ANNEX 4

Statistical procedures for nutritional surveys

Introduction

This annex provides guidelines for statistical procedures, including sampling methods and determination of sample size, to be used in nutritional surveys. It fulfils the need — unmet by most handbooks, which deal more with surveys of communicable diseases — for guidance on the type of community-based survey essential for nutritional assessment.

The essential procedures for anthropometric surveys are covered in Chapter 3 and Annex 3; Chapter 2 outlines the parameters and criteria (mostly clinical and biochemical) used in assessing micronutrient deficiencies. In practice, a survey that combines clinical, anthropometric, and biochemical elements is required. Different types of nutrient are usually assessed in different age groups or among individuals of different physiological status, and few manuals provide guidance on how such assessments should be combined or integrated. Table A4.1 shows the suggested age/sex groups to be examined — usually on the basis of the household-selection procedure described in this annex.

Involvement of a statistician right at the start of the survey design process is important, to ensure that sample sizes are appropriate (neither too large nor too small) and will produce results from which valid comparisons can be made between different populations and in the same population over time. The sample size is usually similar for anthropometry and for assessment of the different types of nutrient, but the design factor (increase in size of cluster sample required because of patchy distribution of the deficiency) is generally recommended to be larger for micronutrient surveys (3) than for anthropometric surveys (2).

The first part of the annex deals with the principles of random sampling and with sample size, and the second part presents various sampling procedures.

Principles of random sample surveys

Basic concepts

When dealing with large population groups it is not feasible to survey all individuals. However, valid conclusions can be drawn from measurements made on only a limited number of individuals within the population, provided that this "sample" is representative of the population as a whole.

The sampling techniques described in this annex are designed to ensure this essential representativeness through randomization in selection and elimination of observer bias. Data obtained only from health services, for example, are

***Table A4.1** Examples of appropriate age/sex groups for nutritional assessments*

Age/sex group	Type of assessment
Children <5 years	Anthropometry Anaemia, vitamin A deficiency, beriberi, scurvy (if any cases seen)
Children of school age (6–12 years) and adolescents	Goitre prevalence; urinary iodine; anaemia/iron deficiency
Women of reproductive age, or pregnant women	Anaemia/iron deficiency, beriberi, scurvy
Adults	Anthropometry Beriberi, pellagra, scurvy (if any cases seen)

unlikely to be representative of the population as a whole; data collected only in the most accessible villages, or in camps that are reported to be in a bad state, will be similarly unrepresentative. Strict procedures must be followed in selecting individuals to be included in a sample to ensure that it is representative. Moreover, if the objective of a survey is to compare the nutritional status of two groups, representative data must be collected from the two groups separately.

The techniques, and the methods of analysing the results, recognize and allow for the fact that there may be some inaccuracy. Data gathered from a sample of a population provide only an estimate of what the results would be if measurements were made on the entire population. Whenever a sample is drawn, there is a risk that it may not be truly representative and therefore yield data that do not reflect the true situation. Inevitably, therefore, if a second sample is drawn from the same population, slightly different results are likely be obtained.

From a sample it is possible to calculate not only an estimate of malnutrition (or other variable of interest) but also the *range* of values within which the actual rate of malnutrition in the entire population almost certainly lies. The confidence interval is strictly not symmetrical, but as the sample size increases it becomes more and more symmetrical. For example, the 95% confidence limits for a 10% estimate of malnutrition based on a randomly selected sample of 30 children are 2% and 26%. However the confidence limits for a 10% estimate based on a sample size of 2000 are 9% and 11%. See Table A4.3.

A 95% confidence level[1] is usually considered to be appropriate for nutritional surveys. The precision of the result and the size of the confidence interval depend on the sample size and the actual prevalence of malnutrition (or other variable of interest) in the population.

Basic sampling procedure

Three main sampling methods can be used — random, systematic, and cluster. Cluster sampling is the most widely used and often the only feasible method in emergencies involving large population groups. In all cases, estimates are required of the total population and of any subgroups to be distinguished within the total. The essential steps in obtaining a sample are as follows:

[1] A 95% confidence level represents an *error risk* of 5%, meaning that, out of 100 surveys, as many as 5 may give results that do not reflect the true situation purely by chance.

1. *Obtain available population data.* Census data and a list of all settlements in the area might be obtained from departments of planning, statistics, or malaria control, for example. If no data are available, as may be the case for refugees or displaced persons, a rough population estimate should be made by counting the dwellings and estimating the number of people in each dwelling.
2. *Divide the total population into groups relevant to the information to be collected.* In the case of camp populations, it may be desirable to distinguish between different camps, different sections of camps, or between long-term residents and new arrivals. Among rural populations it is generally appropriate to distinguish pastoralists (such as nomadic herders), subsistence farmers, and others (including artisans and traders). If different groups are not distinguished, the survey findings may be difficult to interpret.
3. *Choose the sampling methodology to be used.* The required precision should be identified and the necessary sample size determined accordingly.
4. *Select the households or individuals to be examined.* The relevant sampling procedures should be followed carefully.

Defining sample size

The sample size is the number of individuals to be included in the survey to "represent" each population of interest. The sample size required depends on the following factors:

- *Required precision and confidence level.* The greater the precision required, the larger the sample needed.
- *Expected prevalence of malnutrition* (or other variable being estimated). The smaller the expected proportion of people presenting malnutrition, the greater the size of the sample required for a particular level of precision.
- *Time and resources available.* The time, personnel, equipment, transport, and funds available for the survey may limit the number of individuals or households that can be visited.

In practice, selection of sample size almost always involves a trade-off between the ideal and the feasible. A sample that is too small gives results of limited precision and therefore of questionable usefulness. For example, a result of 10% wasting (below median −2 SD weight-for-height) in a sample of 100 children would give a confidence interval ranging from approximately 4% to 16% — a result that cannot be interpreted usefully. Beyond a certain level, however, increases in sample size produce only small improvements in precision but involve disproportionate increases in costs. The formulae for calculating sample size (n) are as follows

- *for simple random sampling*

$$n = \frac{1.96^2 \times (1 - p)}{p \times r^2}$$

• *for cluster sampling*

$$n = \frac{k \times 1.96^2 \times (1-p)}{p \times e^2}$$

where:

n = sample size required

p = expected prevalence of malnutrition in the population; as the prevalence of malnutrition is not known before the survey is done, an estimate must be used — this is usually an experienced guess, or derived from a small pilot survey

e = relative precision required

1.96 is a statistical parameter corresponding to the confidence level of 95% (an error risk of 5%).

k = "clustering" factor, or design factor, which is a measure of the clustering of the characteristic being measured.[1]

p and e can be expressed either as percentages or as fractions of 1 (10% = 0.10), but must both be expressed in the same terms.

The sample size for a cluster survey is likely to be larger than that for a random sample for the same precision. This is because the units within a cluster tend to be similar in their characteristics. Poor (and therefore malnourished) people, for instance, are likely to be found living together in the same areas.

Example

Expected prevalence of malnutrition 15%: $p = 0.15$

Relative precision required (e) 20% of the estimated prevalence

Design factor $k = 2$.

For random sampling:

$$n = \frac{1.96^2 \times 0.85}{0.15 \times 0.20^2} = 544$$

For cluster sampling:

$$n = \frac{2 \times 1.96^2 \times 0.85}{0.15 \times 0.20^2} = 1088$$

Table A4.2 shows the sample sizes required for particular levels of expected prevalence and required precision with a fixed error risk of 5%. To take another example, if the expected malnutrition rate is 15%, and a relative precision of 3% is required, a sample size of 24 188 obtained by *simple random sampling* will be needed. For *cluster samples*, the figures in Table A4.2 should be multiplied by the appropriate design factor for the "clustering" of the characteristic being measured within sample clusters.

Table A4.3 shows confidence intervals at the 95% level (5% error risk) corresponding to various sample sizes and observed rates when random sampling is

[1] According to studies analysed by CDC, the design factor k usually has a value of approximately 2 in anthropometric studies among children under 5 years of age, with 30 clusters.

Table A4.2 Sample sizes for estimating a population proportion with specified relative precision (95% confidence level)[a]

ε[c] \ P[b]	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95
0.01	729904	345744	217091	153664	115248	89637	71344	57624	46953	38416	31431	25611	20686	16464	12805	9604	6779	4268	2022
0.02	182476	86436	54423	38416	28812	22409	17836	14406	11738	9604	7858	6403	5171	4116	3201	2401	1695	1067	505
0.03	81100	38416	24188	17074	12805	9960	7927	6403	5217	4268	3492	2846	2298	1829	1423	1067	753	474	225
0.04	45619	21609	13606	9604	7203	5602	4459	3602	2935	2401	1964	1601	1293	1029	800	600	424	267	126
0.05	29195	13830	8708	6147	4610	3585	2854	2305	1878	1537	1257	1024	827	659	512	384	271	171	81
0.06	20275	9604	6047	4268	3201	2490	1982	1601	1304	1067	873	711	575	457	356	267	188	119	56
0.07	14896	7056	4443	3136	2352	1829	1456	1176	958	784	641	523	422	336	261	196	138	87	41
0.08	11405	5402	3401	2401	1801	1401	1115	900	734	600	491	400	323	257	200	150	106	67	32
0.09	9011	4268	2688	1897	1423	1107	881	711	580	474	388	316	255	203	158	119	84	53	25
0.10	7299	3457	2177	1537	1152	896	713	576	470	384	314	256	207	165	128	96	68	43	20
0.15	3244	1537	968	683	512	398	317	256	209	171	140	114	92	73	57	43	30	19	9
0.20	1825	864	544	384	288	224	178	144	117	96	79	64	52	41	32	24	17	11	5
0.25	1168	553	348	246	184	143	114	92	75	61	50	41	33	26	20	15	11	7	—[d]
0.30	811	384	242	171	128	100	79	64	52	43	35	28	23	18	14	11	8	5	—[d]
0.35	596	282	178	125	94	73	58	47	38	31	26	21	17	13	10	8	6	—[d]	—[d]
0.40	456	216	136	96	72	56	45	36	29	24	20	16	13	10	8	6	—[d]	—[d]	—[d]
0.50	292	138	87	61	46	36	29	23	19	15	13	10	8	7	5	—[d]	—[d]	—[d]	—[d]

[a] $n = (Z_{1-\alpha/2})^2(1-P)/\varepsilon^2 P$, where $Z_{1-\alpha/2}$ represents the number of standard errors from the mean, and α is the significance level of a test.

[b] P = anticipated population proportion (prevalence).

[c] ε = relative precision.

[d] Sample size less than 5.

Table A4.3 Confidence intervals at 95% probability level corresponding to various sample sizes and sample percentages

Sample size	Percentage observed in sample					
	5%	10%	20%	30%	40%	50%
30	1–18	2–26	8–39	15–49	23–59	31–69
40	1–17	3–24	9–36	17–47	25–57	34–66
50	1–15	3–22	10–34	18–45	26–55	36–65
60	1–14	4–20	11–32	19–43	28–54	37–63
80	1–12	4–19	12–31	20–41	29–52	39–61
100	2–11	5–18	13–29	21–40	30–50	40–60
200	2–9	6–15	15–26	24–37	33–47	43–57
300	3–8	7–14	16–25	25–36	35–46	41–56
400	3–8	7–13	16–24	26–35	35–45	45–55
500	3–7	8–13	17–24	26–34	36–45	46–55
1000	4–7	8–12	18–23	27–33	37–43	47–53
2000	4–6	9–11	18–22	28–32	38–42	48–52

used. For cluster sampling, the sample sizes must be multiplied by the appropriate design factor to take into account the clustering of the characteristic being measured.

If, for example, the observed malnutrition rate is about 20%, a total sample size of 100 will make it possible to estimate the true rate somewhere between 13% and 29%, assuming random sampling. If greater accuracy is required, for instance 18–22%, a sample size of 2000 would be needed.

In nutrition surveys in emergencies, the expected prevalence of severe malnutrition usually ranges between 5% and 20%, and the precision must be defined accordingly; a relative precision of 20–25% is generally appropriate.

The size of the total population does not normally affect the size of the sample required. However, if the population is small and the calculated sample size turns out to be greater than 10% of the total population, a correcting factor (finite population factor) can be applied as follows:

$$n_f = \frac{n}{1+f}$$

where

n_f = adjusted sample size for small (finite) population

n = sample size for large (infinite) population (for example, as set out in Table A4.2)

N = population size

$f = n/N$.

Calculating results and confidence intervals

When results have been calculated, the corresponding confidence interval, d, should be calculated as follows and reported:

- *for random sampling*:

$$d = \frac{1.96 \times (1 - p)}{np}$$

- *for cluster sampling* the following formula can be used to give an approximate result:

$$d = \frac{k \times 1.96 \times (1 - p)}{np}$$

Using a random number table

A set of random numbers is presented in Table A4.4. Numbers can be read in any direction — from left to right, right to left, top to bottom, or bottom to top.

Table A4.4 Random numbers

13118	50901	57493	96647	46146	65512	97571	49079	92251	36599
81111	33653	61544	90072	61635	94254	95222	49594	99403	56052
07124	56894	00475	09815	05299	17082	80775	11320	98562	68057
55155	23168	83063	80324	51450	68094	71844	68302	49552	12082
46406	44641	45461	75174	33268	86032	40355	58288	05532	29419
10616	17092	76614	04950	67982	28515	16782	80129	44391	64449
38497	57435	46124	37302	10783	93043	06903	77158	49638	26211
83203	45640	75843	75843	74567	75971	97779	98047	68916	35038
19236	62703	12863	14452	72228	55022	07024	43615	74802	02110
79024	60592	93692	29737	09314	26191	52484	11588	14078	85947
76073	57262	52795	67673	62267	29552	08244	49280	58583	42190
50568	66590	38807	30061	26336	46147	04554	41562	72604	63031
11838	73906	55981	23668	22627	88438	96686	73645	81410	10942
57618	30523	16757	11956	58411	41647	67884	30084	14500	66958
61846	47265	09508	11030	10462	93922	17022	71031	07827	94722
60935	25351	11687	07679	73455	55617	24415	56921	89450	50471
63328	21749	74262	77143	55995	50707	91516	38002	60552	00634
75937	07127	11014	00738	46150	09866	87587	41648	36538	24398
11981	89485	54065	08300	67724	24919	65682	50101	45470	07232
12311	17067	42758	04557	46297	28414	93801	81180	12176	08536
45160	76932	00433	42228	73696	27478	65321	22979	30198	86709
26427	48280	53441	44543	95231	39939	09251	09755	26671	89392
54568	17774	95705	28018	26507	63504	98872	22449	56423	59133
80855	94883	08969	16949	86045	68398	46161	57147	35104	37262
96203	73918	77875	48444	08167	58460	87945	52146	20330	77172
91210	89152	93904	27666	51060	00487	12073	41639	28717	33909
37808	11431	03351	82979	96677	41588	17592	5111x	64657	25427
47738	40686	00948	46598	99095	67011	05786	05642	26282	97456
03255	71561	78549	15611	49097	58375	70087	10066	63530	26684
92658	11755	39005	72396	20601	49630	85266	78939	89931	99674
80040	46908	88153	05616	91381	88378	28263	34725	80739	15251
67806	60615	14520	04557	72939	71060	10650	58769	07497	00808
46138	03111	47053	89391	83636	05877	17980	63940	23003	23737
81514	46994	77869	72054	22819	89316	77195	20194	65043	27706
28419	60216	07640	80670	84427	98368	99656	10214	04023	39899
99109	64711	06962	56790	96313	54470	18568	04319	31880	39507
15045	85129	03531	06107	93785	38290	00911	68388	68686	53357
61398	94861	90462	09438	53920	59096	91957	39255	86563	20781
58455	18205	39389	18266	22994	78421	22241	04228	86679	47840
81025	70374	79493	39366	41707	57491	35647	43409	37182	73435

Numbers can be read off with any required total number of digits. The steps involved in using this, or any other, set of random numbers are:

1. Decide on the direction in which numbers will be read; e.g. left to right going down the page.
2. Specify the required number of digits. If a random number is required in the interval 0001 to 1342, 4 digits are needed (any of which may be zero).
3. Close your eyes and stick a pin (or other sharply pointed object) in the table. Read off the required number of digits in the direction chosen in step 1, starting with the first digit to the left of the point. If the resulting number falls within the required interval, use this number. If not, repeat the process until an eligible number is drawn or move to the next number.

Sampling methods

All sampling methods involve a highly ordered form of selection designed to eliminate observer bias; each can be adapted in various ways depending on the situation. The paragraphs that follow provide a general description of each method and how it can be applied.

In all cases, each selected individual, or every child under 5 years old belonging to each selected household, must be seen and (for an anthropometric survey) measured. The survey team, with the help of the community, must find the individuals concerned, wherever they are. If necessary, the team must return later to see and measure an individual missed on the first visit. No substitutions can be allowed and no one can be missed (unless they have died or left the community being surveyed).

Random sampling

Random sampling is the best method — when it can be used — since it is the only one that ensures representativeness. An up-to-date list of all individuals in the population is needed, with enough information to allow them to be located. Individuals are randomly drawn from the list using a random number table (see above and Table A4.4). For a nutritional survey the sample would be restricted to children aged 6–59 months or 65–110cm in length or height.

In practice, a reliable population list is rarely available, and it is sometimes practical to use the following alternative procedure:

1. Go to the area and make a list of *all* households included in the area of interest.
2. Assign each household on the list an identification number.
3. Select the required number of households using a random number table. Otherwise, pick household identification numbers out of a hat or a large box. (If this type of selection is done in public, the community can see how households are selected.) A number corresponding to each household is written on a small piece of paper, which is placed in the hat or box. The pieces of paper are shuffled and the required number of papers are then picked out (blindly). The households selected in this way become the sample for the survey.

4. Visit all of these (and only these) households. No households may be excluded or substituted for any reason. In a nutritional survey, all children in the specified age group belonging to each selected household must be measured.

Systematic sampling

Systematic sampling eliminates the need for complete, up-to-date population registers, but requires:

- a reasonably accurate plan or map showing all households; and
- an orderly layout, or site plan, which makes it possible to go systematically through the whole site.

This technique has been used in well-organized refugee camps, where households are arranged in blocks and lines. The procedure is as follows:

1. Either list all households and assign each one an identification number, or trace a continuous route on the map, which passes in front of every household.
2. Calculate the number of households to be visited in order to obtain the required sample. If the required sample size is 544 and there are, on average, 15 children (aged 6–59 months) per 10 households, the number of households to be visited is 544/1.5 = 362.6, or 363 (*round up* to the nearest whole number in this calculation).
3. Calculate the "sampling interval" by dividing the total number of households by the number that must be visited. If the total number of households is 5000, and 363 are to be visited, the sampling interval is 5000/363 = 13.8, or 13 (*round down* to the nearest whole number in this calculation).
4. Select the first household to be visited within the first sampling interval at the beginning of the list (or route) by drawing a random number which is smaller than the sampling interval. If the number drawn is 7, start with the seventh house.
5. Select the next household by adding the sampling interval to the first household identification number (or counting that number of households along the prescribed route), e.g. 7 + 13 = 20.
6. Continue in this way (e.g. 7, 20, 33, 46, etc.) until the number of households required for the survey has been systematically selected.
7. Visit all of these (and only these) households. No selected household may be excluded or substituted for any reason.

Two-stage cluster sampling

Two-stage cluster sampling is used in large populations, when no register is available and households cannot be visited systematically. Sampling is done in two stages:

1. Clusters, or sampling sites, within the total population are selected randomly. (Clusters may be natural groupings such as villages or, in a camp, blocks of a

few houses. Where natural groupings do not exist, artificial clusters may be defined by imposing a grid on a map of the area.)

2. Within each selected cluster, an appropriate number of individuals or households are randomly selected.

This process is applied separately to each population of interest. For instance, if a comparison is to be made between two separate, large refugee camps, the same number of clusters must be surveyed in each camp.

The larger the number of clusters, the higher is the probability of good representativeness of the population under study. In practice, physical constraints will limit the number of subjects who can be conveniently studied in a cluster; 30 subjects may often be the maximum to which easy access is possible in a community. The number of clusters to be examined is then derived by dividing the desired sample size, as determined below, by 30. It should be remembered that the sample size for clusters is larger than that for simple random samples.

Stage 1: selecting the clusters

Where feasible, the population is divided into a large number of clusters (at least 100) containing similar numbers of people using administrative, physical, or geographical boundaries. For this purpose, a map and a list of all separate identifiable units will be needed. Well defined villages of similar size are examples of possible clusters. Larger villages can be divided into two or more clusters. In a refugee camp, existing or imposed "sections" can be used. These clusters are numbered and then, using a random number table or systematic sampling, 30 are selected.

Alternatively, and more usually, the following procedure can be used:

1. Prepare a list of all existing units or zones with their estimated populations. (A unit or zone may comprise a village, camp, defined neighbourhood, or "section" within a camp.)

2. Add two more columns. In the first, record the cumulative population figures obtained by adding the population of each unit or zone to the combined population of all the preceding units or zones on the list, as shown in Table A4.5.

3. Calculate the sampling interval by dividing the total population by the number of clusters required (30). For example, if the population is 18 600, the interval will be 18 600/30 = 620.

4. Using a random number table, obtain a number between 1 and the sampling interval to define the unit or zone where the first cluster will be drawn. In the example in Table A4.5, a random number of 510 places the first cluster in unit 1.

5. Add the sampling interval repeatedly to the original random number (e.g. 510, 1130, 1750, 2370 . . .) to locate additional clusters up to the required total of 30, as shown in Table A4.5. Note that large population units are likely to be assigned more than one cluster; small units (with populations less than the sampling interval) may have none.

6. Within each unit to which more than one cluster is assigned (e.g. unit 3 in Table A4.5) further sampling is undertaken to locate the required number of

Table A4.5 Example of first stage of cluster sampling

Geographical units/zones	Estimated population	Cumulative population	Attributed numbers	Location of clusters
Unit 1	800	800	1–800	1
Unit 2	310	1110	801–1110	
Unit 3	1220	2330	1111–2330	2, 3
Unit 4	550	2880	2331–2880	4
etc ...	...	...	...	...
...	...	...	...	...
Total	18600	18600	18600	(30)

Note: See page opposite for an explanation.

clusters within the unit. Make a sketch map of the unit or zone and subdivide the whole into subunits of roughly equal population (or numbers of households), as illustrated in Fig. A4.1. Randomly select from these the required number of clusters using a random number table or by drawing numbers out of a hat.

Never change a sampling site because it is too remote or is close to a bigger and "worse affected" place that someone feels should be surveyed in preference to the randomly selected "unimportant" site.

Strictly speaking, clusters for nutritional surveys should be defined on the basis of the numbers of children aged 6–59 months. In most situations, the proportion of children is relatively uniform, and figures for the population as a whole can be used, as indicated above. However, if there are known to be wide variations in the proportion of children in the populations of different areas, the numbers of children aged 6–59 months should be estimated and used as a basis for defining clusters. On the other hand, where reliable population figures are not available, clusters may have to be defined on the basis of estimates of the numbers of households in different units or zones.

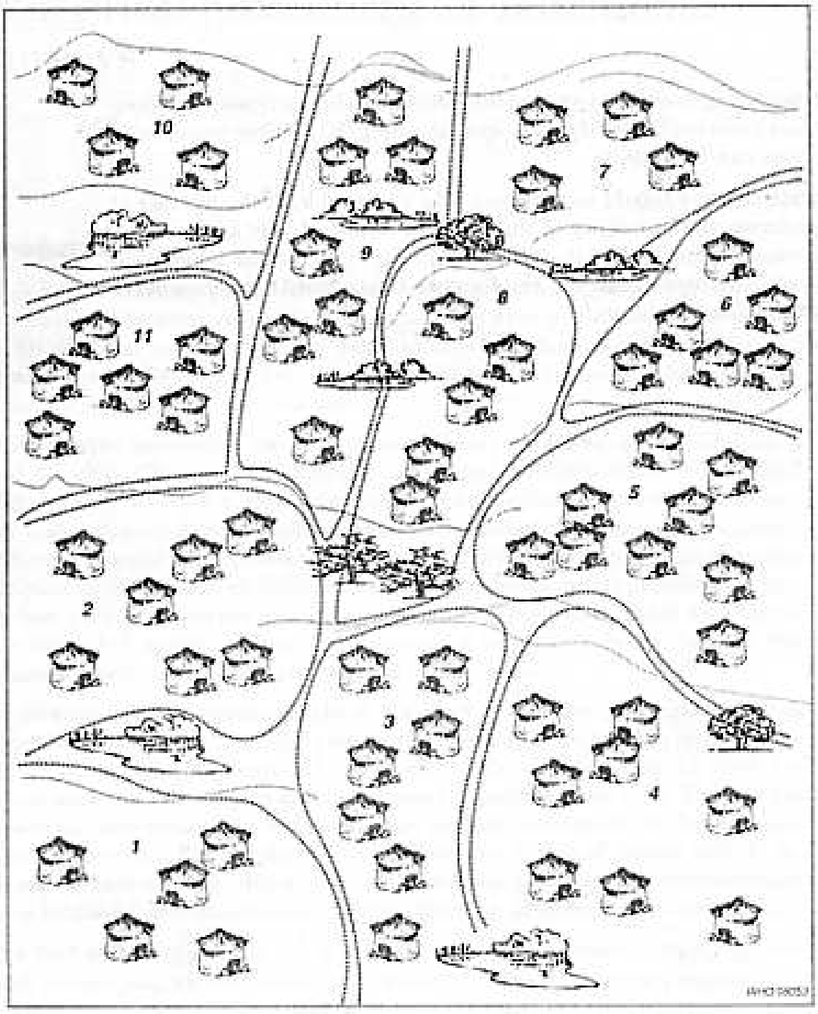
Stage 2: selecting individuals within each cluster

Once the survey team is on site, the required number of children (usually 30) can be selected by systematic sampling, as described above, if the site layout permits. Alternatively, a sketch map of the area should be drawn, the houses numbered, and households selected using a random number table. In many situations, neither of these methods is feasible and the following procedure is adopted:

1. Go to the centre of the selected unit or cluster.
2. Randomly choose a direction by spinning a pencil (pen, bottle) on the ground (or a flat surface) and noting the direction in which it points when it stops.
3. Walk in that direction from the centre to the outer perimeter of the unit or cluster, counting the number of households along this line.
4. Using a random number table, obtain a number between 1 and the number of households counted.
5. Go to the household indicated and examine all children belonging to that household (e.g. if the number is 5, go to the fifth household along the randomly chosen line).

Fig. A4.1 Division of a unit or zone for the selection of clusters

Note: In most cases a population will be divided into at least 100 clusters, of which 30 will be selected.



6. Go to the next nearest house, the one with the door nearest to the last house surveyed.

7. Continue the process until the required number of children (probably 30) has been completed.

The method to be used must be decided in advance and used consistently throughout the survey. It is important that there be no element of deliberate choice by the survey team in selecting the sample houses.

All children belonging to each selected household should be surveyed, including those in the last household (even if this means exceeding the number "required"). No substitutions can be made.

Thirty separate clusters should be surveyed if at all possible. If the number of clusters is reduced, the reliability of the estimate obtained may be poor and provide an inaccurate picture of the true nutritional status of the population being surveyed. A greater number of children per cluster does not compensate for a reduced number of clusters.[1]