

Sampling Methods and Sample Size Calculation for the SMART Methodology

June 2012

Table of Contents

Sampling Methods and Sample Size Calculation for the SMART Methodology	1
April 2012	1
Table of Contents	2
1. Sampling Design	4
1.1 What is Sampling?	4
1.1.1 Principles of Sampling	5
1.1.2 Definition of Concepts	5
1.2 Sampling Methods	6
1.2.1 Simple Random Sampling	6
1.2.2 Systematic Random Sampling	8
1.2.3 Cluster Sampling	11
2. Sample Size Calculation	13
2.1 Sample size expressed in terms of children vs. households	13
2.1.1 SMART Recommendation	13
2.1.2 Conversion of Sample Size	14
2.2 Accuracy, Precision and Confidence Interval	15
2.2.1 Accuracy	15
2.2.2 Precision	15
2.2.3 Confidence Interval and Level of Confidence	16
2.3 Parameters for Sample Size Calculation	16
2.3.1 Estimated Prevalence	17
2.3.2 Precision	18
2.3.3 Design Effect	19
2.3.4 Demographic Data	21
2.3.5 Non-Response Rate	22
2.4 Correction for Small Population Size	22
2.5 Combined Surveys	22
2.6 Determining the Number of Clusters	23
2.7 Assigning Clusters	24
2.7.1 Small Villages	24

2.7.2	Large Villages	25
2.8	Reserve Clusters	25
3.	Second Stage Sampling	25
3.1	Preparation for household selection	26
3.1.1	List of Households Available	26
3.1.2	List of Households Not Available	26
3.1.3	Segmentation.....	27
3.2	Household Selection Techniques	28
3.2.1	Simple random sampling	30
3.2.2	Systematic random sampling.....	30
3.2.3	Modified EPI.....	30
4.	Special Cases.....	31
4.1	Polygamous Families/Compounds	31
4.2	Impossible to Visit a Selected Household	33
4.3	No children in the Household	33
4.4	Absent Household	33
4.5	Abandoned Household.....	33
4.6	Absent Child	34
4.7	Child with Disability.....	34
4.8	Child in a Center or a Clinic	34

1. Sampling Design

1.1 What is Sampling?

When we want to collect data on a population, there are different ways to proceed. In some cases, it is possible to measure the *entire* population to get an accurate picture of the nutritional status or mortality rates for the population at hand. This is called an **exhaustive survey**. These cases are however *very rare* as they are feasible only for target populations that are geographically concentrated and not exceeding 1000 households, as seen in camps or very small areas (Figure 1). For larger populations, an exhaustive survey will be long, expensive, difficult and unnecessary. In addition, the results of such a survey cannot be extrapolated to other regions.

Figure 1: Camp Stankovic II, Skopje. Macedonia, 1999



Example: An organization would like to know the malnutrition rate in a small camp. There are between 2,000 and 3,000 people in this camp. The camp is composed of roughly 400 to 600 children less than 5 years old. In this case, *ALL* the children can be measured. The results will therefore represent the exact nutritional situation of the children in the camp.

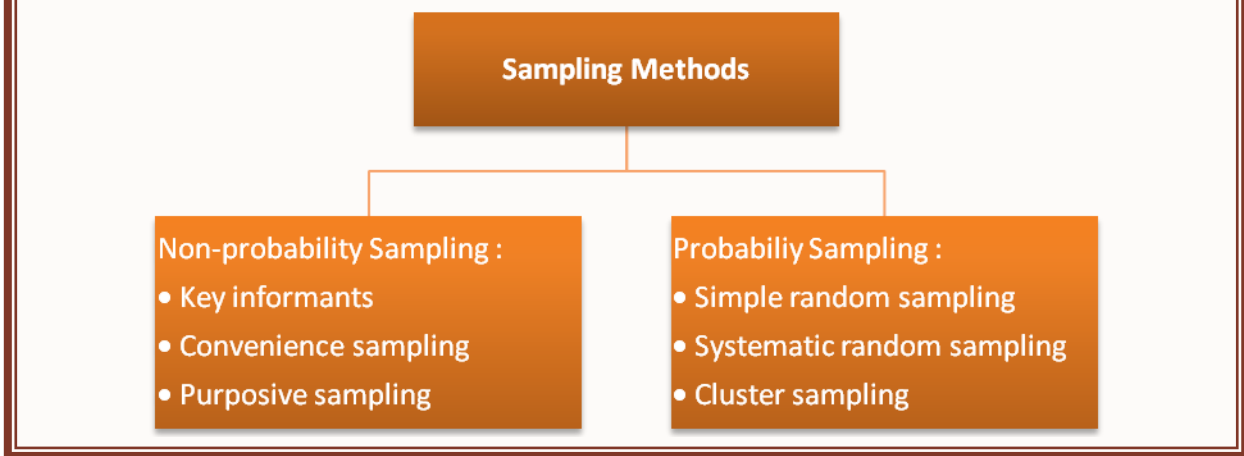
When it's impossible to collect data on the whole population because it is large and/or geographically dispersed, we would choose a sub-group of this population, which will be called a **sample**. This sub-group, or sample, can be chosen following 2 broad groups of methods: probability and non-probability¹ sampling methods.

Figure 2 shows the sampling methods that belong to each of those two groups. An example of the non-probability sampling would be for surveyors to go to a village and call all mothers to bring their 6-59 month-old children to the health center and collect data only on those children. We can easily assume that some mothers might only bring sick children, thinking that they will receive free treatment in the health center. On the other hand, we can also assume that mothers might bring only healthy children to avoid being judged by others as being "bad mothers", and we can continue with other assumptions regarding the "type" of children that will be taken to that health center for data collection on that day (sex/age of the children, access to the health center, etc.) Therefore, to avoid these problems, we generally choose a **representative sample** using a probability sampling method². This paper will only cover this type of sampling.

¹ In general, non-probability sampling is often used for collecting qualitative information.

² Different probability sampling methods will be discussed into more details in section 3.

Figure 2: General Types of Sampling



1.1.1 Principles of Sampling

A representative sample needs to have the **same characteristics** as the target population. For example, if our target population are children 6-59 months living in a pastoralist population, our sample needs to contain children 6-59 months who are living in this same population. The distribution of age and sex and other characteristics in our sample should be very similar to the distribution seen in the target population. Having a representative sample also means that:

1. Each individual or sampling unit in the population has a **known, non-zero chance or probability of being selected**.
2. The selection of one individual is **independent** from the selection of another.

Since we are only collecting data on a sub-group of the population when sampling, it is important to remember that the result obtained will only be an **estimate** of the indicator that needed to be measured. As discussed above, to get the true value, it would be necessary to conduct an exhaustive survey.

1.1.2 Definition of Concepts

It is important to understand the terminology used when talking about sampling. Below are some of those concepts with their definition.

Sampling universe: population from which we are sampling.

Sampling frame: description of the sampling universe, usually in the form of the *list of sampling units* (for example, villages, households or individuals). Sometimes, it may be outdated or otherwise not accurate, and thus would not provide an accurate description of the sampling universe (census data not recent, recent population movements, etc.)

Sampling unit: the unit selected during the process of sampling. If you are selecting districts during the first stage of cluster sampling, the sampling unit (also called primary sampling unit) at the first sampling stage is therefore the district. If you select households from a list of all households in the population, the sampling unit is in this case the household.

Basic sampling unit or elementary unit: the sampling unit selected at the last stage of sampling. In a multi-stage cluster survey (see section 1.2.3), if you first select villages and then select households within those selected villages, the basic sampling unit would be the household.

Respondent: person who's responding to your questionnaires on the field.

Survey subject: entity or person from whom you're collecting data.

1.2 Sampling Methods

Three methods of sampling will be covered in this manual:

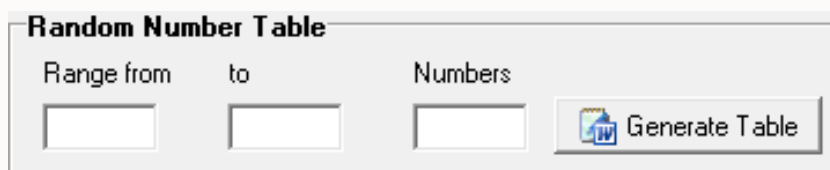
- Simple random sampling.
- Systematic random sampling.
- Cluster sampling.

The sampling method is selected based on the **spatial distribution of households** and **population size**. The basic principle for selecting households to visit is that each individual in the target population must have a known (and preferably equal) chance of being selected for the survey.

1.2.1 Simple Random Sampling

Simple random sampling can be used for small populations that contain more than 1000 sampling units (or households). This method requires a complete and updated list of sampling units, which may be difficult to obtain in emergency situations. Since in most contexts the number of basic sampling units (BSU) is large, simple random sampling procedure can be conducted by numbering each basic sampling unit and then choosing the desired number of units randomly from the random number table (Figure 3). Measures will then be taken based on these units only.

Figure 3: ENA Random Number Generator



The image shows a software interface titled "Random Number Table". It contains three input fields labeled "Range from", "to", and "Numbers". Each field has a small rectangular box next to it for entering a value. To the right of these fields is a button with a blue icon and the text "Generate Table".

Scenario:

You are requested to conduct a nutrition survey in a refugee camp that contains about 10000 people. All births are registered, enabling you to have the list of all <5 years-old children with their birth dates. This list contains 1700 children and after completing your calculations, you find that your sample needs to include 450 children. Here are the procedures to follow, using the *Random Number Table* in ENA to select your basic sampling units (which are children in that case):

- Number each of the 1700 children of the camp from 1 to 1700.
- Enter 1 in the box named *Range from* and 1700 in the box named *To* (Figure 3).
- Enter the number of children you need for your sample (450) in the box named *Numbers*.
- Click on the *Generate Table* button.
- A word document will open displaying the 450 numbers selected randomly (Figure 4).

Figure 4: List of random numbers in a Word document generated by ENA

Random Number table	
Range: 1 to 1700, Number: 450	
501	1601 795 458 327 962 1094 631 714 893 1070 1175 328
614	3750 455 1465 1068 1650 1452 1296 1468 767 635 185
1165	432 1480 58 1057 832 149 1167 915 687 1038 844
1305	785 856 422 518 1275 897 207 1148 1276 309
1419	542 1307 84 530 1124 980 316 951 212 96 899
1110	1351 1323 963 49 917 1080 1283 821 1206 1053 686
961	1351 968 54 467 889 129 707 417 1072 1511
1389	1406 1357 1641 221 1273 1231 1415 1568 463 982 652
1045	36 1393 682 469 1519 789 588 1312 1269 328 375
425	1285 1372 921 336 1085 118 1228 937 255 1464 715
1022	911 201 1073 1136 1442 209 1002 1337 557 28
1508	292 1092 884 337 1056 122 384 1477 1483 741 972
48	668 1542 854 1555 1404 1618 552 1505 382 553
1409	38 916 1016 385 50 883 909 1075 754 788 571
281	332 364 227 568 1140 1358 1431 305 490 175
1164	1066 771 975 269 220 1501 1189 150 1154 124 925
1220	1597 1054 298 746 75 1041 737 434 431 713
1346	459 1173 1646 1007 1586 355 1532 833 1350 683
1338	1553 148 76 791 981 869 513 1293 166 823
1084	541 1018 251 1226 811 1557 12 438 1293 1650 386
1420	1078 850 1554 1382 875 855 922 788 749 450
1142	814 812 1446 1069 842 1592 236 989 894 1652 789
1394	1121 532 1655 1394 1052 141 1218 1393 514 1189
1640	214 545 941 1541 393 176 165 794 1416 1602 934
1334	990 773 1144 1682 1364 771 752 1455 1014 517
1599	1280 816 1513 611 152 1161 1255 245 843 1319 493
1546	830 449 242 618 465 510 1438 632 716 1210 733
1520	388 1176 43 858 585 285 834 1318 564 280
1521	681 1326 1010 1549 330 1234 523 1382 847 1403
1388	1527 246 1264 1182 147 1397 1090 579 312 566
1506	155 1526 413 878 873 171 1700 360 1628 831 283
258	684 1665 638 1115 594 1088 359 1669 426 294
1632	678 404 1425 1678 11 1065 1380 1185 164 1179
1411	218 440 208 1373 1625 1550 137 1106 555 127
1002	252 1563 1605 549 906 931 15 498 135 759
1243	1251 593 1270 740 1135 1174 278 964 1006 502
1234	1414 552 1096 64 380 172 1130 1648 1699 1082 177
696	1298 563 1507 1677 282 706 1324 650 1432 235 63
750	1277 1049 428 971 1384 675 699 519 1498 703
1611	

NB: This same procedure can be applied if the basic sampling units are households instead of children.

Two scenarios are possible for simple random sampling depending on the demographic data available:

1. **List of target population is available:** if there is an updated list of 6-59 children for example, the number of children needed for the sample can be drawn randomly from this list.
2. **List of households is available:** if there is only a list of all households within the area to survey, the desired number of households (calculated by ENA) can be drawn randomly from this list. Once those households are selected, all children present in those households will be measured.

1.2.2 Systematic Random Sampling

Systematic random sampling is based on selection of units situated at a certain predetermined interval called the **sampling interval**. It is applicable for small scale surveys (approximately 1000 to 5000 BSU) and one of its main advantages is that it can also be used without having a list of basic sampling units, as in situations where dwellings are well organized in rows, blocks, or along a river or main road, for example.

Sampling Interval Calculation

$$\text{Sampling interval} = \frac{\text{total number of basic sampling units (BSU) in the population}}{\text{Number of sampling units needed for the sample}}$$

Example:

You need to conduct a mortality survey in a camp which contains 2400 households. According to your calculations, your sample should contain 300 households. The sampling interval you will be using is:

$$\text{Sampling interval} = \frac{2400 \text{ households in total population}}{300 \text{ households for your sample}} = 8$$

Procedures for Systematic Random Sampling:

1. After calculating your sampling interval, you are ready to select your first BSU.
2. Choose a random number³ between 1 and the sampling interval you calculated.
3. Starting at number 1 of your list or survey area, count the BSUs until you get to the one that has the number you've picked randomly. For example, if you have picked number 5, start counting from the start of the list/camp until you get to the 5th BSU. This will be the first BSU in your sample.
4. To choose the 2nd BSU, add the sampling interval to the number of the 1st BSU. In the example above, your 2nd BSU to survey will be 5+8 (if 8 is the sampling interval), which results in 13. Therefore, after coming out of BSU 5, you keep walking and counting BSUs until you get to number 13 and this will be the 2nd BSU of your sample.
5. Subsequent HH are selected in the same manner, by adding the sampling interval to the number found for the previous BSU.

Three scenarios are possible for systematic random sampling depending on the demographic data available (see scenarios 1 & 2 on page 10):

1. **List of target population available:** a sample of individuals (from target population) can be drawn directly from this list using ENA random number table function.
2. **List of households available:** a sample of households can be drawn randomly from this list. Once those households are selected, all children in those households will be measured.

³ This number can be chosen by picking a number out of bag or hat or using a random number table.

3. **No list is available but households are arranged in an easy geometric pattern:** you need to know the boundaries of your survey area and to determine what could be considered as its start point and end point and which direction you will be following when walking or driving through the area.

In some contexts, the calculated sampling interval may be a decimal number. In that case, the procedure to find the numbers of HH to visit is slightly different. The rounding will be done after adding the sampling interval to the decimal number found in the previous step (see example below for more details).

Example:

The village of Kavar is composed of 3400 households numbered sequentially from 1 to 3400. You need a sample of 250 household.

The sampling interval: $s = 3400 / 250 = 13.6$

You ask someone to randomly choose a number between 1 and 13 and they choose **11**. This number is the equivalent to the 1st household to survey and the following are chosen as follows:

Table 1: Calculation of HH numbers using a decimal sampling interval

HH Rank	Calculations	Total	HH number
1 st household	11	11	11
2 nd household	11+13.6	24.6	25
3 rd household	24.6+13.6	38.2	38
4 th household	38.2+13.6	51.8	52
5 th household	51.8+13.6	65.4	65
6 th household	65.4+13.6	79	79
7 th household	79+13.6	92.6	93
8 th household	92.6+13.6	106.2	106
9 th household	106.2+13.6	119.8	120
Etc.	Etc.	Etc.	Etc.

Scenario 1: Systematic random sampling using a list of children

You want to conduct a nutrition survey in a camp which has a list of 1500 children 6-59 months-old. Your sample size is 300 children.

- First, make sure the list is not ordered in a way that would introduce a bias in the selection (according to age, mother's level of education, etc.)
- Calculate your sampling interval: $1500 \div 300 = 5$
- Choose a random number between 1 and 5. Let's say, you choose 3.
- Count from the top of the list down to child number 3 on the list. This child will be the 1st child in your sample.
- The 2nd child in the sample will be: $3+5 = 8$; the 8th child on the list.
- The 3rd will be $8 + 5 = 13$, and so on, until you get the 300 children needed for your sample.

Scenario 2: Systematic random sampling without a list.

You want to conduct a nutrition survey in an area that has a relatively small population, which is arranged in rows (as seen Figure 5). Your basic sampling units are households. The area contains 3000 households, but you only need 250 for your sample.

- First determine the start and end of this survey area and the direction you will be following.
- Calculate your sampling interval: $3000 \div 250 = 12$.
- Choose a random number between 1 and 12; let's say you have picked the number 4.
- From the start of the survey area, count dwellings until you get to the 4th one. This will be your 1st HH in the sample (see Figure 5).
- To find the 2nd HH in your sample, add 4 to the sampling interval (12): $4+12 = 16$. Your 2nd HH of the sample will be the 16th in the village.
- Continue with this same process until you get the 250 HH needed for your

Figure 5: Example of a geometric setting of dwellings (adapted from MSF)



1.2.3 Cluster Sampling

In cases where your survey area is too large (a country, province or district for example) and/or the population is dispersed, cluster sampling can be used. This method is the **most frequently used in the field**. In cluster sampling, basic sampling units are selected within groups named clusters (villages, administrative areas, camps, etc.) The objective of this method is to choose a limited number of smaller geographic areas in which simple or systematic random sampling can be conducted. It is therefore a multi-stage sampling method. Very often, it's completed in 2 stages:

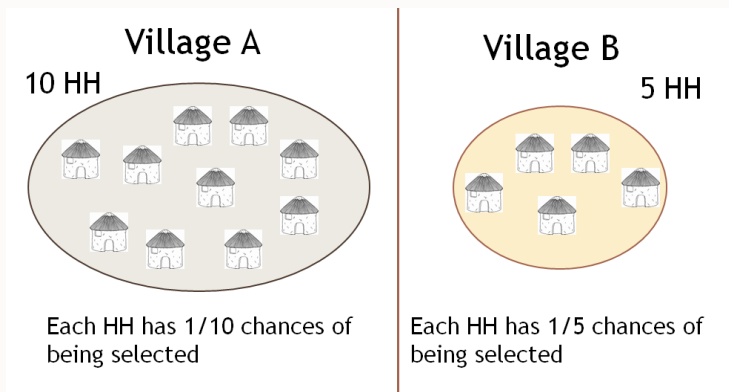
- **1st stage = random selection of clusters:** the entire population of interest is divided into small distinct geographic areas, such as villages, camps, etc. You then need to find an approximate size of the population for each “village”. At this stage, the primary sampling unit is the village. Afterwards, clusters could be assigned randomly to villages using ENA software.
- **2nd stage = random selection of households within clusters:** households are chosen randomly within each cluster using simple or systematic random sampling.

This section will discuss in further details the **1st stage of sampling**.

Large survey areas often contain geographical units or settlements of various sizes. Therefore, it is important that each individual in each of those settlements has **an equal chance** of being selected whether he/she lives in a large or small settlement. To illustrate this, imagine that you have to survey an area which contains 2 villages A and B (Figure5

). A contains 10 HH and B has 5 HH. Village B is half the size of the village A. Let's assume that we choose randomly one of the 2 villages by tossing a coin for example. In village B, each HH has double chances of being selected than in village A (1/5 compared to 1/10). Therefore, chances of HH in each village to be selected will not be equal by just tossing a coin. To account for this, the selection of clusters is done by ENA software using **probability proportional to population size (PPS)** method. In PPS, larger settlements have a higher chance of being selected as clusters compared to smaller settlements because the probability of selection is proportional to population size of the settlement.

Figure 6: Example of villages of different sizes



How does ENA assign clusters randomly using PPS?

- As shown in Table 2 below, ENA calculates cumulative population sizes; i.e. for each geographic unit, the cumulative population size will be the size of the population for that unit plus the sum of all the units which come before it on the list.
- A range of households' numbers is also assigned for each geographical unit according to its cumulated population, as shown in the 4th column of Table 2⁴.

Table 2: Example of cluster selection by PPS.

VILLAGES	Estimated Total Population	Cumulative Population	N° Allocated	CLUSTERS
N° 1	500	500	1 - 500	1; 2; 3
N° 2	400	900	501 - 900	4; 5; 6;
N° 3	160	1060	901 - 1060	7
N° 4	650	1710	1061 - 1710	8; 9; 10; 11; 12
N° 5	520	2230	1711 - 2230	13; 14; 15; 16
N°6	640	2870	2231 - 2870	17; 18; 19; 20
N°7	700	3570	2871 - 3570	21; 22; 23; 24; 25
N°8	100	3670	3571 - 3670	26
N°9	470	4140	3671 - 4140	27; 28; 29
N°10	60	4200	4141 - 4200	30

- ENA then calculates the sampling interval, which is the total population of all the geographic units divided by the number of clusters needed. If the total population is 4200 and you need 30 clusters; then, the sampling interval will be $4200/30 = 140$.
- Sampling will begin at a randomly selected starting point; therefore, we choose a random number as our starting point between 1 and the sampling interval (140). The geographic unit where this number lies will be the cluster number 1. Let's assume that the random start number was 100. This falls into village 1.
- The second cluster will be: $100 + 140 = 240$, which also falls into the range defined for village 1 (1 to 500) and this will be cluster 2.
- For cluster 3, it will be: $240 + 140 = 380$, which is again in village 1. However, if we add again the sampling interval to 380, you will get 520, which falls into village 2. ENA continues this way automatically until it assigns the 30 clusters (Figure 7).

⁴ In this example, we ended up having a lot of clusters in the same villages because the number of villages was limited. In real situations, you might have villages that will have no clusters assigned, others that will have only 1 or 2, and yet others very large that might have 3 or more.

Figure 7: ENA table for cluster selection

Table for Cluster sampling

Population Size:

Geographical unit	Population size	Cluster
BUTAGANZWA	52746	1,2,3,4
GAHOMBO	41851	5,6,7
GATARA	66963	8,9,10,11,1...
KABARORE	35187	14,15
KAYANZA	86154	16,17,18,1...
MATONGO	34551	23,24,25
MUHANGA	74124	26,27,28,2...
MURUTA	27955	31,32
RANGO	72440	33,34,35,3...

*RC: Reserve Cluster

Once those villages to contain clusters are chosen, second-stage sampling will be conducted in each village to randomly select households to be a part of the sample. This will be discussed in more detail in section 3 of this paper.

2. Sample Size Calculation

2.1 Sample size expressed in terms of children vs. households

2.1.1 SMART Recommendation

There are two ways to express the final sample size required, either with a fixed number of households or with a quota that is, a fixed number of children.

The quota sampling method means that teams keep selecting households in the cluster until a specific number of children are reached, regardless of the number of households visited. For example, if each cluster should contain 20 children, teams will be selecting as many households as necessary in the village to reach that number of children by the end of the day, and sometimes, they have to visit a neighbouring village to complete their cluster.

In contrast, when using the fixed household method, only a certain number of households, selected randomly on the field, will be visited. At the end of the survey, some clusters will have more children than others and the total should not differ significantly from what was planned. Therefore, survey teams should be given **only the target number of households** per cluster (and not the number of children per cluster) in order to avoid confusion and unnecessary stress and errors.

SMART recommends using the fixed number of households method for 3 main reasons:

- Household selection technique: SMART recommends using simple or systematic random sampling methods to choose households **within the clusters** (in the 2nd stage

sampling) since they are better than modified EPI in terms of representativeness of the sample and introduce less bias. These 2 methods are based on the selection of households either from a list (simple) or with a sampling interval (systematic). Therefore, when using one of these 2 sampling techniques, it is more logical to have a fixed number of households as a target to reach for each cluster. In other words, since it is only possible to estimate the approximate number of eligible children per household prior to data collection (which might not reflect the actual number found in selected houses), it will be impossible to know in advance the number of HH to select that will contain the exact number of children⁵.

- b. Inclusion of HH with no children: When survey teams have a target number of children to reach, they may have a tendency to skip households that don't have children. Other indicators collected during the same survey (such as mortality, water and sanitation, food security, etc.) are measured at the household level. Measuring these indicators only in households with small children and excluding all other households will create a serious bias.
- c. Sample size calculation may be done not only for anthropometry, but also for mortality or other indicators measured on the household level. To compare and reconcile these sample size requirements both sample sizes (for example, anthropometry and mortality sample sizes) should be expressed in the same units (i.e., households).

2.1.2 Conversion of Sample Size

On the planning screen, ENA will automatically convert the number of children found for the sample size into a number of household, following the formula below⁶:

Equation 1: Conversion of sample size from number of children into number of households

$$N_{HH} = \frac{N_{Children}}{(HH\ size \times \% \ of\ Under\ 5 \times 0.9)}$$

Where: N_{HH} = sample size in terms of households
 $N_{Children}$ = sample size in terms of children
 $HH\ size$ = average household size
 $\% \ of\ under\ 5$ = proportion of under 5 years-old children in the population
 0.9 = fraction of 6-59 months children within the under 5 age category

Example:

If $N_{Children} = 500$, average HH size = 6.2, and % of under 5 = 15% (that is, 0.15 when expressed as a proportion),

Then the sample size in HH is $N_{HH} = 500 / (6.2 \times 0.15 \times 0.9) = 598\ HH$

⁵Quota sampling was more adapted to the EPI method for household selection, since it was not necessary to determine the number of households to visit in advance as it's the case for simple or systematic methods. However, SMART does not recommend the use of EPI as a preferred method.

⁶Note that once N_{HH} is found, ENA adjusts it using the non-response rate (see 2.3.5 for more details).

2.2 Accuracy, Precision and Confidence Interval

When choosing a sample within a total population, we agree that we will get only an *estimate* of the indicator of interest and not the true population value. Therefore, as portrayed in Figure 8, we need to consider:

1. How close this estimate is likely to be to the true value (level of **accuracy**);
2. How similar will be the results if several surveys are conducted with the same methodology (level of **precision**).

Figure 8: Precision and accuracy



2.2.1 Accuracy

Accuracy measures the **validity** of the estimate, and a lack of it is related to **bias**, which reflects the difference between the sample estimate and the true population value due to error in measurement, selection of a non-representative sample or factors other than sample size. ***We cannot control bias once it is present.*** However, we can try to reduce its occurrence by standardizing survey procedures, providing adequate training to every member of the survey team, and closely supervising all survey work.

2.2.2 Precision

Precision measures the **consistency** of the results and is related to **sampling error**, which is the difference between the sample estimate and the population value due to measuring a sample instead of the whole population: ***we can control sampling error by controlling sample size.*** In essence, sampling error is the measure of uncertainty stemming from the fact that our estimate is derived from measuring only a relatively small proportion of the target population, and therefore it will never be exactly the same as the true population value, even if the sample is selected randomly and representatively. The larger the proportion of the target population that is measured, the lower this uncertainty becomes. Therefore, the **higher the sample size, the higher the precision.**

A larger sample size increases the precision of the results but does not guarantee the absence of bias. When the sample size is very large, the quality control becomes difficult because of the high number of teams to train and supervise, and there may be a higher risk of bias.

REMEMBER: **preventing bias is critical; never try to achieve higher precision at the expense of introducing bias.** It might be better to have a smaller sample size with less precision but much less bias.

2.2.3 Confidence Interval and Level of Confidence

As ENA calculates the sample size needed, it will determine by convention a 95% confidence interval. This means that in absence of bias, once you have the estimate, you can be 95% sure⁷ (level of confidence) that the true population value of your indicator is within the limits of the interval calculated. The more representative is the sample and the more accurate are the measurements, the more certain you can be about the absence of bias and of having the true population value for the indicator within the confidence interval.

2.3 Parameters for Sample Size Calculation

The calculation of sample size depends on the sampling design and the indicator.

In most cases, to calculate sample size for **simple or systematic sampling** you need to come up with 2 numbers:

1. *Expected prevalence of the indicator you are measuring*
2. *The width of the confidence you have to achieve to meaningfully interpret your estimate*

Equation 1: Sample size formula for anthropometry in SRS

$$n = \left[z^2 \times \frac{p \times q}{d^2} \right]$$

Where:

- n = sample size
- z = linked to 95% confidence interval (use 1.96)
- p = expected prevalence (as fraction of 1)
- q = 1- p (expected non-prevalence)
- d = relative desired precision

In cluster sampling, the sample size calculated for simple or systematic random sampling is multiplied by a factor called the *design effect* to account for the heterogeneity between clusters with regard to the measured indicator (see 2.3.3). Therefore, if you are doing cluster survey, you will need to come up with the 3rd number:

3. Expected Design Effect

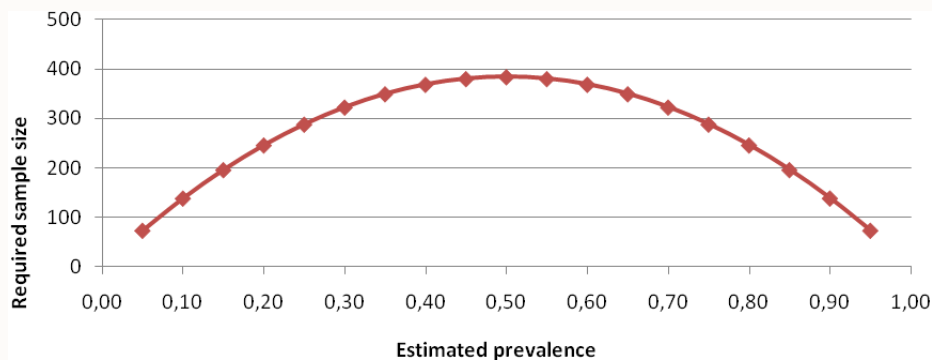
The following sections will provide guidance on how best determine these numbers:

⁷ The level of confidence used in ENA is always 95%.

2.3.1 Estimated Prevalence

Information to determine the estimated prevalence can be obtained from various sources such as previous surveys, surveillance data or rapid assessment results. Health workers can also provide interesting information about whether they see more thin children than in the past. If there is a range of possible values, it is always better to choose the one which is closer to 50%, which is the most conservative value (producing the largest sample size possible given the other parameters held constant) as shown in Figure 9.

Figure 9: Effect of Changing the Estimated Prevalence; (assume 95% CI, +/- .05, large population)



To make the best possible “guesstimate” of expected prevalence:

1. Look at results of previous surveys conducted in this area or in similar areas, and consider confidence intervals reported in these surveys. If no surveys exist, try to estimate the prevalence from using data from rapid assessments, anecdotal reports, feeding program admissions’ trends, etc.
2. Adjust previous estimates for seasonality, and according to your understanding of how and if the situation changed since last surveys were conducted.
3. After going through points (1) and (2), you may come up with the range of values where you think the current prevalence might be. To be on the safe side, use the higher limit of this range to plug into the formula for sample size calculation. For example, if you think the current prevalence is likely to be between 15% and 19%, use 19% for your calculations

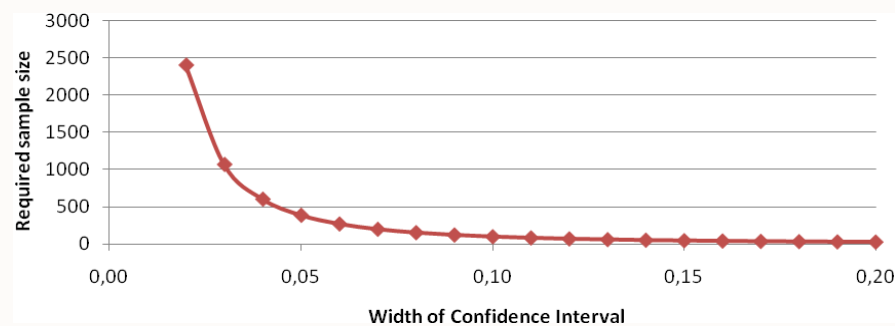
2.3.2 Precision

There is no standard precision to use when planning a survey. Desired precision depends on the objectives of the survey, estimated prevalence or rate, and resources available. Other things kept equal, the higher your desired precision⁸, the larger is the sample size. This effect can be seen through Equation 3 where squared precision is in the denominator, and therefore is inversely quadratically proportional to sample size. In other terms, to decrease the width of your confidence interval by 2 times, you will need to increase your sample size 4 times (2^2). Therefore, logistical or financial resources might also be taken into consideration when calculating sample size. If you get a sample size of 384 HH with a precision of $\pm 3\%$ (estimated prevalence: 10%), increasing this precision to $\pm 2\%$ will result in a sample size of 864 HH (2.25 times higher). Much more resources will then be needed to survey that many HH (Table 3 and Figure 10).

Table 3: Effect of changing precision on the sample size

Malnutrition Prevalence %	Desired Precision \pm %	Sample Size (children)
10	1.5	1537
10	2.0	864
10	2.5	553
10	3.0	384

Figure 10: Effect of Changing the Desired Precision; (assume 95% CI, 50% prevalence, large population) (source: CDC)



⁸ Note that a higher precision for measurements means a smaller value associated to that precision. For example, changing from a precision of 10% to 5%, the value is decreased, but this means achieving a higher precision.

In general, to meaningfully interpret GAM estimate in your survey, a precision of $\pm 3\%$ should be sufficient in most cases, even if expected prevalence is low. If expected prevalence of GAM is higher, for example 15-20%, the precision of $\pm 4-5\%$ would likely be sufficient. As shown in Figure 10, requiring confidence interval to be narrower than $\pm 2.5\%$ would result in very large sample sizes that would be very difficult to achieve, and in most cases would be unnecessary. If the estimated prevalence is very high, precision does not need to be very high, since the limits of the confidence interval might not make a big difference in decision making. For example, if the prevalence of GAM is of 45%, having a precision of $\pm 10\%$ might be enough, because the confidence interval will be from 35% to 55% and at that stage, it shouldn't make much difference on whether or not to intervene (Table 4)⁹.

Table 4: Variation of precision needed with different malnutrition prevalences.

Malnutrition prevalence %	Desired Precision $\pm \%$	Sample size (children)
5	3.0	203
10	3.5	282
15	4.0	306
20	4.5	304
30	5.0	323
40	6.0	256

Precision may need to be higher if survey results are to be used to monitor the situation on a yearly basis, if you are trying to detect statistical difference between two survey estimates. In that case, a different formula for survey size calculation is used.

2.3.3 Design Effect

The design effect (DEFF) is a “correction factor” to account for the heterogeneity between clusters with regard to the measured indicator. Therefore, it is only used to determine sample size in cluster sampling (Equation 3).

⁹ Values in tables 3 and 4 are only suggestions. Depending on the context, you may want to choose a precision that is more adapted to your survey.

Equation 3: Sample size formula for anthropometry in cluster design

$$n = \left[t^2 \times \frac{p \times q}{d^2} \right] \times DEFF$$

Where: n = sample size
 t = linked to 95% confidence interval for cluster sampling (2.045)
 p = expected prevalence (fraction of 1)
 q = 1- p (expected non-prevalence)
 d = relative desired precision
 DEFF = Design Effect

If the prevalence of GAM among the clusters in your survey is very different, the number of GAM cases per cluster will likely vary widely, meaning that the GAM distribution is heterogeneous, and DEFF is likely to be high. On the other hand, if the prevalence of GAM in all clusters is about the same, the number of GAM cases per cluster will also be similar, meaning that GAM distribution in the survey area is fairly homogenous, and DEFF is likely to be low.

The design effect to use when planning the survey can be taken from previous surveys in the same area, if there is no reason to think that there has been any change that might have increased or decreased the heterogeneity among clusters.

Example: A measles outbreak in one or two clusters may increase prevalence of malnutrition in these clusters compared to others with no outbreak. This high cluster prevalence is not representative of the whole area but will increase the general prevalence estimate which will not reflect the reality of the majority of clusters. To compensate for this, we will increase our sample size by using a correction factor called design effect.

Generally, if there is no previous information about design effect, 1.5 can be used as a default for GAM.

DEFF depends on the prevalence and the size of the clusters:

1. The higher the expected prevalence, the higher would be DEFF. For example, if your expected prevalence is around 10%, expected DEFF may be 1.5, whereas if expected prevalence is around 25-30% you would increase your expected DEFF to 1.7-1.8.
2. The smaller the number of children per cluster, the smaller the DEFF. For example, if you are measuring 15 children per cluster, your DEFF may be 1.5, whereas if you plan to measure 25-30 children per cluster, you would increase your expected DEFF to 1.7-1.8.

Remember, DEFF is a multiplying factor (i.e., sample size is directly proportional to DEFF), so if you increase DEFF 2 times, your calculated sample size will be 2 times higher (Table 5).

Table 5: Effect of changing design effect on sample size assuming a 3% precision

Malnutrition prevalence (%)	Design Effect	Sample Size (Children)
10	1.0	418
10	1.5	625
10	2.0	836
10	3.0	1 255

The 3 components above: estimated prevalence, desired precision and design effect allow ENA to calculate sample size in terms of number children. To convert this number into number of households, more elements need to be determined, as described below.

2.3.4 Demographic Data

Two types of demographic data are needed to convert the number of children into the number of households:

- Average HH size
- Proportion of children under 5 in the population

This information can be obtained from a census or a past mortality survey that was conducted in the same area.

In many contexts of developing countries, the proportion of children under 5 does not reach 20%. Over-estimating this proportion and/or the number of persons per HH will result in having a small sample size in terms of HH (other parameters kept equal), and therefore smaller than expected final sample size in terms of children.

Even if your target population is children 6-59 months, you only need to enter the under-5 proportion in ENA and it will make the calculations automatically assuming that 90% of the children under 5 are aged between 6-59 months.

Household Definition

Definition of household is a key part of the survey planning since most of the time; households will be the basic sampling unit at the last stage of sampling. There is no clear consensus about a definition. It is recommended to use the definition that is used by the country's departments. This definition needs to be agreed upon **before** the survey starts.

The most frequently used definition is: "**People who slept in the house last night and ate from the same cooking pot.**"

In country where polygamy exists, a clear definition of a household is even more crucial. By having a specific definition, you'll be able to determine how many households there are in a polygamous family. For example, if all wives cook together, eat together and live in the same compound, this would be 1 household. However, if each wife has her own kitchen and prepares food for her own children, those would be separate households.

2.3.5 Non-Response Rate

Non-Response Rate (NRR) accounts for households that could be either absent, not accessible, refuse to be surveyed, or any other reason that prevent survey teams from surveying a selected household. ENA uses this rate to inflate sample size using the following formula:

Equation 2: Non-response rate calculation

$$Final\ N = \frac{Number\ of\ HH\ needed}{1 - NRR}$$

Example: If according to your estimated prevalence, desired precision, and design effect, number of persons per household and % of under 5, you get 695 HH as the sample size, and you expect to have about 5% of non-response rate, the final sample size calculated by ENA should be:

$$Final\ N = 695 / (1 - 0.05) = 732\ HH$$

2.4 Correction for Small Population Size

As you may have noticed, sample size formula does not include the size of the population of your sampling universe. However, if the **target population** is below approximately 10 000, you have to check the box in ENA planning screen called *Correction small population size* in order to adjust the sample size. If your target population is small, smaller sample size would be needed to achieve required precision.

ENA calculates the target population from the total number entered in the cluster selection table and % of Under-5 children entered into the calculator. For example, if the total population size in the cluster selection window is 40,000, and % of under-5 is 15%, ENA would assume that there are $40,000 * 0.15 * 0.9 = 5,400$ children aged 6-59 months in this sampling universe, and use this number for adjustment for small population size.

2.5 Combined Surveys

In some contexts, you might want to conduct a mortality and anthropometric survey simultaneously in the same population. ENA can be used to calculate sample size for both¹⁰. However, in most cases, sample sizes for anthropometry and mortality will not be the same.

¹⁰ For more information about sample size calculation for mortality, please refer to mortality module in the SMART manual or the SMART Standardized Training Package on www.smartmethodology.org.

There are 2 scenarios possible:

- Small difference between the 2 sample sizes: the total number of households to survey for both components (anthropometry and mortality) will be determined by the higher sample size found through calculations. For example, if sample size is 497 HH for anthropometry and 425 HH for mortality, the higher value (497 HH) will be considered as the sample size to survey for both indicators. This will add about 2-3 households to each cluster in a survey containing 30 clusters.
- Large difference between the 2 sample sizes: every indicator should be surveyed with more or less the sample size calculated for that indicator; you will therefore have to set an interval at which teams will be collecting data for that indicator that requires smaller sample size. For example, if the sample size for anthropometry is 200 HH and the one for mortality is 400 HH, teams will be conducting anthropometry measurements only in children found in every 2nd HH selected in each cluster.

Scenarios to Avoid in Combined Surveys

1. Measuring one indicator only in some of the *clusters*, and measuring both indicators in other clusters. This would create a selection bias because the assignment of clusters at the planning stage is done according to the number of clusters needed for the whole survey. By doing an extra selection within those clusters, not every individual in the population will have had the same chance of being selected for that second indicator.
2. Measuring both indicators only in the first households of every cluster *when doing systematic random sampling*. This might also create a selection bias as it alters representativeness of the sample if the last households in the village are different from the first ones (access to road, market, etc.)

2.6 Determining the Number of Clusters

Once you have calculated your sample size in terms of households, you will need to determine the number of households to include in each cluster in order to find the number of clusters that will be visited. This is done by taking into consideration the reality of the team work on the field during data collection. This allows survey managers to not overload teams by requesting to survey too many households per day.

If we suppose that one cluster will be completed each day by each team, the calculation for the number of households to include in each cluster is based on:

- a. Travel time.
- b. Number of work hours per day.
- c. Number of hours spent on the field (excluding transportation).
- d. Number of hours spent on surveying in the households and getting from one selected household to the next (excluding break times and time spent to introduce teams and select households.)

However, **there is a minimum of clusters that should be included in each survey for the survey to be considered valid**¹¹. 25 clusters are considered a minimum, but normally nutrition surveys include at least 30 or more clusters. Therefore, in most situations, please consider 30 clusters as a default. If there is a lot of heterogeneity in the distribution of measured indicator, consider including more clusters and measuring fewer children per cluster to decrease DEFF.

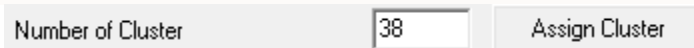
2.7 Assigning Clusters

After defining your geographical area at the planning stage and taking into account travel times, security issues, and any other factor that could influence your ability to get to the cluster site, you would need to choose the smallest administrative unit that has population data and a known name.

At this stage, each geographical unit should have at least the number of HH required to complete a cluster. Otherwise, you should pair it with the neighbouring unit when constructing your sampling frame for cluster selection.

After that, the sampling frame is pasted into the cluster selection window on the Planning page of ENA. The sampling frame is a list of all geographic units (or settlements) with the population size of each geographical unit (expressed in terms of total number of people, not households). The order of settlements on the list is not important, it can be sorted alphabetically or geographically. It is very important that **all** settlements in the survey area be included. If some areas are omitted at this step, they will not be part of the surveyed population and this should be clearly described in the report. You can then enter the number of clusters to select from this sampling frame and press the button *Assign Clusters* (Figure 11). The software will automatically select the areas or “villages” that contain a cluster. This is assigned randomly using PPS procedure described above (see Section 1.2.3).

Figure 11: ENA functionality for assigning clusters



Number of Cluster 38 Assign Cluster

2.7.1 Small Villages

If a village does not have enough households for 1 cluster, it should be combined with the closest one. If this combination of villages is chosen by ENA to contain a cluster, allocate HH proportionally amongst the two villages.

Example:

Village A has 15 HH; village B has 60 HH. Since village B is 4 times bigger than village A, we will allocate 4 times more HH to village B. If the cluster should contain 20 HH:

- Village A: 4 HH
- Village B: 16 HH

¹¹ Binkin N. *Rapid nutrition surveys: how many clusters are enough?* Disasters 16(2): 97-103.

2.7.2 Large Villages

If very large village is chosen to contain only 1 cluster, use the segmentation as explained in section 3 of this paper.

If a large village is chosen to contain 2 clusters or more, two options are possible:

- a. Segmentation: divide the village into 2 or more segments and assign a cluster randomly to each segment if equal segment or using PPS if segments are unequal.
- b. Conduct the clusters one after the other. However, in order to have a proper data analysis, each of those 2 clusters needs to have a different number. Therefore, teams have to change the cluster number on the questionnaires after finishing each one.

2.8 Reserve Clusters

As discussed earlier, the sampling frame should only include sampling units that are accessible and can be visited if selected. However, in some cases surveyors could be prevented from accessing previously selected clusters after survey starts. Clusters might be impossible to survey because of insecurity, accessibility, refusal of the community or leaders, etc. Examples can be a conflict that starts in one village that contains an original cluster, or a flood might occur and a cluster becomes inaccessible.

At the planning stage, when assigning clusters, ENA automatically chooses additional clusters called **reserve clusters** or **replacement clusters**. The number of reserve clusters (RC) selected by ENA is as follows:

If 25-29 clusters in the survey → 3 RC; if 30-39 clusters → 4 RC, if 40-49 clusters → 5 RC, etc.

Reserve clusters should **only** be used if the 10% or more of original clusters were impossible to reach during the survey. Example: you plan to survey 30 clusters. As your data collection goes on, you find out that there are 3 or more clusters that became impossible to survey. You can therefore include the replacement clusters.

All reserve clusters should be surveyed if you haven't reached 10% or more of your clusters. Example: if your target is to survey 40 clusters and you only could access 36 clusters, you should include all replacement clusters in your survey, even if there are 5 of them.

3. Second Stage Sampling

If you have decided at the planning stage to use cluster sampling, ENA will select randomly villages to contain each cluster. This is called the first stage of sampling. The next step is to select randomly basic sampling units within those clusters, which is called the **second stage (or last stage) sampling**.

3.1 Preparation for household selection

As mentioned above, one of the reasons to use cluster sampling is to divide the survey area into smaller geographical units where simple or systematic random sampling will be feasible. However, in some cases, even villages that are chosen to contain a cluster might be too large and in most cases, no list of households is available at the village level. Therefore, when arriving to the village, some preparatory work needs to be done before being able to select the actual households that will be included in your survey.

3.1.1 List of Households Available

After introducing the team and explaining the objectives of the survey to the local authorities in the village, it is important to get an **updated** list of the households for that village. This list should not be based on some socio-economic indicators that would exclude part of the populations in the village; for example, some lists only include mothers who have children and are followed by the female health workers. Using only this list will introduce an important selection bias. If the list is not updated or complete, the team leader can use the local knowledge to do so. It is very important that *no household* is excluded from that list.

Listing Households

It is very important to explain the definition of households to the local authorities when asking them for lists. Very often, people might confuse families and households. Hence, survey teams should make sure that village leader and people present during those preparatory steps understand the definition of household that was used in the planning stage of the survey.

3.1.2 List of Households Not Available

3.1.2.1 Drawing a Map

If there is no list in the village, survey teams should try to make one with the local authorities. The easiest procedure to make this list is to:

1. Ask the village leader or the most knowledgeable person present to draw the boundaries of the village on a paper.
2. Identify the natural and community landmarks on the map (e.g., rivers, roads, church or mosque, school, health center, market, etc.) This will allow you to define internal boundaries and consider separate sections of the village for the next step.
3. Ask the village leader to list households that live in each section limited by specific landmarks; for example, survey team can ask the village leader to list all households that live in the area between the market and the river. Go through all sections of the village writing the head of household's name and giving each household an ID.
4. If the village is not too large, walk around with the village leader and confirm the number of households for each section.

3.1.2.2 Numbering the households

In some cases, villages are not very large but village leaders are still unable to list all households. In such situations, survey team members can walk around the village and identify all households, by writing a number (starting at 1 to the total number of households in the village) with a chalk on their door, for example.

3.1.3 Segmentation

In some cases, villages selected randomly to contain a cluster might be very large or households very dispersed and sample selection can then become very tedious; teams will have long distances to walk and not enough time to complete one cluster per day. In those scenarios (approximately more than 100 households in the village), **segmentation** can be used in order to reduce the area that will be covered by the survey teams.

The objective of this procedure is to divide the village into smaller segments and choose one segment randomly to include the cluster. This division can be done based on existing administrative units (neighbourhoods, etc.), natural landmarks (river, road, mountains, etc.) or public places (market, schools, churches, mosques, temples, etc.)

Segmentation can be done into equal or unequal parts.

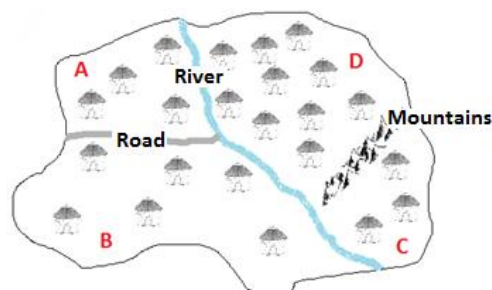
3.1.3.1 Segmentation into equal parts

If the village can be divided into 2 or more **approximately equal** parts each containing less than 100 households, the survey team leader can write the name of those parts on pieces of paper that he folds and put into a bag or hat and have the village leader or his representative choose one part randomly. Team will therefore go to that part of the village to conduct the survey for that cluster.

3.1.3.2 Segmentation into unequal parts

In some cases, it might be impossible to divide the village into equal parts, as shown on Figure 12. You should therefore try to find some natural landmarks that can help divide the village into separate clearly defined segments. Once those segments are defined with an approximate population size, one segment will be selected randomly using PPS as shown in the example below.

Figure 12: Segmentation into unequal parts



Example: You're conducting a survey where each cluster should contain 15 households.

Figure 12 shows one of the villages selected to contain a cluster. It has segments that have unequal number of households.

- A = 70 HH
- B = 100 HH
- C = 30 HH
- D = 190 HH
- TOTAL = **390** HH

To choose one segment randomly using the PPS, we have to create a table with population intervals for each segment (Table):

Table 6: PPS for segment selection

Segment	Population (HH)	Cumulative population	Intervals
A	70	70	1-70
B	100	170	71-170
C	30	200	171-200
D	190	390	201-390

After filling out the cumulative population numbers and setting intervals for each segment, pick a 3-digit number from the random number table between 1 and the cumulative total of all the segments.

The segment containing the selected number will be the one to survey.

In this example, a 3-digit number must be picked from: 001- 390. E.g. we picked: 167. This number is within the segment B. **Survey will therefore be conducted in segment B**

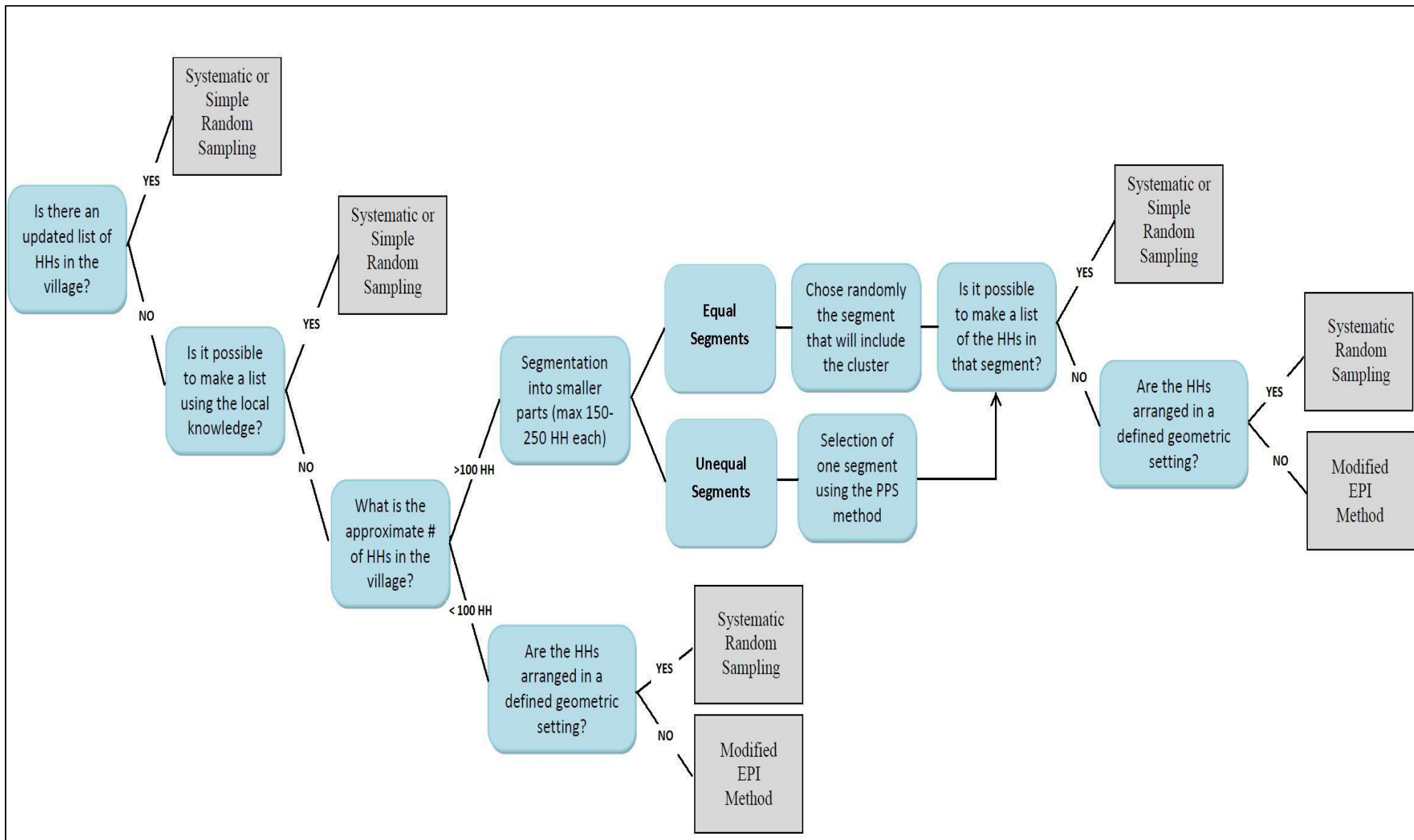
3.2 Household Selection Techniques

Once a segment is selected, survey team would need to go through the same process of getting a list of households as explained in sections 3.1.1 and 3.1.2. After the preparatory steps mentioned in previous section (segmentation, getting a list of households), selection of households within clusters can start using the most recommended methods: simple or systematic random sampling. Another method that was used in the past is the modified EPI. The choice of the method depends on:

- The population size of the village or segment.
- The distribution of households.
- The possibility to get or build an updated list of households.

Figure 13 shows the different pathways to help decide which household selection technique to use depending on the context.

Figure 13: Decision tree for household selection at the last stage of sampling



3.2.1 Simple random sampling

As discussed in section 1.2.1, simple random sampling can be conducted when an updated and exhaustive list of households is either available or possible to make. Picking numbers from a hat/bag or using a random number table can allow choosing the number of households needed for the cluster.

3.2.2 Systematic random sampling

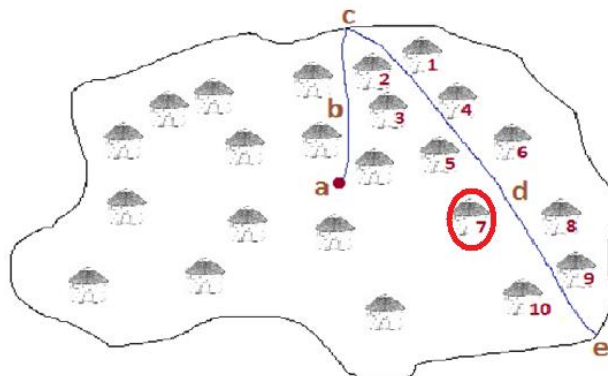
Systematic random sampling has been discussed in section 1.2.2 and relies on the calculation of a sampling interval that will allow finding the first household of the cluster and selecting the following households until the required number is reached.

3.2.3 Modified EPI

The Expanded Programme on Immunization recommended a procedure in the late 1970s to select samples in surveys measuring immunization coverage. This procedure has been slightly modified and used by some teams for nutrition surveys. However, as mentioned earlier, it should only be used when it's really impossible to use either simple or systematic random sampling. To apply this method, you need to (Figure 14):

- First, find the center of the village or segment where you are conducting your survey and choose a random direction by spinning a pen or any sharp object.
- Walk in a straight line until you reach the boundary of the village or cluster.
- Spin the pen or the sharp object once again.
- Walk along this 2nd direction, numbering households on both sides (left and right); for example, N= 10.
- Choose a random number between 1 and N (e.g., between 1 and 10; let's say 7).
- Go back to household number 7. This is the first household to survey.
- Continue to next household on the right until the cluster is completed.

Figure 14: Steps for modified EPI method



If the team arrives at the edge of the village without reaching the number of households needed to complete the cluster, they have to go back to the center of the village and repeat steps *a* through *g*. If all the households of a village have been selected and yet the cluster has

not been completed, survey team has to go to the nearest village and repeat the steps *a* through *g* until all households needed for the clusters are surveyed. This situation can be partly avoided if there are good population data before selecting clusters, so that villages that don't contain the minimum number of households (needed to complete a cluster) are grouped with the nearest ones, and this group is entered in the table for cluster sampling of ENA as one sampling unit (0).

Modified EPI Limitations

This method is the least recommended one because it induces certain biases in the sampling:

- Since we always select the closest household on the right of the previous one, or the one that has the closest door, it creates a bias because the selection of every household is not independent from the selection of the previous one (and this is one important criterion for probability sampling).
- Also, by choosing households by proximity, you are more likely to stay within the same area of the village, while households close to each other are more likely to have the same characteristics, and hence, you might not be getting a representative sample of your cluster.

4. Special Cases

On the field, teams might encounter various difficulties some of which are addressed in this section. Teams should therefore use a tool such as a cluster control form (see Figure 15 below) to summarize what has been done in each cluster and keep track of which household refused to participate, which had eligible household members absent and need to be re-visited, etc.

4.1 Polygamous Families/Compounds

As mentioned earlier, a clear definition of household has to be agreed upon before starting the survey. This will allow determining the number of households within a compound or a polygamous family.

If a compound/polygamous family contains more than 1 household, each HH should be included separately in the list for household selection.

If this information was not communicated when the list of households was constructed with the village leader, and the team finds out about it once in the compound, random selection can be done among those households to choose the one that will be included in the sample. Even if this step will facilitate selection for the teams, this scenario is not ideal since it adds an extra stage of sampling and hence, decreases the chances of each household within a compound of being selected compared to the rest of the sample.

Figure 15: Example of a Cluster Control Form.

CLUSTER CONTROL FORM							
Province : _____		District : _____		Village/Camp : _____		Cluster # : _____ Team # : _____	
Date of Sampling : ____ / ____ / ____							
HH no	Head of HH name	Visit Outcome 1 = completed 2 = partly completed 3 = refused 4 = absent*	Number of eligible children	Number of eligible children measured	Household needs to be revisited Yes or No	Household re-visited Yes or No	Comments
1							
2							
3							
4							
5							
6							
7							
8							
9							
10							

* If household, eligible child or PLW is absent, team should re-visit the household once before leaving the village to conduct the interview and/or measure the child or PLW.

Team leader's signature:

4.2 Impossible to Visit a Selected Household

Many reasons can prevent a team from visiting a household that was selected during the sampling procedure: refusal, fear of dog, access, conflicts with village leader, etc. In this case, the teams should visit the next selected household according to the sampling procedure that was used, **without replacing this household**. The final cluster size of that day will then be decreased by the number of households that were impossible to visit, but this was accounted for in the planning stage by inflating the sample size with the non-response rate.

4.3 No children in the Household

If the team arrives at a household that only has adults, they should still complete the rest of the questionnaire with a respondent and then move to the next household according to the sampling method chosen. This household should not be replaced. It is normal that not all households in the sample contain children and this was also planned for by entering the proportion of expected children in the population. Excluding households without children from selection will introduce serious selection bias in measuring household-level and other non-child variables (e.g., mortality, haemoglobin in women, access to safe water, etc.)

4.4 Absent Household

Survey team can arrive at a household and find no one because all members are absent. To simplify the procedures, an absent household is defined as the household where some members have slept there the previous night but were absent on the day of the survey.

Survey teams can ask neighbours why the house is empty, but **should not** fill out household survey questionnaire by interviewing neighbours about the absent household. Afterwards, they could move to the following household according to the sampling procedure.

Survey teams should return to absent households before leaving the village, to see if residents are back. If not, this should be reported on the questionnaire. Absent households **should not be replaced**, as discussed in section 4.2 above.

If more than 5% of the households in a cluster are not found, the team should revisit the area at another time to see if they can complete the sample.

4.5 Abandoned Household

An abandoned household is either a house that had no one living in it for a long period, or one that is empty but does not fall in the absent household category (inhabitants left more than 1 day ago and will not be back at the end of the day, etc.)

Those households should not be included in the list of households used for household selection at the preparatory steps when survey team arrives to the village. If such empty house is encountered during systematic selection, it should be ignored, and only inhabited houses should be considered in selection procedure.

4.6 Absent Child

If a child lives in the household but is not present at the time of the survey, this should be recorded on the household questionnaire and on the cluster control sheet when the household is visited; the rest of his/her information (age, sex, feeding practices, immunizations, etc.) and rest of the questionnaire can still be filled out with the respondent such as mother or caregiver.

The team should tell the mother that they will return later on that day to measure the child. If the team returns later but child is not present, this should be recorded as such in the questionnaire and on cluster control sheet, and total number of absentees has to be included in the final report. Absent child in selected household **should not be replaced**.

4.7 Child with Disability

If the target population of your survey is 6-59 months children, those with a disability falling into that age group are eligible for your survey. However, some disabilities might not allow you to take all anthropometric measurements needed or might lead to a biased measure. For example, the weight of a child missing a limb will not be very meaningful when comparing it with the standard population.

All other data that is not influenced by the disability should be collected such as sex, age, edema (if the child has both feet), etc.

4.8 Child in a Center or a Clinic

When a team finds that one of the eligible children is in a feeding center or a clinic, they should first try to find out the type of disease: malnutrition or not. This child should then be recorded as absent with the reason the absence. The team should attempt to find and measure the child in a feeding centre if it is nearby, or re-visit the household later in the day if the child returns home in the evening.

If a lot of children are being reported as being in the health centre or rehabilitation centre, this may introduce important selection bias. This should be described and discussed in the discussion part of the final report.