



WHO/CDS/LEP/86.1 ✓  
 ORIGINAL: ENGLISH  
 1906

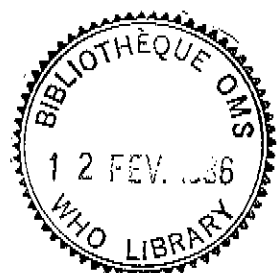
- o e e

SAMPLE SURVEYS IN LEPROSY - AN INTRODUCTORY MANUAL

by

*Data collection - methods*

T.K. Sundaresan, WHO Consultant,  
 in collaboration with Dr H. Sansarricq (former Medical Officer, Leprosy),  
 and Dr S.K. Noordeen, Chief Medical Officer, Leprosy,  
 Division of Communicable Diseases, WHO, Geneva

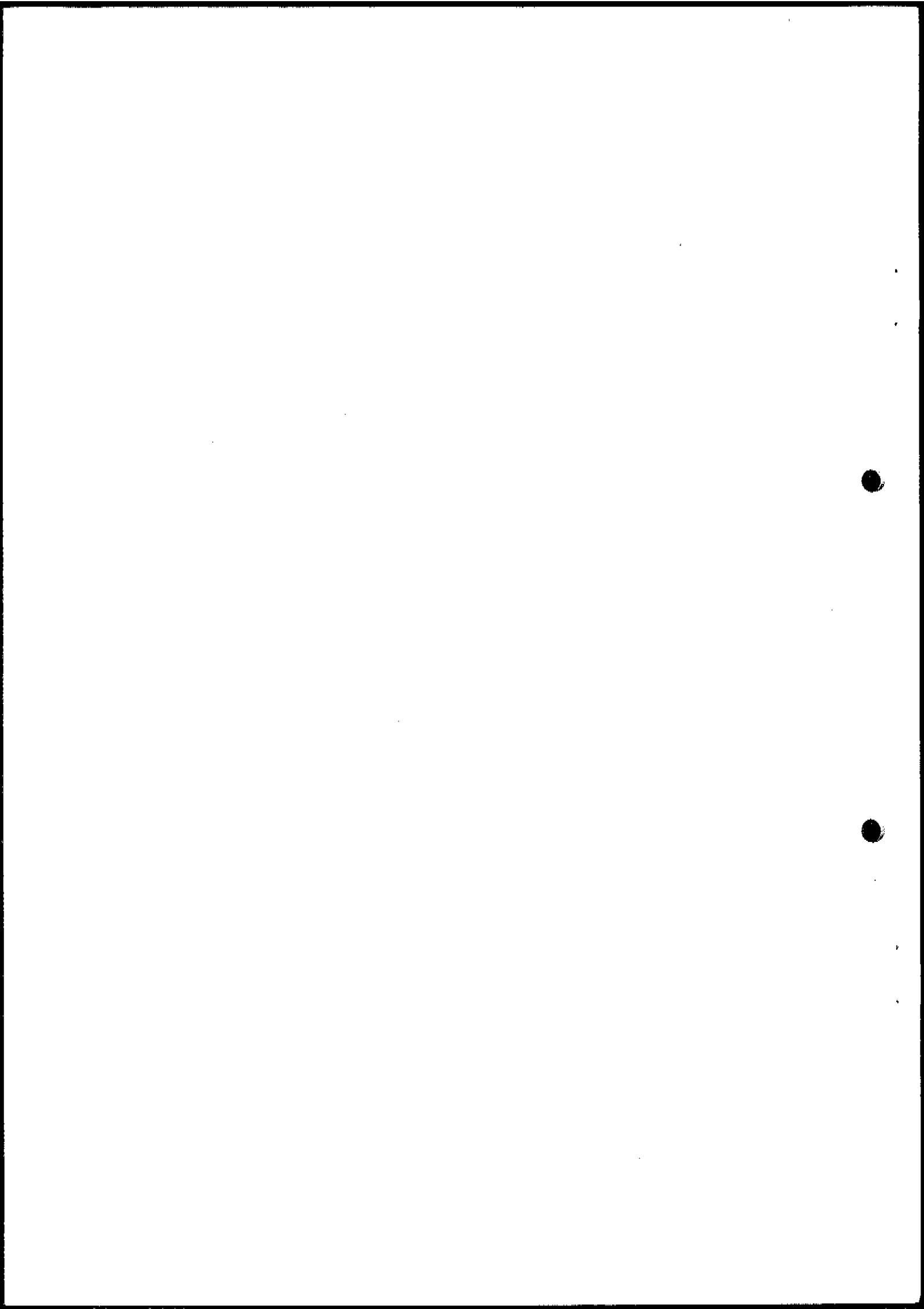


CONTENTS

	<u>Page</u>
1. Introduction . . . . .	1
2. The rationale behind sample surveys . . . . .	2
3. Some basic considerations in the choice of a sample design	2
4. Techniques of sampling . . . . .	3
5. Certain essential concepts in sampling . . . . .	5
6. The use of a table of random numbers and practical examples of sample selection . . . . .	8
7. Experience from the LAT surveys (surveys carried out by the WHO Leprosy Advisory Team). . . . .	18
8. Some important practical considerations . . . . .	19
9. Organization of a sample survey . . . . .	22
10. Procedures and criteria for diagnosis . . . . .	24
11. Non-sampling errors . . . . .	26
12. Periodic surveys for evaluation . . . . .	29
 ANNEX I: Formulae for Standard Errors - Cluster Sampling . .	 32

The issue of this document does not constitute formal publication. It should not be reviewed, abstracted, quoted or translated without the agreement of the World Health Organization. Authors alone are responsible for views expressed in signed articles.

Ce document ne constitue pas une publication. Il ne doit faire l'objet d'aucun compte rendu ou résumé ni d'aucune citation ou traduction sans l'autorisation de l'Organisation mondiale de la Santé. Les opinions exprimées dans les articles signés n'engagent que leurs auteurs.



## 1. INTRODUCTION

The following is a quotation from "A Guide to Leprosy Control" (WHO, 1980)<sup>1</sup>:

"The data on the prevalence of leprosy in most countries are unreliable, because neither case-finding nor reporting is at a sufficiently high level... The surveys of the WHO Leprosy Advisory Team showed that, even in countries with a satisfactory case-finding programme, new cases could still be found amounting to as many as three-quarters of the number registered...".

Many countries have given high priority to the control of leprosy. It is evident that: 1) reliable information should be obtained on the magnitude of the leprosy problem in all its important facets, viz. infectious forms, disabilities, etc., to enable the formulation of a national plan for leprosy control and appropriate allocation of resources, and 2) an information system should be established for the periodic evaluation and monitoring of any intervention programme that may have been initiated.

The WHO Expert Committee on Leprosy in its fifth report (Technical Report Series No. 607, 1977)<sup>2</sup> concluded that:

"Incidence, or the number of new cases occurring during a given period (generally one year) in relation to the population, is the only index for measuring the efficacy of the measures taken, i.e. the reduction of transmission.

The continued occurrence of new infections in children is of great epidemiological significance as it indicates continuing transmission of the disease. In such cases, it is useful to separate the incidence rates for children and adults.

Changes in prevalence are only indirectly indicative of the impact of a programme on the epidemiological situation. However, the total number of patients to be treated needs to be known for planning and organizational purposes."

At present there is a need in many control programmes for evaluation of the existing epidemiological situation. The Committee recommended that such an evaluation be undertaken by means of sample surveys that include an assessment of the number of still undetected cases of leprosy.

The Committee further recommended that the information collected for evaluation should be simple and capable of leading to decisions and should be of a kind that can be collected directly by the staff responsible for health activities at the peripheral level, who are not necessarily specialized in leprosy.

Investigations using sampling methods have an appeal because of the resulting economy in resources and because information can be obtained more rapidly. However, in order to undertake a sampling investigation it is important to have a knowledge of the basic concepts of sampling and the way to interpret the results derived from the sample.

The purpose of this guide is to introduce leprosy workers to these concepts and help them to undertake sample surveys routinely in limited areas:

- (i) to estimate the prevalence of leprosy in the areas, the case-load and the proportion of different types of leprosy;
- (ii) to assess the impact of a programme in terms of prevalence and
- (iii) to evaluate certain operational aspects of leprosy control programmes.

For any large undertaking such as a national survey, it is best to associate a statistician. The application of more elaborate statistical techniques may be indicated in such circumstances, and the cost of a statistical consultancy will be more than offset by the overall saving effected.

Sample surveys to indicate case-load are indicated only when the prevalence rates are at least of the order of 10 per 1000. If the expected rates are lower the sample size required would be quite large. In such circumstances methods of epidemiological surveillance would be more appropriate. Since leprosy is not uniformly spread in any country, prevalence could be limited to areas of high endemicity. While in general, sample surveys indicate the limits within which the true prevalence in the population should be, sometimes when the prevalence is low sample surveys can be useful for indicating an upper limit for the population value.

Although the illustrations in this manual are for prevalence surveys the basic principles of sampling are of general application.

## 2. THE RATIONALE BEHIND SAMPLE SURVEYS

In situations where no information exists on the extent and magnitude of the leprosy problem in countries, but where there are grounds for believing that leprosy is a problem of considerable magnitude, it is most useful to conduct a sample survey to assess more precisely the magnitude of the problem. Such information is necessary for planning purposes. Complete examination of the whole population has to be ruled out as it will be unfeasible and far too expensive.

There is yet another and perhaps even more important use for sample surveys. This is for monitoring the progress and impact of a control programme that is in operation. Sample surveys conducted periodically in an area where some control operations are going on will provide a relatively inexpensive means of measuring the progress, highlighting operational deficiencies, and where indicated suggesting alternative strategies for producing greater effectiveness.

Sample surveys on a national scale are useful for providing planners with adequate information to be used for appropriate allocation of resources. The periodical surveys provide useful information for monitoring the programme. In either case a sample survey is not an end in itself but only a means of obtaining information vital for certain objectives. Consequently the costs of, and other resources used in sample surveys should be minimal and should be a small fraction of the leprosy budget.

## 3. SOME BASIC CONSIDERATIONS IN THE CHOICE OF A SAMPLE DESIGN

Sampling investigations provide the best estimates of the true values of the variables or characteristics (e.g. prevalence rates) in the general population from which the samples are drawn. However, they can only be "estimates" and seldom be identical with the unknown "true" value in the population. Nevertheless it is possible to say, with a high degree of confidence, that the true population value is within a range. For example, if it is found from the sample that 2% have a disease, we might, depending on the sample design, be able to declare (with 95% confidence) that the true population prevalence is between 1.6% and 2.4%. By increasing the size of the sample, the range (e.g. 1.6 to 2.4) can be made as narrow as may be desirable. Technically this range is known as the "confidence interval" and reflects the "precision" of the estimate obtained from the sample. Thus in the above example if the sample had been very large we might have been able to state that the "true population value is between 1.8 and 2.2". In this case the second estimate, obtained with the larger sample, is more precise than the first.

While a high degree of precision is desirable on scientific grounds, it may prove unduly expensive. For example to double the precision one may need to quadruple the sample size. Thus it is important to consider carefully, and well in advance, the use to be made of the result of the sample survey and arrive at a minimum acceptable precision.

Closely allied to the above consideration is the need to state the objectives of the sampling investigation in as much detail as possible, specifying the categories of persons to be surveyed.

To be more specific, let us say that there exists only very scanty information on the prevalence of the disease and the planners require an estimate of the national case-load to effect necessary budgetary appropriation. Very often one is confronted with a situation of scarce resources and it is adequate to have an idea of the order of magnitude of the problem; e.g. whether the prevalence is around the 20 per 1000 level or 10 per 1000 level. In such a situation a rapid survey with a sample of moderate size may provide the information needed at that stage. On the other hand, in surveys of the "before" and "after" type where it is important to detect significant improvements as a result of intervention measures, different orders of precision may dictate the choice of sample size and design.

Leprosy has many facets drawing different degrees of public attention and stimulating governmental action. There are deformities attributable to the disease, active cases with manifest lesions and, at the other extreme, tiny and not easily noticeable patches that can be detected only after complete examination. While information on all these aspects will be obtained in a sample survey, the order of precision required for each one of these variables could be different as action programmes for the three variables could have widely varying budgetary implications.

The variation of sample values around the true population value is referred to as "sampling error" and can be controlled by an efficient sample design. On the other hand there are a number of other factors that could give rise to divergence from the true population value. For example, non-coverage of part of the selected sample population, incomplete examination of the individuals, errors in diagnosis etc. can all contribute to estimation errors. They are referred to as non-sampling errors, and because of their importance, especially in leprosy surveys, they are dealt with in a separate section.

In this connexion a distinction must be made between "precision" and "accuracy". The concept of "precision" is dealt with in section 5. It may be mentioned at this stage that precision refers to the closeness of the value obtained from the sample to the true population value and can generally be controlled by suitable sample size. Accuracy on the other hand, takes into account also the non-sampling errors. It is the total of both the sampling and non-sampling errors. Above all the sample must be truly representative of the population from which it is drawn. All persons in the sample must have the possibility of being represented in the sample.

#### 4. TECHNIQUES OF SAMPLING

A variety of sampling techniques are currently available, their main objective being to secure maximum efficiency and operational convenience. It may be relevant to make reference at this stage to a few of these techniques that have particular applicability to leprosy surveys:

One of the first pre-requisites of a sampling investigation is the existence or construction of a "sampling frame". Examples of sampling frames are: census tracts, lists of households, lists of villages with population in each. Usually such lists are not up-to-date. However, such frames, if they are not too old, can still be used at least as a first stage in the selection of a sample. This aspect is dealt with in more detail in section 9 under "Organization of a sample survey". It is mentioned here because the construction of a suitable sampling frame is the first step in the sampling process and the nature of the frame is relevant to considerations in the choice of the most appropriate technique of sampling.

#### 4.1 Simple random sampling

This is the name given to a design "in which the sampling units are independently selected with equal probability, the sampling being without replacement and without restrictions".<sup>3</sup> Thus, if the population is 1 000 000 and it is desired to examine a representative sample of 10 000, this design ensures that every one of the million population has the same chance of becoming a member of the sample selected. In practice this would mean first listing all the million individuals in the population and choosing 10 000 out of them at random - by adopting methods used in lotteries or by the use of random sampling numbers (see Section 6). The procedure is however extremely cumbersome and unfeasible in practice for population surveys.

#### 4.2 Stratified sampling (see also Section 8)

Sometimes information is required on the prevalence or incidence of the disease for different segments or "strata" of the population. This may be necessary for administrative purposes or for a better understanding of the epidemiology of the disease. The characteristics (such as urban or rural, endemicity level, hilly regions, special population groups, etc) on which stratification is desired should first be decided and the population in each stratum delineated. From each stratum a separate sample is drawn. The percentage of the population that constitutes the sample need not be the same from one stratum to another. Stratification may also be resorted to for increasing the efficiency of the sampling procedure. For example, if the population is stratified according to different endemic levels (known or suspected), it pays to take different sample sizes from different strata and derive an estimate for the total population by combining the estimates from the individual strata after weighting them suitably.

#### 4.3 Cluster sampling

Any group of individuals, e.g. members of a household, people living in a village, etc, can constitute a cluster. Once a cluster is defined, such as a household, then a sample of households is selected from a complete list of households in the community and all individuals within the selected households examined. If the village is defined as a cluster, then a sample of villages is selected from a list of villages and all the population in the selected village examined. When the clusters are "naturally" defined, e.g. household, village, etc., the cluster sizes, i.e. the number of individuals in each of the clusters, would normally not be equal. However, for operational and statistical purposes it may be convenient to create clusters of approximately equal size by grouping, e.g. small villages could be grouped to constitute one cluster with a total population equal to that of a big village.

When a communicable disease like leprosy is being investigated, it is likely that the individuals in one cluster are more similar to one another as regards the risk of the disease than the individuals in another cluster. Such similarity within clusters is called "intra-cluster correlation". When such correlation exists, the estimates from the sample will have a wider variation and consequently be less precise. (See next section on "Certain essential concepts in sampling".) The sample size, computed theoretically on the basis of a simple random sample of individuals, will have to be increased to allow for such intra-cluster correlation.

The cluster sample design is the one most widely used for estimating disease characteristics, such as prevalence. Any increase in sample size that may be necessary is usually offset by the operational convenience. In any case, for leprosy surveys, the random selection of individuals from a population can be ruled out as unfeasible.

#### 4.4 Multi-stage sampling

When a large population, such as the population of the country is sampled, the sampling is usually done by stages. For example, if the country is divided into districts, a sample of districts may first be selected. From within each selected

district a sample of subdistricts may be selected. From within each selected subdistrict a sample of villages may be selected. Such a design is called multi-stage sampling, the selection of districts, subdistricts, villages, etc., representing the different stages. At each stage the probability of selecting a subdivision may be made proportional to its population; thus a large subdistrict will have relatively greater chance of being selected than a smaller one.

In all cases, it should be borne in mind that the results of analysis (i.e. inferences from the sample) apply in all cases only to the universe from which the sample was taken. For example, if it is decided to exclude certain population groups, such as nomads, from the sample, the results from the sample are only applicable to the general population, excluding the nomads.

## 5. CERTAIN ESSENTIAL CONCEPTS IN SAMPLING

### 5.1 Standard error

However good the coverage, however representative the sample and however perfect the diagnostic tools, the fact remains that in sampling investigations one is examining a part of the population and the result of the sample survey (such as prevalence rate) can seldom be identical with the result of examining the entire population. On the other hand, if a number of samples are taken, the sample values have a tendency to cluster around the true value in the population, sometimes a little less and sometimes a little more. This variation around the true population value is measurable and is expressed by the standard error. The following formulae are for the Simple Random Sampling design and are presented for their simplicity. Formulae for cluster sampling are given in the annex. If:

- (1)  $n$  is the sample size;
- (2)  $p\%$  is the observed prevalence rate of leprosy;
- (3)  $q\%$  ( $= 100 - p$ ) is the percentage without leprosy;

then the standard error (s.e) of  $p$  is  $\sqrt{\frac{pq}{n}}$ \*

We also know from statistical theory that the true prevalence in the population ( $p\%$ ) is most likely to be within  $p \pm 1.96$  (s.e). The probability that the true prevalence will be outside these limits is 5%.

When we say "most likely" in this context we mean that, when we generalize from a sample to the population as above, i.e. when we say that the true prevalence is between  $p \pm 1.96$  s.e, then in the long run we will be correct on 95% of the occasions. In practice we may say we are 95% sure that the true prevalence in the population is between  $p \pm 1.96$  s.e. For this reason the two values  $p + 1.96$  s.e and  $p - 1.96$  s.e are the 95% "confidence limits". The confidence interval can be wide or narrow depending on the value of the standard error, which therefore determines the precision of the estimate.

---

\*The exact formula is  $s.e. = \sqrt{1-f} \sqrt{\frac{pq}{n}}$  where  $f =$  sampling fraction which is  $n/N$

where  $n$  is sample size and  $N$  is the population. However, for most purposes,  $n/N$  is small and the multiplication factor  $\sqrt{1-f}$  will be very nearly equal to 1 and can be ignored.

As an example, suppose there are two samples of different sizes both giving a sample estimate of 2% prevalence. In one case we may be able to say that the true prevalence is between 2.5% and 1.5% and in the other case (the larger sample) we might be in a position to state that the true prevalence is between 2.2% and 1.8%. In the latter case the estimate is said to be more "precise" than in the former.

Obviously an estimate that gives a very wide confidence interval is not of much use. On the other hand if one wants the interval to be very narrow, i.e. if one wants an extremely precise estimate, it may mean an enormous increase in sample size and consequently prove expensive. A balance has to be drawn and one must arrive a priori, at an acceptable level of precision. The idea will be found to be further elaborated in the next section on sample size.

### 5.2 The coefficient of variation

It will be seen from the above that the variable determining precision is the standard error. Now we may have two surveys with different prevalence rates but with about the same standard error. For example one survey may give a prevalence rate of 2% with a standard error of 0.2% and another with a prevalence rate of 4% but again with the same standard error, viz 0.2%. The 95% confidence intervals in the two cases would be 1.6 to 2.4, and 3.6 to 4.4. It is apparent that in the second case the estimation appears more precise because in the first case the confidence limits are 2%  $\pm$  20% of 2, and in the second case they are 4%  $\pm$  10% of 4. Thus with varying degrees of prevalence it is useful to think of an index which relates the standard error to the estimate of prevalence. Such an index is provided by the coefficient of variation, which is defined as the ratio of the standard error to the estimate expressed as a percentage.

### 5.3 Coefficient of variation (c.v) = $\frac{s.e}{p} \times 100$

Thus in the above example the prevalence of 2% is associated with a coefficient of variation of  $\frac{0.2}{2} \times 100$  or 10%, and in the second case the prevalence of 4% is associated with a coefficient of variation of  $\frac{0.2}{4} \times 100$  or 5%

### 5.4 Sample size

When a sample survey is being planned the first question that comes to mind is 'how big should the sample be?'. In order to answer this question satisfactorily, indications are required on the following points:

1. What is the expected magnitude of the variable being estimated? For example, if it is a survey to estimate the prevalence we should have an idea of the order of magnitude of the prevalence, i.e. whether it is around 1 or 2% or between 10 and 15% etc. Only an approximate range of values is required at this stage to serve as a starting point. Usually this information can be based on past records, well-informed guesses etc.
2. The second point to be indicated is the magnitude of tolerable variations. For example, is it enough to have a coefficient of variation of 20% or is this required to be smaller etc. To provide a sound answer to this question it is important to have in mind the use that is going to be made of the survey result. For planning broad strategies on a national scale a high degree of precision may not be essential whereas to evaluate programme impacts in smaller communities a higher degree of precision may be called for. It must be

remembered that the primary objective of a sample survey is to obtain an estimate acceptably close to the true value. It is not a case-finding exercise. Asking for a high degree of precision may result in an enormous sample and the consequent expenses may not be warranted by the purpose of the survey itself.

3. The third point to be considered is the sampling technique that is operationally feasible. For example, is Cluster sampling operational?; if so how big should the clusters be? Whether stratification is indicated either for administrative reasons or for statistical efficiency etc.

Although simple random sampling is rarely resorted to in population surveys on leprosy, it is still the basis for most other techniques of sampling. The usual practice is to compute a sample size assuming simple random sampling and then multiply the size so derived by a suitable factor.

In the case of a simple random sample the general formula for determining the sample size  $n$  is :

$$n = (z^2 pq)/d^2 \text{ where:}$$

$p$  = the expected percentage with the disease. As mentioned before this is an approximate figure based on past experience. If by past experience it is suspected that the prevalence is between two values, say 2% and 4% the lower value should be taken. This is because the sample size determined with a lower value will be sufficiently large and adequate to cover sample estimates if they turn out to be higher; on the other hand if the larger value is chosen for determining sample size, and if the actual survey result yields a lower figure, the estimate will not be precise enough.

$q = 100 - p$  (i.e. the percentage without the disease).

$z$  = a value that could be obtained from statistical tables to correspond to the probability level of the confidence limits (95%, 99% etc, - see section 5). For most practical purposes the 95% probability level is adequate and the corresponding value of  $z$  is 1.96.

$d$  = the expression for the precision required. It is the difference between the actual estimate and its confidence limit (or half-confidence interval). In practice it is usual to specify that this difference should not exceed  $x\%$  (say 20%) of the estimate  $p$ ; in which case  $d$  will be taken as  $xp/100$  (0.2 $p$  in the example).

The following example makes the use of the formula clearer.

Suppose it is desired to estimate the prevalence of all forms of leprosy in a community; from previous experience it is suspected that the prevalence is at least 2% and not more than 4% and further the estimate is required with a precision which would lead to the 95% confidence limit, not to exceed 20% of whatever prevalence value may be obtained when the actual survey is done; this means that if the survey when done gives an estimate of 3%, the confidence limits should be within  $3 \pm 20\%$  of 3 (=0.6) - i.e. we should be able to state that the true population prevalence is between 2.4% and 3.6%.

To estimate the sample size we would take the lower of the two suspected estimates (viz 2%). The specified precision should lead in this case to a value of  $d = 20\%$  of 2 (i.e. 0.4).

Substituting these values in the formula:

$$\begin{aligned} n &= 1.96^2 \cdot 2 \cdot (100 - 2)/0.4^2 \\ &= 4705.96 \text{ or } 4706 \text{ individuals.} \end{aligned}$$

Thus we would need to examine 4706 individuals by simple random sampling. Since in practice we would be using cluster sampling this figure has to be multiplied by an appropriate factor reflecting the degree of homogeneity to be expected within each cluster as regards the risk of the disease.

If in the above example we are in a position to accept a lower order of precision, e.g. we could accept a value of  $d = 30\%$  of  $p$  ( $= 2\%$  in this case)  $n$  would be  $= 1.96^2 \cdot 2 \cdot (100 - 2) / 0.6^2 = 2091.54$ . Thus a simple random sample of about 2100 would have been adequate. In the first case we would be able to say that the estimate is  $2\% \pm 0.4\%$ ; in the second  $2\% \pm 0.6\%$ . It will be noteworthy that the small gain in precision (0.6 to 0.4) requires more than doubling the sample size. In general, for increasing the precision  $y$  times we need to increase the sample size  $y^2$  times. This underlines the importance of deciding a priori on the minimum acceptable precision.

## 6. THE USE OF A TABLE OF RANDOM NUMBERS AND PRACTICAL EXAMPLES OF SAMPLE SELECTION

The process of choosing at random is equivalent to choosing numbers as in a lottery. There are practical problems in organizing a "lottery" each time a random selection has to be made. However, random numbers have been selected many times and there are several published tables which could conveniently be used. A specimen page from one such publication is reproduced here<sup>4</sup>.

In order to use the table it must be realised that each number is (single-digit) independent. The printing in blocks of five (major blocks of 5 and smaller blocks of 2) has been arranged solely to relieve eye-strain in reading. Thus it will be observed that there are fifty rows and fifty columns in the attached table, comprising altogether 2500 independently selected single digit numbers (i.e. numbers from 0 to 9). The process of random selection may be illustrated by the following hypothetical example.

Suppose there are twenty-five villages (of approximately equal size) and it is desired to select 6 of these at random. Let us imagine the villages to be numbered 01, 02, 03, ... 25. We may now start from any row and column in the attached table and look for two-digit numbers. Let us suppose we start with row 14 and column 16. Since the interest is in two digit numbers, we have to read columns 16 and 17 together. In this case the first number is 31. Since the total number of villages is only 25 this number 31 has to be rejected. The next two numbers looking downwards are 46 and 73 which have to be rejected for the same reason. The next one is 19 and hence village 19 is selected. Proceeding in the same way it will be noticed that the next five numbers are greater than 25 and have to be rejected, the next admissible number being 10. Hence village 10 is selected. The next number is "07" which means village 7 is also selected. The next four numbers are greater than 25. The fifth is 24 and hence village 24 is also selected. The next six numbers are greater than 25. The seventh number is 10 which is admissible but village 10 has already been selected; hence we have to look for other numbers. Proceeding in this way reading downwards the next two admissible numbers will be seen to be 14 and 15, indicating that the villages bearing these numbers should be selected. Thus we have selected at random the six villages bearing numbers: 19, 10, 07, 24, 14 and 15. Putting the numbers in sequence we could say that villages 7, 10, 14, 15, 19 and 24 have been selected.

TABLE XXXIII. RANDOM NUMBERS (IV)<sup>4</sup>

10 27 53 96 23	71 50 54 36 23	54 31 04 82 98	04 14 12 15 09	26 78 25 47 47
28 41 59 61 88	64 85 27 20 18	83 36 36 05 56	39 71 65 09 62	94 76 62 11 89
34 21 42 57 02	59 19 18 97 48	80 30 03 30 98	05 24 67 70 07	84 97 50 87 46
61 81 77 23 23	82 82 11 54 08	53 28 70 58 96	44 07 39 55 43	42 34 43 39 28
61 15 18 13 54	16 86 20 26 88	90 74 80 55 09	14 53 90 51 17	52 01 63 01 59
91 76 21 64 64	44 91 13 32 97	75 31 62 66 54	84 80 32 75 77	56 08 25 70 29
00 97 79 08 06	37 30 28 59 85	53 56 68 53 40	01 74 39 59 73	30 19 99 85 48
36 46 18 34 94	75 20 80 27 77	78 91 69 16 00	08 43 18 73 68	67 69 61 34 25
88 98 99 60 50	65 95 79 42 94	93 62 40 89 96	43 56 47 71 66	46 76 29 67 02
04 37 59 87 21	05 02 03 24 17	47 97 81 56 51	92 34 86 01 82	55 51 33 12 91
63 62 06 34 41	94 21 78 55 09	72 76 45 16 94	29 95 81 83 83	79 88 01 97 30
78 47 23 53 90	34 41 92 45 71	09 23 70 70 07	12 38 92 79 43	14 85 11 47 23
87 68 62 15 43	53 14 36 59 25	54 47 33 70 15	59 24 48 40 35	50 03 42 99 36
47 60 92 10 77	88 59 53 11 52	66 25 69 07 04	48 68 64 71 06	61 65 70 22 12
56 88 87 59 41	65 28 04 67 53	95 79 88 37 31	50 41 06 94 76	81 83 17 16 33
02 57 45 86 67	73 43 07 34 48	44 26 87 93 29	77 09 61 67 84	06 69 44 77 75
31 54 14 13 17	48 62 11 90 60	68 12 93 64 28	46 24 79 16 76	14 60 25 51 01
28 50 16 43 36	28 97 85 58 99	67 22 52 76 23	24 70 36 54 54	59 28 61 71 96
63 29 62 66 50	02 63 45 52 38	67 63 47 54 75	83 24 78 43 20	92 63 13 47 48
45 65 58 26 51	76 96 59 38 72	86 57 45 71 46	44 67 76 14 55	44 88 01 62 12
39 65 36 63 70	77 45 85 50 51	74 13 39 35 22	30 53 36 02 95	49 34 88 73 61
73 71 98 16 04	29 18 94 51 23	76 51 94 84 86	79 93 96 38 63	08 58 25 58 94
72 20 56 20 11	72 65 71 08 86	79 57 95 13 91	97 48 72 66 48	09 71 17 24 89
75 17 26 99 76	89 37 20 70 01	77 31 61 95 46	26 97 05 73 51	53 33 18 72 87
37 48 60 82 29	81 30 15 39 14	48 38 75 93 29	06 87 37 78 48	45 56 00 84 47
68 08 02 80 72	83 71 46 30 49	89 17 95 88 29	02 39 56 03 46	97 74 06 56 17
14 23 98 61 67	70 52 85 01 50	01 84 02 78 43	10 62 98 19 41	18 83 99 47 99
49 08 96 21 44	25 27 99 41 28	07 41 08 34 66	19 42 74 39 91	41 96 53 78 72
78 37 06 08 43	63 61 62 42 29	39 68 95 10 96	09 24 23 00 62	56 12 80 73 16
37 21 34 17 68	68 96 83 23 56	32 84 60 15 31	44 73 67 34 77	91 15 79 74 58
14 29 09 34 04	87 83 07 55 07	76 58 30 83 64	87 29 25 58 84	86 50 60 00 25
58 43 28 06 36	49 52 83 51 14	47 56 91 29 34	05 87 31 06 95	12 45 57 09 09
10 43 67 29 70	80 62 80 03 42	10 80 21 38 84	90 56 35 03 09	43 12 74 49 14
44 38 88 39 54	86 97 37 44 22	00 95 01 31 76	17 16 29 56 63	38 78 94 49 81
90 69 59 19 51	85 39 52 85 13	07 28 37 07 61	11 16 36 27 03	78 86 72 04 95
41 47 10 25 62	97 05 31 03 61	20 26 36 31 62	68 69 86 95 44	84 95 48 46 45
91 94 14 63 19	75 89 11 47 11	31 56 34 19 09	79 57 92 36 59	14 93 87 81 40
80 06 54 18 66	09 18 94 06 19	98 40 07 17 81	22 45 44 84 11	24 62 20 42 31
67 72 77 63 48	84 08 31 55 58	24 33 45 77 58	80 45 67 93 82	75 70 16 08 24
59 40 24 13 27	79 26 88 86 30	01 31 60 10 39	53 58 47 70 93	85 81 56 39 38
05 90 35 89 95	01 61 16 96 94	50 78 13 69 36	37 68 53 37 31	71 26 35 03 71
44 43 80 69 98	46 68 05 14 82	90 78 50 05 62	77 79 13 57 44	59 60 10 39 66
61 81 31 96 82	00 57 25 60 59	46 72 60 18 77	55 66 12 62 11	08 99 55 64 57
42 88 07 10 05	24 98 65 63 21	47 21 61 88 32	27 80 30 21 60	10 92 35 36 12
77 94 30 05 39	28 10 99 00 27	12 73 73 99 12	49 99 57 94 82	96 88 57 17 91
78 83 19 76 16	94 11 68 84 26	23 54 20 86 85	23 86 66 99 07	36 37 34 92 09
87 76 59 61 81	43 63 64 61 61	65 76 36 95 90	18 48 27 45 68	27 23 65 30 72
91 43 05 96 47	55 78 99 95 24	37 55 85 78 78	01 48 41 19 10	35 19 54 07 73
84 97 77 72 73	09 62 06 65 72	87 12 49 03 60	41 15 20 76 27	50 47 02 29 16
87 41 60 76 83	44 88 96 07 80	83 05 83 38 96	73 70 66 81 90	30 56 10 48 59

The procedure outlined above ignores the size of the villages; i.e. each of the 25 villages, irrespective of its size, has an equal chance (6/25) of being selected. The procedure is acceptable if the villages are of approximately equal size. If however, they are very unequal it is preferable to make the selection in such a way that the probability of selecting a village is proportional to its size. The method of selection is best illustrated by the following example.

Suppose the population of the 25 villages is as indicated in the following table:

VILLAGE NO.	POPULATION	CUM. TOTAL	NUMBERS IN VILLAGE
1	150	150	0001-0150
2	270	420	0151-0420
3	220	640	0421-0640
4	380	1020	0641-1020
5	125	1145	1021-1145
6	420	1565	1146-1565
7	318	1883	1566-1883
8	118	2001	1884-2001
9	520	2521	2002-2521
10	126	2547	2522-2547
11	315	2962	2548-2962
12	278	3240	2963-3240
13	514	3754	3241-3754
14	245	3999	3755-3999
15	337	4336	4000-4336
16	267	4603	4337-4603
17	188	4791	4604-4791
18	265	5056	4792-5056
19	484	5540	5057-5540
20	196	5736	5541-5736
21	284	6020	5737-6020
22	503	6523	6021-6523
23	96	6619	6524-6619
24	121	6470	6620-6740
25	276	7016	6741-7016
	7016		

The total population of all the twenty-five villages is 7016. If we imagine that the individuals, starting from village 1 are numbers from 0001-7016, the process consists of selecting six individuals (i.e. six random numbers between 0001 and 7016), identifying the villages where these individuals are located and regarding these identified villages as the ones selected. The operation is facilitated by constructing a third column in the above table called 'cumulative totals'. The successive figures in this column are population of village 1, population of villages 1 and 2, population of villages 1,2 and 3 etc., the last figure entered against village 25 is the total of all the villages 1, 2 etc. up to and including the population of village 25. A fourth column is constructed to readily indicate the numbers that would point to the corresponding village.

We could now choose six random numbers from 0001 and 7016. Since we are interested in choosing six separate villages, if two numbers point to the same village the second number has to be rejected.

As an example we could start with row 36 and col. 31 of the attached table of random numbers and consider all four-digit numbers from this starting point. The first number is seen to be 6869. If we look at the cumulative totals it will be seen that numbers 6741 to 7016 (and thus the selected number 6869) are located in village 25, which is consequently chosen. The next number 7957 is clearly in excess of 7016 and has to be rejected. The next number is 2245. From the table of cumulative totals it will be seen that numbers 2002 to 2521 (which includes 2245) are located in village 9, which shall therefore be the second village selected.

The third number, 8045, exceeds 7016 and has to be rejected. The next one is 5348, which can be seen to be in village 19. Thus the third village selected is village 19. The next number is 3768, located in village 14. The next is 7779 to be rejected and the next 5566 is seen to be located in village 20. The next number, 2780, is in village

11. Thus the six villages selected in order are 19, 25, 9, 14, 20 and 11. Putting the numbers in sequence, the villages selected are 9, 11, 14, 19, 20 and 25. (If after choosing a village, e.g. 14, we had again hit a number such as 3801 pointing to the same village, such a number has to be rejected).

In practice, for reasons explained in section 4.3 it will be most convenient to use a cluster sampling design for estimating the prevalence of leprosy. The practical procedure is best explained by the following hypothetical example:

Suppose it is desired to estimate the prevalence of all forms of leprosy in an area which consists of 10 sub-districts with population as follows:

Sub-district No.	Estimated population
1	15 678
2	10 879
3	7 670
4	3 654
5	8 987
6	14 356
7	7 987
8	11 889
9	12 678
10	5 889
	-----
Total	99 667

The determination of an adequate sample size will be dealt with in another section. Let us assume for the present that the optimal design is to choose as our sample 20 clusters of size 300 each, i.e. a total population of 6000. It is further assumed that one estimate for the entire area is all that is required and that prevalence estimates for individual sub-districts are not needed.

In this case the problem resolves to one of selecting 20 locations at random among the 10 sub-districts. It could turn out that some of the sub-districts may not appear in the sample and that more than one cluster will be located among some sub-districts.

The first stage in the process of selection is to (1) list the sub-districts; (2) note the population against each; (3) construct the cumulative totals as in the previous example, and (4) indicate the numbers falling with each sub-district. This is illustrated in the following table:

Sub-district number	Estimated population	Cumulative totals	Numbers pointing to the sub-district
1	15 678	15 678	00 001-15 678
2	10 879	26 557	15 679-26 557
3	7 670	34 227	26 558-34 227
4	3 654	37 881	34 228-37 881
5	8 987	46 868	37 882-46 868
6	14 356	61 224	46 869-61 224
7	7 987	69 211	61 225-69 211
8	11 889	81 100	69 212-81 100
9	12 678	93 778	81 101-93 778
10	<u>5 889</u>	99 667	93 779-99 667
Total	99 667		

The next step is to choose 20 independent five digit random numbers between 00001 and 99667. Looking at the attached table of random numbers and starting at random, say row 9, column 19, we find the first number is 949362 and this number will be found to be located in sub-district 10. The next numbers and their location are indicated in the following table:

<u>Random number</u>	<u>Sub-district number</u>
17479	2
09727	1
71092	8
25544	2
52662	6
53957	6
48442	6
60681	6
99672	number to be rejected being more than 99667
38676	5
72865	8
51741	6
23765	2
86795	9
01773	1
14483	1
49891	6
50018	6
28074	3
29396	3
56328	6

Thus twenty numbers have been chosen, and on this basis it would seem that sub-districts 1, 2, 3, 5, 6, 8 and 9 have been selected, and the number of clusters to be examined in each are as follows:

<u>Sub-district</u>	<u>No. of clusters to be examined</u>
1	3
2	3
3	2
5	1
6	8
8	2
9	1
<hr/>	
TOTAL	20

What has been done is to select, at the first stage, sub-districts and allocate the twenty clusters to be examined among these. The selection and allocation has been done in such a way that the probability of selecting a sub-district and the number of clusters to be allocated in the selected ones are made proportional to the population of the sub-districts.

If a sub-district selected is large and divided into smaller sub-divisions, the above process may be repeated for choosing the sub-divisions (and the number of clusters in each). This will be the second stage and the process may be repeated to cover a third, fourth etc. stage selection till a sub-division is reached with a reasonably small population. So far, for the selection at every stage, the estimated populations have been used. This may be based on an old census data or other estimations. However, once a stage is reached with sufficiently small populations (e.g. village groups) a census must be undertaken at this level, to get a more up-to-date and precise estimate of the population.

As an example, let us take sub-district 3. Two clusters are to be selected from this sub-division. Let us say that this sub-district consists of 200 villages, some small and some large.

It is better at this stage to group small adjoining villages into one so that eventually we have 25 village groups of approximately equal size (viz. population of 300).

The next step is to choose two random numbers between 01 and 25 and completely examine the village-groups thus indicated.

#### 6.1 Villages with large populations

In the above practical example we have been assuming that the villages are small with a population of 500 or less. It could happen in many situations that some or all of the villages have larger populations. If the cluster size has been fixed at 500, the question arises as to how the 500 individuals should be selected.

If the village can be conveniently divided into small sections by well defined boundaries such as a canal, field etc. it will be simplest to choose an appropriate number of such sections at random in the manner described above for the choice of villages.

Alternatively, the following procedure is suggested. Suppose the population of the village is 2000 living in 300 houses. If the houses are not already numbered they should be numbered from 001 to 300.

From a table of random numbers choose a number between 001 and 300. Suppose number 056 is chosen. Then this will be the starting point. All individuals in houses, 056, 057, 058 etc. will be examined until the required number of 500 is reached. All members in the last of the houses falling into the sample should be examined. For example house numbers 056-155 may contain 497 individuals and house number 156 eight individuals. The principle is to examine every one in the 156th house. The cluster will then consist of 505 individuals. If instead of house number 056 a house number nearer to 300 was selected, say number 290, the sample will consist of members living in house 290, 291...300 and 001, 002 ...etc. until a cluster size of approximately 500 persons have been obtained.

#### 6.2 Selection of clusters in urban areas

The procedure is best illustrated by the following hypothetical example. Suppose the township has an estimated population of 90 000 according to the latest census (which may be a few years old), and that the township is divided into five localities A, B, C, D and E with estimated populations as follows:

<u>Locality</u>	<u>Estimated population</u>
A	15 500
B	23 400
C	12 000
D	26 000
E	<u>13 100</u>
TOTAL	<u>90 000</u>

It happens quite often that a majority of the more well-to-do among the population live in clearly distinguishable areas (say locality E). It may be found desirable to stratify the population at least into two strata to distinguish the well-to-do from the others and to choose separate clusters from each of the strata. Suppose that on theoretical and practical grounds it is best to take one cluster from the well-to-do area E and two clusters from the rest of the town.

One cluster will, therefore be chosen from locality E and two other clusters selected as follows from localities A, B, C, and D:

Locality	Estimated population	Cumulative total	Numbers pointing to locality
A	15 500	15 500	00 001-15 500
B	23 400	38 900	15 501-38 900
C	12 000	50 900	38 901-50 900
D	26 000	76 900	50 901-76 900

As explained in the previous sections a table is set up as above. Let us assume now, that it is desired to have the two clusters not in the same locality.

We can now look for a five-digit number (because the total is 76 900) from the table of random numbers. Taking the table provided in this guide, and starting from some point at random, row 15 and column 17, we will hit the five-digit number 67539. This number can be seen to be located in locality D which is consequently selected. The next number is 34484 and is located in locality B. Thus localities B and D are selected.

The selection of the cluster will be illustrated for locality B.

The project leader (or the statistician if one is available), in collaboration with the local authorities, draws a map of the locality dividing it into blocks, estimates the number of inhabitants in each block and allocates block numbers. Suppose the locality is divided into about 100 blocks. The blocks should be arranged in serial order together with the estimated population. A random number is chosen as before to determine the starting point (i.e. the block). It is practical at this stage to examine all persons living in the selected block and the next blocks until the required number (cluster size) is reached.

### 6.3 Selection of clusters in widely scattered populations and in shanty towns

There are two situations in practice where the selection of clusters as indicated in the preceding sections becomes impracticable. These are:

- (1) when the community is small and lives in widely scattered areas;
- (2) when the community consists of the "marginal population" that is gathering in urban areas and not entering the census. These populations usually live in shanty towns on the outskirts of urban localities, in improvised shelters.


In these instances the best that can be done is to resort to "Area sampling".


The procedure is first to draw a map outlining the area. An easily distinguishable landmark, such as the chief's hut, a local school etc., is chosen and marked on the map as accurately as possible. A set of coordinates is drawn through this point and the map is then blocked into equal squares (see illustration). The numbering of the squares can proceed in any convenient order.

Suppose there results 64 such squares. A random number between 01 and 64 is now chosen from a table of random numbers. Suppose 09 is selected. Square 9 is found on the landscape by using a compass and pacing out the distances (or a tape measure or the trip mileage recorder of a motorised vehicle). The inhabitants of "square 9" are now registered beginning from the part of the square closest to square 8. Registration continues in squares 10, 11 etc. until a cluster size of 500 is reached.

The squares included in the sample group should be marked by corner-flags, sticks, stones etc., to facilitate the identification of the group during examinations. With this method the population living in the border areas will have a greater chance of being selected than the others. This can be remedied by assigning different weights (based upon estimated population etc.) to the different squares and choosing a starting point as before for the selection of villages.

28							
	13	12	61	36	37	52	53
27	14	11	62	35	38	51	54
26	15	10	63	34	39	50	55
25	16	9	64	33	40	49	56
24	17	8	1	32	41	48	57
23	18	7	2	31	42	47	58
22	19	6	3	30	43	46	59
21	20	5	4	29	44	45	60

 Landmark  
(starting point)

 boundary  
of area

258158 DMK

7. EXPERIENCE FROM THE LAT SURVEYS

In the early 1960s the WHO Leprosy Advisory Team (LAT) was responsible for a number of surveys. Most of the surveys were confined to parts of the countries and the estimates provided cannot be taken as national estimates for leprosy. Nevertheless the pattern observed may serve as a useful guide in planning future surveys.

The following table\* gives the numbers examined, the number of leprosy cases in the sample, the prevalence rate and the 95% confidence limits. It will be seen that the prevalence ranges from 0.14% (in Parana and San Jose in Argentina) to 3.86% (in Shwebo and Myingyan in Burma). In Burma, where more than 19 000 people have been examined the precision is highest - the confidence limits being within + 20% of the estimate. Elsewhere, except in Argentina, the limits are mostly within + 30% of the sample estimates. The survey carried out in Khon Kaen and Lampang provinces in Thailand provides perhaps an indication of the size of a sample survey. The sample estimate was 1.22% with confidence limits 1.55 and 0.88 - an estimate that could be accepted as reasonably precise for most purposes. The sample design consisted of examining 33 clusters of approximately 500 in each cluster. The total number examined was 16 862.

Leprosy Prevalence - some selected countries

	EXAM.	CLUSTERS	POSITIVE	POS.	CONF.	
					UPPER	LOWER
Argentina Parana & San José	<u>5169</u>	<u>50</u>	<u>7</u>	<u>0.14</u>	<u>0.23</u>	<u>1.04</u>
Argentina Chaco I & II	<u>1635</u>	<u>15</u>	<u>42</u>	<u>2.57</u>	<u>4.10</u>	<u>1.04</u>
Liberia	5327	35	87	1.63	2.09	1.18
Cameroun	14473	30	374	2.58	3.26	1.91
Western Nigeria	6496	16	101	1.55	2.04	1.07
Thailand	16862	33	205	1.22	1.55	0.88
Burma	19493	38	753	3.86	4.58	3.15
Philippines	10361	20	69	0.67	0.87	0.46

\* The actual sample designs used in these surveys were varying and somewhat more complex. In this section, merely to illustrate the effects of cluster sampling, the data relating to different clusters in the different surveys are reproduced and computations are made assuming a simple cluster design as described in this guide.

The following additional features observed from the LAT surveys are suggestive for future surveys, planned either for evaluation of obtaining baseline information.

	% of pop. over 14 years of age	% of all leprosy cases within this age-group	Prev. (%) (in age-group 14)
Cameroun	62.4	93	3.9
W. Nigeria	51.6	89	2.7
Thailand	49.6	93	2.3
Burma			
Shwebo	55.3	71	5.1
Myingyan	59.8	74	5.5
Philippines	57.5	87	1.01
Argentina			
Parana & San José	65.5	91	0.14
Chaco	57.0	93	0.9

\* Generally, between 40 and 60% of the cases were found to have disabilities.

Thus, in general in many hyper-endemic areas (in developing countries), about 70-90% of the cases are likely to be found in the age-groups beyond 14, which comprise a little over 50% of the population. Hence, one possibility is to confine the population type survey (Household visits etc.) to the adult population (aged over 14 years). The generalization based on the sample will then apply only to the adult population of the community or country. In many situations the prevalence was found to be quite high already in the age-group 10-14 (e.g. 2.9% in the Singu area of upper Burma). In such circumstances the design could consider age-group 10 and above. The children examined in these households could provide information on the lower age-groups but with lower precision. However the overall prevalence for all age-groups as obtained in this way will not be unduly affected. This is further discussed in the following paragraph.

#### 8. SOME IMPORTANT PRACTICAL CONSIDERATIONS

The following general points regarding sample surveys deserve emphasis:

1. Once the sample has been selected it is imperative that every one (100% coverage) in the sample is examined. In practice this is often impossible. Efforts should be made to achieve as high a coverage as possible and make intensive efforts to examine a random sub-sample of the sample population not examined in the first instance. Information from the sub-sample will provide a means of adjusting the estimate from the sample population actually covered. The overall precision may be reduced to some extent as we are adding the sampling variation of the sub-sample to the sampling variation of the original sample. However, the estimate thus obtained will be unbiased. Bias is one of the most important non-sampling errors and should be guarded against.

2. As mentioned in the previous section on the experience with LAT surveys, the majority of cases are in the adult age-group, but it is known from practice that children are easier to cover. Most of the surveys report only an overall coverage. A good overall coverage may consist of an exceedingly good coverage of children coupled with a bad coverage of adults. For this reason the following methodology is proposed:

- (a) Treat the sampling investigation as one stratified according to children and adults.
- (b) Choose a sample size suitable for the estimation in children and another sample size suitable for the estimation in adults.
- (c) Operationally the same sets of villages chosen for adults will be used for children. It is expected that a larger sample will be needed for children as the disease prevalence is rather low among them. Additional sets of villages may be selected at random to provide the required number of children. In these additional villages the examinations can be confined to the children alone. If it is necessary to examine some adults to suit operational convenience, this can be done but will not be accounted for in the analysis.
- (d) While under coverage, especially among the adults, may introduce serious bias in the estimates, examination of more adults from additional villages to make up the numbers may also again introduce serious bias.

It must be remembered that a sample survey is not the same as mass case detection. The only purpose of the sample survey is to provide unbiased and reasonably precise prevalence estimates for the entire community from which the samples are drawn. The time, effort and resources spent on following up a sub-sample of the uncovered sample population is far more meaningful than the examination of a large number of individuals who are not in the original sample selected.

A practical example will illustrate the points mentioned above. Let us assume that the expected prevalence in all age-groups is around 1.2% and the prevalence in the age-group 15 and above 2.3%. Further the population in this adult age-group is 50% of the total. Let us stipulate that the precision of the estimates should be such that the coefficient of variation for adults should be within 15% and that for children within 20%.

The expected age-specific prevalence in a situation similar to the above is usually as follows:

Age-group under 15	-	0.18%
Age-group over 15	-	2.3%

Under simple random sampling conditions the adequate sample sizes would be 1835 adults and 13 863 children. Under the cluster sampling design, the sample size would need to have been increased four times. This would mean a sample size of 7340 adults and about 55 000 children. Using a typical cluster design, (viz. villages or village clusters of size 500) and assuming 50% of the population is 15 years of age and over, this would mean examining a total population of 14 680 or around 29 clusters for the adults and a child population of 55 500 that will be obtained from a total population of 110 000 or 220 clusters.

The procedure will be to choose 220 clusters at random and choose 29 from these again at random. From the 29 clusters the total population will be examined and from the remaining 191 only children under 15 will be examined.

If for some reason only 60% of the selected adults could be examined, i.e. around 4400, it is suggested that a 10% sample of the remaining, i.e. 10% of 2940 or 294 adults be selected at random and every effort made to trace them.

The total child population of 55 000 in the above example seems unusually large. This is because the prevalence in that age-group is very low. The desired degree of precision means that if the sample prevalence is in fact 0.18%, we should be in a position to state that the 95% confidence limits are  $0.18 \pm 0.07$  (i.e. 40% of 0.18) or between 0.25% and 0.11%. It may be enough if we could demonstrate that the child prevalence is below 0.3%. In this case we will be asking for a precision expressed by the coefficient of variation as being  $0.06/0.18$  or 33.33%. This would give rise to a sample size of 4941 by simple random sampling. Allowing for effect of clustering this would mean examining four times\* this number, or about 20 000 children. This would be obtained from a total population of 40 000 or 80 clusters.

The 29 clusters selected primarily for adults will in fact yield 7340 children. With the observed prevalence 0.18%, this gives rise to a 95% upper confidence limit of 0.38%.

The sample sizes required for children could perhaps be reduced if a relationship could be established between the prevalence among adults and that among children. If such a relationship is perceived, even during the survey (as for example in the selected clusters) more refined methods of estimation are possible and the sample size for children could be considerably reduced. Special statistical consultation is desirable at this stage.

#### Stratification

Stratified sampling may be necessary in leprosy surveys for the following reasons:

1. Countries are usually divided into provinces or States and leprosy control is often under the State health authorities. Information on the leprosy status will therefore be required separately by States, or perhaps even by smaller administrative sub-divisions.
2. Usually information is required separately for urban and rural areas as the organization for leprosy control can be different for urban and rural areas.
3. From statistical considerations, stratification of the country or State by endemic levels will contribute to greater efficiency in the sampling design. For this purpose, the area can be divided, on the basis of past experience, into hyperendemic, moderately endemic and other areas.

There are a number of other epidemiological variables known to be associated with the disease such as age, sex, household contact status etc. Stratification on the basis of these variables can theoretically contribute to greater efficiency, but operationally may be inconvenient.

Once the strata are decided on, random samples are taken in each of these independently. The size of the sample (or the fraction of the population which it represents) can be and usually is quite different for the different strata. The global estimate (i.e. the estimate for the country etc) is the combination of the estimates for the individual strata suitably weighed (i.e. after allowing for differences in sample size).

---

\*with the experience in Thailand.

## 9. ORGANIZATION OF A SAMPLE SURVEY

When a sample survey is planned to be carried out in an area, the project leader should visit the area approximately one month before the planned date of commencement of the survey. At this time he should identify the "key" person/persons in the area whose cooperation should be sought for the successful conduct of the survey. The objectives and method of the sample survey should be explained to him in detail. He should be convinced that the information from the surveys is solely for planning purposes and for the ultimate benefit of the community. Names of individuals will be kept confidential and publication will only be of aggregated numbers. The "key" persons will themselves give reasons why individuals may not cooperate. Means will have to be mutually settled on ways to get over these reasons. Usually they are related to social stigma and fear of losing jobs. Perhaps, if the resulting non-cooperation is serious, the word "leprosy" can be deleted from the name of the survey. If convenient it could be simply called a "dermatological survey". The leprosy diagnosis could be kept a secret between the patient and the leprosy worker. If necessary the completion of the diagnostic portion of the record form could be coded.

One important problem is often related to examination of women. The ways by which this problem is overcome should be decided at an early stage. For instance a local midwife or a trusted woman out of the community can be invited to join the survey team locally and refer to a team member all the skin abnormalities and deformities. This should be discussed with the key person who then would be in a position to obtain agreement from their people.

Opportunity should be taken at this time to obtain a detailed map of the area, the latest census information in as much detail as possible (age, sex, etc. by villages or town-blocks, list of households, etc.). If census information is not available recourse must be had to other means of getting population estimates. Usually there exists some indication of the size of the population, either through censuses (though old) or sample surveys or surveys done in some other connection, at least in parts of the area. In one instance a list of tax-payers was available. This was multiplied by a factor reflecting the estimated proportion of tax-payers in the community to give an approximate first indication of the size of the population. In principle, if not strictly necessary, tax-payer lists should not be used as any association between a survey and taxes should be avoided. If, after an intensive search no information is available, one may have to resort to area sampling.

At the time of the first visit, information should be obtained on ethnic groups, geographical areas etc. that should be excluded from the sample survey. Such exclusion relates for example to nomads who cannot be reached, areas of unrest and others where the survey will not be permitted to enter. For all such groups, the size and location of the group, should be ascertained and information obtained as to whether the group, if included in the census, is well enough defined to be removed from the population figures by district and community.

Other information that would be of value is:

- how the population is distributed over rural and urban areas;
- extent to which they are scattered over large areas;
- accessibility of communities generally and during seasons;
- periods when people move out for work etc.

In general, the time of the year, week and day when one can hope to meet the population easily.

Since cluster sampling will be the recommended design, this will be the time to decide on the size of clusters. It is practical to regard as a cluster a village or a group of villages in the rural areas. The average number of persons that can be examined in a specified (say three-week) period of time must be determined after taking into account the time for travel and preparation of work.

### Registration of the sample population

Before the actual commencement of the examinations the team leader and/or a member of the team responsible for statistics, should start registering the sample population. Accompanied by a person of authority in the community he should visit all living quarters (including hospitals, rest houses, boarding schools etc. as well as homes) in the selected clusters and register the persons residing there. The interval between such visits and the actual examinations should be as short as possible as in some instances the population can be mobile.

Each household should be given a number, which is painted at the entrance of the house or compound. An individual card should be filled in for every person belonging to the household, that is, who habitually sleeps in the house or has habitually slept there and intends to return.

The following information should be recorded on the card:

Cluster number:

Name of village:

Household number:

Serial number:

Name: (first name and father's/husband's name)

Age: (if age is not stated, it should be estimated. For children under 1 year of age, the age should be given in months)

Sex:

Relationship to head of household:

Cards should also be prepared for absent members of the household and for those temporarily present in the house. Temporarily present may for convenience be defined as "those who expect to sleep in the house on the night preceding the first day of examinations but do not belong to the household". If an entire household is away at registration time, the number of persons in the household is estimated (or ascertained from neighbours etc.) and a corresponding number of cards filled with the cluster, household and serial numbers only.

The coverage and consequently the reliability of the estimates depend very much on the accuracy of registration and hence great care must be taken at this stage. The main principle is to include every member of the household included in the sample and exclude others. If the correct status of an individual is difficult to establish, a careful interview of the person and the head of the household is indicated and if doubt still persists, the "doubtful" status should be recorded on the card.

A person is classified as "temporarily absent"(TA) if he belongs to the household but is absent from the community and is expected to remain absent on the night preceding the examinations. A person who is simply away from home working in the fields or in a factory or trading in the village or a nearby town but is expected to return to the household at night is to be included in the examinations and must not be classified as TA. Should every effort to locate and examine him fail, he is still not classified as TA but is simply recorded as being absent during the examinations.

A person is classified as temporarily present (TP) if he does not belong to the household but is present at the registration and is expected to sleep in the house the night before the examination starts. No cards need be made out for casual visitors to the house who do not intend to sleep there.

If for some reason it is necessary to include in the examinations individuals who are not in the selected sample (e.g. at the request of the village chief etc.) a card may be completed but the sentence "Does not belong to Sample" written on the card in bold characters diagonally across the card.

Information on the card should be checked on the day of examination. This is particularly important for the items such as TA and TP, as these refer, as stated above, to the absence or presence of the person in the house on the night preceding the examinations.

Persons appearing for examination who have not been registered are issued cards and included in the examination only after the team has carefully checked that they do in fact belong to the sample group.

The work may be organized by house-to-house visits or by instructing all members of the sample group to attend the examinations at a central place.

House-to-house visits are generally to be preferred in leprosy surveys. One of the reasons is that this procedure enables the team to become familiar with the sample group and thereby to check completeness of registration and examinations. However there are areas where it is difficult to perform a proper examination of the skin in the houses when these have small windows and the illumination is not sufficient. In such situations one should look into the possibility of examining people at home but outside the house or in a central place with proper illumination. Examination at a central site allows much less control of attendance and identification and should only be used when the population is so scattered that the team cannot reasonably be expected to cover the area on foot. There may be considerable difficulty when examining at a central site in identifying an individual and finding his card, especially if he does not attend with his household.

Irrespective of how the work is organized the identity of the individual is checked at each examination by asking his name and father's name (in some cultures, and the mother's name) checking the information against that registered on the individual card. The examinee should be asked to give his or her name; the team member does not suggest the name or ask whether it is correct.

Every effort must be made to include in the examination everyone registered in the sample group. Constant check should be kept on the coverage so that those who are not appearing for examination can be found and special appointments made during the evening or early morning hours when they can be expected to be home. In many cases repeat visits to the home may be necessary before the individual is able to or can be persuaded to attend the examinations.

## 10. PROCEDURES AND CRITERIA FOR DIAGNOSIS

### 10.1 Introduction

The procedures and criteria for diagnosis of leprosy in sample surveys should be generally limited to what is possible in the field and to methods which can be applied in a reasonably standard and consistent manner.

The diagnostic procedures are mainly clinical, and to some extent bacteriological, and therefore the examiners should have had adequate training in the specific methods and should also have had reasonable experience. In certain situations it may be useful to control this through verification of individuals identified as cases, suspects, and normals by an independent reliable senior examiner, possibly, a physician.

Problems in diagnosis arise mainly in two areas. (a) Early lesions, particularly when some of them have only minimal or marginal evidence of leprosy. This would mean that examiners should try to be as objective as possible. (b) Inactive lesions which have become so either due to treatment or spontaneously. As far as possible these should be identified and excluded from the list of active cases.

#### 10.2 Procedures for diagnosis

Procedures for diagnosis should include the following:

Inspection: Inspection of total body surface in good light and in a systematic manner. Where cultural situations demand, women examiners may be required. Minimum time for inspection should be 2 minutes. Asking the examinee about any patch or sensory loss noticed by him may be of value. Inspection involves looking for macules, papules, diffuse infiltration, nodules, dry areas of skin, plantar ulcers, and deformities.

Palpation: Palpation of nerve trunks should always include ulnar and lateral popliteal nerves. Other nerves should be examined as and when indicated.

Testing for sensory loss: Testing for sensory loss should be carried out (a) on all skin lesions the appearance of which is compatible with those of leprosy, (b) when a patient complains of numbness and (c) when there is a nerve trunk thickening even in the absence of any skin lesion. Testing for sensation will be confined to light touch (touch) and pin prick (pain). In some situations only pin prick will be possible.

Skin smear examination: Skin smear examination should be carried out on all individuals identified as having definite or doubtful leprosy. The examination, including reading of smears, should be carried out in the standard manner. Refer: Smear Technique; measurement of the Bacteriological Index (BI): Doc. TDR/THELEP/Protocol/82.1 Appendix 4.

Skin biopsies: At the planning stage a decision should be taken as to whether they will be used. In general it seems appropriate to take biopsies only on doubtful cases or when it is difficult to distinguish between active and inactive states.

#### 10.3 Diagnostic criteria and categories:

For the purpose of the sample survey the examined population could be categorized as below:

I. Multibacillary leprosy: Skin lesions compatible with leprosy with AFB positive smears.

II. Paucibacillary leprosy-active: (a) Anaesthetic skin lesions compatible with leprosy and with AFB negative smears, or (b) Area of sensory loss over apparently normal skin with thickened and tender nerve and with AFB negative smears.

III. Leprosy-inactive: Scarred anaesthetic skin lesions which would otherwise be compatible with leprosy with AFB negative smears.

IV. Leprosy-activity status doubtful: (a) Anaesthetic skin lesions compatible with leprosy, with neither erythema, nor infiltration, nor scarring, and with AFB negative smears, or (b) Area of sensory loss over apparently normal skin with non-tender thickened nerve and with AFB negative smears.

V. Doubtful leprosy: (a) Non anaesthetic smear negative skin lesions otherwise compatible with leprosy or (b) Nerve trunk thickening with no area of sensory loss and with AFB negative smears or (c) Area of sensory loss over apparently normal skin with no nerve trunk thickening and with AFB negative smears.

VI. No leprosy: No skin lesion, no nerve thickening, and no sensory loss compatible with leprosy.

The examination procedures are illustrated schematically in the adjoining chart.

#### 11. NON-SAMPLING ERRORS

Statistical theory is sufficiently advanced to control and minimise sampling variations. On the other hand the control of non-sampling errors in medical surveys is bound up with the precision in diagnosis and the coverage one obtains. Statistical principles can be in-built into the survey operations to control or at least give estimates of the variations that could be expected due to variations in diagnosis. The ability to achieve a high coverage depends to a large extent on the socio-cultural setting and the leadership qualities of the survey personnel.

Coverage plays an unusually important role in leprosy sample surveys. The disease is very often closely related to the characteristics of the population not covered (hereinafter referred to as "no-respondents") and this will seriously upset the estimates based on the population covered and offset any advantages gained from a sophisticated sampling design to control sampling errors. For this reason a considerable part of the total expenditure for the sample survey must be set aside for the control and management of non-response.

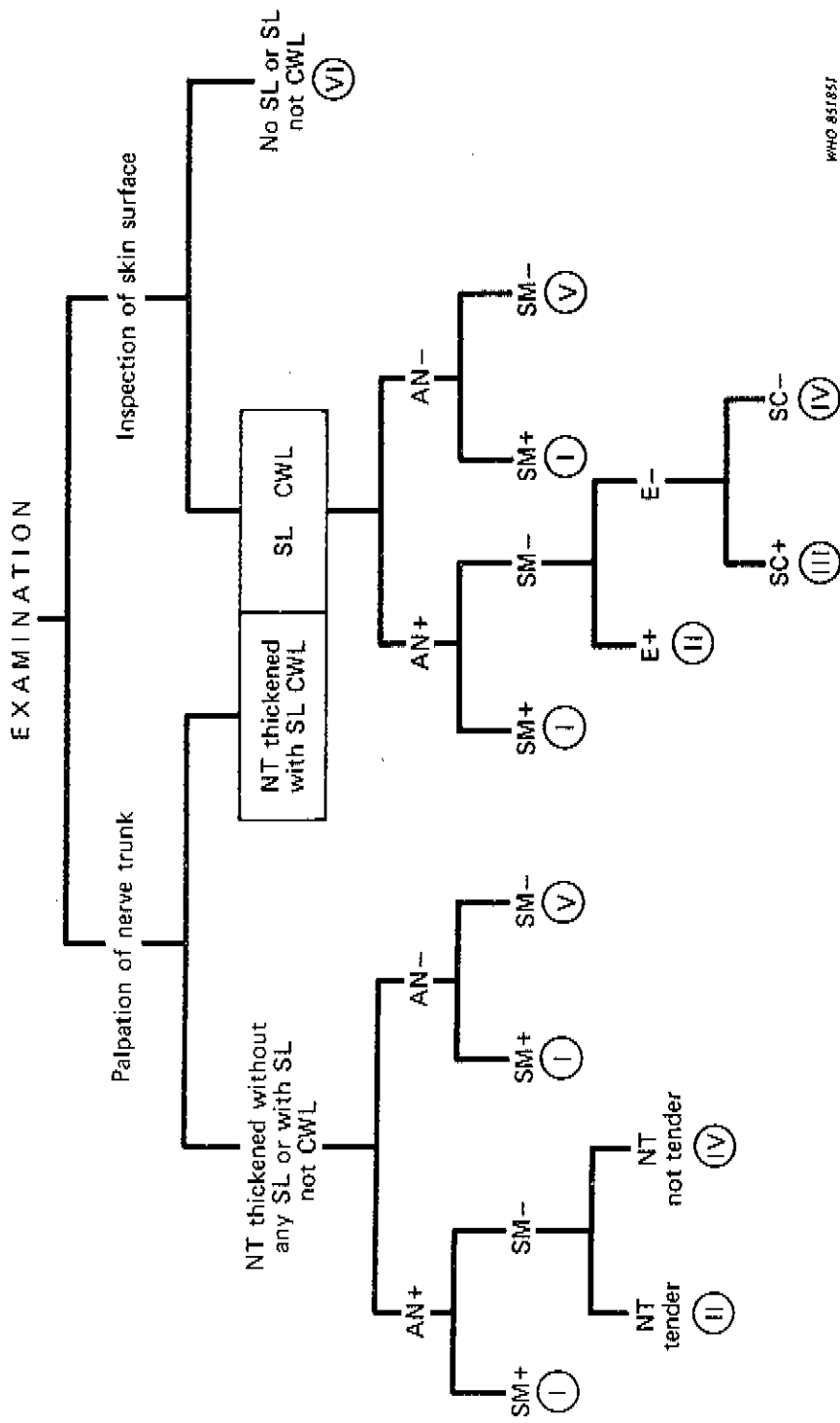
While it is hard to generalise and establish methods for coping with non-response, it will nevertheless be useful to list below some of the commonly observed causes for non-response and suggest some measures to minimise these.

##### 11.1 Stigma attached to the disease

In most parts of the world there still exists a considerable amount of stigma attached to the disease and the individuals are afraid to be identified as leprosy patients. This could be got around to some extent by an intensive public information exercise assuring the population that the information will be kept strictly confidential and the publication of the results will only mention numbers in aggregate. Further, if an individual is suspected, in the general examination for patches, etc., his subsequent recall for more detailed examination if any, will be arranged with the utmost discretion so that no one else will be aware of the recall. Further an undertaking should be obtained from the health agencies, employers, etc., that the information on the surveys will be treated as confidential and will be used only for statistical analysis, and this undertaking should be extensively publicized.

##### 11.2 Incomplete examination of females

This is again a common phenomenon, as it can be delicate and embarrassing or totally unproductive to ask women to expose completely. This could be mitigated by employing women workers (preferably elderly women or midwives) and ensuring that the skin areas normally covered by clothes, etc, are seen.



WHO 85185J

- I Multibacillary leprosy
  - II Paucibacillary leprosy-active
  - III Leprosy-inactive
  - IV Leprosy-active/doubtful
  - V Doubtful leprosy
  - VI No leprosy
- 
- SL = Skin lesion
  - CWL = Compatible with leprosy
  - NT = Nerve trunk
  - AN = Anaesthesia
  - E = Erythema
  - SM = Skin smear
  - SC = Scarring

### 11.3 Ignorance of the disease and disregarding mild lesions

Sample surveys should be preceded by an intensive health education exercise acquainting the sample population about leprosy and the importance of diagnosing early lesions. During the health education exercise, all individuals should be encouraged to attend the examination sessions. The population should be informed that this is important not only for those who have noticed they have lesions and suspect the disease, but that attendance is equally important for those who have no such suspicion because of possibilities of signs having escaped their notice.

### 11.4 Unwillingness of certain classes of the population, such as beggars, to be brought under registration

This is a case for treating sample survey as something different from general case-detection in a control programme. The initial publicity should stress the fact that newly discovered patients will be given treatment or referred to the general programme only on a voluntary basis.

### 11.5 Unsympathetic attitude of leprosy workers

This is again a matter of educating the leprosy workers and perhaps eliciting the cooperation of a social worker or a community health worker.

### 11.6 Apathy towards surveys

The importance of sample surveys for planning or monitoring purposes is not readily seen or appreciated by most people. For many it has the appearance of an empty academic exercise. Any public information programme to explain the objective of the sample survey has to be supplemented by an assurance that any cases discovered will be treated if required and also hold up an attraction that any other ailment discovered or reported will be attended to. It is most desirable for the team to carry with it a stock of some of the common and inexpensive drugs and first-aid equipment. Nothing elicits public sympathy and cooperation as much as the age-old tradition of the "healer", i.e. the willingness to listen with sympathy to the ailments of the individuals and if possible help to alleviate them.

### 11.7 Inaccessibility of selected villages, disappearance of villages listed in census frames, absence of seasonal workers

These are common to most health surveys and unrelated to the disease in question. They underline the importance of initial organization to check census frames, understand seasonal peculiarities such as monsoons which render places inaccessible. Planting and harvesting seasons, market days and festivals should be avoided. In short, the timetable of the visits should be well planned so as to ensure that the sample population is, in fact, present in the selected villages when the team visits them.

The survey should finish not too late in the day (e.g. 12.00 - 1.00 p.m.) in order to enable people to attend to their daily business. If the survey really has to continue during the whole day the sample area should be clearly divided into two or three parts and the time when these parts will be visited should be clearly stated; otherwise many people will become tired of waiting and will move. Temporary absentees should be asked to come for examination in the afternoon.

It could happen that, in spite of the best preparations, a village selected from an old list does not exist anymore or for some unforeseen reason a village is not accessible. Usually another village is chosen at random to complete the work schedule. In all such situations, it should be recognized that a certain "bias" could have been introduced. A judgement should be based on other epidemiological evidence.

In some situations, such as in Savanna Africa, the adult males will be out working for long periods of time.

#### 11.8 Identification of "key" persons

Identifying a key person who carried influence in the village, eliciting his cooperation and giving him some responsibility and a sense of importance, will greatly help secure the cooperation of the population. The time and effort spent in cultivating such a key person will be well worth-while.

11.9. Once all the phases of the preparation of a survey have been completed it will be most desirable to undertake a pre-survey before the actual survey starts. Such a pre-survey will enable one to: examine the adequacy of the record-forms in relation to the objectives of the survey and if necessary, to modify the same; ensure that the methods to be applied are clear to the personnel undertaking the survey; review the survey manual prepared for the use of these personnel and help complete their training.

Such a pre-survey should be carried out on a limited population, typical of the population selected but not in the villages (or localities in the case of urban populations). Apart from helping to perfect the statistical procedures referred to above such pre-surveys can provide relatively accurate indications on the personnel and material requirements, the speed of the work and global estimates of the survey cost.

#### 12. PERIODIC SURVEYS FOR EVALUATION

The base-line survey establishes the magnitude of the problem in all its facets and helps organize appropriate control activities. Sample surveys conducted in the area periodically (say every 3 or 4 years) will help to monitor the effectiveness of the control programme and help to initiate the alternative strategies should the current strategy prove insufficiently effective.

The following indicators are suggested for measuring the effectiveness of a programme:

- (1) age and sex distribution of the cases;
- (2) proportion of undetected cases (discovered in the survey) to the total registered cases;
- (3) proportion of active cases to the total;
- (4) proportion of multibacillary cases to the total;
- (5) proportion of patients discovered in the survey who are regularly taking drugs;
- (6) proportion with disabilities.

Bias is likely to be introduced in the estimates if the same sample is used for each periodic survey. It is therefore proposed that a system of panel sampling with partial replacement be used. In this system a part of the original sample is retained and another part chosen afresh. An example will make the application clear.

Let us suppose that a rural area, with a population of one million, has been earmarked for an intensified leprosy control activity. Twenty-five clusters (villages or groups of villages) with a size of approximately 500 may have been chosen for the base-line survey. In the next survey to be conducted after, say 3 years later, a new set of fifteen clusters may be selected from the same area. These together with ten clusters selected randomly from the original twenty-five could constitute the new sample.

Of course one could design more sophisticated methods. For example, a fixed number of the original clusters, selected at random, say 10, could be retained in the second survey, the remainder (15 in the example) selected at random from the population after excluding the clusters originally selected. A design of this nature over several periods may facilitate the cohort type of analysis. Depending on the type of analysis one is interested in, the sampling procedures can be suitably designed. Since the object of this guide is to provide a simple procedure for monitoring the impact of programmes on the community, the procedure mentioned in the previous paragraph would be the most suitable, and refined procedures intended for in-depth epidemiological analysis will not be dealt with.

It has been the past experience that where a good control programme has been started, reduction in the incidence has occurred over relatively short periods of time. However, on theoretical grounds, considering the long incubation period of the disease, the reduction in incidence among adults may not appear to be commensurate with the intensity of a leprosy control programme. The reduction among children, especially infants, should *prima facie* be substantial and directly reflecting the control effort. However, incidence in children is not high and recourse should be made to methods other than population based sample surveys. Schools and MCH centres are possibilities, whenever these services cover the population in these age-groups reasonably well.

#### ACKNOWLEDGEMENTS

The completion of this manual was greatly helped by valuable comments from:

Dr Jair Ferreira, Porto Alegre, Brazil; Dr M.D. Gupte, Indian Council of Medical Research, India; Dr R. Kersauze, Ecole Nationale des Ingénieurs des Travaux Ruraux, Strasbourg, France; Dr M. Lechat, School of Public Health, Catholic University of Louvain, Brussels, Belgium; Dr D.L. Leiker, Royal Tropical Institute, Amsterdam, Netherlands; Mr K. Uemura, Director, Division of Health Statistics, WHO, Geneva; Dr B. Zuiderhoek, WHO Leprologist, Ujung Pandang, Indonesia; the WHO Regional Office for the Americas and the WHO Regional Office for the Western Pacific.

REFERENCES

1. World Health Organization, A Guide to leprosy control. Geneva, 1980.
2. WHO Technical Report Series, No. 607, 1977 (Fifth report of the Expert Committee on Leprosy).
3. United Nations Department of Economic and Social Affairs, Statistical Office of the United Nations. Recommendations for the preparation of sample survey reports (Provisional issue). Statistical Papers, Series C, No. 1, Rev. 2 (New York, 1964).
4. FISHER, R.A. & YATES, F. Statistical tables for biological, agricultural and medical research, 6th ed., revised and enlarged. Edinburgh, Oliver and Boyd, 1963, p. 137.
5. MARTINEZ DOMINGUEZ, V. ET AL. Epidemiological information on leprosy in the Singu area of Upper Burma. Bulletin of the World Health Organization, 58: 81-89 (1980)

FORMULAE FOR STANDARD ERRORS - CLUSTER SAMPLING

I. All clusters of equal size

a = number of clusters

b = cluster size

$y_i$  = number with disease in cluster i

f = sampling fraction\*

$r_i = y_i/b$

The estimate of proportion with disease  $r = \frac{\sum y_i}{a \cdot b}$

$$v = \text{variance of } r = \frac{(1-f)^*}{a(a-1)b^2} \left\{ \sum y_i^2 - \frac{(\sum y_i)^2}{a} \right\}$$

s.e = standard error of  $r = \sqrt{v}$

95% confidence limits of  $r$  are  $r \pm 1.96$  (s.e)

II. Unequal clusters

a = number of clusters

$x_i$  = the size of cluster i

$y_i$  = the number with disease in cluster i

r = estimate of proportion positive (i.e. with disease)

x =  $\sum x_i$  = total sample size

$r = \sum y_i / \sum x_i$  ; f = sampling fraction\*

$$v = \text{variance of } r = \frac{(1-f)^*}{x^2} \frac{a}{(a-1)} \left\{ r^2 \sum x_i^2 + \sum y_i^2 - 2r \sum x_i y_i \right\}$$

s.e = standard error on  $r = \sqrt{v}$

95% confidence limits of  $r$  are  $r \pm 1.96$  (s.e)

---

\*may be ignored if the sample represents a small proportion (e.g. less than 5%) of the population.