

DRAFT - DO NOT QUOTE OR CITE - DRAFT

## DRAFT “NON-BINDING BEST PRACTICES EXAMPLES TO ILLUSTRATE THE APPLICATION OF SAMPLING GUIDELINES”

### A. Purpose of the document

1. The purpose of the non-binding best practice examples for sampling and survey is to demonstrate practical implementation of sampling and to provide illustration of the applications of the “General Guidance for sampling and surveys for small-scale CDM project activities”.<sup>1</sup>

### B. Determining the sample size

2. The calculation of the sample size for a sample survey for determining the value of a specific parameter of a subject population is dependent on the following:

- (a) Type of the parameter of interest, for example:
  - (i) A percentage, such as the proportion of annually operating cook stoves;
  - (ii) An absolute value (numerical data), 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 minimum quality);
  - (iii) There are other parameters, e.g. ratios, but this guide cover only proportions and means;
- (b) Value that the parameter is expected to take, for example:
  - (i) 80% of the installed cook stoves are still in operation;
  - (ii) The mean value of operating hours per day of the CFLs are 3.5;
- (c) Amount of variation, and the distribution of that variation that can be expected in that parameter within the subject population, as well as the level of precision (e.g.  $\pm 10\%$  of relative or absolute value of the parameter’s true value) and confidence (e.g. 90% or 95%) in that precision that is desired for determining the parameter.

### C. Preparing the sample design

3. The design of a survey requires consideration of the sampling units and their variability as well as the survey objectives (e.g. to estimate the proportion of annually operating cook stoves, the mean value of the operating hours of CFLs etc). This means that a clear description of the proposed sampling design can therefore be quite detailed. As shown in paragraph 33 of the “General Guidelines for sampling and surveys for small-scale CDM project activities”, the project proponent should present details on the sampling design.

---

<sup>1</sup> While approving the “General Guidelines for sampling and surveys for small-scale CDM project activities”, the CDM Executive Board at its fiftieth meeting requested SSC WG to work on examples to illustrate the application of sampling methods for small-scale CDM project activities.

4. The typical surveys involve determining parameters that define one or more critical characteristics of large populations – households using the project equipments (e.g. cook stoves, solar water heaters, biogas digesters, CFLs). These households will have different physical attribute – some will be larger than other, have more animals than others, be in rural or urban areas etc. There will be many aspects of the population that will make the parameter vary for individual members of the population. Thus, clearly defining the parameter of interest and the population are critical first steps for preparing a sampling design.

5. The next sections provide examples of the five common sampling methods detailed in the “General Guidelines for sampling and surveys for small-scale CDM project activities”, namely: (a) Simple Random Sampling; (b) Systematic Sampling; (c) Stratified Random Sampling; (d) Cluster Sampling; and (e) Multi-stage Sampling. The advantages and disadvantages of each sampling method are summarized in the table below.

Sample selection method	Advantage	Disadvantage
Simple Random Sampling	<ul style="list-style-type: none"> <li>This is the easiest method to understand and therefore use. It is suitable if there is little heterogeneity amongst the units being sampled</li> </ul>	<ul style="list-style-type: none"> <li>Requires knowledge of entire population before a sample can be selected;</li> <li>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;</li> <li>It is suitable only if the population being studied is relatively homogeneous with respect to the parameter being studied</li> </ul>
Systematic Sampling	<ul style="list-style-type: none"> <li>It is easy to apply. It is commonly used as it ensures there is always sufficient distance between samples</li> </ul>	<ul style="list-style-type: none"> <li>It will lead to units being spread out over a large geographic area. Such a geographic distribution can often be costly</li> </ul>
Stratified Random Sampling	<ul style="list-style-type: none"> <li>It improves the precision of the estimate (compared to simple random sampling) if there are differences between the strata</li> </ul>	<ul style="list-style-type: none"> <li>More complicated to calculate. It is not always straightforward to work out what the stratification factors should be</li> </ul>

Sample selection method	Advantage	Disadvantage
Cluster Sampling	<ul style="list-style-type: none"> <li>It is the most economical form of sampling as units are all grouped according to one criteria (often geographical). It is also 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</li> </ul>	<ul style="list-style-type: none"> <li>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]</li> </ul>
Multi-stage Sampling (e.g. two-stages cluster sampling)	<ul style="list-style-type: none"> <li>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</li> </ul>	<ul style="list-style-type: none"> <li>Analysis is harder and so is the calculation of sample size</li> </ul>

**D. Simple Random Sampling**

**Formulae that include a finite population correction factor**

**For percentage data (e.g. the proportion of annually operating project equipments such as cook stoves, biogas digesters, CFLs, etc)**

6. Suppose efficient cook stoves have been distributed to 640,000 households, and we want to determine the proportion of project cook stoves that are still in operation at the end of the third year after all the stoves have been distributed.

7. In order for this sample selection approach to be applicable one must first understand the population and be able to assume that the population is homogenous with respect to the continued use of the cook stoves.

8. In order to estimate the required sample size the following have to be pre-determined:

- (a) Value that the parameter is expected to take. In this example, assume that 50% ( $p = 0.5$ ) of the cook stoves is still operational at the end of the 3<sup>rd</sup> year of the crediting period;
- (b) Level of precision and confidence in that precision that is desired for determining the parameter. In this example, we assume 90% confidence and 10% precision.

9. Thus, we want our estimate to have a margin of error of no more than  $\pm 10\%$  in absolute terms. We want to be 90% confident that the margin of error in our estimate is not more than 0.1. To derive the required sample size, we want:

$$1.645 * SE \leq 0.1 \tag{1}$$

Where:

*SE* The standard error of the estimate

1.645 Is the *z* value for a 90% confidence level

10. Inputting the standard error and rearranging we have:<sup>2</sup>

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

Where:

*n* Sample size with finite population correction

*N* Total number of households

*p* Our expected proportion

0.1 represents the 10% precision

11. In this case, we get:

$$n \geq \frac{1.645^2 \times 640,000 \times 0.5 \times (1-0.5)}{0.1^2 \times (640,000-1) + 1.645^2 \times 0.5 \times (1-0.5)} = 67.6 \tag{3}$$

12. Therefore the required sample size is at least 68 households. This assumed a homogenous population and 50% of the cook stoves would be operating. However if we changed our prior belief of the underlying true percentage of working stoves *p*, this sample size would need recalculating.

13. Note that if we expected the response rate from the sampled households to be only 80% then we would need to scale up the number obtained above accordingly. Thus we would decide to sample  $68/0.8 = 85$  households.

14. In here we must assume that 1 household = 1 stove only.

**For numerical data (e.g. mean operating hours of project equipments)**

15. In the same case as above, suppose we want to determine the mean value of operating hours of the cook stoves. How many samples are needed to achieve the 90:10 confidence : precision?

16. Unlike proportions, here the desired reliability is expressed in relative terms to the mean.

17. For calculation, we need to know the expected mean and standard deviation. We may refer to the result of previous studies to get estimates of the mean and standard deviation. In case we do not have these yet, we need to take a preliminary sample as a pilot, in order to get estimates of the mean and standard deviation. In this example, assume that the expected mean usage hour is 3.5 hours and expected standard deviation is 2 hours.

---

<sup>2</sup> Using the fact that the SE of a proportion is always  $\sqrt{\frac{p(1-p)}{n}} \sqrt{\frac{N-n}{N-1}}$ .

18. Then, the following formula with finite population correction can be used:

$$n = \frac{1.645^2 N \left(\frac{SD}{mean}\right)^2}{(N - 1) \times 0.1^2 + 1.645^2 \left(\frac{SD}{mean}\right)^2}$$

Where:

*n* Sample size with finite population correction

*N* Total number of households

*mean* Our expected mean (3.5 as mentioned above)

*SD* Our expected standard deviation (2.0 as mentioned above)

0.1 Represents the 10% precision, expressed in relative terms to the mean

$$n = \frac{1.645^2 \times 640,000 \times (2.0/3.5)^2}{(640,000 - 1) \times 0.1^2 + 1.645^2 \times (2.0/3.5)^2} = 88.3$$

19. Therefore the required sample size is at least 89 households.

20. Note that if we expected the response rate from the sampled households to be only 80% then we would need to scale up the number obtained above accordingly. Thus we would decide to sample  $89/0.8 = 139$  households.

**Approximate formula for percentage data and numerical data**

21. The formulae used in these two examples are exact formula derived from simple random sampling theory. When population sizes are large (or infinite), then approximate formulae can be used, which ignore the actual size of the population.

22. These approximate formulae are as follows for the 90:10 confidence : precision guideline:

	Formula	Sample size for the above examples
Percentage data	$n = \frac{1.645^2 p(1-p)}{0.1^2}$	68 $\left( = \frac{1.645^2 \times 0.5 \times (1-0.5)}{0.1^2} \right)$
Numerical data	$n = \frac{1.645^2 \left(\frac{SD}{mean}\right)^2}{0.1^2}$	89 $\left( = \frac{1.645^2 \left(\frac{2.0}{3.5}\right)^2}{0.1^2} \right)$

23. In both these cases the sample size from the approximate formula is very similar to the one from the exact formula, but that is because the population is very large. When the population size is smaller there will be differences between the two formulae.

24. Since the exact formula can be easily calculated using Excel, it is recommend that the exact formula 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 formula.

25. Easy to see that the non-response rate applies here too.

**E. Systematic Sampling**

26. Suppose we want to check if project bricks produced in the brick production facility (640,000 brick production per year) meet or exceed the performance level of the baseline bricks (e.g. dry compressive strength, wet compressive strength, density) as per “AMS-III.Z.: Fuel Switch, process improvement and energy efficiency in brick manufacture --- Version 3”. For this purpose, we want to determine the mean value of dry compressive strength. How many samples are needed to achieve the 90:10 confidence : precision?

27. We take a preliminary sample of 8, say, in order to get estimates of the mean and standard deviation (unless we already have these). Our values for density may be 2.5, 5.3, 1.0, 8.0, 7.6, 5.5, 3.9, 6.3 which has a mean of 5.0125 and a standard deviation of 2.43.

28. We want our estimate to be 90% accurate to within ±10%, therefore we want:<sup>3</sup>

$$n \geq \frac{1.645^2 \left(\frac{SD}{mean}\right)^2}{0.1^2} = 63.6$$

29. Therefore, we want 64 samples. So we want to sample one brick for every  $N/n$ , in this case one brick every  $640,000/64 = 10,000$ .

30. We randomly choose the starting individual between 1 and 10,000 as our first sample and then we sample every 10,000th brick after it, for example sample the 3,944th brick, then the 13,944th, then the 23,944th etc.

**F. Stratified Random Sampling**

31. The key assumption for this example, is that unlike the random sample example, it is not assumed that the population is homogeneous and that different parts of the population will have different values for the parameter of “the mean value of operating hours of the cook stoves”.

32. The stoves were distributed in different districts that each have different economic backgrounds and we assume that because of this the mean value of operating hours of the cook stoves may be different between the districts, thus, we are now interested in sampling users of cook stoves from all the districts to ensure all areas are well represented.

33. We wish to determine the number of samples needed in total as well as within each district. The example shown assumes proportional allocation.

34. Each district has the following number of households and mean and standard deviation:

District	Number of households	Mean	Standard deviation
A	246,050	4.6	2.31
B	69,541	5.3	1.94
C	104,933	4.4	1.68
D	88,239	3.8	2.6

<sup>3</sup> We want our 90% confidence interval,  $1.645 * \text{standard error} < 0.1 * 5.0125$ , where standard error  $= \frac{SD}{\sqrt{n}}$ . Rearranging this gives the formula cited.

District	Number of households	Mean	Standard deviation
E	74,248	4.2	2.15
F	56,989	5.1	2.37

35. We need the mean and standard deviation for each district in order to calculate the estimated sample size needed. If you do not know them, or do not have any previous information that you can use, then you will need to undertake a small initial pilot study to estimate them. Here let's assume we have the information in each district as shown in the table above.

36. From here we can estimate the overall mean and standard deviation. Note that the formulae for the mean averages the values, whereas the formula for the standard deviation averages the squared values of the six different standard deviations, i.e. the variances. Both formulae are weighted according to the total number of households in each district.

$$\begin{aligned}
 OverallMean &= \sqrt{\frac{\sum_{region} region\ size \times mean}{N}} \\
 &= \sqrt{\frac{(246,050 \times 4.6 + 69,542 \times 5.3 + 104,933 \times 4.4 + 88,239 \times 3.8 + 74,248 \times 4.2 + 56,989 \times 5.1)}{640000}} \\
 &= 4.531 \\
 OverallSD &= \sqrt{\frac{\sum_{region} region\ size \times SD^2}{N}} \\
 &= \sqrt{\frac{246,050 \times 2.31^2 + 69,541 \times 1.94^2 + 104,933 \times 1.68^2 + 88,239 \times 2.6^2 + 74,248 \times 2.15^2 + 56,989 \times 2.37^2}{640000}} \\
 &= 2.212
 \end{aligned}$$

Having obtained these values, the total sample size of households we should take which will give us 90:10 confidence/precision is:

$$n = \frac{N \times 1.645^2 \left(\frac{SD}{mean}\right)^2}{N \times 0.1^2 + 1.645^2 \left(\frac{SD}{mean}\right)^2} = \frac{640,000 \times 1.645^2 \times \frac{2.212^2}{4.531^2}}{640,000 \times 0.1^2 + 1.645^2 \times \frac{2.212^2}{4.531^2}} = 64.5$$

37. How many households in each district? You multiply the proportion of households within that district found above.

District	Number in each district
A	$\frac{246,050}{640,000} \times 64.5 = 24.8$
B	$\frac{69,541}{640,000} \times 64.5 = 7.0$
C	$\frac{104,933}{640,000} \times 64.5 = 10.6$
D	$\frac{88,239}{640,000} \times 64.5 = 8.9$

District	Number in each district
E	$\frac{74,248}{640,000} \times 64.5 = 7.5$
F	$\frac{56,989}{640,000} \times 64.5 = 5.7$

38. Rounding these figures would give 25 households in A district, 8 in B district, 11 in C district, 9 in D district, 8 in E district and 6 in F district, making a total of 67.

39. Again this can easily be set up on an Excel worksheet – illustrated below for this example. Simply change the numbers for the mean etc. for each district, and the rest of the worksheet will update itself.

District	Mean	SD	District size	Sample size	Rounded up sample size
A	4.6	2.31	246,050	24.8	25
B	5.3	1.94	69,541	7.0	8
C	4.4	1.68	104,933	10.6	11
D	3.8	2.6	88,239	8.9	9
E	4.2	2.15	74,248	7.5	8
F	5.1	2.37	56,989	5.7	6
	Overall mean	Overall SD	Overall size	Overall sample size	
	4.531	2.212	640,000	64.5	

### G. Cluster Sampling

40. In this approach, the total population is divided into sub-groups (clusters), and the sub-groups are sampled, rather than the individual elements to be studied. For example, suppose a project which installs efficient cook stoves in 50 villages. If we are interested in estimating the operating hours of the cook stoves, one might take a sample of villages instead of the cook stoves, and then meter all of cook stoves in the selected villages. How many clusters do we need to sample?

41. Ideally, we would have estimates of the mean and variation between the villages, but if we didn't we might take information on 3 such villages. From this, we find the average operating hours for each village, say 3.4, 4.9 and 5.0. Calculating the mean and standard deviation gives us

$$\text{mean} = 4.43 \quad \text{standard deviation} = 0.90$$

42. The number of clusters, i.e. villages  $c$  we need is:

$$c \geq \frac{1.645^2 * M * \frac{SD^2}{mean^2}}{(M - 1) * 0.1^2 + 1.645^2 * \frac{SD^2}{mean^2}} = \frac{1.645^2 * 50 * (\frac{0.90}{4.43})^2}{(50 - 1) * 0.1^2 + 1.645^2 * (\frac{0.90}{4.43})^2} = 9.2$$

Where:

$M$  Is the total number of villages

43. So we need to sample 10 villages to satisfy the 90:10 confidence: precision criterion. Once village is selected, all households in a selected village will be sampled.

#### **H. Multi-stage Sampling**

44. Now let's move to a different example. We have been testing light bulbs at different research centres, and there are many different scientists at each centre who undertake the testing.

45. We now want to carry out two-stage sampling where we randomly select some centres and then randomly select some testers within the centres. The response of interest is the total number of light bulbs that are tested.

46. Let us start by assuming that we want to sample 10 testers at each centre. In general terms we will call this number *ppl*.

47. From a small pilot study involving five centres and three testers per centre we already know the following information.

Centre	Number of light bulbs tested			Total tested	Standard deviation (between testers within centres)
	Tester 1	Tester 2	Tester 3		
1	58	44	18	120	20.30
2	42	53	10	105	22.34
3	13	18	37	68	12.66
4	16	32	10	58	11.37
5	25	23	23	71	1.15
Average within centre standard deviation ( $SD_w$ )					13.57
Standard deviation between centres ( $SD_b$ )				26.633	

48. From the above table we also have:

- (a) The mean number of lightbulbs that were tested by a tester = 28.1 (overall mean);
- (b) The mean total number of lightbulbs that were tested at a centre = 84.4 (centre mean).

49. Then the required number of centres that are required for 90:10 confidence : precision criterion is:

$$c \geq \frac{\left(\frac{SD_b}{CentreMean}\right)^2 \times \left(\frac{M}{M-1}\right) + \left(\frac{1}{ppl}\right) \times \left(\frac{SD_w}{OverallMean}\right)^2 \left(\frac{\bar{N}-ppl}{\bar{N}-1}\right)}{\left(\frac{0.1}{1.645}\right)^2 + \frac{1}{M-1} \left(\frac{SD_b}{CentreMean}\right)^2}$$

Where:

- M* Total number of research centres (in our case 50)
- $\bar{N}$  Average number of testers per centre (prior knowledge tells us it is about 60)
- c* 30.99, therefore we should take 10 people from 31 centres
- ppl* Number of testers we have decided to sample in each one of the selected centres

50. It may be easier to have this calculation automated and then one can consider sampling a different number of testers from each centre in order to see how the number of centres changes. This is not difficult to do in an Excel spreadsheet and if the values in the first row under the header are changed, so the required minimum number of clusters will change too.

SDb	centre mean	M	Ppl	SDw	N	overall mean	c
26.63	84.4	50	5	26.63	60	28.13	47
26.63	84.4	50	8	26.63	60	28.13	35
26.63	84.4	50	10	26.63	60	28.13	31
26.63	84.4	50	15	26.63	60	28.13	26
26.63	84.4	50	20	26.63	60	28.13	23

51. In this example, we can see that if we sample just five testers in each centre, we would need to visit nearly all the centres. This would be far more work than sampling 10 people and visiting our 31 centres. By doubling the number of testers to be sampled in each centre from 10 to 20, we would need to visit 23 centres instead of 31.

52. Having a table like the one above will help decide the practicalities of allocating limited resources, while still satisfying the 90:10 confidence : precision criterion.

**I. Sample Size Adjustment when the expected proportions are above 90% or below 10%**

53. When the parameter of interest for sampling is a proportion or percentage (e.g. estimating the proportion of equipment functioning properly in that year) and if the expected proportions are above 90% or below 10%, project proponents may take the following approach to keep the bounds of confidence interval within the range of zero to one.

- (a) Take precision to be equal to a half of the expected proportion (*p*) if *p* is below 0.1 (10%), and take the precision to be 0.5\*(1-*p*) if *p* is above 0.9 (90%);
  - (i) When the *p* is below 0.1 (10%), precision = 0.5\**p*;

- (ii) When  $p$  is above 0.9 (90%), precision =  $0.5*(1-p)$ ;
- (b) For example, if  $p$  is 0.04, project participants may use 0.02 precision, and if  $p$  is 0.98, use 0.01 precision.