Notes on required sample sizes for CSAS and similar coverage surveys

VALID International Ltd.

Version 0.32 • July 2006

Sample size considerations for overall estimates

The typical sample size requirement for a *centric systematic area sample* (CSAS) survey of CTC coverage is about 80 cases. This seems small compared to those used with more familiar survey techniques such as the two-stage cluster sampled survey method, commonly used to estimate the prevalence of acute undernutrition (the "30-by-30" survey), and the *Expanded Program of Immunization* (EPI) coverage survey method, commonly used to estimate the coverage of immunisation programs. These survey methods are usually applied to large populations and sample sizes are inflated by applying a *design effect* in order to account for the loss of sampling independence that is introduced by the cluster sample design. In contrast, CSAS (and similar) surveys are usually applied to small populations and sample sizes do not need to be inflated to account for the sample design.

The sample size calculation formulae that most people are familiar with use what is called an *infinite population assumption* or *large population assumption*. The term *infinite* means, in this context, about 50,000 or more. The problem with these formulae is that if the population is small then the required sample size returned by standard formulae will be larger than is needed. In some cases it may even be larger than the population under survey. It is important, therefore, to define the size of the population being surveyed and to correct sample size calculations accordingly.

Throughout these notes we will use an example population of:

Overall population :	94,500
Proportion of overall population aged between 6 and 59 months :	17.3%
Prevalence of global acute undernutrition :	9.6%
Prevalence of moderate acute undernutrition :	9.1%
Prevalence of severe acute undernutrition :	0.5%

If we want to estimate the prevalence of acute malnutrition in this population we must first make an initial estimate of prevalence (e.g. 10%) and an estimate of the population size. The estimated survey population size in the example population is:

$$94{,}500 \times \frac{17.3}{100} \approx 16{,}349$$

This is because the survey will only sample children between 6 and 59 months of age (who account for 17.3% of the overall population). The general rule is:

$$\text{Survey population} = \text{Overall population} \times \text{Proportion elligible for inclusion in the survey}$$

We must also specify the level of precision (i.e. the width of the 95% confidence interval) that we want from the survey. In this example we will specify a precision of ± 3%. The required sample size returned by the standard formulae is 385. If we use formulae that account for the population size by applying a *finite population correction* we find that a sample size of 376 is required. The effect of the *finite population correction* is small in this example.

If we wanted to estimate the prevalence of acute undernutrition in a small internally displaced persons (IDP) camp we still need an initial estimate of prevalence and an estimate of the population size. To get an estimate of population size we might (e.g.) conduct a *roof count* (or *door count*) as well as a small survey to find the average number of children aged between between 6 and 59 months in each household. If we estimate that there are 450 households in the IDP camp each containing, on average, 1.5 children aged between between 6 and 59 months then the estimated survey population is:

$$450 \times 1.5 = 675$$

If we initially estimate prevalence to be 10% and specify a precision of ± 3% (i.e. as in the previous example), and calculate a sample size using the *finite population correction,* the required sample is 246.

The required sample size is reduced (i.e. from 385 without correction to 246 with correction) because the *sampling proportion* has increased. In the population of 16,349 we would have sampled, weighed, and measured:

$$\frac{376}{16{,}349}\times 100 \approx 2.3\%$$

of the population but in the population of 675 we we would have sampled, weighed, and measured:

$$\frac{246}{675}\times 100 \approx 36.4\%$$

of the population.

In the previous examples we have looked at the problem of estimating *prevalence*. The method used to estimate the survey population size for surveys estimating the *coverage* of ***non-selective*** programs is exactly the same. The method is, however, slightly more complicated for ***selective*** programs because we also need to account for program entry criteria. For example, if we wanted to estimate the coverage of a selective feeding program for children with severe acute undernutrition we would need to account for the prevalence of severe acute undernutrition.

In the example population:

Overall population :	94,500
Proportion of overall population aged between 6 and 59 months :	17.3%
Prevalence of global acute undernutrition :	9.6%
Prevalence of moderate acute undernutrition :	9.1%
Prevalence of severe acute undernutrition :	0.5%

The estimated survey population is:

$$94{,}500\times\frac{17.3}{100}\times\frac{0.5}{100}\approx 82$$

This is because the survey will only sample children between 6 and 59 months of age (who account for 17.3% of the overall population) who are also cases of severe acute undernutrition (who account for 0.5% of the population between 6 and 59 months of age). The general rule:

$$\text{Survey population} = \text{Overall population} \times \text{Proportion elligible for inclusion in the survey}$$

has not changed.

If we use standard formulae to estimate a prevalence (in this case the ***coverage proportion***) of 50% with a precision of ± 10% we find that we will need a sample size of 97. We cannot sample 97 children from a population of only 82 children! If we apply the finite population correction we find that we need sample only 45 children.

All of the above examples assume a *simple random sample*. If we were to use *cluster sampling* then we would need to inflate the required sample size to account for loss of sampling independence. This effect is called the *design effect* and, for sample size calculations, is usually assumed to be 2.0. The required sample size is calculated assuming a simple random sample and is then multiplied by the design effect. In EPI coverage surveys, for example, the sample size is calculated assuming simple random sampling from an infinite population to estimate a coverage proportion of 50% with a precision of ± 10%. Standard formulae return a required sample size of 97. This is multiplied by the expected design effect of 2.0 to give a required sample size of 194. This is usually collected as thirty clusters of seven children.

Sample size calculations for CSAS surveys

CSAS samples can safely be treated as *simple random samples* and required sample sizes can be calculated assuming a *simple random sample*. This means that no *design effect* needs to be specified.

CSAS surveys are typically used to estimate the coverage of ***selective*** feeding programs with small eligible populations. This means that required sample sizes should always be calculated using methods that apply a *finite population correction*.

The steps required to calculate the required sample size for a CSAS sample are:

1. **Collect demographic data** including:

   The overall population for the survey area : You can usually get this from census reports or reports produced by UN organisation such as UNICEF. Make sure that you correct population figures for factors such as population growth, migration, displacement, and mortality.

   The proportion of the overall population eligible for a non-selective program : For CTC programs this will probably be the proportion of children aged between 6 and 59 months. It you cannot get this data then it is usually safe to assume that 20% of the population are aged between 6 and 59 months.

2. **Make an initial estimate of the prevalence of the condition being treated :** For CTC programs, this will usually be the prevalence of severe acute undernutrition. The best source of data for this is a recent nutritional anthropometry survey for the program area. If recent data is unavailable then you could use more recent data from neighbouring districts or correct older data using (e.g.) food security data such as *agricultural calendars*.

3. **Estimate the size of the survey population** (this is is the same as estimating the number of eligible individuals in the survey population) using the data collected for steps (1) and (2) above:

	Description	***Calculation***	***Example***
a	Overall population for the survey area	NA	94,500
b	Proportion eligible for *non-selective* intervention	NA	17.3%
c	Prevalence of condition being treated	NA	0.5%
d	Size of the survey population	$a \times \frac{b}{100} \times \frac{c}{100}$	$94{,}500 \times \frac{17.3}{100} \times \frac{0.5}{100} \approx 82$

4. **Make an initial estimate of overall program coverage :** If you have no idea of program coverage then you should use 50% since this will return the largest required sample size for a given level of precision.

5. **Specify the precision required for the overall program coverage estimate :** Coverage surveys for child survival programs such as EPI usually specify precision as ± 10%. Better precision requires a larger sample size.

6. **Calculate the required sample size** using your estimate of the size of the survey population, your initial estimate of overall program coverage, and the required precision using formulae (or software that implement formulae) that use a *finite population correction* or using a lookup table and chart (see below). If you need to specify a design effect in sample size calculation software (e.g. **`SampleXS`**) then you should specify 1.0 (i.e. no design effect).

Very small survey populations

The example population is an extreme example and illustrates that, with small survey populations, standard formulae may yield grossly inappropriate sample sizes. In low prevalence situations you may find it difficult to meet the required sample size without sampling a very large number of communities. The sampling proportion for the required sample size in the example population is:

$$\frac{45}{82} \times 100 \approx 55\%$$

This means that you may need to sample more than half of the communities in the program area in order to meet the required sample size. You should not come across such situations in emergency contexts, where the prevalence of severe acute undernutrition is likely to be high, unless you survey towards the close of a program when prevalence has returned to acceptable levels. You may, however, come across low prevalence situations in developmental contexts.

With very small survey populations and high sampling proportions, the best you can do is to collect as much data as is feasible within time and cost constraints. Such a survey will still provide useful data on the reasons for coverage failures and return a relatively imprecise estimate of overall coverage.

Representativeness of CSAS samples

When considering the representativeness of the sample taken by a CSAS survey you should keep in mind:

- The survey is sampling from a *small population* using a survey design that does not introduce a *design effect* that will inflate the size of the required sample.
- The CSAS method returns an even spatial sample from a wide range of communities whereas survey methods that use cluster sampling tend to concentrate data collection in the most populous communities. Cluster sampled surveys tend to leave areas of low population density unsampled (i.e. those areas consisting of communities likely to be distant from health facilities, feeding centres, and distribution points). This may cause cluster sampled surveys to evaluate coverage as being adequate even when coverage is poor or non-existent in areas outside of urban centres.
- CSAS surveys tend to sample considerably more communities than competing methods. Cluster sampled surveys tend to sample slightly fewer than thirty communities. CSAS surveys usually sample more than 100 communities.
- Coverage estimates from CSAS surveys are usually based on much larger sample sizes than surveys that attempt to "nest" coverage estimation within a nutritional anthropometry ("30-by30") survey. In the example population, a "30-by-30" survey would find only four or five cases on which to base a coverage estimate.

Sample size considerations for per-quadrat estimates

The sample sizes required for per-quadrat coverage estimates may be calculated using the method outlined above. It is, however, usually not possible to do this because population and prevalence estimates are often unavailable at the quadrat level and because coverage is likely to vary considerably between quadrats. Instead, CSAS (and similar) coverage survey methods rely on taking a *census sample* (i.e. a sample that includes all, or nearly all, eligible individuals) from a reasonably large proportion of the communities in each quadrat. To ensure that this is the case, you must use a high sensitivity case-finding method (ideally this should be 100%) and a case-finding method that is rapid enough to allow you to sample a reasonably large proportion of the communities in each quadrat. If you use a case-finding method (e.g. door-to-door screening) that restricts your ability to sample a large proportion of communities in each quadrat then you should consider reducing the size of each quadrat so as to increase the sampling proportion of communities sampled from each quadrat.

Sample size table

The table below shows the required sample sizes for different sizes of survey population assuming a coverage proportion of 50% and a precision of ± 10%:

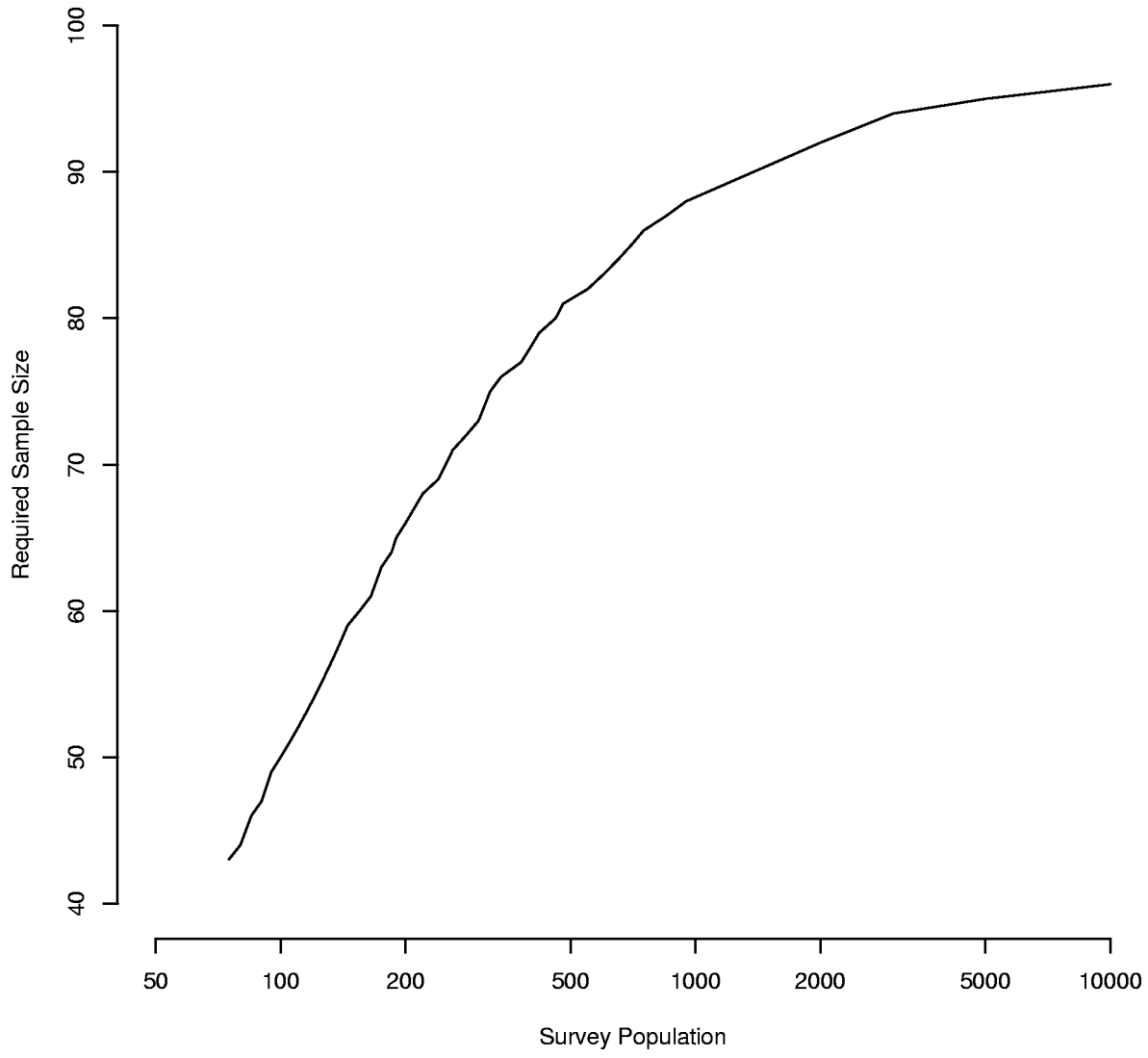
Survey Population	*Required Sample Size*
75	43
100	50
125	55
150	59
175	63
200	66
225	68
250	70
275	72
300	73
350	76
400	78
450	80

Survey Population	*Required Sample Size*
500	81
550	82
600	83
650	84
700	85
800	86
900	87
1000	88
1500	91
2000	92
3000	94
5000	95
10000	96

You can use this table to find required sample sizes.

Sample size chart

The chart below shows the required sample sizes for different sizes of survey population assuming a coverage proportion of 50% and a precision of ± 10%:



You can use this table to find required sample sizes.

The sample size table and chart present the same data. Use whichever tool you feel most comfortable with.

Sample size software

Free software is available for calculating required sample sizes using a *finite population correction*. Microsoft Windows™ users can use **SampleXS** which can be downloaded from:

> http://www.brixtonhealth.com

Linux and UNIX™ users can use **sampsize** which can be downloaded from:

> http://sampsize.sourceforge.net

There is also a web interface to **sampsize** at:

> http://sampsize.sourceforge.net/iface/index.html

which can be used on any computer (Microsoft Windows™, UNIX™, Linux, Mac, &c.) connected to the Internet.