performed while the author was employed at Tacoma Power, Tacoma, Washington.
[2] A study that is designed and sampled to meet a 95% Confidence Interval and a 5% Margin of Error.

associated with evaluation and research often increase exponentially in parallel to increases in required rigor, so the need for rigor must be carefully weighed against cost and purpose.

Organizations of various types may undertake energy efficiency research and evaluation data for very different purposes. Some non-profit consortia conduct research to inform their efforts to transform a market, while many utilities, now pursuing energy conservation as a resource in lieu of the purchase of other, more expensive and environmentally costly energy forms, use data to inform their investment. These utilities may require studies conducted to yield greater statistical confidence and precision to support the design of cost-effective programs that will consistently achieve expected, and sometimes required, savings. In the United States, the Energy Independence Act[3] (also known as I-937) passed by Washington State voters in 2006 mandates that qualifying Washington utilities (those with more than 25,000 customers) pursue all available cost-effective conservation, and that these utilities offer proof that they have met the very aggressive savings targets established for each biennium. (Utilities are subject to a significant monetary penalty for each megawatt hour they fall short of targets.) Also, because energy efficiency is considered a resource, utilities advocate that more rigorous research and evaluation are needed – not only to insure that required savings targets are met, but that new equipment/programs will succeed (in terms of customer satisfaction and cost-effectiveness) and to reduce the risk that they will be forced to buy expensive emergency power.

It is not surprising, then, that utilities are apt to demand that their evaluation consultants and contractors faultlessly design well-powered research on which to base the design and management of their energy efficiency programs. (Foremost in the minds of utilities managers are the potential costs and liabilities that may arise from efficiency programs founded on poorly designed, biased, inaccurate, or underpowered research.) Energy efficiency, much like other fields, has become increasingly reliant on statistically sound data, and it is critical that organizations funding research and evaluation possess the statistical expertise to ensure that the parameters of their research requests (i.e., level of rigor, expected sample sizes, funding) are achievable. Lacking statistical sophistication, funders may risk setting conditions in their research requests that proposers cannot possibly meet; although consultants that also lack that statistical expertise may insist they can secure information with a level of rigor that, in fact, is unattainable.

Consultants, hired to conduct research/evaluation primarily are concerned with meeting funders' requirements while providing a high-quality product, meeting deadlines, managing costs, achieving adequate profit margins, and encouraging repeat business. They do not bear the research-related risks utilities do; having met a funder's stipulations, they are far less reliant on the outcome or applicability of the research/evaluation they conduct. Consultants view rigor primarily as a contractual (and statistical) requirement, within which they must plan and reconcile the expense of conducting the work profitably. It is critical that consultants have the statistical expertise to determine whether a funder's request regarding rigor, sample size, and budget can be met.

Sample sizes may be a contentious issue for a number of reasons, all of which revolve around increased costs to both funding agencies and consultants; larger samples increase costs for funders and may limit the profit margin a consultant might hope to achieve. Some consulting firms may prefer to divert funds into overall project execution in order to deliver a superior product, believing statistically required samples to be unnecessarily large at the expense of what they may consider to be "more important" aspects of the work. Related to this, the energy efficiency field appears to be in the midst of "growing pains," as some long-practicing research and evaluation professionals assert that their experience qualifies them to determine sufficient sample sizes, whether or not these meet statistical requirements. Others have embraced statistical software packages and technologies that have advanced the field's ability to determine and meet statistical

[3] The Washington State Energy Independence Act as codified at RCW 19.285, from http://apps.leg.wa.gov/rcw/default.aspx?cite=19.285

rules. Whatever the etiology, some consultants' and utilities' fundamental objectives seem to be in conflict almost as often as they are in sync. Recently, we observed that, when a large-scale study had to meet 95/5 criteria while keeping sample sizes within budget parameters, the coefficient of variation, which is the element of sampling calculations most susceptible to manipulation, was biased downward.

Statistical Considerations

Although no study is "perfect," since there is always the element of chance involved, studies that are powered to meet a 95/5 level of confidence and precision have long been regarded as the gold standard in much of the academic and research world. In the Northwestern United States, it seems that some practitioners in the energy efficiency field remain convinced that energy conservation research and evaluation must achieve this standard.

Energy efficiency is critical to our world, but a much more practical approach to powering studies and selecting sample sizes for energy research and evaluation might better serve this industry. In truth, it is less likely that this standard need be imposed for most types of efficiency evaluation/research, and due to the large sample sizes associated with this level of rigor, 95/5 studies often are nearly impossible to fund. Unlike medical research, energy evaluation and research conducted to a lower degree of confidence and precision likely will not cause any deaths. In many cases, a 95/7, or even a 90/10 study that offers +/- 10% accuracy and confidence, provides enough rigor for the purposes of energy efficiency, although it can be surprisingly difficult to overcome the notion that a study is valuable only if it meets the 95/5 level of confidence and precision.

The practical challenges of rigor are revealed when one first attempts to balance considerations of precision, risk, sample size, and expense. To many who are unfamiliar with the rules and practices of statistical sampling, the larger sample sizes required when rigor increases seem extreme, particularly when one considers that, in the absence of a stratification/clustering plan, the unadjusted sample is the total of the sample calculated, multiplied by the number of domains within the population. Table 1 displays the sample required at differing levels of precision and confidence for a study in which the population is comprised of a single domain (one that is absent distinguishable layers of variability within the characteristic of interest).

Table 1. Comparison of Rigor and Required Sample Sizes*

Coefficient of Variation	**90/10 (Z =1.64)**	**95/5 (Z =1.96)**
0.2	11	62
0.3	24	138
0.4	43	246
0.5	67	384
0.6	97	553
0.7	132	753
0.8	172	983
0.9	218	1,245

*Assumes a large population and that the population is comprised of a single domain

The number of units required for a sample is a function of the degree of variability in the population of interest. The pressure on consultants to meet 95/5 criteria sometimes can lead to the application of creatively derived CVs, resulting in methodological and interpretive problems. To better understand the importance of the coefficient of variation, we first briefly review sample size calculation.

Sample Size Calculation

There are multiple approaches to determining sample size, which are specific to the type of study being conducted. Most approaches are based either on the margin of error that can be tolerated, or on the precision required for implementing the study's findings. To calculate the required sample size in a descriptive study, one first selects the **desired levels of confidence and precision** and then determines the degree of variability in the population on the attribute/s of interest. For energy efficiency research involving large populations, one of the most frequently employed equations is Cochran's formula (Cochran, 1977):

$$n = \frac{Z^2 \sigma^2}{e^2}$$

where n is the sample size, Z is the abscissa of the normal curve that cuts off an area at the distribution tails (desired confidence level), e is the desired level of precision, and σ is the variance within the population (expressed as the coefficient of variation).

Confidence

The confidence level, or level of risk associated with a study is derived from the Central Limit Theorem. The theorem is founded on the idea that when a population is repeatedly sampled, the average value of the attribute obtained by those samples is equal to the true population mean (average). The values obtained by these samples are distributed normally about the true value, with some measurements having a higher value and some obtaining a lower value than that of the true population mean. In a normal distribution, approximately 95% of sample values are within two standard deviations of the true population mean. If a 95% confidence level is specified, a study is designed with the intent that 95 of every 100 samples will be representative of the true population value (e.g., 95 of 100 residences, meter readings, etc., are representative of what would be found in the population as a whole).

Precision

The level of precision (*e*), sometimes referred to as sampling error, is the range in which the true mean value of the population is expected to lie. Precision is expressed in percentage points (e.g., +/- 5%). For example, an energy study might report that 60% of residents have adopted a recommended practice with a precision rate of +/-5%, meaning that between 55% and 65% adopted the practice.

The principle challenge of sizing a sample via Cochran's equation (based on the population mean) is that a measure of the population variance (σ) is required. Absent available measures of population variability, estimates must be employed.

The Coefficient of Variation

The third criterion, **coefficient of variation (CV),** is a measure of the variability of the population. The CV is calculated by dividing the standard deviation of a population by its mean. A heterogeneous population — one not evenly divided by the presence or absence of an attribute – will be more difficult to

measure precisely. In general, the more variability one expects in the population, the larger the sample required.

The challenge we face in energy efficiency is that we often do not know, and frequently are unable to measure, the variability within a target population. As a result, the CV to be input into a sample calculation is unknown. When the CV is unknown, energy efficiency researchers/evaluators may implement a default value ranging from not less than 0.5 for homogenous populations (which are uniform per the classification criteria), to 1.0 for samples that are heterogeneous. This method is frequently employed until the time that a CV can be estimated from the project sample population (PJM 2010).

Where possible, CVs (or even sample sizes) are perpetuated from one energy efficiency study to the next. This strategy is valid only if the works referenced are identical for precision, confidence, and variability; use the same frame of reference (e.g., existing housing vs. new construction); measure the same attributes as those considered for the current study; and under-sampling did not occur in the referenced studies. When such references are unavailable, some contractors use their experience and judgment to estimate CV, although this method is least preferable.

Intuitive assessments of the variability of a population are nearly always are wrong (and can lead to under-sampled studies that deplete funding before the confidence and precision levels are met). That there is nearly always a tendency to overlook or underestimate chance, variability is but one instance of a pervasive "overconfidence bias" (Fischoff, Slovic, & Lichtenstein 1977). This is demonstrated regularly when experienced professionals (including statisticians) attempt to intuitively judge population variability, and are unable to do so accurately (Tversky and Kahneman 1971).

In the next section, we describe practical applications of this theory.

An Assessment of Residential Building Stock in the Northwestern United States

In 2010, a non-profit organization working to maximize energy efficiency market transformation in the Northwestern United States (in collaboration with utilities and the Bonneville Power Administration, a federal power marketing agency), released a request for proposals (RFP) seeking an inventory of existing residential building stock in the region, based on data obtained from energy audits of existing homes. This project required that proposers estimate the variability within the conditioned square footages of single-family, multifamily, and manufactured residences throughout the Northwest in order to calculate the samples required for each type of housing.

This paper addresses the differences between the consultant's original estimates of variability in the population, the CVs demonstrated in the incoming audit data, and the strategies upon which all parties (funder, consultants, and oversight committee) agreed, which served to ensure that data obtained by the project met the needs of the sponsoring organizations and regional utilities.

The data obtained by this expensive, large-scale building stock assessment was intended to not only support the primary funder's market transformation strategies, but also inform utility savings potentials, including the savings targets public utilities in Washington State would be required to meet. The study proposed intensive, on-site audits (requiring roughly three hours per residence), during which auditors would collect detailed information about the residence's energy-related aspects, including heating and cooling equipment, envelope and glazing characteristics, appliance types, and plug load characteristics. Auditors also would perform a socket saturation survey, collect detailed data on consumer electronics, and conduct flow tests of showerheads. Blower door tests and duct blasting would be performed later on a percentage of the final sample to determine the envelope characteristics of residence types. Utility billing records were to be obtained for these residences, from which weather-normalized whole-house Energy Use Indices (EUI) would be developed for each building and heating type. The database would be developed from audit and billing analyses and would be made available to funders and participating utilities.

The funders invited a range of interested parties from throughout the region to review the proposals and to oversee the study throughout its duration. The RFP sought a statistically representative sample for the Northwestern United States and set precision at a +/- 5% margin of error and a 95% level of confidence. Proposers were directed to first stratify by region and then by each of the four Northwest states (Washington, Oregon, Idaho and Montana). The RFP also requested that proposers' designs stratify the regional sample to meet the required 95/5 statistical significance for the seven public power sub-regions served by Bonneville Power Administration (BPA) – for a total of 11sampling domains (Table 2).

Table 2. Sampling Domains for All Residential Customers

Regional Sampling Domains	
NWR	All residential utility customers in WA, OR, ID and MT (BPA region w/ NW Energy)
WA	All residential utility customers in Washington
OR	All residential utility customers in Oregon
ID	All residential utility customers in Idaho
MT	All residential utility customers in Montana (BPA region w/ NW Energy)
Public Power Sampling Domains	
NWP	All public residential utility customers in the region
Western WA	All public utility residential customers in Western Washington, excluding Puget Sound
Puget Sound	All public utility residential customers in Puget Sound
Western OR	All public utility residential customers in Western Oregon
Eastern OR/WA	All public utility residential customers in Eastern Oregon and Eastern Washington
ID/MT	All public utility residential customers in Idaho and Montana

Four finalists offered a variety of sampling plans to meet the 95/5 criteria; some specified both stratification plans, and most provided the source for their estimation of CVs.

One firm calculated sample sizes based on the minimum required CV for a heterogeneous population, 0.5 (PJM 2010), and stated that the statistically required sample of 384 sites for each state domain (1,536 residences total) could be reduced in number (via stratification or clustering) to1,400 - 1,500 sites across the three housing types. Given the fixed budget, this firm ultimately advocated reducing the full sample size to a fixed number of 1,200 (the top of the range specified in the RFP) and reducing the level of precision required to 10%.

Another finalist proposed the same CVs for estimates of variance and also calculated the unstructured sample required (without benefit of clustering or stratification) at 384 sites per domain. This firm similarly proposed reducing this number by clustering and reducing the required level of precision to 7%. Both of these proposers pointed out that, at 95/5, the unstructured sample exceeded the number of sites that could be funded, and strongly suggested lowering the required level of precision.

The winning firm did not specify the CVs used in deriving its samples and offered no sampling details beyond the total number of residences of each type the firm would audit (roughly equivalent to the 1,200 maximum suggested in the RFP). This firm assured the funder that the proposed sample would meet

the 95/5 requirement for every domain, although the proposal did not provide details of how it would achieve them. The proposed sample was fixed at 1,192 residences, allocated as shown in Table 3.

Table 3. Selected Consultant's Sample Design

Housing Type	Total Sample
Single-Family	745
Multifamily	347
Manufactured	100
Totals	1,192

Referring back to Table 1, note that, absent a stratification/clustering plan and the adjustment of precision (given the 11 domains specified), this sample could not possibly meet the 95/5 criteria for each housing type and domain (e.g., 11*384 = 4224) unless the CVs estimated for each domain were unexpectedly lower than those the other proposers recommended.

The oversight committee invited to oversee this study included statisticians; and when the committee became actively involved in the study review process, members requested more detail and asked questions about the winning consultant's sampling calculations, CVs, and schema. In response to the committee's inquiries, the consultant released a sampling plan that provided the CV attributed to each domain, and allowed that, "based on expert judgment, and past survey results, we posit a CV for each of these domains" (Table 4).

Table 4. Coefficient of Variation by Sample Division

Domain Name	House Type		
	Single-Family	Multifamily	Manufactured
Regional	0.45	0.3	0.25
State	0.3	0.3	0.25
Sub-Region	0.3	0.2	0.25

Note that this table shows only the final CVs estimated for the sampling of this project. (The CV representing regional variability for single-family housing originally was 0.4, and the CV for each housing type at the state level originally was 0.25. Both CV's were raised after questioning by the oversight committee and funders.) These CVs were significantly different from the 0.5 CV the other finalists had recommended for all domains.

To investigate the variability in the region, members of the oversight committee conducted an extensive literature search. In addition, efficiency staff at Tacoma Power calculated the CV for the >90,000 single-family homes in the utility's territory – 0.47 – which demonstrated that there was more variability in this utility's territory than the consultant posited for the four-state region. Despite the consultant's opposition, fears of meeting the 95/5 criteria prompted the oversight committee to ask the funders to lower the study's requirements for rigor. Still, without stratification or a clustering plan available, the oversight

committee convinced the funding agency to lower the study criteria to 90/10. At this lower confidence level, even a straightforward random sample would meet the requirement.

Results

After further consideration by the oversight committee, funders, and the consultant, it was agreed that manufactured housing would be dropped from the study in order to enlarge the sample for single-family housing. Further, the difficulty experienced by contractors in recruiting multifamily residences influenced the decision to focus the study entirely on single-family residences, increasing the final sample to approximately 1,056 residences. The study is just concluding at this time, and approximate final samples sizes, along and the CV's calculated for each domain and the mean square footage, are shown in Table 5.

Table 5. Variation of *Single-Family* Square Footage in the Sample

Domain	Sample Size	Variability in Square Footage (CV)		Mean Square Footage
		Expected	Actual	
NWR Total Sample	1,056	0.45	0.47	2,054
WA	42	0.3	0.41	2,071
OR	283	0.3	0.45	2,043
ID	182	0.3	0.48	2,232
MT	166	0.3	0.55	2,295
Western WA	137	0.3	0.41	1,863
Puget Sound	178	0.3	0.43	1,979
Western OR	227	0.3	0.46	1,883
Eastern WA	110	0.3	0.38	2,245
Eastern OR	56	0.3	0.39	1,962
ID/MT	**	**	**	**

* Data for approximately 150 residences is currently unavailable due to extended quality reviews

**Data for this domain unavailable at the time of publication

The CVs for each domain were larger than originally postulated (refer to Table 4). However, because the oversight committee had suggested that statistical precision and confidence levels be lowered to 90/10, the statistical criteria for sampling were met for the region and for each of the domains, as an unstructured random sample.

Conclusion

For many energy efficiency studies, a required 95/5 level of confidence and precision frequently requires sample sizes that are unattainable under most budgetary constraints, and this degree of rigor exceeds what is required for most applications of the data attained. This mistake is common and involves confusing statistical significance with scientific significance. The first task in designing a study or evaluation always should be targeting statistical rigor based on the specific applications planned for the data obtained. Then, funders, consultants, and contractors must balance the rigor and funding, always seeking the most reliable information within a project's budget.

In the case of the recent assessment of residential building stock in the Northwestern United States, it was preferable to both funders and contractors to accept the more realistically attainable sampling criteria of a 90/10 level of confidence and precision, rather than risk under-sampling the study at the initially required 95/5 levels. Under-sampling in studies frequently is not recognized as such, which raises the risk that conclusions supported by such studies will be applied as if the study were as rigorous as the levels of confidence and precision that were specified. This is particularly problematic for utilities whose energy savings expectations and required savings targets are informed by energy efficiency research and evaluation.

The data yielded by the study discussed in this paper will support and inform future expectations for energy savings in the Northwestern United States. The research has been conducted in an exemplary fashion by a superior team of contractors, who are among the best in the energy efficiency business. Funders and contractors selected a well-informed oversight committee, which included experienced evaluators and statisticians, and then allocated a good deal of time throughout the project's design and data collection phases for discussions of sampling strategies and alternative ways to improve the quality of the data yielded by this research.

Lessons Learned

The lessons learned during the conduct of this project include:

1. Well-informed, experienced, and active oversight committees are extremely valuable to the conduct of even smaller-scale energy efficiency studies, and are invaluable to large-scale research and evaluation efforts.
2. A well-sampled study that meets a 90/10 level of confidence and precision is of much more value than a study that under-samples while attempting to achieve a 95/5 level. Consultants should not hesitate to propose alternative levels of rigor to funders, particularly when budget constraints limit sample sizes.
3. Statisticians should be involved early, and often, and should always be consulted about issues of sampling strategy and methodological design. Utilities, funders, and contractors that are involved in energy research and evaluation benefit from this statistical expertise.
4. Even seasoned energy professionals and statisticians can underestimate chance variability in a population based on experience and intuition. Those who select CVs for sample calculations should consult a variety of resources, including published study results, previous research, and pilot studies, as well as energy efficiency research and evaluation guides.
5. When the CV is unknown for a sampling characteristic, researchers should implement a value of not less than 0.5 for homogenous populations, to 1.0 for samples that are heterogeneous.

References

Cochran, W.G. 1977. *Sampling Techniques.* New York, NY: John Wiley & Sons: pp.74-76.

Fischoff, B., Slovic, P., and S. Lichtenstein. 1977. “Knowing with Certainty: The Appropriateness of Extreme Confidence.” Journal of Experimental Psychology: Human Perception and Performance (3): 522-564.

PJM. 2010. *PJM Manual 18B: Energy Efficiency Measurement & Verification.* Revision: 01. Pennsylvania, Jersey & Maryland: PJM Forward Market Operations.

Tversky, A., and D. Kahneman. 1971. “Belief in the Law of Small Numbers.” Psychological Bulletin (76): 105-110.