Annex 5

GUIDELINES FOR SAMPLING AND SURVEYS FOR CDM PROJECT ACTIVITIES AND PROGRAMME OF ACTIVITIES

(Version 02.0)

CONTENTS

I. Introduction

1. This document describes common types of sampling approaches and includes recommended outline for a sampling plan, recommended practices for unbiased estimates of sampled parameters and recommended evaluation criteria for DOE validation besides several best practice examples covering large and small-scale project activities. It also provides examples for checking reliability of data collected through sample surveys.

2. Furthermore, it covers the following items:

(a) Methods, if any, to deal with missed reliability targets without compromising conservative estimates for emission reduction;

(b) Best practice examples for DOE validation/verification for sampling and surveys.

II. Common types of sampling approaches

3. This section provides a summary[1] of some of the most common types of sampling approaches and typical situations where each is recommended. Formulas for calculating standard errors of estimates from each sampling technique, confidence intervals and associated sample sizes are provided in the reference texts cited at the end of this report. The provided sampling information primarily relates to determining point estimates of average (mean) values of a parameter.

A. Simple random sampling

4. A *simple random sample* is a subset of a population (e.g. villages, individuals, buildings, pieces of equipment) chosen randomly, such that each element (or unit) of the population has the same probability of being selected. The sample-based estimate (mean or proportion) is an unbiased estimate of the population parameter.

5. Simple random sampling is conceptually straightforward and easy to implement – provided that a sampling frame of all elements of the population exists. Its simplicity makes it relatively easy to analyze the collected data. It is also appropriate when only minimum information of the population is known in advance of the data collection.

6. Simple random sampling is suited to populations that are homogeneous. In many instances a large population size and dispersed nature of population may cause a lack of homogeneity, while in some cases those factors may have relatively low impact on homogeneity (e.g. a large number of biogas digesters located in varying altitudes and temperature zones may be less conducive for simple random sampling to determine the average amount of biogas production per digester, while the usage hours of light bulbs across wide geographic areas and among large populations with similar socioeconomic circumstances connected to a single or similar grid/s may be sufficiently homogeneous for simple random sampling). The costs of data collection under simple random sampling could be higher than other sampling approaches when the population is large and geographically dispersed.

B. Stratified random sampling

7. When the population under study is not homogeneous but instead consists of several sub-populations which are known (or thought) to vary, then it is better to take a simple random sample from each of these sub-populations separately. This is called *stratified random sampling*. The sub-populations are called the strata. When considering stratified random sampling it is important to note that when identifying the strata no population element can be excluded and every element must be assigned to only one stratum. For example, the population of participants in a commercial lighting programme might be grouped according to building type (e.g. restaurants, food stores, and offices).

8. Stratified random sampling is most applicable to situations where there are obvious groupings of population elements whose characteristics are more similar within groups than across groups (e.g. restaurants are likely to be more similar to one another in terms of lighting use than they are to offices or food stores). It requires that the grouping variable be known for all elements

[1] See Table 1 for advantages and disadvantages of each sampling approach.

in the sampling frame. For example, the sampling frame would require information on the building type for each case in the population to allow stratification by that characteristic.

9. Stratification helps to ensure that estimates of a population characteristic are accurate, especially if there are differences amongst the strata. For example, if lighting use within office buildings tends to be lower (on average) than in food stores then this can be taken into account when estimating the overall average number of hours of operation. Equally, if the cases within each stratum are more homogeneous than across strata, then the estimated number of hours of operation will be more precise than if a simple random sample of the same size had been taken.

C. Systematic sampling

10. *Systematic sampling* is a statistical method involving the selection of elements from an ordered sampling frame. The most common form of systematic sampling is an equal-probability method, in which every k^{th} element in the frame is selected, where k, the sampling interval (sometimes known as the “skip”), is calculated as:

$$k = \text{population size } (N) \text{ / sample size } (n)$$

11. Using this procedure, each element in the population has a known and equal probability of selection. The project participant shall ensure that the chosen sampling interval does not hide a pattern. Any pattern would threaten randomness. A random starting point must also be selected. Systematic sampling is to be applied only if the given population is logically homogeneous, because systematic sample units are uniformly distributed over the population.

12. Systematic sampling is applicable in a number of situations. If there is a natural ordering or flow of subjects in the population, such as output of bricks in a manufacturing process, then it is typically easier to sample every k^{th} unit to test for quality as they are produced. In all cases, it is important that the list of subjects or the process is naturally random, in the sense that there is no pattern to its order.

D. Cluster sampling

13. *Clustered sampling* refers to a technique where the population is divided into sub-groups (clusters), and the sub-groups are randomly selected (sampled), rather than the individual elements which are to be studied. The data are then collected on all the individual elements in the selected sub-groups.

14. Cluster sampling is used when “hierarchical” groupings are evident in a population, such as villages and households within villages, or buildings and appliances within buildings. For example, suppose a project installs high-efficiency motors in new apartment buildings, with several motors typically in each building. In order to estimate the operating hours of the motors, one might take a sample of the buildings instead of the motors, and then meter all of the motors in the selected buildings.

15. In contrast to stratified sampling, where the equipment of interest is grouped into a relatively small number of homogeneous segments, there are many clusters of motors (i.e. apartment buildings), and there is no expectation that the motors in each building are more homogeneous than the overall population of efficient motors.

16. Cluster sampling is useful when there is no sampling frame at the lowest level of the hierarchy but there is one at the cluster level, as in the case above where a ready list of all motors would not be available, but a list of all new apartment buildings would be.

17. In many applications to monitor efficient equipment, the units occur naturally in clusters, with a different number of elements per cluster. For example, a building or plant location might constitute a natural cluster, with varying numbers of pieces of equipment per location.

18. A cluster sampling approach can offer cost advantages. For instance, if a significant component of the cost of data collection is travel time between buildings, but there is minimal cost to collect data on units within a building, then it is more cost-effective to collect data on all units within a sample of buildings than to take a simple random sample across all units in the study. It will, however, usually be necessary to meter more pieces of equipment (sample more clusters) to achieve the same level of precision as the simple random sampling, but the reduction in cost and other benefits may more than offset this apparent increase in effort.

E. Multi-stage sampling

19. *Multi-stage sampling* is a more complex form of cluster sampling. Measuring all the elements in the selected clusters may be prohibitively expensive, or not even necessary. In multi-stage sampling, the cluster units are often referred to as primary sampling units and the elements within the clusters secondary sampling units. In contrast to cluster sampling where all of the secondary units are measured, in multi-stage sampling data are collected for only a sample of the secondary units.

20. For example, in a study of efficient lighting, if the operation hours of motors within any one building are thought likely to be similar across all motors then – especially if the cost of measuring them is relatively high – there is not much to be gained by metering all of them. It might be better to draw a sample of buildings, and then only measure a sample of motors from within each selected building. On the other hand, if the measurements are inexpensive once a technician is on-site, then it may make sense to monitor all of the fixtures.

21. Multi-stage sampling can be extended further to three or more stages. For example, one might group the population into building complexes, then buildings, and finally fixtures.

22. So far, most of the methods above have been based on simple random sampling. Another option is to sample with probability proportional to size, and this is sometimes used in cluster sampling where clusters are of different sizes, or in multi-stage sampling.

23. There are therefore many variations in methods in applying multi-stage sampling. If the number of secondary units in each primary unit is not known in the sampling frame, then one approach is to draw a sample of primary units at random, count the number of secondary units in each selected primary unit, and then take detailed measurements for a sample of secondary units. Another option is to sample the primary units with probability proportional to size, and to draw a random sample of the secondary units in the selected primary units. The relative performance of these alternatives depends on the population characteristics, the costs of data collection, and the availability of information on the primary and secondary units in the sample frame.

24. Table 1 below indicates advantages and disadvantages of various sampling schemes in a summary form.

Table 1: Advantages and disadvantages of different sampling schemes

Sampling Scheme	Advantages	Disadvantages
Simple Random Sampling: Taking a random sample from the whole population.	Easiest method to understand and therefore use. Suitable if there is little heterogeneity amongst the units being sampled	Requires knowledge of entire population before a sample can be selected. If the population covers a large geographical area, then it can often lead to sampling units that are spread out over the area. Such a situation can often be costly. Only suitable if the population being studied is relatively homogeneous with respect to the parameter being studied
Systematic Sampling: Taking a sample every *n* units	Easy to apply. Commonly used as it ensures there is always sufficient distance between samples	Leads to units being spread out over a large geographic area. Such a geographic distribution can often be costly
Stratified Random Sampling: Randomly sampling a different number of units from each strata according to the weight of each strata in the population	Improves the precision of the estimate (compared to simple random sampling) if there are differences between the strata	Complicated to calculate. What the stratification factors should be is not always obvious
Cluster Sampling: Sampling every unit in a sample of *n* clusters from the population	The most economical form of sampling as units are all grouped according to one criterion (often geographical). Sometimes the only approach, since a list of all households may not be available, only a list of villages. Once the villages have been selected, the households can be sampled. It saves time at a management level	Results are not normally so ‘good’ (i.e. standard errors of estimates tend to be high due to homogeneity of characteristics in the subgroup sampled). But a larger sample can help to compensate for this

Sampling Scheme	Advantages	Disadvantages
Multi-stage Sampling: Randomly sampling a number of units within a number of randomly selected clusters	Enables sampling approach at two levels. Can compare different scenarios – number of clusters and number of units within the clusters – in order to find most cost-efficient and reliable scenario	Analysis and the sample size calculation are more difficult

III. Recommended outline for a sampling plan

25. The sampling plan should contain information relating to: (A) sampling design; (B) data to be collected; and (C) implementation plan.

A. Sampling design

26. *Objectives and reliability requirements* describes the objective of the sampling effort, the timeframe, and the estimated parameter value(s). Identify the sampling requirements (applicable CDM methodology or sampling standards) and the confidence/precision criteria to be met. For example, the objective is determining the mean monthly value of parameter "X" during the crediting period, and with a 90/10 confidence/precision.

27. *Target population* defines the target population, and describes any particular features associated with it.

28. *Sampling method* selects and describes the sampling method, e.g. simple random sampling, stratified sampling, cluster sampling. Strata or clusters shall be clearly identified if sampling other than simple random sampling is to be used.

29. *Sample size* addresses and justifies the estimated target number of "units" – pieces of equipment, solar cookers, buildings, motors, log-books, etc. – which are to be studied (i.e. the sample size). The justification shall include the parameter of interest, the value it is expected to take and an estimate of the variance associated with the data, as well as the level of confidence and precision (note that if the parameter of interest is a proportion, or a percentage, then there is no need to specify a variance estimate).

30. *Sampling frame* identifies or describes the sampling frame to be used. This shall agree with the information about the Target Population and Sampling Design above. For instance, if cluster sampling is to be used in a study of equipment in buildings, then the frame should be a listing of the buildings from which the sample will be selected.

B. Data to be collected

31. *Field measurements* identifies all the variables to be measured and determine appropriate timing and frequency of the measurements. When the measurements are conducted only during limited time periods and are to be scaled up to the whole year, demonstrate that the parameter of interest is not subject to seasonal fluctuations or the time period selected is conservative or the necessary corrections are applied. Methods of measurement shall be described as appropriate;

32. *Quality assurance/Quality control* describes how to achieve good quality data, for example describe the procedures for conducting the data collection and/or field measurements including training of field personnel, provisions for maximizing response rates, documenting out-of-population cases, refusals and other sources of non-response, and related issues. An overall quality control and assurance strategy shall be documented in the plan. This shall include a procedure for defining outliers and under what circumstances outlier data/measurements may be excluded and/or replaced.

33. *Analysis* describes how the data will be used.

C. Implementation

34. *Implementation plan* defines the schedule for implementing the sampling effort and identify the skills and resources[2] required for data collection and the analyses.

[2] A general description of qualifications and experience of personnel who will be engaged should be provided, not necessarily listing specific names, qualification and experience.

IV. Recommended practices for unbiased estimates of sampled parameters

35. Practitioners are expected to observe sound practices in designing samples and administering surveys and field measurements.[3] Those practices include:

(a) **Defining precisely the sampling objectives and target population and the measurements to be taken and/or data collected.** The sampling objectives will, for the most part, be concerned with estimation, i.e. estimating a characteristic (e.g. mean or percentage) of a population. Occasionally, the objective will be one of comparison, for example to compare the uptake in rural areas with that of urban areas:

(i) The target population is the “greater entity” to which the results from the survey sample are to be generalized, for example all new light fittings that are installed in new buildings in country X;

(ii) The information that will be collected will depend on the objectives, for example if a project needs to estimate the average number of hours of operation of a new efficient motor, then the data to be collected on each sampling unit is its number of hours of operation. Other measurements to be taken may relate to the characteristics of the strata, or clusters, or any other variable that may be relevant to the project objectives.

(b) **Deciding on the sampling design and the size of the sample.** This decision is based on the information provided above;

(c) **Developing the sampling frame.** A *sampling frame* is a complete listing of all individual units (elements, members) that can be considered as a representation of the whole population, and which can be used as a basis for selecting a sample, such as a list of all households in an area that have had solar cookers installed.[4] In the case of cluster sampling or multi-stage sampling, the sampling frame is a complete listing of sub-groups of the study area/population[5] which constitutes all the clusters or primary sampling units.

Without such a frame, or its equivalent, methods of sampling with assured properties such as unbiasedness are not available. The implementer of the survey effort shall compile a clear description of the target population, including those characteristics of the population which define membership. From the description and characteristic the implementer can then select a sampling frame;

(d) **Randomizing cases and drawing sample.** The implementer should ensure that the sample is drawn at random from the sampling frame. This can be done using random number tables or using the random number generator of appropriate

[3] For a very comprehensive treatment of issues surrounding sample/survey design, see Household Sample Surveys in Developing and Transition Countries, United Nations, 2005, ISBN 92-1-161481-3.

[4] Such a listing shall be available for check during validation/verification but not necessarily included in the PDD documents.

[5] A suitable map with the sampling units marked on it and properly delineated may also be regarded as a sampling frame and used in drawing samples.

software. If a systematic sampling is chosen, then the ordering of subjects on the sample should be random and free of any trend or cyclical pattern;

(e) **Selecting the most effective information-gathering method.** The implementer should decide on what would be the most reliable and cost-effective method for collecting the data, depending on the variables of interest. Alternative methods include visual inspections, physical measurements, respondent self-reports, and operational logs. For example, equipment retention rates may be determined by inspections or self-reports. Estimates of electric consumption could be based on different metering technologies depending on the characteristics of the equipment. Vehicle travel miles or equipment operating schedules could be drawn from odometers or operation logs;

(f) **Conducting surveys/measurements.** The project implementer is expected to establish and implement procedures to ensure that the field data collection is performed properly and that any potential intentional errors or unintentional errors are minimized and documented. Such procedures include: developing field measurement protocols; training personnel; establishing contact procedures; documenting coverage problems, missing cases, and non-response; minimizing non-sampling measurement errors; and quality control for data coding errors;

(g) **Minimizing non-response and adjusting for its effects**. The project implementer is expected to make all reasonable efforts to minimize non-response, to analyze potential bias arising from non-response, and to correct for any detected biases or losses in precision due to non-response. Field data collection protocols should specify procedures for multiple contacts to minimize non-response, require documentation of reasons for non-response, and prescribe corrective measures to compensate for its occurrence. Corrective measures may include over-sampling, replacing non-respondents with similar subjects, applying “correction factors” and imputing responses.

V. Recommended evaluation criteria for DOE validation

36. The following questions and evaluation criteria serve as examples and should be utilized by DOEs to validate the proposed sampling plans:

(a) Does the sampling plan present a reasonable approach for obtaining unbiased, reliable estimates of the variables?

(i) In terms of assessing reliability, are the elements of Objectives and Reliability Requirements complete? Do the requirements specified agree with those stated in the appropriate standards? If not, is there a reason why they are not met?

(ii) From all the different elements of the Design, is there any reason to suspect that the results from the activity will be biased? For instance, is the population under consideration only urban households? What about rural households? Might this cause a bias when the data are extrapolated to emission reductions?

(b) Is the population clearly defined, and how well does the proposed approach to developing the sampling frame represent that population?

(i) The population should be clear from the Target Population description. Whether or not the sampling frame is possible or appropriate will depend on the detail and the particular situation, for example if a map is going to be used, a question would be whether a map already exists, and how reliable it is. If a map does not exist, then who is going to create it?

(c) Is the proposed sampling approach clear?

(i) Is it clear which sampling method is being proposed? For example, is it simple random sampling, or some other method of sampling?

(ii) Does the method agree with the description of the population? Are there clusters or strata, and if so does it state what they are? For example, are they buildings, villages, etc.?

(d) Is the proposed sample size adequate to achieve the minimum confidence/precision requirements? Is the ex ante estimate of the population variance needed for the calculation of the sample size adequately justified?

(i) All of the information set out in the sampling plans should help answer this question. If not all information is provided then the question cannot be answered;

(ii) Is the target value for the population parameter reasonably anticipated?

(iii) Does the estimate of variability seem reasonable?

(e) Is the sample representative?

(i) Is it clear how the sample is to be selected? For example, is it to be selected randomly?

(ii) Does the Plan indicate that the sampling frame will be kept (e.g. in hard copy or a computer file of screen shot copy), and that random numbers will be generated and these random numbers will then be used to select the sample?

(f) Is the data collection/measurement method likely to provide reliable data given the nature of the parameters of interest and project, or is it subject to measurement errors?

(i) Are the methods of data collection clear and unambiguous? Are there questions which could be subject to respondent error due to sensitivity (e.g. "How much money do you spend on heating?"), lack of recall (e.g. "How many times did you buy fuel last year?"), and the like?

(ii) Are there questions that could be subject to measurement error? For example, is a particular measurement method known to under-record key data, such as the weight of bricks?

(g) Are the procedures for the data measurements well defined and do they adequately provide for minimizing non-sampling errors?

(i) Is the quality control and assurance strategy adequate?

(ii) Are there mechanisms[6] for avoiding bias in the answer?

(h) Does the frame contain the information necessary to implement the sampling approach?

(i) Are the proposed skill sets, qualifications and experience of the personnel to be engaged to conduct sampling adequate?

[6] Mechanisms for avoiding non-sampling errors (bias) include good questionnaire design, well-tested questionnaires, possibly pilot testing the data collection.

Appendix A

Best practice examples for sample size calculations

1. Introductory notes on sample size calculations

37. There are different equations to calculate a required sample size for different situations. Most of the examples in this document are for finite populations such as cook stoves or compact fluorescent lamps (CFLs), but there is also one example for a wastewater treatment plant where the measurements are done for the continuous flow of wastewater.

38. Which equation to use depends on the following:

(a) Parameter of interest, for example:

(i) A percentage, such as the proportion of annually operating cook stoves;

(ii) A numeric value, such as the mean value of operating hours of CFLs, the mean value of dry compressive strength (to check whether the manufactured bricks are of a certain quality).

39. There are other parameters, that is ratios, but this guide only covers proportions and means:

(a) Sampling scheme. This document contains the equations and examples using the following five sampling schemes:

(iii) Simple random sampling;

(iv) Systematic sampling;

(v) Stratified random sampling;

(vi) Cluster sampling;

(vii) Multi-stage sampling.

40. There are a number of factors that affect the sample size required, and these are described below:

(a) The value that the parameter is expected to take, for example:

(i) Sampling to see whether 80% of installed cook stoves are still in operation will give a different sample size required than 65%. The same is true for mean values;

(b) The amount of variation affects the sample size required. The larger the variation associated with the parameter of interest the larger the sample size required for the same level of confidence and precision;

(c) The level of precision (e.g. ±10% of relative value of the parameter's true value) and confidence (e.g. 90% or 95%) in that precision which is desired for determining the parameter also determines the sample size. The higher the required confidence and the narrower the precision the more samples are required.

41. Estimates of the parameter of interest (proportion, mean and standard deviation) are required for sample size calculations. There are different ways to obtain these:

(a) We may refer to the result of previous studies and use these results;

(b) In a situation where we do not have any information from previous studies, we could take a preliminary sample as a pilot and use that sample to provide our estimates;

(c) We could use "best guesses" based on the researcher's own experiences.

42. Note that if the standard deviation is unknown but the range (maximum – minimum) is known then a rough "rule of thumb" is that the standard deviation can be estimated as the range divided by 4.

43. Also, for different sampling schemes additional information is required, such as strata estimates rather than just the population information.

44. There are three additional points to make in relation to these sample size calculations:

(a) If sample size calculations are being performed manually, it is important to retain as many decimal places as relevant, until the final calculated figure is reached. Only then should rounding be carried out. In this document, however, for clarity of presentation the detailed calculations are shown with only a small number of decimal places, although the actual calculations themselves used more than is shown;

(b) Researchers are encouraged to carry out more than one sample size calculation. It is highly unlikely that accurate estimates of the parameters will be available, and so the calculation should be performed for a range of possible estimates (e.g. proportion, or mean and standard deviation), and the largest sample size chosen. This should help to ensure that the sample selected will meet the required reliability criteria;

(c) The pilot studies that are included in the examples here are deliberately small so that calculations can be illustrated fairly easily. In real-life situations they should be larger than those used here.

2. Sample size calculations - Small-scale examples

45. For all of the small-scale examples below, we require 90% confidence that the margin of error in our estimate is not more than ±10% in relative terms.

Proportional parameter of interest (Cook stove project)

46. This section covers sample size calculations based on a proportion (or percentage) of interest being the objective of the project, under four different sampling schemes. Regardless of the sampling scheme used, the following have to be pre-determined in order to estimate the sample size:

(a) The value that the proportion is expected to take;

(b) The level of precision, and confidence in that precision (90/10 for all small-scale examples).

47. For all of the cook stove examples below, the proportion of interest is the number of project cook stoves that are still in operation at the end of the third year after the stoves were

distributed; it is thought that this proportion is 0.5 (50%). The cook stoves were distributed to 640,000 households, and it has been assumed that 1 household = 1 cook stove.

Example 1 – Simple random sampling

48. Suppose that the population is homogenous with respect to the continued use of the cook stoves. Then simple random sampling would be an appropriate method to estimate the proportion of cook stoves still in operation.

The equation to give us the required sample size is:

$$n \geq \frac{1.645^2 N \times p(1-p)}{(N-1) \times 0.1^2 \times p^2 + 1.645^2 p(1-p)} \tag{1}$$

Where:

n	Sample size
N	Total number of households (640,000)
p	Our expected proportion (0.50)
1.645	Represents the 90% confidence required
0.1	Represents the 10% relative precision ($0.1 \times 0.5 = 0.05$ = 5% points either side of *p*)

Substituting in our values gives:

$$n \geq \frac{1.645^2 \times 640{,}000 \times 0.5 \times 0.5}{(640{,}000 - 1) \times 0.1^2 \times 0.5^2 + 1.645^2 \times 0.5 \times 0.5} = 270.4 \tag{2}$$

49. Therefore the required sample size is at least 271 households. This assumes that 50% of the cook stoves would be operating. If we changed our prior belief of the underlying true percentage of working stoves *p*, this sample size would need recalculating.

50. Note that the figure of 271 households means 271 households with data for analysis. If we expected the response rate from the sampled households to be only 80% then we would need to scale up this number accordingly. Thus we would decide to sample 271/0.8 = 339 households.

51. If we did not scale up our sample size and experienced a response rate of 80% then we would only have 216 (271×0.8 = 216) households/cookers with data, and consequently the level of precision would be detrimentally affected. We can calculate the actual level of precision by substituting n = 216 into the equation:

$$\frac{1.645^2 \times 640{,}000 \times 0.5 \times 0.5}{(640{,}000 - 1) \times precision^2 \times 0.5^2 + 1.645^2 \times 0.5 \times 0.5} = 216 \tag{3}$$

52. This gives a relative precision of 0.1119, or 11.2% – not the 10% required. So by not adjusting our estimated sample size to take into account the expected response rate we have an increased margin of error.

53. One solution to this could be to take an additional sample of households. This additional sample would need to recruit 69 households which would then, again assuming a response rate of 80%, provide data on 55 households (80% of the 69). Adding these to the existing 216 gives us data for 271 households, the number required to achieve 90/10 reliability.

Approximate equation

54. The equation used above is the exact equation derived from simple random sampling theory. When population sizes are large (or infinite), then an approximate equation can be used, which ignores the actual size of the population (N). The approximate equation for the 90/10 confidence/precision guideline is:

	Approximate Equation	Sample size for the above example
Proportion data	$n = \frac{1.645^2(1-p)}{0.1^2 \times p}$	271 $\left(= \frac{1.645^2 \times (1-0.5)}{0.1^2 \times 0.5}\right)$

Notes on approximate equations

55. As the sample size in this example is large, there is no difference between the sample sizes derived from the exact and approximate equations. However, for smaller populations (N<5000) and small p's (less than 0.5) there will be a difference.

56. Since the exact equation can be easily calculated, it is recommended that the exact equation be used in preference to the approximate one. It avoids having to decide whether the population size is large enough for it to be possible to use the approximate equation.

57. The scaling-up of the sample size due to non-response will also apply to the approximate equation.

Example 2 – Stratified random sampling

58. This time we know that stoves were distributed in four different districts and that the cook stoves are more likely to be still in operation in certain districts compared to others.[7] In this situation we want to take our knowledge about the district differences into account when we do the sampling, and sample separately from each district. Estimates of the proportion of cook stoves still in operation in each district, as well as the population size of each district are required.

District	**Number of households with cook stove in district*** **(*g*)**	**Proportion of cook stoves still in operation in district** **(*p*)**
A	76,021	0.20
B	286,541	0.46
C	103,668	0.57
D	173,770	0.33

[7] If the proportions were expected to be the same in each district then simple random sampling should be used.

** Note that the districts cover all of the population (sum of district populations = total population)*

59. The equation for the total sample size is:

$$n \geq \frac{1.645^2 NV}{(N-1)\times 0.1^2 + 1.645^2 V} \quad \textbf{(4)}$$

Where: $V = \dfrac{SD^2}{\overline{p}^2} = \dfrac{\text{overall variance}}{\overline{p}^2}$ and $\overline{p}$ is the overall proportion.

60. To then decide on the number of households in the sample that come from each district we could use proportional allocation, where the proportions of units from the different districts in the sample are the same as the proportions in the population. This gives $n_i = \dfrac{g_i}{N}\times n$ where i=1,…,k and k is the number of districts in the area (in this case 4).

Where:

g_i	Size of the ith group (district) where i=1,…,k
N	Population total

61. We use the figures from the table above to calculate the overall variance,[8] and proportion of cook stoves still in operation.

$$SD^2 = \frac{(g_a \times p_a(1-p_a)) + (g_b \times p_b(1-p_b)) + (g_c \times p_c(1-p_c)) + ... + (g_k \times p_k(1-p_k))}{N} \quad \textbf{(5)}$$

$$\overline{p} = \frac{(g_a \times p_a) + (g_b \times p_b) + (g_c \times p_c) + ... + (g_k \times p_k)}{N} \quad \textbf{(6)}$$

Where g_i and N are as above and p_i is the proportion for the ith group (district); i=1,…,k

Substituting the values from the table into the above equations for SD^2 and $\overline{p}$ gives:

$$SD^2 = \frac{(76021\times 0.20\times 0.8) + ... + (173770\times 0.33\times 0.66)}{640000} = 0.23 \quad \textbf{(7)}$$

$$\overline{p} = \frac{(76021\times 0.20) + (286541\times 0.46) + (103668\times 0.57) + (173770\times 0.33)}{640000} = 0.41 \quad \textbf{(8)}$$

Therefore:

$$V = \frac{SD^2}{\overline{p}^2} = \frac{0.23}{0.41^2} = 1.37 \quad \textbf{(9)}$$

[8] The variance of a proportion is calculated as: p(1-p).

Substituting in V into our sample size equation gives:

$$n \geq \frac{1.645^2 \times 640000 \times 1.37}{(640000-1) \times 0.1^2 + 1.645^2 \times 1.37} = 367.0 \qquad \textbf{(10)}$$

62. The total sample size required is 367 households. This then needs to be divided up according to the size of each district to get the number of households that should be sampled in each district.

General equation: $n_i = \dfrac{g_i}{N} \times n$ **(11)**

District A: $n_a = \dfrac{76021}{640000} \times 367 = 43.7$

District B: $n_b = \dfrac{286541}{640000} \times 367 = 164.8$

District C: $n_c = \dfrac{103668}{640000} \times 367 = 59.6$

District D: $n_d = \dfrac{173770}{640000} \times 367 = 99.9$

63. Rounding up the district samples sizes gives the number of households to be sampled in each district, 44 in A, 165 in B, 60 in C, and 100 in D (the sum of these is slightly greater than the total required sample size due to the rounding up of households within each district).

64. Note that these sample sizes do not take into account non-response. If the expected level of response is 75% across all districts then divide each district sample size by 0.75; this will result in larger sample sizes allowing for the non-responders.

Example 3 – Cluster sampling

65. Now consider a different scenario. The households are not located in different districts. Instead they are ‘clustered’ or grouped into lots of villages. Instead of going to numerous individual households, we want to go to a number of villages and sample every household within each village.

66. For this example the population comprises 120 villages, all of approximately similar size. In order to have some understanding of the proportion of cook stoves still operating and the variation in this proportion between villages, a small preliminary sample has been taken:

Village	Estimated proportion of cook stoves operating in each village
1	0.37
2	0.48
3	0.50
4	0.27
5	0.68
Average ($\overline{p}$)	0.46
Variance $\mathrm{SD_B}^2$	0.024

67. The average $(\overline{p})$ is just $\frac{0.37+0.48+0.50+0.27+0.68}{5}=\frac{2.3}{5}=0.46$ and the variance between the clusters is:

$$\mathrm{SD_B}^2=\frac{1}{n-1}\sum_{i=1}^{n=5}(p_i-\overline{p})^2=\frac{(0.37-0.46)^2+(0.48-0.46)^2+...+(0.68-0.46)^2}{4}=\frac{0.0946}{4}=0.0237 \quad \textbf{(12)}$$

The equation for the number of villages that need to be sampled is:

$$c \geq \frac{1.645^2 MV}{(M-1)\times 0.1^2+1.645^2 V} \quad \textbf{(13)}$$

Where:

$$V=\frac{SD_B{}^2}{\overline{p}^2}=\frac{\text{variance between clusters (villages)}}{\text{average proportion}}$$

C	Number of clusters to be sampled (villages)
M	Total number of clusters (villages) – this must encompass the entire population
1.645	Represents the 90% confidence required
0.1	Represents the 10% relative precision required

68. Substituting our values into the above equation gives the number of villages that are required to be sampled as:

$$V=\frac{SD_B^2}{\overline{p}^2}=\frac{0.0237}{0.46^2}=0.11 \quad \textbf{(14)}$$

$$c \geq \frac{1.645^2\times 120\times 0.11}{(120-1)\times 0.1^2+1.645^2\times 0.11}=24.3 \quad \textbf{(15)}$$

69. Therefore we would have to sample every household within 25 randomly selected villages. This approach to sampling assumes that the villages are homogenous. In this example this means that the proportion of cook stoves still operating in a village is independent of any other factors such as district (see example 2 – stratified sampling), economic status, etc. If the proportions are not independent of another factor then cluster sampling within each strata of the factor can be used.

70. Since cluster sampling is dealing with data from whole clusters (villages in this example), non-response at the within-village level (household in this case) is less likely to be an issue, unless there is a high percentage of non-responses within a village. If there are only one or two missing values in a village it is still possible to obtain a usable proportion for that village based on all the other households that did provide data.

Example 4 – Multi-stage sampling

71. Multi-stage sampling can be thought of as sampling from a number of groups, and then going on to sample units within each group. Continuing with the cook stove example, we want to sample a number of villages and then a number of households within each sampled village.

72. We know that there are 120 villages and there are on average 50 households within each village, of which we plan to sample 10. From a small pilot study we already know the following:

Village	**Proportion of cook stoves in operation**
A	0.37
B	0.48
C	0.50
D	0.27
E	0.68

73. The equation for the number of villages to be sampled is:

$$c \geq \frac{\frac{SD_B^2}{\bar{p}^2} \times \frac{M}{M-1} + \frac{1}{\bar{u}} \times \frac{SD_w^2}{\bar{p}^2} \times \frac{(\bar{N}-\bar{u})}{(\bar{N}-1)}}{\frac{0.1^2}{1.645^2} + \frac{1}{M-1} \frac{SD_B^2}{\bar{p}^2}} \qquad \textbf{(16)}$$

Where:

C	Number of groups that should be sampled
M	Total number of groups in the population (120 villages)
$\bar{u}$	Number of units to be sampled within each group (pre-specified as 10 households)
$\bar{N}$	Average units per group (50 households per village)
${SD_B}^2$	Unit variance (variance between villages)
${SD_W}^2$	Average of the group variances (average within village variation)
p	Overall proportion
1.645	Represents the 90% confidence required
0.1	Represents the 10% relative precision

74. Using our table of pilot information we can calculate the unknown quantities for the equation above.

Village	Proportion of cook stoves in operation (p_i)	Variance within village ($p_i(1-p_i)$)
A	0.37	0.2331
B	0.48	0.2496
C	0.50	0.2500
D	0.27	0.1971
E	0.68	0.2176
Average	$\overline{p} = 0.46$	$SD_W{}^2 = 0.2295$
Variance	$SD_B{}^2 = 0.0237$	

Where:

$\overline{p}$ is the average proportion of cook stoves, i.e. $\frac{0.37+...+0.68}{5} = 0.46$

$SD_W{}^2$ is the average variance within the villages, i.e. $SD_W{}^2 = \frac{0.2331+...+0.2176}{5} = 0.2295$

$SD_B{}^2$ is the variance between the village proportions, i.e. the variance between 0.37, 0.48 etc. This can be calculated in the usual way for calculating a variance i.e. using the equation

$SD_B{}^2 = \frac{\sum_{i=1}^{n}(p_i - \overline{p})^2}{n-1}$ which gives $SD_B{}^2 = 0.0237$

75. Substituting our values into the group sample size equation gives:

$$c \geq \frac{\frac{0.0237}{0.46^2} \times \frac{120}{(120-1)} + \frac{1}{10} \times \frac{0.2295}{0.46^2} \times \frac{(50-10)}{(50-1)}}{\frac{0.1^2}{1.645^2} + \left(\frac{1}{(120-1)} \times \frac{0.0237}{0.46^2}\right)} = 43.4 \qquad \textbf{(17)}$$

76. Therefore if we were to sample 10 households from each village we should sample 44 villages for the required confidence/precision.

77. It is usually useful to have this calculation automated so that a series of different *u* values (the number of units to be sampled in each group) can be used and the effect that this has on the number of groups to be sampled can be observed.

Number of households sampled in each village *u*	Required number of villages *c*
5	68
10	44
15	36
20	32
30	28
50	25

78. In this example, by doubling the number of households within each village to be sampled from 10 to 20, we reduce the number of villages that need to be visited from 44 to 32.

79. Note that when u = the average number of households in a village (50), the required sample size is the same as that from cluster sampling as everyone within each village would be sampled. When u is smaller than the average number of households, the number of villages that need to be sampled under multi-stage sampling is greater than that from cluster sampling as not everyone within each village is being sampled.

Mean value parameter of interest (CFL project)

80. This section covers sample size calculations where the objective of the project relates to a mean value of interest, under four different sampling schemes. For the sample size calculations, regardless of the sampling scheme, we need to know:

(a) The expected mean (the desired reliability is expressed in relative terms to the mean);

(b) The standard deviation;

(c) The level of precision, and confidence in that precision (90/10 for all small-scale examples).

81. The examples below are based on the parameter of interest being average daily CFL usage, which is thought to be 3.5 hours, with a standard deviation of 2.5 hours. The population consists of 420,000 households to which CFLs were distributed; we are assuming that 1 household = 1 CFL.

Example 5 – Simple random sampling

82. For simple random sampling to be appropriate we are assuming that CFL usage is homogenous amongst the households.

83. The following equation can be used to calculate the sample size:

$$n \geq \frac{1.645^2 NV}{(N-1)\times 0.1^2 + 1.645^2 V} \tag{18}$$

Where:

$$V = \left(\frac{SD}{mean}\right)^2$$

n	Sample size
N	Total number of households
Mean	Our expected mean (3.5 hours)
SD	Our expected standard deviation (2.5 hours)
1.645	Represents the 90% confidence required
0.1	Represents the 10% relative precision

$$V = \left(\frac{2.5}{3.5}\right)^2 = 0.51 \qquad \textbf{(19)}$$

$$n = \frac{1.645^2 \times 420{,}000 \times 0.51}{(420{,}000 - 1) \times 0.1^2 + 1.645^2 \times 0.51} = 138.0 \qquad \textbf{(20)}$$

84. Therefore the required sample size is at least 138 households.

85. Note that if we expected the response rate from the sampled households to be only 70% then we would need to scale up the number obtained above accordingly. Thus we would decide to sample 138/0.7 = 198 households.

Approximate equation

86. The equation used above is the exact equation. When population sizes are large (or infinite), then approximate equations can be used, which ignore the actual size of the population (N).

87. The approximate equation follows the 90/10 confidence/precision guideline:

	Approximate Equation	Sample size for the above example
Mean value data	$n = \frac{1.645^2 V}{0.1^2}$ Where: $V = \left(\frac{SD}{mean}\right)^2$	$138\left(= \frac{1.645^2 \times 0.51}{0.1^2}\right)$

88. Please see the 'Approximate equation' section under **i) Cook Stove Project - Proportional parameter of interest**, **Example 1: Simple random sampling** for notes relating to approximate equations.

Example 6 – Stratified random sampling

89. The key to this example is that, unlike under simple random sampling, it is not assumed that the population is homogeneous – different parts of the population are expected to have different CFL usage averages.

90. Suppose that the CFLs were distributed in different districts in which each has a different CFL usage pattern (due to district economic backgrounds). We are now interested in sampling users of CFLs from all the districts to ensure all areas are well represented.

91. Each district has the following number of households, and mean and standard deviation CFL usage:

District	Number of households in district given a CFL	Mean (hours)	Standard deviation (hours)
A	146,050	3.2	1.9
B	104,474	2.4	0.8
C	38,239	4.5	1.6
D	74,248	1.6	1.7
E	56,989	2.3	0.7

92. The total sample size of households across all five districts is:

$$n \geq \frac{1.645^2 \times NV}{(N-1)\times 0.1^2 + 1.645^2 V} \quad \textbf{(21)}$$

Where:

$$V = \left(\frac{SD}{mean}\right)^2$$

SD Is the overall standard deviation, and

Mean Is the overall mean.

93. Using the data in the table above we can estimate the overall mean and standard deviation. Both equations are weighted according to the total number of households in each district.

94. Overall Standard Deviation:

$$SD = \sqrt{\frac{(g_a \times SD_a^2) + (g_b \times SD_b^2) + (g_c \times SD_c^2) + ... + (g_k \times SD_k^2)}{N}} \quad \textbf{(22)}$$

Where:

SD Weighted overall standard deviation
SD_i Standard deviation of the ith group where i=1,…,k, (note that these are all squared – so the group size is actually being multiplied by the group variance)

g_i Size of the ith group where i=1,…,k

N Population total

$$mean = \frac{(g_a \times m_a) + (g_b \times m_b) + (g_c \times m_c) + ... + (g_k \times m_k)}{N} \quad \textbf{(23)}$$

Where:

Mean Weighted overall mean

m_i Mean of the ith group where i=1,…,k

95. Substituting the values from our example into the above expressions gives:

$$SD = \sqrt{\frac{(146050 \times 1.9^2) + (104474 \times 0.8^2) + ... + (56989 \times 0.7^2)}{420000}} = 1.49 \tag{24}$$

$$mean = \frac{(146050 \times 3.2) + ... + (56989 \times 2.3)}{420000} = 2.71 \tag{25}$$

96. Substituting these values into the equation for *V* gives:

$$V = \left(\frac{SD}{mean}\right)^2 = \left(\frac{1.49}{2.71}\right)^2 = 0.3 \tag{26}$$

And hence, for the sample size:

$$n = \frac{1.645^2 \times 420{,}000 \times 0.3}{(420{,}000 - 1) \times 0.1^2 + 1.645^2 \times 0.3} = 81.7 \tag{27}$$

97. This example assumes proportional allocation, which means that the number of households we want to sample from each district is proportional to the size of the district within the population.

The equation for each district sample size is: $n_i = \frac{g_i}{N} \times n$ **(28)**

District A: $n_a = \frac{146050}{420000} \times 82 = 29$

District B: $n_b = \frac{104474}{420000} \times 82 = 21$

District C: $n_c = \frac{38239}{420000} \times 82 = 8$

District D: $n_d = \frac{74248}{420000} \times 82 = 15$

District E: $n_e = \frac{56989}{420000} \times 82 = 12$

98. The summation of these district sample sizes (29+21+8+15+12=85) is slightly greater than that calculated from the total sample size equation (82) above due to rounding.

99. As with previous examples, the sample sizes above need to be scaled up to take into account any non-response expected.

Example 7 – Cluster sampling

100. Suppose CFLs were distributed to households in 50 villages. Instead of sampling from the whole population of households with CFLs, we sample a number of villages (villages=clusters), and then collect data from all households within the villages.

101. The equation used to give us the required number of clusters, *c*, to sample is:

$$c \geq \frac{1.645^2 MV}{(M-1)\times 0.1^2 + 1.645^2 V} \qquad \textbf{(29)}$$

Where:

$$V = \left(\frac{SD}{Cluster\ mean}\right)^2$$

M	Total number of clusters (50 villages)
1.645	Represents the 90% confidence required
0.1	Required precision

102. To perform the calculations we need information about CFL usage at the village level, rather than at the household level. If such information does not already exist, we could possibly collect it in a pilot study. The example here assumes that data are available from a pilot study on five villages.

Village	**Total usage** across all households in the village[9]
A	30458
B	27667
C	31500
D	28350
E	19125

103. Calculating the mean and standard deviation of these figures gives:

$$\text{Cluster mean } (\bar{y}) = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{30458 + 27667 + ... + 19125}{5} = 27420 \qquad \textbf{(30)}$$

$SD = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}$ where the y_i are the total usages for the villages.

$SD_B{}^2 = 23902660$ and so $SD_B = 4889$

104. These statistics (i.e. mean and SD of data) are easily produced using statistical software.

[9] In the pilot study these totals may be derived from collecting data on all households in the village, or else by taking a sample of households in the village and scaling up from the sample to all households.

105. Substituting these values into the equation gives the required number of clusters, i.e. villages as:

$$V = \left(\frac{4889}{27420}\right)^2 = 0.03 \qquad \textbf{(31)}$$

$$c \geq \frac{1.645^2 \times 50 \times 0.03}{\left(50-1\right) \times 0.1^2 + 1.645^2 \times 0.03} = 7.5 \qquad \textbf{(32)}$$

106. So we need to sample eight villages to satisfy the 90/10 confidence/precision criterion. Once a village is selected, all households in the selected village should be sampled.

107. The above equation assumes that CFL usage in a village is independent of any other factors, such as economic status. If CFL usage was expected to vary according to another factor then cluster sampling can be used within each level of the factor.

Example 8 – Multi-stage sampling

108. Multi-stage sampling combines the cluster and simple random sampling approaches in a two-stage sampling scheme which enables us to randomly select some groups (villages) and then randomly sample some units (households) within those groups (villages). As with simple random sampling and cluster sampling, we are assuming homogeneity across villages in the usage of CFLs. We know that the 420,000 households are in 50 villages.

109. Let us start by assuming that we want to sample 10 households in each village. In general terms we will call this number *u* (for units).

110. In order to perform a sample size calculation we need information on:

(a) The variation between households within the villages;

(b) The variation between villages;

(c) The average household usage;

(d) The average usage at the village level.

111. A previous study had provided data for households in five villages, and the results are summarized below. Note that not all villages in this example are exactly the same size.

CFL average daily usage (hours)				
Village	**Number of households**	**Mean usage**[10] per household in a village	**Total usage** across all households	**Standard deviation**[11] (between households within villages)
A	8500	3.58	30458	2.60
B	8300	3.33	27667	2.70
C	8400	3.75	31500	0.66
D	8100	3.50	28350	0.75
E	8500	2.25	19125	1.50
Total number of households	**41800**			
Overall mean usage per household		3.28		
Mean usage per village			27420	
SD$_B$ = Standard deviation between villages *(SD of the total usage column)*			4889	
SD$_{W =}$ Average within village standard deviation				1.86

112. In the table above, the overall mean CFL usage is the average usage for a household, i.e.

$$Overall\ mean = \frac{30458+27667+...+19125}{41800} = 3.28$$

113. The cluster or village mean CFL usage is the average usage for village, i.e.

$$Cluster\ mean = \frac{30458+27667+...+19125}{5} = 27420$$

114. $SD_W{}^2$ is the average of the variances between households within the villages. Its square root (*SD$_W$*) is the average within village standard deviation. The equation for $SD_W{}^2$ is:

$$SD_W{}^2 = \frac{8500\times 2.60^2+...+8500\times 1.50^2}{41800} = 3.48 \text{ and so } SD_W = 1.86$$

$SD_B{}^2$ is the variance between the village total usages and its square root is the standard deviation between villages. It can be calculated using the usual equation for a variance, i.e.

$$SD_B{}^2 = \frac{\sum_{i=1}^{n}(y_i - \overline{y})^2}{n-1}$$ where the y_i are the total usages for the villages.

$$SD_B{}^2 = 23902660 \text{ and so } SD_B = 4889$$

[10] This can be a mean from all households or a mean from a sample of households.

[11] And this can be a standard deviation based on all households or a sample of households.

115. We have pre-specified that we want to sample 10 households within each village, so we need to calculate how many villages need to be sampled given a 90/10 confidence/precision criterion is required:

$$c \geq \frac{\left(\frac{SD_B}{Clustermean}\right)^2 \times \left(\frac{M}{M-1}\right) + \left(\frac{1}{u}\right) \times \left(\frac{SD_W}{Overallmean}\right)^2 \left(\frac{\overline{N}-u}{\overline{N}-1}\right)}{\left(\frac{0.1}{1.645}\right)^2 + \frac{1}{M-1}\left(\frac{SD_B}{Clustermean}\right)^2} \quad \textbf{(33)}$$

Where:

M	Total number of groups (50 villages)
$\overline{N}$	Average number of units per group (approximately 8,400 households per village)
U	Number of units that have been pre-specified to be sampled per group (pre-specified number of households to be sampled in each village = 10)
1.645	Represents the 90% confidence required
0.1	Required precision

$$c \geq \frac{\left(\frac{4889}{27420}\right)^2 \times \left(\frac{50}{50-1}\right) + \left(\frac{1}{10}\right) \times \left(\frac{1.86}{3.28}\right)^2 \left(\frac{8400-10}{8400-1}\right)}{\left(\frac{0.1}{1.645}\right)^2 + \frac{1}{50-1}\left(\frac{4889}{27420}\right)^2} = 14.9 \quad \textbf{(34)}$$

116. Therefore if we were to sample 10 households from each village we should sample 15 villages for the required confidence/precision.

117. It is usually useful to have this calculation automated so that a series of different *u* values (the number of units to be sampled in each group) can be used and the effect that this has on the number of groups to be sampled can be observed.

Number of households sampled in each village *u*	**Required number of villages** *c*
5	23
10	15
15	13
20	12
25	11

118. Compared to the cluster sampling example, more villages are required for the multi-stage sampling scheme because fewer households are being sampled within each village.

119. Note that in the above example the villages are of slightly different sizes. In practice this is likely to be the case, although the actual sizes may not always be known. This is not critical to the sample size calculation. What is important is that sensible estimates of the mean and standard

deviation at both the cluster level (village level) and unit level (household level) are used in the calculation.

Mean value parameter of interest (Brick project)

120. This section covers an example sample size calculation based on systematic sampling where the objective of the project relates to a mean value of interest.

As for all mean value parameters of interest examples we need to know:

(a) The expected mean (the desired reliability is expressed in relative terms to the mean);

(b) The standard deviation;

(c) The level of precision, and confidence in that precision (90/10 for all small-scale examples).

121. The following example is based on assessing whether bricks are of a minimum quality after manufacture; dry compressive strength has been identified as a suitable measurement of quality. Prior information gives us a mean dry compressive strength of 158kg/cm^2 with a standard deviation of 65kg/cm^2.

Example 9 – Systematic sampling

122. This example is based on a manufacturing process; we want to systematically sample every nth brick from the production line of 500,000 bricks per year. We wish to know how many bricks should be sampled to ensure an average dry compressive strength of 158kg/cm^2, with 90/10 confidence/precision.

123. The sample size equation for a required 90/10 confidence/precision is:

$$n \geq \frac{1.645^2 V}{0.1^2} \qquad \textbf{(35)}$$

Where:

$$V = \left(\frac{SD}{mean} \right)^2$$

124. Substituting in mean and standard deviation from above gives:

$$V = \left(\frac{65}{158} \right)^2 = 0.17 \qquad \textbf{(36)}$$

$$n \geq \frac{1.645^2 \times 0.17}{0.1^2} = 45.8 \qquad \textbf{(37)}$$

125. Therefore, we should take 46 samples to gain the required levels of confidence and precision. Given that 500,000 bricks are manufactured each year and we want to take 46 samples,

we should sample 1 brick for every *N/n* bricks produced – that is 1 brick for every 10,917 (= 500,000 / 46) bricks.

126. To make sure our sample is random, we randomly choose a starting point (starting brick) between 1 and 10,917 and use this brick as our first sample – for example brick 6505. We continue sampling by taking every 10,917th brick, so our second sample would be brick 6505 + 10,917 = 17,422, the third brick sampled would be 17,422 + 10,917 = 28,339, etc. For the sake of practicality it might be easier to sample every 10,000th brick; instead of every 10,917th, this would give a slightly larger sample size.

Measurements in biogas projects

Example 10

127. A survey will be carried out to estimate the mean chemical oxygen demand (COD) at a wastewater plant. The wastewater is a continuous flow of water that leaves the plant. A 500 ml sample of water will be extracted (from plant inlet) from the continuous flow of wastewater on a regular basis throughout the year and a single measurement of COD (mg/L) made on each sample.

128. This form of sampling, i.e. on a regular basis, possibly with a random start date, is systematic sampling.

129. The wastewater system has been in place for some time, and is considered to be stable in terms of the way it is functioning. The COD for the inlet is thought to be at a constant level throughout the year (apart from random variation)[12].

130. Previous work where measurements were taken on a regular basis suggested that the mean COD for untreated water is likely to be about 31,750 mg/L and the standard deviation (*SD*) in the order of 6,200 mg/L.

131. Since the wastewater is flowing continuously, the study population can be thought of as all possible 500 ml water samples in a whole year – so large as to be almost infinite. The sample size calculation no longer needs inclusion of the finite population size (i.e. N).

132. If the sampling times are sufficiently far apart the data can be regarded as a set of independent observations and treated as a simple random sample. The number of COD measurements that are required to meet the 90/10 reliability is:

$$n = \left(\frac{t_{n-1} \times SD}{0.1 \times mean} \right)^2 \qquad \textbf{(38)}$$

133. Where t_{n-1} is the value of the t-distribution for 90% confidence when the sample size is n.[13] However, the sample size is not yet known, and so a first step is to use the value for 90% confidence when the sample is large, i.e. 1.645, and then refine the calculation.

[12] In reality, temporal fluctuations (daily, weekly, seasonally, etc.) both in the wastewater flow and COD concentration should be taken into account when taking samples.

[13] This is indicated by the subscript (n-1) which is called the degrees of freedom for the t-value.

$$n = \left(\frac{1.645 \times SD}{0.1 \times mean} \right)^2 \tag{39}$$

134. This gives $n = \left(\frac{1.645 \times 6200}{0.1 \times 31750} \right)^2 = 10.3$ which rounds up to 11.

135. The calculation now needs to be repeated using the t-value for 90% confidence and n=11.

136. The exact figure for this t-value can be acquired from any set of general statistical tables or using standard statistical software. For a sample size of 11 the value is 1.812.

137. The calculation now gives $n = \left(\frac{1.812 \times 6200}{0.1 \times 31750} \right)^2 = 12.5$ which rounds up to 13.

138. The process should be iterated until there is no change to the value of n. Here the repeat calculation would have a t-value of 1.782 and the calculation would yield n = 12.11, which would be rounded up to 13. The sample size calculation suggests that sampling every four weeks should be sufficient for 90/10 reliability.

Other calculated sample sizes

139. The above is a relatively simple example, and not all situations will yield values as neat as “once a month” or “once every four weeks”. For instance:

(a) Had the calculation indicated that 48 measurements should be taken, one would most likely decide to sample weekly for the whole year;

(b) If the calculation had indicated 16 samples were required, then one might decide to sample every three weeks. Alternatively, since this may not be an easy schedule to comply with, one might choose to sample every two weeks. This now gives us a total of 26 samples which should ensure that the data, when collected and analysed, have more than adequate precision (assuming, of course, that the figures for the mean and the standard deviation that were used in the sample size calculation were good reflections of the true situation).

140. Instead of trying to follow “unworkable” schedules, it may be more sensible to use the following simplifications:

No. of measurements determined from sample size calculation	Proposed schedule
Less than or equal to 12	Monthly
13–17	Every three weeks
18–26	Every two weeks
26–51	Every week
More than 52	Twice a week

Understanding variation

141. The above example illustrates the sample size calculation using an absolute figure for the standard deviation. However, sometimes researchers have difficulty providing a figure for the standard deviation, but they can express it in relative terms. For instance, when asked about the variation in COD in this wastewater example, the researcher may describe it as 20%.

142. The coefficient of variation (CV) is a summary measure which describes variability in terms of the mean. The actual equation is $CV = \frac{SD}{mean}$. It is sometimes multiplied by 100, in which case it is describing the standard deviation as a percentage of the mean. The sample size equation on the previous page can now be written as $n = \left(\frac{t_{n-1} \times CV}{0.1}\right)^2$ where t $_{n-1}$ is the value of the t-distribution for 90% confidence for a sample of n measurements. Again the value of 1.645 would be used instead of a t-value for the first step in the calculation; and so in this example the first step would be $n = \left(\frac{1.645 \times 0.2}{0.1}\right)^2 = 10.8 \cdot$

3. Sample size calculations - Large-scale examples

143. For the large-scale examples we require 95% confidence that the margin of error in our estimate is not more than ±10% in relative terms.

Proportional parameter of interest (Transport project)

144. This section covers sample size calculations based on a proportion (or percentage) of interest being the objective of the project, under four different sampling schemes. Regardless of the sampling scheme used, the following have to be pre-determined in order to estimate the sample size based on a proportion:

(a) The value that the proportion is expected to take;

(b) The level of precision, and confidence in that precision (95/10 for all large-scale examples).

145. The examples relate to passengers travelling on the transport project in Bogota. It is known that 1,498,630 passengers use the project for transportation every day; the parameter of interest is the proportion of these passengers that would previously have travelled by bus, thought to be 45%.

Example 11 – Simple random sampling

146. Assuming that the proportion of interest is homogenous. The equation for the sample size required under simple random sampling is:

$$n \geq \frac{1.96^2 NV}{(N-1) \times 0.1^2 + 1.96^2 V} \qquad \textbf{(40)}$$

Where:

$$V = \frac{p(1-p)}{p^2}$$

n	Sample size with finite population correction
N	Total number of passengers per day
p	Our estimated proportion (45%)
1.96	Represents the 95% confidence required
0.1	Required precision

147. Substituting our values into the above equation we get:

$$V = \frac{0.45 \times (1-0.45)}{0.45^2} = 1.22 \tag{41}$$

$$n \geq \frac{1.96^2 \times 1{,}498{,}630 \times 1.22}{(1{,}498{,}630-1) \times 0.1^2 + 1.96^2 \times 1.22} = 469.4 \tag{42}$$

148. The required sample size is at least 470 passengers to get an estimated proportion of passengers that would have previously travelled by bus with 95/10 confidence/precision. Note that the sample size will change depending on the estimated proportion value.

149. The above sample size does not take into account any non-responders, that is passengers who do not respond to the question. If we expect that 90% of passengers will respond (and 10% will not) then we should increase the sample size by dividing the sample size calculated above by the expected level of response: 470/0.9 = 523. To account for a non-response of 10% we should sample 523 passengers.

150. Please see the ‘Approximate equation’ section under **i) Cook Stove Project - Proportional parameter of interest**, **Example 1: Simple random sampling** for notes relating to approximate equations. Note that 1.96 should be used in place of 1.645 to account for the increased confidence required for the large-scale projects.

Example 12 – Stratified random sampling

151. Suppose that we believe the proportion of transport project passengers that would have previously used the bus would vary between the eight different zones in which the project operates. We would like to make sure that when we do our sampling, our sample includes a representative proportion of passengers from each zone. To calculate the sample size, estimates of the number of passengers and proportion that would have previously travelled by bus within each zone are required.

Zone	Number of passengers within each zone (*g*)	Estimated proportion of passengers that would have used a bus (*p*)
A	19865	0.43
B	21358	0.57
C	301245	0.4
D	65324	0.71
E	654832	0.32
F	50213	0.46
G	12489	0.26
H	373304	0.68

152. The equation for the total sample size is:

$$n \geq \frac{1.96^2 NV}{(N-1)\times 0.1^2 + 1.96^2 V} \tag{43}$$

Where: $V = \frac{SD^2}{\overline{p}^2} = \frac{\text{overall variance}}{\left(\text{overall proportion}\right)^2}$

153. To then decide on the number of passengers in the sample that come from each zone we could use proportional allocation, where the proportions of units from the different zones in the sample is the same as the proportions in the population. This gives:

$n_i = \frac{g_i}{N}\times n$ where $\mathrm{i} = 1,\ldots,\mathrm{k}$ where k is the number of zones (in this case 8).

Where:

g_i Size of the ith group (district) where i=1,…,k

N Population total

154. Using the figures from the table we can calculate the overall variance,14 and overall proportion:

$$SD^2 = \frac{(g_a \times p_a(1-p_a)) + (g_b \times p_b(1-p_b)) + (g_c \times p_c(1-p_c)) + \ldots + (g_k \times p_k(1-p_k))}{N} \tag{44}$$

$$\overline{p} = \frac{(g_a \times p_a) + (g_b \times p_b) + (g_c \times p_c) + \ldots + (g_k \times p_k)}{N} \tag{45}$$

Where g_i and N are as above and p_i is the proportion of the ith group (district) where i=1,…,k.

[14] The variance of a proportion is calculated as: p(1-p).

$$SD^2 = \frac{(19865 \times 0.43 \times 0.57) + (21358 \times 0.57 \times 0.43) + ... + (373304 \times 0.28 \times 0.72)}{1,498,630} = 0.22 \quad \textbf{(46)}$$

$$\bar{p} = \frac{(19865 \times 0.43) + (21358 \times 0.57) + ... + (373304 \times 0.28)}{1,498,630} = 0.45 \quad \textbf{(47)}$$

155. Therefore:

$$V = \frac{SD^2}{\bar{p}^2} = \frac{0.22}{0.45^2} = 1.09 \quad \textbf{(48)}$$

156. Substituting in our *V* gives:

$$n \geq \frac{1.96^2 \times 1,498,630 \times 1.09}{(1,498,630 - 1) \times 0.1^2 + 1.96^2 \times 1.09} = 419.6 \quad \textbf{(49)}$$

157. The total sample size required is 420 passengers. The next step is to divide this total sample size up according to the size of each zone to get the number of passengers to be sampled within each zone.

General Equation: $n_i = \frac{g_i}{N} \times n$ **(50)**

Zone A: $n_a = \frac{19865}{1,498,630} \times 420 = 5.6$

Zone B: $n_b = \frac{21358}{1,498,630} \times 420 = 6.0$

Zone C: $n_c = \frac{301245}{1,498,630} \times 420 = 84.4$

Zone D: $n_d = \frac{65324}{1,498,630} \times 420 = 18.3$

Zone E: $n_e = \frac{654832}{1,498,630} \times 420 = 183.5$

Zone F: $n_f = \frac{50213}{1,498,630} \times 420 = 14.1$

Zone G: $n_g = \frac{12489}{1,498,630} \times 420 = 3.5$

Zone H: $n_h = \frac{373304}{1,498,630} \times 420 = 104.6$

158. Rounding up the zone sample sizes gives the number of passengers to be sampled in each zone (the sum of these is slightly greater than the required sample size due to the rounding up of passengers within each zone). The sample sizes required vary so much between the zones because the number of passengers in each zone is so different.

159. Note that these sample sizes do not take into account non-response. If the expected level of response is 85% across all zones then divide each zone sample size by 0.85. This will result in larger sample sizes allowing for the non-responders.

Example 13 – Cluster sampling

160. Instead of sampling individual passengers, it has been decided that buses (clusters) are going to be sampled and then every passenger on each of the selected buses will be asked if they travelled by bus prior to the project. To calculate the sample size we require the number of clusters that make up the population, that is the number of buses that carry the transport project passengers; for this example we will assume 12,000 buses. We also need estimated proportions of passengers that would have travelled by bus prior to the project from a number of buses; for this example we have previously sampled four buses and the proportions were:

Bus	Estimated Proportion
1	0.37
2	0.46
3	0.28
4	0.52
Average $(\overline{p})$	0.4075
Variance $({SD_B}^2)$	0.011

161. The equation for the number of buses that need to be sampled is:

$$c \geq \frac{1.96^2 MV}{(M-1)\times 0.1^2 + 1.96^2 V} \tag{51}$$

Where:

$$V = \frac{{SD_B}^2}{\overline{p}^2} = \frac{\text{Variance between clusters(buses)}}{\text{Average proportion over clusters}}$$

162. The average proportion is just $\frac{0.37+0.46+0.28+0.52}{4} = \frac{1.63}{4} = 0.41$ and the variance between the clusters is:

$${SD_B}^2 = \frac{1}{n-1}\sum_{i=1}^{n=5}(p_i - \overline{p})^2 = \frac{(0.37-0.4065)^2 + (0.46-0.4065)^2 + \ldots + (0.52-0.4065)^2}{3} = 0.0110 \tag{52}$$

Where

c	Number of clusters to be sampled (buses)
M	Total number of clusters (buses) - this must encompass the entire population
1.96	Represents the 95% confidence required
0.1	Represents the 10% relative precision

163. Substituting our values into the above equation gives:

$$V = \frac{SD_B^2}{\overline{p}^2} = \frac{0.0110}{0.41^2} = 0.07 \tag{53}$$

$$c \geq \frac{1.96^2 \times 12000 \times 0.0664}{(12000-1) \times 0.1^2 + 1.96^2 \times 0.0664} = 25.5 \tag{54}$$

164. Therefore we would have to sample every passenger on 26 randomly selected buses.

165. This approach to sampling assumes that the population is homogenous. In this example this means that the proportion of passengers that would have previously travelled by bus is independent of any other factors such as zones (see example 12 – stratified sampling), economic status, etc. If the proportion of passengers that would have previously travelled by bus is expected to be different for different zones, then cluster sampling should be used within each zone.

166. Non-response is unlikely to be a problem when using cluster sampling, unless the number of individuals within a cluster could be 0 (a bus with no passengers). If it is thought that it could be a problem then the sample size should be scaled up accordingly.

Example 14 – Multi-stage sampling

167. Instead of sampling every passenger on a number of selected buses, suppose we only want to sample a number of passengers on each bus. This can be thought of as multi-stage sampling as we are sampling a number of buses (groups), and then going on to sample units (passengers) within each group.

168. We know that there are 12,000 buses and there are on average 30 passengers on each bus, of which we plan to sample 15. From a small pilot study we already know the following:

Bus	**Proportion of passengers that would have travelled by bus**
1	0.37
2	0.46
3	0.28
4	0.52

169. The equation for the number of buses to be sampled is:

$$c \geq \frac{\frac{SD_B^2}{\overline{p}^2} \times \frac{M}{M-1} + \frac{1}{\overline{u}} \times \frac{SD_w^2}{\overline{p}^2} \times \frac{(\overline{N}-\overline{u})}{(\overline{N}-1)}}{\frac{0.1^2}{1.96^2} + \frac{1}{M-1}\frac{SD_B^2}{\overline{p}^2}} \tag{55}$$

Where:

c	Number of groups that should be sampled
M	Total number of groups in the population (12,000 buses)
$\bar{u}$	Number of units to be sampled within each group (pre-specified as 15 passengers)
$\bar{N}$	Average units per group (30 passengers on each bus)
$SD_B{}^2$	Unit variance (variance between buses)
$SD_W{}^2$	Average of the group variances (average within bus variation)
$\bar{p}$	Overall proportion
1.96	Represents the 95% confidence required
0.1	Represents the 10% absolute precision

170. Using our table of pilot information we can calculate the unknown quantities for the equation above:

Bus	Proportion of passengers that would have used a bus p_i	Variance within bus $p_i(1\text{-}p_i)$
1	0.37	0.2331
2	0.46	0.2484
3	0.28	0.2016
4	0.52	0.2496
Variance	$SD_B{}^2 = 0.0110$	
Average	$\bar{p} = 0.41$	$SD_W{}^2 = 0.2332$

Where:

$\bar{p}$ is the average proportion of passengers who travel by bus, i.e. $\bar{p} = \dfrac{0.37 + ... + 0.52}{4} = 0.41$

$SD_W{}^2$ is the average variance between passengers on n a bus, i.e.

$$SD_W{}^2 = \frac{0.2331 + ... + 0.2496}{4} = 0.2332$$

$SD_B{}^2$ is the variance between the bus proportions, i.e. the variance between 0.37, 0.48, etc. This can be calculated in the usual way for calculating a variance, i.e. using the equation

$$SD_B{}^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$$ which gives $SD_B{}^2 = 0.0110$

171. Substituting our values into the group sample size equation gives:

$$c \geq \frac{\frac{0.0110}{0.41^2} \times \frac{12000}{(12000-1)} + \frac{1}{15} \times \frac{0.2332}{0.4075^2} \times \frac{(30-15)}{(30-1)}}{\frac{0.1^2}{1.96^2} + \left(\frac{1}{(12000-1)} \times \frac{0.0110}{0.41^2} \right)} = 44.0 \qquad \textbf{(56)}$$

172. Therefore if we were to sample 15 passengers from each bus we should sample 44 buses for the required confidence/precision. The table below gives the number of buses required (c) when choosing to sample different numbers of passengers from each bus (u).

Number of passengers sampled on each bus *u*	**Required number of buses** *c*
5	119
10	63
15	44
20	35
30	26

173. The required sample size from the cluster sampling scheme example was 26; this is the same as the sample size required under the multi-stage sampling scheme when u=30 (the number of passengers sampled from each bus). This is because we assumed an average of 30 passengers on each bus in the calculations, so when we take u=assumed average passengers the two sampling schemes are the same.

Mean value parameter of interest (Transport project)

174. This section covers sample size calculations where the objective of the project relates to a mean value of interest, under four different sampling schemes. For the sample size calculations, we need to know:

(a) The expected mean (the desired reliability is expressed in relative terms to the mean);

(b) The standard deviation;

(c) The level of precision, and confidence in that precision (95/10 for all large-scale examples).

175. The parameter of interest in the examples below is average journey length (km) of people who travel by car, whether this is a domestic car or a taxi, and for people who travel by bus.

Example 15 – Simple random sampling

176. Suppose we are interested in the average distance (km) of car journeys in Bogota in a day (including both domestic cars and taxis), and we assume that the journeys are homogenous. We know that 2,000,000 journeys are made each day and believe that the mean is 8 km with a standard deviation of 3.5 km. Using simple random sampling the sample size equation is:

$$n = \frac{1.96^2 NV}{(N-1)\times 0.1^2 + 1.96^2 V} \qquad \textbf{(57)}$$

Where:

$$V = \left(\frac{SD}{mean}\right)^2$$

n	Sample size
N	Total number of journeys (2,000,000)
mean	Expected mean journey length (8 km)
SD	Expected standard deviation for the journey length (3.5 km)
1.96	Represents the 95% confidence required
0.1	Represents the 10% precision

$$V = \left(\frac{3.5}{8}\right)^2 = 0.19 \qquad \textbf{(58)}$$

$$n = \frac{1.96^2 \times 2{,}000{,}000 \times 0.19}{(2{,}000{,}000-1)\times 0.1^2 + 1.96^2 \times 0.19} = 73.5 \qquad \textbf{(59)}$$

177. Therefore the required sample size is at least 74 journeys to find the average journey length with 95% confidence and a 10% relative margin of error.

178. The calculation above does not take into account non-response. If a level of non-response is expected within the sample then the sample size should be scaled up accordingly. For example if we expected 95% of the people on journeys sampled to respond then we should take this into account and plan to sample 74/0.95 = 78 journeys instead of 74.

179. There is an approximate equation for this sample size calculation. Please see the ‘Approximate equation’ section under **ii) CFL Project – Mean value parameter of interest**, **Example 5: Simple random sampling** for notes relating to approximate equations. Note that 1.96 should be used in place of 1.645 to account for the increased confidence required in the large-scale projects.

Example 16 – Stratified random sampling

180. The fundamental aspect of this sampling scheme is that the average journey length differs between domestic cars and taxis (it is not homogenous as assumed in the previous example). Because we know that the type of vehicle affects the journey distance, we want to make sure that we sample a representative number of domestic cars and taxis. A summary of each stratification group is given below:

Stratification group	Number of journeys per day	Mean (km)	Standard Deviation (km)
Domestic Car	1,595,169	9	3.7
Taxi	982,224	7	2.5

181. Using the data in the table above we can estimate the overall mean and standard deviation:

Overall mean:

$$mean = \frac{(g_a \times m_a) + (g_b \times m_b) + (g_c \times m_c) + ... + (g_k \times m_k)}{N} \quad \textbf{(60)}$$

Where:

mean	Weighted overall mean
g_i	Size of the ith group where i=1,…,k
m_i	Mean of the ith group where i=1,…,k
N	Population total

Substituting the values from our example into the above expression gives:

$$mean = \frac{(1595169 \times 9) + (982224 \times 7)}{(1595169 + 982224)} \quad \textbf{(61)}$$

$$mean = 8.2$$

Overall Standard Deviation:

$$SD = \sqrt{\frac{(g_a \times SD_a^2) + (g_b \times SD_b^2) + (g_c \times SD_c^2) + ... + (g_k \times SD_k^2)}{N}} \quad \textbf{(62)}$$

Where:

SD	Weighted overall standard deviation
SD_i	Standard deviation of the ith group where i=1,…,k, (note that these are all squared – so the group size is actually being multiplied by the group variance)

182. Using the values from our example gives:

$$SD = \sqrt{\frac{(1595169 \times 3.7^2) + (982224 \times 2.5^2)}{(1595169 + 982224)}} \quad \textbf{(63)}$$

$$SD = 3.3$$

183. The sample size equation uses the overall mean and standard deviation calculated above:

$$n \geq \frac{1.96^2 \times NV}{(N-1) \times 0.1^2 + 1.96^2 V} \quad \textbf{(64)}$$

184. Substituting in the values from our examples gives:

$$V = \left(\frac{SD}{mean}\right)^2 = \left(\frac{3.3}{8.2}\right)^2 = 0.16 \tag{65}$$

$$n = \frac{1.96^2 \times 2577393 \times 0.16}{(2577393-1)\times 0.1^2 + 1.96^2 \times 0.16} = 61.4 \tag{66}$$

185. This gives us the total number of journeys that should be sampled across both vehicle types. The section below assumes proportional allocation – which means that the number of journeys we want to sample from each vehicle type is proportional to the number of journeys made by each vehicle type within the population.

General equation: $n_i = \frac{g_i}{N} \times n$ **(67)**

Domestic cars: $n_{Car} = \frac{1592169}{2577393} \times 62 = 38.4$ Taxis: $n_{Taxi} = \frac{982224}{2577393} \times 62 = 23.6$

186. Rounding these figures results in a sample consisting of 39 domestic car journeys, and 24 taxi journeys. The summation of these group sample sizes (39 + 24 = 63) is slightly greater than that calculated from the equation above due to rounding.

187. The sample size calculated above assumes 100% response, and therefore needs to be scaled up where non-response is likely to occur.

188. This sample size is smaller than that from simple random sampling in this example. This is due to the standard deviations within strata being smaller than the standard deviation across the whole population (which is usually the case).

Example 17 – Cluster sampling

189. Now consider a different scenario. The parameter of interest is the average journey length of people who take local buses. Instead of sampling numerous individual passengers we would like to sample everyone from a few buses (clusters). Knowing that there are 12,000 local buses in Bogota each day, how many buses would we have to sample to find the average journey length with 95/10 confidence/precision?

190. The equation used to give us the required number of clusters, *c*, to sample is:

$$c \geq \frac{1.96^2 MV}{(M-1)\times 0.1^2 + 1.96^2 V} \tag{68}$$

Where:

$$V = \left(\frac{SD}{Cluster\ mean} \right)^2$$

c	Number of clusters (buses) to be sampled
M	Total number of clusters (buses)
1.96	Represents the 95% confidence required
0.1	Required precision (the equation takes into account that this is relative)

191. To perform the calculations we need information about journey length at the bus level, i.e. total journey length aggregated across all passengers on a bus. If such information does not already exist, we might collect it in a pilot study. The example here assumes that data are available from four buses.

Bus	**Total journey length (on average)**[15]
A	195
B	96
C	63
D	159

192. Calculating the mean and standard deviation for this total journey length for a bus gives us:

$$Cluster\ mean\ \text{(i.e. } \bar{y}) = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{195+96+63+159}{4} = 128.25 \quad \textbf{(69)}$$

$$SD = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2} = \sqrt{\frac{(196-128.25)^2 + ... + (159-128.25)^2}{3}} = 59.7180 \quad \textbf{(70)}$$

193. These statistics (i.e. mean and SD of data) are easily produced using standard statistical software. Substituting these values into the equation gives the required number of clusters, i.e. buses as:

$$V = \left(\frac{59.7180}{128.25} \right)^2 = 0.22 \quad \textbf{(71)}$$

$$c \geq \frac{1.96^2 \times 12000 \times 0.22}{(12000-1) \times 0.1^2 + 1.96^2 \times 0.22} = 82.7 \quad \textbf{(72)}$$

194. The total number of buses that should be sampled is 83. Asking the journey lengths from everyone on each of the 83 buses sampled will satisfy the 95/10 confidence/precision criterion.

[15] These totals may be derived from collecting data on all individual passengers on a bus, or otherwise by taking a sample of them and scaling up from the sample to all the passengers on the bus.

Example 18 – Multi-stage sampling

195. Continuing the previous example, suppose that we want to sample a number of local buses, but we only want to sample a number of individuals on each bus (unlike cluster sampling where everyone on each selected bus is sampled). This is an example of multi-stage sampling, as we are sampling a number of groups (buses), and we are then going on to sample a number of units (passengers) from each selected group (bus).

196. We start by assuming that we want to sample five passengers on each selected bus. In general terms we will call this number *u* (for units).

197. In order to be able to perform a sample size calculation we need information on:

(a) The variation between individual passengers on the bus;

(b) The variation between buses;

(c) The average journey length for a passenger;

(d) The average journey length at the bus level (when aggregated across all passengers).

198. A previous study had provided data for passengers on three different buses, and the results are summarized below.

199. Note that not all buses will have exactly the same number of passengers.

Journey length (km)				
Bus	**Number of passengers**	**Mean journey length**[16]	**Total journey length** aggregated over all passengers on the bus	**Standard deviation**[17] (between passengers on the same bus)
A	26	6.9	179	3.30
B	21	7.5	157	6.21
C	30	6.7	200	3.78
Total number of passengers	**77**			
Overall mean journey length per passenger		7.0		
Mean total journey length (per bus)			179	
SD_B = Standard deviation between buses *(SD of the total journey length column)*			4889	
$SD_{W =}$ Average between passenger (within bus) standard deviation				4.44

200. In the table above, the overall mean journey length is the average length per passenger, i.e.

[16] This can be a mean from all passengers on the bus or a mean from a sample of passengers.

[17] And this can be a standard deviation based on all passengers or from a sample of passengers.

$$Overall\ mean = \frac{179+157+200}{77} = 7.0 \qquad \textbf{(73)}$$

201. The cluster mean journey length is the mean length per bus, i.e.

$$Cluster\ mean = \frac{179+157+200}{3} = 179 \qquad \textbf{(74)}$$

202. $SD_W{}^2$ is the average variance between passengers within buses. Its square root (SD_W) is the average within bus standard deviation. The equation for $SD_W{}^2$ is:

$$SD_W{}^2 = \frac{26\times 3.30^2 + 21\times 6.21 + 30\times 3.78^2}{77} = 19.75 \text{ and so } SD_W = 4.44 \qquad \textbf{(75)}$$

203. $SD_B{}^2$ is the variance between the mean journey lengths per bus. Its square root is the standard deviation between buses. It can be calculated using the general equation for a variance:

$$SD_B{}^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$$ where the y_i are the journey lengths for the different buses.

$$SD_B{}^2 = 467 \text{ and so } SD_B = 22 \qquad \textbf{(76)}$$

204. As well as the information from the table above we also require the number of buses, and the average number of passengers on each bus for the whole population. For this example we are using 12,000 buses with an average of 30 passengers on each bus.

$$c \geq \frac{\left(\frac{SD_B}{Clustermean}\right)^2 \times \left(\frac{M}{M-1}\right) + \left(\frac{1}{u}\right) \times \left(\frac{SD_w}{Overallmean}\right)^2 \left(\frac{\overline{N}-u}{\overline{N}-1}\right)}{\left(\frac{0.1}{1.96}\right)^2 + \frac{1}{M-1}\left(\frac{SD_B}{Clustermean}\right)^2} \qquad \textbf{(77)}$$

Where:

M	Total number of groups (12,000 buses)
$\overline{N}$	Average number of units per group (30 passengers per bus)
u	Number of units that have been pre-specified to be sampled per group (pre-specified number of passengers to be sampled on each bus = 5)
1.96	Represents the 95% confidence required
0.1	Required precision

$$c \geq \frac{\left(\frac{22}{179}\right)^2 \times \left(\frac{12000}{12000-1}\right) + \left(\frac{1}{5}\right) \times \left(\frac{4.44}{7.0}\right)^2 \left(\frac{30-5}{30-1}\right)}{\left(\frac{0.1}{1.96}\right)^2 + \frac{1}{12000-1}\left(\frac{22}{179}\right)^2} = 32.6 \quad \textbf{(78)}$$

205. Therefore if we were to sample five passengers from each bus we should sample 33 buses for the required confidence/precision.

206. Producing a table such as the one below, with different values of *u*, can help decide the practicalities of allocating limited resources, while still satisfying the 95/10 confidence/precision criterion.

Number of passengers sampled on each bus *u*	**Required number of buses** *c*
5	33
10	17
15	12
20	9
25	7

207. In this example, by doubling the number of passengers on each bus to be sampled from 5 to 10, we substantially reduce the number of buses that need to be sampled from 33 to 17.

208. Note that in the above example the numbers of passengers on a bus were different for the different buses. In practice this is likely to be the case, although the actual numbers may not always be known. This is not critical to the sample size calculation. What is important is that sensible estimates of the mean and standard deviation at both the cluster level (bus level) and unit level (passenger level) are used in the calculation.

Mean value parameter of interest (Transport project)

209. This section covers an example sample size calculation based on systematic sampling where the objective of the project relates to a mean value of interest.

210. As for all absolute parameter of interest examples we need to know:

(a) The expected mean (the desired reliability is expressed in relative terms to the mean);

(b) The standard deviation;

(c) The level of precision, and confidence in that precision (95/10 for all large-scale examples).

211. The parameter of interest for the example below is the average journey time of buses on a specific route. We know that over a month 960 journeys are made on the route of interest, and the average journey time is 18 minutes with a standard deviation of 6 minutes.

Example 19 – Systematic sampling

212. Using systematic sampling we want to sample every n^th journey on that route. The sample size equation for a required 95/10 confidence/precision is:

$$n \geq \frac{1.96^2 V}{0.1^2} \tag{79}$$

Where:

$$V = \left(\frac{SD}{mean} \right)^2$$

213. Substituting in mean and standard deviation from above gives:

$$V = \left(\frac{6}{18} \right)^2 = 0.11 \tag{80}$$

$$n \geq \frac{1.96^2 \times 0.11}{0.1^2} = 42.7 \tag{81}$$

214. In total we should sample 43 journey times. We want to take these samples evenly spread over the 960 journeys made each month, therefore we should sample one journey for every *N/n* – that is one journey every 22 (= 960 / 43).

215. We can make sure that we sample at random by selecting a random starting point between 1 and 22, say journey 18, and then sample every 22nd journey from this point on: 18, 40, 62, 84, 106, etc. up to 960. This would give us a sample evenly spread over the month that is large enough to estimate the average journey time with 95/10 confidence/precision.

216. It may be more practical to sample every 20th journey rather than every 22nd. This would result in more samples being taken than the 43 calculated above – the only effect this would have would be to increase the precision and so would be perfectly acceptable.

Appendix B

Best practice examples for reliability calculations

1. Introductory notes on reliability calculations

217. The two examples presented here illustrate how to estimate a numeric parameter and a proportion and how to check their reliability. The sampling method used in both cases is simple random sampling. Both examples are assumed to be small-scale project activities where the required reliability criteria is 90/10, i.e. 90% confidence and 10% precision.

218. If calculations are being performed manually, it is important to retain as many decimal places as relevant, until the final calculated figure is reached. Rounding can then be carried out. To emphasize this, the calculations presented here use figures to several decimal places.

2. Example 1: CFL Project – Numeric parameter

219. The parameter of interest in this example is the mean average daily usage of a CFL (in hours) for a whole population of CFLs that were distributed in a particular region of a country.

220. The population is the 420,000 households to which CFLs were distributed, one per household. A simple random sample of 140 households was taken, and the average daily usage (in hours) of each CFL was recorded. These are presented in the table below.

Average CFL usage (in hours)

cfl	usage	cfl	usage	cfl	usage	cfl	usage	cfl	usage	cfl	usage	cfl	usage
1	3.78	21	3.63	41	2.81	61	4.17	81	3.62	101	2.24	121	0.58
2	3.12	22	3.17	42	4.57	62	4.68	82	2.46	102	4.79	122	6.09
3	4.42	23	3.26	43	3.56	63	2.99	83	6.14	103	4.59	123	0.39
4	4.09	24	6.97	44	4.41	64	3.34	84	0.67	104	3.27	124	3.69
5	1.15	25	0.48	45	3.26	65	5.37	85	4.73	105	1.86	125	2.04
6	2.87	26	2.50	46	0.30	66	2.17	86	1.03	106	0.00	126	4.51
7	4.79	27	2.92	47	5.48	67	2.36	87	2.34	107	6.70	127	4.39
8	4.20	28	6.82	48	1.75	68	3.12	88	4.66	108	3.36	128	3.58
9	1.13	29	0.92	49	3.38	69	4.69	89	2.40	109	5.39	129	4.23
10	3.68	30	2.35	50	1.24	70	5.40	90	5.28	110	2.04	130	5.28
11	2.91	31	0.19	51	3.62	71	4.22	91	5.90	111	3.58	131	3.71
12	2.47	32	4.19	52	7.41	72	1.27	92	0.60	112	6.27	132	2.41
13	3.46	33	3.15	53	1.74	73	2.93	93	5.85	113	0.41	133	1.58
14	2.19	34	3.19	54	3.60	74	2.17	94	1.22	114	4.55	134	3.96
15	2.25	35	7.15	55	2.18	75	4.24	95	7.76	115	2.61	135	5.86
16	2.37	36	1.70	56	4.12	76	6.07	96	4.50	116	6.37	136	5.46
17	2.38	37	2.98	57	4.88	77	5.26	97	5.68	117	4.30	137	2.90
18	3.23	38	5.00	58	2.92	78	2.46	98	2.81	118	3.08	138	3.17
19	1.78	39	0.99	59	0.82	79	1.33	99	4.03	119	3.17	139	4.17
20	3.57	40	6.54	60	3.16	80	2.55	100	0.24	120	6.24	140	6.93

221. The parameter of interest – the mean average daily usage of a CFL (in hours) for this whole population of CFLs – is estimated from the sample mean. This is often written as $\overline{y}$ (and is equal to $\frac{1}{n}\left(y_1 + y_2 + \ldots + y_n\right)$, or the shorthand form $\frac{1}{n}\sum_{i=1}^{n} y_i$). *n* is the sample size, i.e. 140.

222. The mean average usage for the sample of 140 CFLs is 3.4686 hours. As a simple summary this is rounded to 1 or 2 decimal places, i.e. the mean average usage of the CFLs is estimated to be 3.47 hours.

Confidence, precision and reliability

223. Instead of presenting just a single estimate, it is better to summarize the results of sampling using a confidence interval. In this example the 90% confidence interval is 3.22 to 3.71 hours. We are 90% sure that the true population mean value for average usage of a CFL is between 3.22 hours and 3.71 hours. Whilst the sample mean is the estimate that will be used in calculations, it is always advisable when presenting it in a report to do so along with its confidence interval.

224. The 90% confidence interval for the population mean is given by the equation: sample mean ± t-value × standard error of the mean.

225. The estimate of 3.47 hours is regarded as reliable if the precision of the study – as defined by the t-value × standard error of the mean – is within the pre-specified reliability precision. For small-scale mechanisms this is 10% of the mean.

226. Detailed calculations are presented below. In this example the precision is 7.1% of the mean and so the sample estimate of 3.47 hours is within the required specification.

Checking reliability

(i) Standard error of the mean

227. The equation for the standard error of the mean when data have been collected using simple random sampling is $\sqrt{\left(1-f\right)\frac{s^2}{n}}$.

f is the sampling fraction – the proportion of the population that is sampled.

Here it is $\frac{140}{420000} = 0.00003$.

s^2 is the sample variance (s is the sample standard deviation).

For this sample of 140 CFLs, $s^2 = 3.0826$ and s = 1.7557.

n is the sample size, i.e. 140.

228. Putting all these pieces of information together gives:

$$\sqrt{(1-f)\frac{s^2}{n}} = \sqrt{\left(1-\frac{140}{420000}\right)\times\frac{3.0826}{140}} = \sqrt{0.99967\times\frac{3.0826}{140}} = \sqrt{0.0220} = 0.1484 \qquad \textbf{(82)}$$

and so the standard error of the mean is 0.1484.

(ii) t-value

229. This value depends on (i) the level of confidence and (ii) the size of the sample. The exact figure can be acquired from statistical tables for the t-distribution, or using standard statistical software. The value can also be derived in Microsoft Excel using the TINV[18] function.

For a sample size of 140 the t-value is 1.6559.

(iii) Precision

230. The precision associated with an estimate is: t-value × standard error of the mean.

The precision of the mean average CFL usage (in hours), assuming 90% confidence, in this example is therefore: ± (1.6559 × 0.1484) i.e. ± 0.2457.

The ratio of this relative to the mean CFL usage is $\frac{0.2457}{3.4686} = 0.0708$ and so the relative precision is 7.1%. The data are therefore within the required specification.

Another way of checking reliability

231. The limits of the confidence interval are sample mean ± t-value × standard error of the mean, which can be written more generally as sample mean ± precision, where the lower limit is mean minus precision and the upper limit is mean + precision.

232. Reliability can therefore be checked using the following calculation:

$$\frac{\text{½width of confidence interval}}{\text{mean}}\times 100\% \qquad \textbf{(83)}$$

233. For example, here the mean CFL usage is 3.4686, and the 90% confidence interval is 3.2230 to 3.7143 hours. Reliability is therefore:

$$\frac{\text{½}(3.7143\text{-}3.2230)}{3.4686}\times 100\% = \frac{\text{½} \times 0.4913}{3.4686}\times 100\% = 7.1\% \qquad \textbf{(84)}$$

234. The above approach is likely to be most useful when the data have been analysed using statistical software which produced the relevant confidence interval as well as the sample mean.

[18] TINV(0.10,(sample size minus 1)) will give the t-value associated with 90% confidence. For example here TINV(0.10,139) gives the t-value for a sample size of 140 and 90% confidence.

3. Example 2 : Cook stove project – Proportional parameter

235. The parameter of interest in this example is the proportion (or percentage) of cook stoves in a particular region of a country that were still operational at the end of the third year after the stoves were distributed.

236. The population of interest is the 640,000 households, and there was one cook stove per household. A simple random sample of 274 of these households was taken, and for each of them it was recorded whether or not the cook stove was still operational.

237. The parameter of interest – the proportion (or percentage) of cook stoves that were still operational in the whole population – is estimated from the sample proportion.

238. This is often written as p and is calculated as $p = \frac{r}{n}$ where r is the number of "successes", in this case the number of cook stoves that are still in operation, and n is the total number of cook stoves that are observed in the sample.

239. In this example there were 159 cook stoves out of the 274 that were still in operation. The sample proportion is therefore $p = \frac{159}{274} = 0.5803$. Rounding this to two decimal places gives us a proportion of 0.58. In other words, 58% of the cook stoves were still operational after the third year.

Confidence, precision and reliability

240. Instead of presenting just a single estimate, it is better to summarize the results of sampling using a confidence interval. In this example the 90% confidence interval for the proportion is 0.5313 to 0.6293. We are therefore 90% sure that the percentage of cook stoves in the population that are still operational is between 53% and 63%.

241. The 90% confidence interval for the population proportion is given by the equation: sample proportion ± 1.6449 × standard error of the proportion.[19]

242. The estimate of 58% is regarded as reliable if the precision of the study – as defined by 1.6449 × standard error of the proportion – is within the pre-specified reliability precision. For small-scale mechanisms this is 10% of the proportion. In this case ±0.058 in absolute terms or ± 5.8%.

243. Detailed calculations are presented below. In this example the precision is 8.5% of the sample proportion and so the sample estimate of 58% operational cook stoves is within the required specification.

[19] A confidence interval for a proportion is: sample proportion ± z-value × standard error of the proportion. The z-value depends on the level of confidence. For 90% confidence it is 1.6449.

Checking reliability

(i) Standard error of the proportion

244. The equation for the standard error of the proportion when data have been collected using simple random sampling is $\sqrt{(1-f)\frac{pq}{n}}$

f is the sampling fraction – the proportion of the population that is sampled.

Here it is $\frac{274}{640000} = 0.00043$

p is the sample proportion, i.e. 0.5803. *q = (1-p).* It represents the proportion of cook stoves that are not operational after three years, and is 0.4197. *n* is the sample size, i.e. 274.

245. Putting all these pieces of information together gives:

$$\sqrt{(1-f)\frac{pq}{n}} = \sqrt{(1-0.00043)\frac{0.5803\times 0.4197}{274}} = \sqrt{0.00089} = 0.0298 \qquad \textbf{(85)}$$

246. Note that this standard error could also be calculated using the actual numbers of population size, sample size, number of operational cook stoves etc., i.e.:

$$\sqrt{(1-f)\frac{pq}{n}} = \sqrt{\left(\frac{640000-274}{640000}\right)\frac{\left(\frac{159}{274}\right)\left(\frac{115}{274}\right)}{274}} = \sqrt{0.00089} = 0.0298 \qquad \textbf{(86)}$$

247. The standard error of the proportion is 0.0298. In terms of the standard error of the percentage it is 2.98%.

(ii) Precision

248. The precision associated with a proportion is: z-value × standard error of the proportion. The precision of the proportion of operational cook stoves in this example, assuming 90% confidence, is: ± (1.6449 × 0.0298) i.e. ± 0.0490.

249. The ratio of this relative to the proportion of cook stoves that are still operational is $\frac{0.0490}{0.5803} = 0.0845$ and so the relative precision is 8.5%. The data are within the required specification.

Another way of checking reliability

250. The limits of the confidence interval are sample proportion ± z-value × standard error of the proportion, which can be written more generally as sample proportion ± precision, where the lower limit is the proportion minus precision and the upper limit is the proportion plus precision.

251. Reliability can therefore be checked using the following calculation:

$$\frac{\text{½width of confidence interval}}{\text{proportion}} \times 100\% \tag{87}$$

For example, here the proportion of cook stoves that are still operational is 0.5803, with a 90% confidence interval of 0.5313 to 0.6293.

Reliability is therefore:

$$\frac{½(0.6293 - 0.5313)}{0.5803} \times 100\% \ = \ \frac{½ \ \times \ 0.0980}{0.5803} \times 100 \ = \ 8.5\% \tag{88}$$

252. The above approach is likely to be most useful when the data have been analysed using statistical software which produced the relevant confidence interval and sample proportion.

Comments

253. The equation above assumes that the distribution of the proportion is approximately Normal. That is usually an acceptable assumption provided the proportion of interest is not too small and not too large, and the sample size is not too small;

254. If the sampling fraction *f* is small then the multiplier (1 – *f*) in the above calculation will be very close to 1. In some instances, therefore, the equation that is used for the standard error of the proportion is the conservative equation $\sqrt{\frac{pq}{n}}$;[20]

255. If statistical software is used to undertake the calculation, the software may use the exact equation for calculating the confidence interval (which assumes a Binomial distribution as opposed to the Normal approximation). In this case the reliability would be checked using the equation which is based on the width of the confidence interval.

A. How to deal with failure to achieve reliability

1. Introductory notes on how to deal with failure to achieve reliability

256. This section proposes some steps to follow when the required reliability is not met by sample data. It uses as its scenario a small-scale CDM project activity where the reliability criteria is 90:10 (i.e. 90% confidence and 10% relative precision), and where the parameter of interest is a numeric one.

2. Scenario : CFL project – Numeric parameter

257. The parameter of interest in this example is the mean average daily usage of a CFL (in hours) for a whole population of CFLs that were distributed in a particular region of a country. The

[20] It is conservative because $\sqrt{\frac{pq}{n}}$ will be greater than $\sqrt{(1-f)\frac{pq}{n}}$.

population is the 420,000 households to which CFLs were distributed, one per household. A simple random sample was to be taken.

258. The sample size calculation used a value of 3.5 hours for the expected population mean CFL usage and a value of 2.5 hours for the expected population standard deviation (SD). This gave a required sample size of 138 households.

Example 1

259. A simple random sample of 140 households was taken, and the average daily usage (in hours) of each CFL was recorded. The summaries of the data and of the reliability calculations are presented below.

Summary statistics	Sample data
Population size	420,000
Sample size (n)	140
Mean	3.7230
Standard deviation	3.7838
Standard error of the mean	0.3197
Absolute precision	± 0.5294
Relative precision	± 0.1422 i.e. 14.22%

260. Here the required reliability is not met by the sample data and so we conclude that the mean of 3.72 hours is not sufficiently reliable. It also means that the confidence interval associated with the parameter is wider than required.

261. The 90% confidence interval is 3.19 to 4.25 hours, which is telling us that the population mean CFL usage is likely (90% likely) to be somewhere in the range of 3 hours 11minutes to 4 hours 15 minutes. Since the wide confidence interval is a reflection of lack of precision, it is therefore not advisable to use either the upper limit or the lower limit of the confidence interval as an estimate of the population mean hours of CFL usage. Nor any other “corrective approaches” unless they are supported by documentation detailing their validity.

Possible steps to take when addressing the problem of lack of reliability in sample data include the following, although the first one (scrutinising data) should always be carried out before doing any calculations, in order to ensure that the data are of the highest quality and that the planned analyses are appropriate.

262. The list below offers some order to the approaches. First of all – if it has not been done already – comes scrutinising the data; then there are some possible statistical analysis approaches which could be used on the existing data. Failing that an additional sample could be taken. There is no obvious ordering to the analysis approaches.

1. Scrutinise the raw data;
2. Possible analysis approaches;
    (a) Scrutinise the summary statistics;
    (b) Post-stratification;
3. Take an additional sample.

263. If none of these are successful then the reliability has not been met.[21]

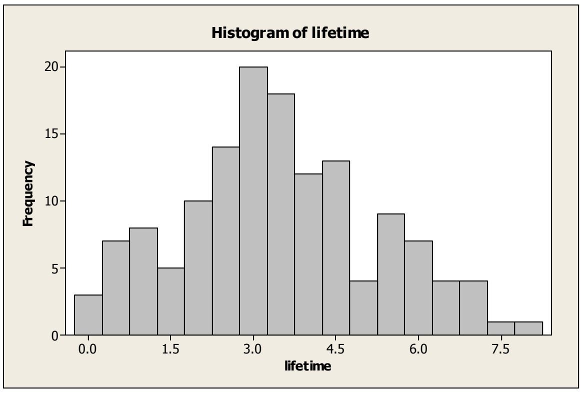
Analysis approach 1: Scrutinise the raw data

264. It is vitally important to scrutinise the raw data carefully prior to estimating the mean and checking its reliability, and this can be done using graphical summaries such as histograms, boxplots, normal probability plots. These plots would show up outliers in the data or any skewness in the distribution of the data.

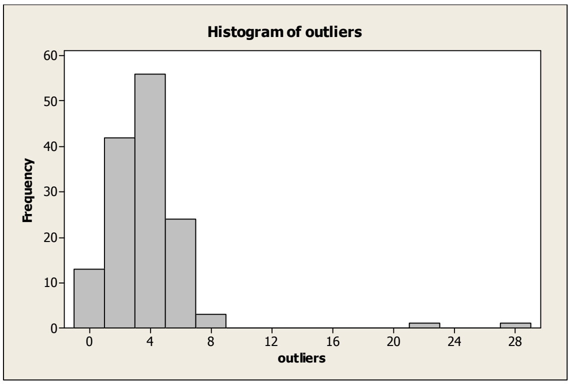
265. An **outlier** can be the result of a mistake (wrongly recorded, or wrongly entered onto the computer in which case it can be corrected); or it could be real value - in which case it must be left as it is and included in the analysis. If data are highly **skewed**, then it may be that they should be transformed prior to the analysis. The reliability would then be determined using the analysis of the transformed data. Example transformations include the logarithm, or the square root.

266. Below shows examples of a histogram for: (i) normal data (ii) data with one outlier and (iii) skewed data.

Normal data

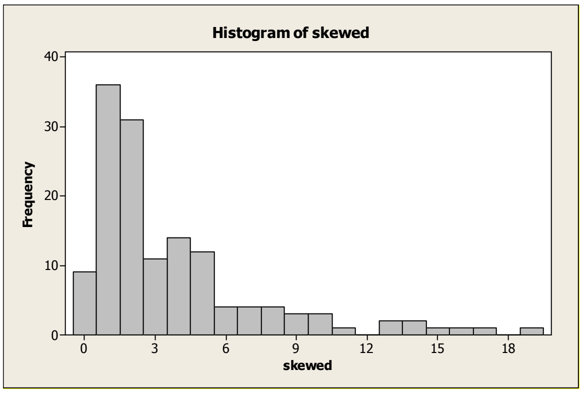


Normal data with two outliers



[21] A generalized method of discounting to account for deficiencies in reliability is not included in this document, however where the project proponents can demonstrate that discounting of emission reduction estimates or taking the lower bound or upper bound of estimates of the parameter are the only recourse available to the project proponents, procedures for request for deviation shall be followed.

Skewed data



Analysis approach 2(a): Scrutinise the summary statistics

267. The two key elements to the sample size calculation (other than the reliability and confidence) are the expected population mean value and the expected population standard deviation. If the sample data fails to meet the required reliability then this could be due to the sample estimates being quite different from the expected values. For instance:

(i) If the sample mean is lower than the expected population mean, but the standard deviation is the same as expected for the population then the reliability will be greater than 10%;

(ii) If the sample mean is the same as the expected population mean but the standard deviation is larger than expected then the reliability will be greater than 10%.

268. This is demonstrated in the table below, where the sample size calculation used is the same as described in the Scenario on page 1 (mean CFL usage of 3.5 hours, SD of 2.5 hours giving a required sample size of 138 households).

	Population size (N)	**Sample size (n)**	**Sample mean**	**Sample SD**	**Absolute precision**	**Relative precision**
(i)	420000	140	3.0 hours	2.5	0.35	11.68%
(ii)	420000	140	3.5 hours	3.0	0.42	11.99%

269. In (i) the absolute precision is as planned but, because the sample mean is lower than expected, the relative precision is larger than the required 10%. In (ii) because the SD is larger than expected both the absolute and relative precision are larger than required by the sample size specification.

270. It seems not unreasonable to try to accept data that come from scenarios of type (i). The following rule is therefore proposed.

- Provided the sample standard deviation is not more than 10% greater than the standard deviation that was used in the sample size calculation, and the sample size is at least 100 units, the sample data could be accepted.

271. In Example 1, the sample size is over 100 but the standard deviation is considerably larger than was used in the sample size calculation (3.78 as opposed to 2.5; the sample standard deviation is more than 50% more than the one used in the sample size calculation). The sample data cannot be accepted.

272. Rationale behind the conclusions in the above paragraphs is explained below:

The distribution of the sample variance s^2 is as follows:

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{(n-1)}$$

273. From probability tables for the Chi-squared distribution we can see that, for samples of size 100, one would expect 95% of sample variances (s^2) to be less than 1.24 × the population variance (i.e. less than 1.24 σ^2). One could then argue that, provided it was not more than 24% more than the expected population variance, the sample variance can be regarded as an acceptable estimate of the population variance.

274. In other words, provided that the sample standard deviation was not more than 12% more than the population standard deviation then the sample standard deviation can be regarded as an acceptable estimate of the population standard deviation.

275. The table below extends this argument to other sample sizes. As the sample size becomes smaller the multiplier becomes larger, which is to be expected. The smaller the sample size the more variability in the estimates of the population variance and consequently the wider the distribution.

Sample size and Ratio of sample SD to population SD	
Sample size	**95% of ratios less than this value**
30	1.21
40	1.18
50	1.16
60	1.15
70	1.14
80	1.13

Sample size and Ratio of sample SD to population SD	
Sample size	**95% of ratios less than this value**
90	1.12
100	1.12
125	1.10
150	1.09
175	1.09
200	1.08
250	1.07
300	1.07

276. Balancing the requirement for a reasonable amount of information with a realisation that the sample standard deviation is unlikely to ever be exactly the same as the population one, then it seems not unreasonable to have a rule of thumb which states that:

(a) Provided the sample is large enough; and
(b) Provided the sample standard deviation is only a fraction larger than the population one;

then the sample data could be accepted. It is therefore proposed that a minimum sample size be 100 units and maximum “overage” be 10%.

Analysis approach 2(b): Post-stratification

277. If there appears to be some characteristic of the population that is responsible for the apparent increase in the variability of the data – e.g. households in the urban areas are all using their CFLs for longer than the households in rural areas - then this characteristic can be regarded as a stratification variable, and the mean CFL usage recalculated using post-stratification techniques. In this example geographical area (rural or urban) is the stratification variable. Other examples might include the different times of the year when the data are collected, if usage varies according to time of year.

Taking the stratification variable into account should improve the precision of the estimate of the mean usage for the whole population, and hence its reliability. It might also give a more accurate estimate of the population mean usage. An example of post-stratification is explained below.

278. This uses the same example as Example 1 where the following data were collected.

Summary statistics	
Population size	420,000
Sample size (n)	140
Mean	3.7230
Standard deviation	3.7838

279. The total population of 420,000 in this part of the country are distributed across the rural and urban areas in a ratio of 60% to 40% respectively; and it is thought that CFL usage is lower in the rural areas than in the urban. Of the 140 households in the sample, it transpires that 104 of them were from rural areas and 36 from urban areas. The proportions in the sample from the rural and urban areas at 74% and 26% respectively do not match those of the population, and so the estimate of average usage over all 140 CFLs may be slightly lower than it should be for the whole population.

280. Summary statistics for the two sub-groups are as follows:

Summary statistics	Rural	Urban
Population size	252,000	168,000
Sample size (n)	104	36
Mean	1.96	8.816
Standard deviation	0.55	3.875

281. We can use post-stratification to estimate the mean average CFL usage to reflect the rural:urban proportions in the population by using a weighted average of the stratum means as follows:

282. Post-stratification mean $\overline{y}_{post\_st} = \frac{1}{N}\sum_{h=1}^{L} N_h \overline{y}_h$ where N is the total population size; N_h is the population size in each stratum; y-h bar is the mean for each stratum; L is the number of strata.

i.e. $\frac{(252000 \times 1.96 + 168000 \times 8.816)}{420000} = 4.7024$

283. The standard error of this estimate which can then be used to determine the precision is:

$se = \sqrt{\sum_{h=1}^{L}\left(\frac{N_h}{N}\right)^2\left(\frac{N_h - n_h}{N_h}\right)\frac{s_h^{\ 2}}{n_h}}$ where N, N_h are as above; and s_h is the stratum standard deviation.

i.e. $\sqrt{\left(\frac{252000}{420000}\right)^2 \times \left(\frac{252000-104}{252000}\right) \times \frac{0.55^2}{104} + \left(\frac{168000}{420000}\right)^2 \times \left(\frac{168000-36}{168000}\right) \times \frac{3.875^2}{36}}$

and equals 0.2616.

284. Consequently the absolute precision (assuming a 90% confidence interval) is 1.645×0.2616 = 0.4304 and so the relative precision is 0.4304 divided by the mean of 4.7024 i.e. 9.15%. The reliability has been met.

Analysis approach 3: Take an additional sample

285. Another option to improve the precision of the study data is to take an additional sample. The formula for the size of this additional sample would be the same as for the actual study. However, the value(s) that should be used for the standard deviation (SD) and the mean should be different. It is also advisable to try different combinations of the values (sample and population ones) in order to identify a total sample size which will be large enough to address the reliability concern. The additional sample size will then be the difference between this figure and the originally planned sample size.

286. For example in Example 1 the mean and standard deviation used in the original sample size calculation were 3.5 and 2.5 hours respectively. The sample of 140 CFLs gives a mean of 3.72 hours and a standard deviation of 3.78. These figures could therefore be more realistic estimates of either the mean daily hours of CFL usage or the standard deviation, and so can be used in new sample size calculations. The table below presents a few different combinations of mean and standard deviation and their resulting sample sizes. The first line in the table (in italics) is the original calculation.

Sample size for 90:10 reliability		
(90% confidence and 10% precision)		
sd	**mean**	**sample size**
2.5	*3.5*	*138*
2.5	3.72	123
3.78	3.5	316
3.78	3.72	280

287. If the PP believes that the population mean really is about 3.5 hours, but the standard deviation was underestimated originally and is likely to be closer to 3.78 then the required total sample size for the study should have been 316. An additional sample of 316-140 = 176 is therefore required. However, if they think that the mean was also too low originally and it is likely to be about 3.72 hours then the additional sample size would be 280-140 = 140.

288. Note that the above illustration only uses two figures for the mean and two for the standard deviation. In practice a range of different possibly relevant means and standard deviations should be used. This is no different to the practice recommended when the sample size calculation is being performed in the first place!

Appendix C

Best practice examples – acceptance sampling

1. Introductory notes on acceptance sampling

289. The aim of this section is to demonstrate the process of selecting a validation/verification sample and using it to decide whether or not the PP's data are valid. The methodology is based on a procedure commonly known as Acceptance Sampling. It is illustrated using a survey example where the aim is to estimate a numeric parameter.

2. Example: CFL project – Numeric parameter

290. The parameter of interest is the mean average daily usage of a CFL (in hours) for the whole population of CFLs that were distributed in a particular region of a country. The population is the 420,000 households to which CFLs were distributed, one per household. A simple random sample of 140 households was taken by the PP, and the average daily usage (in hours) of each CFL was recorded. From these data the PP determines a mean average daily usage.

291. To validate/verify the PP's sample the DOE needs to take, and observe, a simple random sample of households from the PPs sample. The decision about whether or not the PPs data are valid will depend on the number of discrepancies there are between the DOE's data and the PP's data.[22] The DOE therefore needs to set up, using their own professional judgement, criteria for deciding what constitutes a discrepancy. The DOE also needs to decide – again using their own professional judgement - the following:

(a) The proportion of discrepancies between the PP's data and DOE's data that can be considered acceptable in their sample. This is referred to as the AQL (Acceptable Quality Level);

(b) The proportion of discrepancies between the PP's data and DOE's data that would be considered unacceptable in their sample. This is the UQL (Unacceptable Quality Level).

292. In this example we will assume that the AQL is 1% and the UQL is 10%, though these could be different for different types of study.

293. The process of determining the size of the DOE's sample also requires what are referred to as the producer's risk and the consumer's risk,[23] both of which are set at 5% according to the current sampling standard.[24]

[22] If the data for a household in the DOE's sample is different from the data for that same household in the PP's sample that constitutes a discrepancy.

[23] Producer's risk is the chance that the DOE will wrongly reject the PP's dataset (i.e. reject a dataset of acceptable quality). Consumer's risk is the chance that the DOE will wrongly accept the PP's dataset i.e. accept a dataset which is unacceptable as defined above in (b).

[24] The calculations require a value of acceptable quality associated with the producer's risk and a value of unacceptable quality associated with the consumer's risk. The values between the AQL and UQL need to be seen in this context.

294. All this information then determines:

(i) The size of the DOEs sample; and

(ii) The acceptance number.

295. The acceptance number is the number of acceptable discrepancies. The DOE needs to observe no more than this number of discrepant records for the PP's data to be valid.

296. The calculation can be carried out using reliable statistical software, published tables of acceptance sampling standards, or hand calculations. Table 1 below may not be complete for every study that requires validation/verification, but it provides the required sample size and acceptance number for several different scenarios.

Table 1: Sample size and acceptance number[25]

AQL	UQL	Sample Size	Acceptance number
1%	10%	61	2
1%	15%	30	1
1%	20%	22	1
0.5%	10%	46	1
0.5%	15%	30	1
0.5%	20%	22	1

297. In this case the size of the DOE's sample is 61 households, and the acceptance number is 2. Hence, if it transpires that there are more than 2 records in the DOE's sample that do not agree with the PP's then the PP's data are not accepted. If there are none, 1 or 2 discrepant records then the PPs data are valid.

Worked example

298. The calculation can be seen in terms of the operating curve – i.e. the probability of accepting the PPs data for different proportions of discrepancies between their data and the DOEs – where we want:

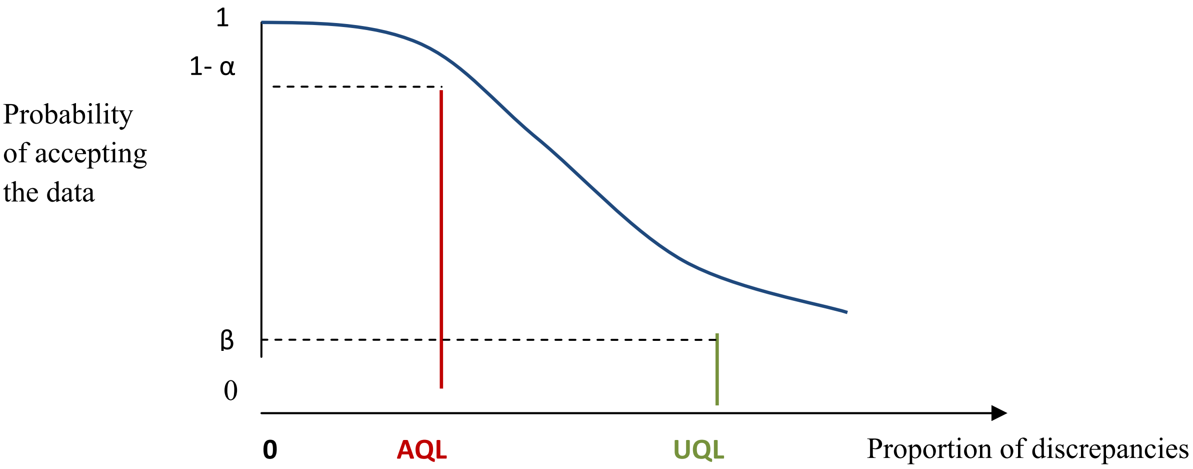
(i) A high chance of accepting the PPs data when it is of acceptable quality, i.e. Probability (Proportion of discrepancies with DOEs data is less than the AQL) $\geq 1-\alpha$;

(ii) A low chance of accepting the PPs data when it is of unacceptable quality, i.e. Probability (Proportion of discrepancies with the DOEs is more than the UQL) $\geq \beta$.

299. Note that the *producer's risk* is the opposite of (i); i.e. it is the chance of not accepting the PPs data when it is of acceptable quality, and is equal to α. The *consumer's risk* is (ii), i.e. β.

[25] The table is based on both the Producer's risk and the Consumer's risk being 5%.

300. In this example the AQL is 1% (or 0.01 in proportion terms) with a producer's risk α = 0.05; and the UQL is 10% (or 0.1 in proportion terms) with a consumer's risk β = 0.05.



301. We need to find a sample size and acceptance number that satisfy (or nearly satisfy) these probability statements. Owing to the discreteness of the data it may not be possible to satisfy them exactly.

302. The approach is based on the Chi-square distribution with 2(c+1) degrees of freedom. It determines the acceptance number first of all and then the sample size.

Step 1: Acceptance number

303. Let

$$r(c) = \frac{\chi^2_{1-\beta}}{\chi^2_{\alpha}} \qquad \textbf{(89)}$$

where χ^2_{α} is the100α percentile and $\chi^2_{1-\beta}$ the 100(1- β) percentile of the χ^2 distribution with 2(c+1) degrees of freedom.

304. Then *c* is the smallest value satisfying:

$$r(c-1) > \frac{UQL}{AQL} > r(c) \qquad \textbf{(90)}$$

305. Here the ratio of *UQL/AQL* is 10 and so we need to find a value of *c* that satisfies the above.

306. Since acceptance numbers are going to be small we can construct a table of Chi-square values and the corresponding ratios for different values of c from c = 0, 1, 2, 3, etc. as in Table 2 below.

Table 2

Tabulated values of a Chi-square distribution with 2(c+1) degrees of freedom where α=0.05 and β=0.05			
c	**α**	**1-β**	**ratio**
0	0.1026	5.9915	58.40
1	0.7107	9.4877	13.35
2	1.6354	12.5916	7.70
3	2.7326	15.5073	5.67
4	3.9403	18.3070	4.65
5	5.2260	21.0261	4.02

307. Our *UQL/AQL* ratio of 10 falls between 7.70 and 13.35. So *c* = 2 is the smallest value which satisfies the above.

Step 2: Sample size

308. The required sample size, *n*, is such that:

$$\frac{\chi^2_{1-\beta}}{2\times UQL} \leq n \leq \frac{\chi^2_{\alpha}}{2\times AQL} \qquad \textbf{(91)}$$

where χ^2_α and $\chi^2_{1-\beta}$ are defined as before, but now c = 2, and the χ^2 distribution has 2(c+1) = 6 degrees of freedom.

With c = 2 this is:

$$\frac{12.59159}{2\times 0.1} \leq n \leq \frac{1.635383}{2\times 0.01} \qquad \textbf{(92)}$$

i.e. 62.96 ≤ n ≤ 81.77.

309. So we have a sample size of 63 and an acceptance number of 2.

Step 3: Refining the calculation

310. The above steps used a Chi-square approximation, but the data actually have a Binomial distribution. The calculations can now be refined to see if the value of n could be modified. Table 3 shows the exact values of α and β, for an acceptance number of 2, and different sample sizes around 63. Whilst the above calculation showed that we needed a sample size of 63, the table shows that sample sizes of 62 and 61 would both also have α and β below 0.05.
The required sample size is therefore 61 with an acceptance number of 2.

Table 3

Acceptance number (c)	Sample size (n)	Exact probabilities based on Binomial distribution	
		alpha	beta
2	60	0.022	0.053
2	61	0.023	0.049
2	62	0.024	0.045
2	63	0.025	0.042
2	64	0.027	0.039

Excel functions

311. The Excel function used in Table 2 is CHIINV(PROB, DF). It returns the value X, for a Chi-square distribution with DF degrees of freedom, where the probability of being greater than X is PROB.

312. The Excel function used in Table 3 is BINOMDIST(acceptance number, sample size, PROB, TRUE) where PROB is either the AQL or the UQL.

- - - - -

History of the document

Version	Date	Nature of revision(s)
02.0	13 September 2012	EB 69, Annex 5 Revisions are: • To include examples to illustrate acceptance sampling as one means of verifying the sampling records of the project proponent; • To include examples for possible methods to deal with failure to achieve reliability including scrutiny of the data, post stratification, adding more sample units; • To delete examples of how to deal with very small or very large proportions in the context of simple random sampling due to the new proposed requirement on the minimum sample size; • To change the title of this draft from “Best practice examples focusing on sample size and reliability calculations and sampling for validation/verification” to “Guidelines for sampling and surveys for CDM project activities and programme of activities”.
01.1	16 May 2012	Editorially revised to remove “draft” from paragraph 3 and amended title for readability.
01.0	11 May 2012	EB 67, Annex 6. Initial adoption.
Decision Class: Regulatory **Document Type**: Guideline **Business Function**: Methodology		