53145

# UNITED NATIONS
# ECONOMIC AND SOCIAL COUNCIL

Distr.
LIMITED

E/CN.14/SM/7
14 May 1968

Original: ENGLISH

ECONOMIC COMMISSION FOR AFRICA
Seminar on Sampling Methods
Addis Ababa, 3-14 June 1968

## APPLICATIONS OF SAMPLING TO STATISTICS OF DISTRIBUTION

(Prepared by the Statistical Office of the United Nations)

M68-694

# APPLICATIONS OF SAMPLING TO STATISTICS OF DISTRIBUTION

## CONTENTS


| | | Paragraphs |
|---|---|---|
| I. | INTRODUCTION | 1 - 2 |
| II. | CHARACTERISTICS OF THE POPULATION TO BE SAMPLED | 3 - 4 |
| III. | RESOURCES AND FACILITIES NEEDED | 5 - 8 |
| IV. | COVERAGE AND REPRESENTATIVENESS IN INQUIRIES OF DISTRIBUTION | 9 - 16 |
| | Retail Establishments | 10 - 14 |
| | Wholesale Establishments | 15 - 16 |
| V. | SAMPLING PLAN | 17 - 72 |
| | General Observations | 17 - 27 |
| | Division of Retail Trade into Two Sub-Populations and Allocation of Sample to Each | 28 - 35 |
| | Sample from the Sub-Population of Listed Large Establishments | 36 - 54 |
| | Sample from the Sub-Population of the Remaining Establishments | 55 - 63 |
| | The Use of Rotating Samples and Ratio Estimation | 64 - 72 |
| VI. | HOUSEHOLD RETAIL TRADE | 73 - 84 |
| VII. | NON-SAMPLING ERRORS | 85 - 94 |
| | Response Errors in Sales | 87 |
| | Listing Errors | 88 |
| | Scope Differences | 89 - 90 |
| | Kind-of-Business Differences | 91 |
| | Additional Remarks on Non-Sampling Errors | 92 - 94 |

# I. INTRODUCTION

1. There are advantages in using sampling techniques in the collection of comprehensive distribution statistics once every five or ten years as well as in the collection of these data monthly or quarterly and annually. Sampling is valuable in gathering data from small establishments and in collecting figures on such items of data as capital expenditures and the classification of sales by kind of commodity, kind of customer and method of payment. Countries will find sampling useful for many of the items of data that are placed in the category of second priority. Samples of establishment returns would also help to tabulate rapidly preliminary data in comprehensive inquiries. In doing this, care must be taken that the sample selected is representative of all establishments covered in the inquiry. Countries engaged in taking annual and monthly or quarterly distribution inquiries would find it useful to use sampling to reduce the number of respondents and make it possible to issue results rapidly and to keep the cost of making such inquiries within reasonable bounds.

2. To ensure that the units enumerated in a survey are representative of all the establishments to which the survey relates, probability sampling or random sampling needs to be utilized. In these sampling methods, the probability that an establishment is included in the sample is known. With probability samples one can state an objective basis for choosing from among the alternative methods of sampling and methods of estimation. The sample can be designed in such a way that it will yield the required precision at a minimum cost or, conversely, at a fixed cost it will yield estimates of the characteristics desired with the maximum precision possible. A method of selecting a sample often employed in place of random or probability sampling is to choose a sample of establishments which is representative with respect to certain known characteristics of the population. However, in such samples one cannot have an objective measure of the reliability of the sample results, because the various establishments may have differing and unknown chances of being drawn in the sample. Although many countries utilize purposive or judgement samples, an increasing number are turning to the use of probability samples. The purposive selection is economical for obtaining a sample of given size, in both time and money, but there is always the possibility that it will lead to faulty conclusions.

## II CHARACTERISTICS OF THE POPULATION TO BE SAMPLED

3. The characteristics of statistical units in the distributive trade sector can be studied from different angles and taking different factors into consideration. To ensure an efficient sampling design it is essential to have some knowledge of the distribution pattern of the population of establishments. Information about the spread, size or volume of sales, etc., by geographical areas, kind of activity and, if possible, by kind of business could make an effective contribution in assigning statistical units to different strata and allocating the sample units to these strata.

4. A particularly important characteristic of the distribution of business establishments is that it is highly skewed. A large proportion of small establishments contribute a small proportion of the total turnover. Whereas a small proportion of large establishments claim a sizable share of the total turnover. The large establishments are concentrated in large places, namely, cities whereas the small establishments are spread throughout the country. These features of the statistical units could be used with advantage in stratifying the units according to size and geographical areas. The advantage of the skewness of the population can be utilized only if approximate measure of size are available in advance of taking the survey. The larger ones which could be listed could be considered separately from the small or the non-listed units. In many instances substantial information, is available for the larger establishments, whereas an up-to-date and complete list of all establishments is lacking.

## III RESOURCES AND FACILITIES NEEDED

5. In order to carry out a sample survey of distribution establishments, one or more of the following sources can be used:

(i) Census of distribution or establishments, covering most of the distributive sector and yielding comprehensive and detailed data. These records may not be up to date on the date of the sample inquiry; however, a census of distribution provides extensive statistical information on distributive trade as of the census date that can be effectively used;

(ii) Administrative records, supplemented by information available in business or telephone directories, etc. The records pertaining to sales tax, shops and establishment acts, income tax and social security laws can be profitably employed in constructing the frame for the sampling inquiry. One should take into account, however, the limitations of these sources. In several cases, even the initial compilation suffers from omission of new units, retention of dead units and incomplete geographical coverage;

(iii) Inquiries generally covering limited parts of the distributive sector and yielding aggregate data;

(iv) Inquiries other than establishment censuses that yield data on distribution as part of a wider inquiry into other fields of statistics, for example, population censuses and labour force, employment and capital expenditure surveys.

6. Whenever above-mentioned sampling frames are employed, especially the last three it is necessary to ensure that additional sources are used to include smaller establishments which may have missed one or all of the above sources as well as pedlars and hawkers.

7. In addition to the above-mentioned sources of information, on which a sample survey of distribution may be based, a highly desirable element in the sampling plan is an already existing nation-wide field organization operating in several different sample areas of the country. Such a field organization, even if originally established for other purposes such as demographic or labour force surveys, can be used and extended for surveys of distributive trade.

8. Another element of importance in the sampling plan is the availability of unambiguous maps covering large cities. These maps ought to be brought up to date annually and ought to show, for the most part, the location and general type of use of industrial structures. Such maps are often acquired by census and survey authorities as an aid in sampling and census taking, and can be effectively utilized in sample surveys of both population and distribution.

## IV. COVERAGE AND REPRESENTATIVENESS IN INQUIRIES OF DISTRIBUTION

9. When one examines the sampling practices of various countries one finds considerable differences among the countries with regard to questions of coverage and representativeness of the sample. As regards the coverage, the Statistical Commission has enumerated the field to be covered in distributive trade statistics.1/ We shall mainly be concerned with the wholesale and retail trade establishments in this paper.

### Retail establishments

10. In respect of coverage of retail establishments, differences exist on three main points: the geographical coverage, the types of organization and of business included. The sampling fraction varies greatly between various types. In certain cases all or a considerable number of large stores are included in the sample, but small independent traders are sometimes omitted, or seriously under-represented. In several cases one observes that the rural regions are always seriously under-represented, and hence the sample results may be representative for urban regions only.

11. The exclusion or under-representation of small independent traders, who in most countries cover together a substantial part of total retail trade, especially in some lines of business, is likely to be the cause of more or less serious bias. A further source of error, which is of particular importance in respect of these small traders, is the fact that generally no account or inadequate account is taken of "births" and "deaths" of establishments.

12. The deficiencies in coverage mentioned in the preceding paragraphs may affect the representativeness of the sample unfavourably. A low sampling fraction for certain geographical areas, types of organization or kinds of business would, of course, not necessarily imply that the total sample is not representative for a characteristic such as a total sales. As long as the sampling fraction is high enough for each stratum of the population covered (geographical area, type of organization, kind of business) to yield representative results for that stratum, representative results would be obtained for the population as a whole if proper weighing coefficients were applied. When, however, a sample for a particular geographical area, type of organization or line of business is not representative, this must necessarily affect the representativeness of the results for the whole, even when adequate weights are applied.

13. The question of representativeness is raised not only in obtaining structural data as above but also in obtaining current statistics. For example, when the indices for the total retail sales include unrepresentative sub-indices for certain lines of business, the combined index will be biased to some extent. This has been found to be the

---

1/ International Standard Industrial Classification of all Economic Activities (Statistical Papers, Series M, No. 4, Rev. 1).

case in certain cases. More common, however, is the fact that certain lines of business are entirely omitted from the indices. This does not affect the representativeness of the index for the line of business which is covered, but it does affect the international comparability of statistics on total retail sales.

14. Another reason for lack of representativeness of the indices is deficiencies in the weighting systems. In some cases, the weights are derived from inquiries covering total retail sales in a given (base) period, e.g., from a distribution census. In many such cases, censuses of distribution are very infrequent and the information may be seriously out of date. In some cases where a census of distribution has never been taken, the weights for the index of retail sales are based on sources of varying reliability, such as tax records, or an estimates of total consumption in a given year, or on the number of establishments of different kinds; or the indices for total sales are sometimes obtained by simply adding the turnover figures of the establishments included in the sample. In such instances, the weighting system may introduce a bias into the index, especially when the movements of the sub-indices are divergent, as in instances where co-operative organizations or large establishments have grown rapidly.

Wholesale establishments

15. The geographical coverage is better in the case of wholesale establishments than for retail trade for characteristics such as sales, since for the lines of wholesale a business included, the statistics generally cover the entire country and quite often all types of traders.

16. In many cases, where current statistics on sales of wholesale establishments are obtained from sampling inquiries, the sampling fractions for the lines of business covered appear to be generally higher than in the case of retail establishments. Since, normally, differences in type of organization are not important in respect of wholesale trade in given line of business, the problem of representativeness of the sample for different types of organization, and of the adequacy of the weights of each type of organization in the combined index for the line of business concerned are less important than in retail trade. In some cases, where co-operative buying of independent retailers is developing rapidly, under-representation of this type of organization may affect the representativeness of the sample.

## V. SAMPLING PLAN

General observations

17. It is not possible to present here a detailed description of the sampling problems, as many of the details would be modified depending on the circumstances peculiar to a given country.

18. The major difficulties faced by countries wanting to take surveys of distribution are that often little is known about the characteristics of the population of wholesale and retail traders, and that, particularly in respect of retail trade, the population changes rather frequently because of births and deaths of establishments and seasonality of trades. An adequate frame which is both comprehensive and up to date, is therefore, not easily obtained.

19. To design, in a scientific way, a sample which is likely to be fully representative of traders is difficult and may be costly. Many countries, therefore, have apparently found it necessary to follow less scientific lines, e.g., by including in their samples all establishments which it is relatively simple to locate and which can readily provide data (such as large establishments in cities), and only a proportion, often quite small, of other establishments.

20. A method often used in surveys of wholesale and retail trade is to mail questionnaires to all or a sample of establishments included on available lists and to depend primarily on a voluntary mail response to the mailed inquiry. In using such returns it is usually assumed that, if sales for two periods of time are collected from an identical set of establishments, the change in sales between the two periods for the identical establishments will, in fact, reflect the change in total sales for the specified kind of business. This type of sample estimate may be subject to two sources of bias:

(1) The mail responses do not reflect the change in sales of the non-respondents to the mail questionnaires; and

(2) the available mailing lists are usually out of date and hence the sample drawn from the lists does not provide the information for estimation of the net effect of the turnover in business establishments.

21. In most of the developing countries the mail survey method especially for small-scale statistical units is out of the question. The second method of leaving a blank questionnaire with small retailers and collecting it after a certain period or sending it by mail and having the filled-in questionnaire collected by a field investigator may be considered. The main hurdle here is the lack of adequate or proper accounts in the small-scale statistical units. Because of this, the owner of the small-scale statistical unit may be unwilling to build

up the data for the reference period even if time is given to him. If the reference period is not too long, it is likely that the owner will fill up the details to the best of his knowledge. Even here, it may be necessary to establish a rapport between the field investigator and the owner by a first visit. In any case, the investigator has to go from unit at least once. If no further sampling has to be done within the selected segment, he can deliver the blank questionnaire to all of them at the first visit after explaining the purpose of the survey and the items of data sought. He should also clarify the confidential nature of the data and point out that information for less than three questionnaires together will not be revealed to any party. It is of the utmost importance in this sector to make greater efforts to reduce non-response than in the case of listed establishments, as the non-response rate is likely to be higher here.

22. Where illiteracy is at a high level, however, as it is in many developing countries, where respondents are not well educated in realizing the importance of a survey and are not quite co-operative in carrying out field surveys, and where there is an attitude of apathy towards accounts and figures, even this method may not work. In such a situation, it may be best for the investigator to go to the owners of the units, talk to them and try to elicit the required information as accurately as possible, making use of intelligent cross-questioning to check the veracity of the figures without annoying the respondent and without taking too much of his time. The need for expertise in this field is thus obvious. It is, incidentally, also one of the advantages of staggering the sample over twelve months (or any other period) of a year that less staff is required than if the whole field operation is done at the end of the year within a very small span of time. This staff can be more easily trained and better qualified staff can be employed.

23. Censuses of distributive trades yield useful data for the conduct of sample surveys. These data can serve to build up a classificatory frame. If full censuses of distribution have been taken, furnishing information on the structure of trade, they would provide not only a list of establishments from which a sample can be selected but also information on the distribution of establishments according to size and other characteristics. This type of information is very valuable since it enables one to select establishments with probability proportional to their sales or to stratify the sample according to the size of establishments, thereby increasing the accuracy of the results. It is clear, therefore, that in designing a sample the fullest possible use should be made of census information where it exists.

24. However, even a full census of distribution does not provide an entirely satisfactory sampling frame for surveys of distribution. The turnover of establishments operating in the field of distribution, and in particular in retail trade, as mentioned above, may be considerable,

and a constant sample of establishments, which does not take into account births and deaths of establishments, is very likely to develop bias.

25. When frequent benchmark data with full coverage are available, the sample design is easier and the representativeness of current statistics can be tested more frequently. Even when the current statistics are obtained from a constant sample, and births and deaths of establishments are not (or not fully) taken into account, remedial action can be taken by adjusting the results periodically.

26. A more complete frame than a census (especially an out of date census) may therefore be provided by registers prepared as by-product of administration for example by the authorities responsible for taxation, licensing or regulation of business, or insurance schemes if the administrative authority covers all traders and if the registers are kept up to date. Such registers are more useful if they contain, in addition to establishment addresses, also some indication of their size, e.g., sales, numbers employed, payrolls. There may still be some practical difficulties in taking direct account of births and deaths of establishments; but such registers may enable one to make reasonable estimates of the effect of births and deaths on total sales, even when a constant sample of establishments is employed.

27. A brief outline of a generally applicable sampling plan is described in the following sections, stating problems which are common to sample surveys in many countries.

Division of retail trade into two sub-populations and allocation of sample to each

28. It is convenient to regard the population of all retail establishments in a country as divided into two sub-populations:

(1) those that are included on a list of large establishments; and

(2) those that are not. One can proceed separately to design a sample for each of these sub-populations.

29. Listed establishments will include all those which find a place on some list which may be readily available. Large-scale establishments which may not find a place on the list can also be included in this category as their location will be known by virtue of their size. In a survey, it is useful to treat large and small establishments separately as their characteristics are likely to be basically different.

30. A question that must be considered concerning the sub-division of the population into a sub-population of listed large establishments and a sub-population of all other establishments is: How many establishments should be included in the large establishment list? Certain

optimum sampling principles, which are described later, dictate that the very large establishments must be on the list and suggest that the more establishments included on the list the better it is. On the other hand, certain administrative considerations, such as the difficulty of identifying and matching establishments in order to avoid including some in both sub-populations, together with problems of dealing with births and deaths of establishments, operate to hold the number of establishments included on the list at a minimum.

31. In countries where taxation records are available, a list of large establishments may be prepared with total sales indicated for each establishment. In other countries such a list may be prepared based on social insurance records. Each establishment on the list has an indication of size based on payrolls during a previous period. Of course, the actual current size of each establishment listed is not known, but the wages and salaries paid during a prior period provide an approximate measure of size.

32. The use of past payroll information to distinguish large from small establishments means thatthere will be a number of establishments not on the list that will be larger in current sales than many of those that are on the list. Despite this fact, the use of a sub-population of listed large establishments will be highly effective in increasing the reliability of sample results. With the methods to be followed, if the measures of size used are highly correlated with the actual sales during the current periods under consideration, then the use of the measures of size in classifying the establishments will be highly effective in reducing the sampling error. This statement is true only for a single type of business since the amount of sales per person engaged for an establishment varies a great deal depending on the type of business a given establishment is engaged in; therefore, when following such methods, it may be necessary to stratify establishments by type of business. On the other hand, even if the measures of size are not highly correlated with actual sales, no bias would be introduced into the sample. Thus, the design ought to be such that effective use has been made of available resources, so that their use increases the sample efficiency without in any way biasing the sample.

33. Considerations of optimum allocation make it clear that the sub-population of listed large establishments should be sampled more intensively than those not on the list.

34. It is such considerations that determine the relative allocation of the sample between the sub-population of listed large establishments and the remaining establishments. Moreover, it will be found that by using the optimum allocation method as described above, the sampling error may be reduced very substantially below what would result from proportional sampling from each sub-population. Of course, not all the facts necessary to determine an optimum allocation are available in

advance, and adjustments need to be made in the sample with further experience. However, knowledge of the principles and of the general characteristics of the population to be sampled should make it possible to approximate the optimum allocation in the initial design.

35. It is advisable to settle the details of definition and procedures of determining the listed establishments, and finalizing the actual listing before tackling the sampling design for non-listed establishments. Once the list is evolved, a large establishment coming to notice later on should not be put on the list and should be treated as a non-listed establishment.

Sample from the sub-population of listed large establishments

36. The variation among the sizes of even those establishments on the list of large establishments is very great and again, according to the principles of optimum allocation, gains might be achieved by additional stratification by size of establishments and by using increasingly higher sampling fractions as the size of the establishment increases. Such considerations suggest, also, that all establishments larger than a particular size are of sufficient importance to be included in the sample. There are, however, serious disadvantages in having too many different sampling fractions. The more sampling fractions imposed, the larger the number of different weights that must be used. Moreover, experience has proved that most of the gains of optimum allocation to size groups would be achieved if one sets up only two classes of large establishments, one consisting of the establishments which would be included with certainty, and the other consisting of the remaining listed large establishments. Therefore, once the list of large establishments is prepared, it may be further divided into a very large establishment list from which one makes every effort to obtain returns 100 per cent, and a medium large establishment list from which typically, say, 20 per cent returns may be required.

37. The criteria for dividing the listed establishments into large and medium can be the sales turnover or employment. Listed establishments accounting for a substantial amount of sales or employment may be regarded as large and the rest as medium. What is to be regarded as substantial is a matter for local judgement. Large establishments will be few in number as the distribution of establishments by sales turnover or employment is generally highly skewed and a small percentage of establishments which are large in size account for a large percentage of sales. In view of this, the cost of follow-up or multiple visits to listed establishments will be reduced. The medium establishments will be large in number and more widely scattered and area sampling for these will help to reduce the cost of the survey.

38. Area sampling can be used when no lists are available. It consists of selecting areas - metropolitan area, towns, villages - through probability sampling and make a complete initial listing of all establishments or only distributive trades establishments within these selected areas. Where large areas, e.g., a metropolitan area, and towns and large villages are selected, further area sampling will be needed to save time in complete listing. Generally, population census or local authorities delineate the sub-divisions or segments of these places as wards, census blocks, etc., and a certain number of sub-divisions or segments can be selected. Alternatively, if time permits, mapping of each large area, such as metropolitan area or town, can be initially done, showing principal concentrations of commercial localities, so that by judicious sub-stratification, a more efficient design can be evolved. Discernible areas with a more or less equal number of houses/households or equal area segments can be marked out on these maps and a certain number of these may be chosen for further complete listing of establishments. Establishments in these selected area units can be completely enumerated or a sample of them may be chosen on a probability basis for detailed canvass. This area sampling helps to reduce the cost of the survey.

Stratification

39. In any scheme of sampling, stratification plays a key role. Stratification has to be done to segregate statistical units into different strata in such a way that among themselves they are as heterogeneous as possible and within each stratum, the units are as homogeneous in character as possible. This is mainly achieved by using the available knowledge of the characteristics of a universe of sampling for the grouping of statistical units - primary to ultimate - before the actual sample selection process commences. This leads to efficient estimates, and the payoff in efficienoy will depend on the skill with which this can be achieved. How far such skill, even if available, can be utilized will depend on the extent of the auxiliary data available. Too many strata will, however, involve a lot of additional computational work in estimation as well as variance computations. This is particularly so when the sampling designs are not self-weighting at all stages up to stratum level. Even for separating out the establishments according to a multi-fold classification, more time and effort will be necessary than for a simple stratification scheme. A balance has, therefore, to be struck in deciding how deep stratification should be. It is also important to note that a stratification considered efficient for one characteristic may not be equally efficient or may be even inefficient for another characteristic when several of them are to be estimated through one survey.

40. With this background, the type of stratification may be considered. It needs, however, to be made clear that there is nothing sacrosanct in the suggested scheme and that it is only one possible way of stratifying a large number of stratistical units. A first observation in this regard is the skew distribution of trading establishments, which means that a small number of large-sized establishments account for a sizable proportion of the turnover and other values of related items. Again, the skewness also exists in spatial terms, which means that the distributive trade is concentrated in a few large places and trading centres, although even in small places trading establishments are to be found. A large proportion of the large establishments will be found in large area units such as big cities or towns or big villages. If it is possible to ensure that this small number of large trading establishments and a small number of large area units are separated out and given over-representation in the sampling design as compared to medium establishment and small area units, much of the job of ensuring estimates with greater precision than would otherwise be the case can be said to have been completed. For skew distributions, it is precisely this function which proper stratification can perform. Stratification by area will be necessary when estimates are to be presented for such area breakdowns. A possible mode of stratification is to first have space stratification. In a country, each state can be considered to be a primary stratum. When estimates are required by these geographical area breakdowns, they constitute the domain of study. This will also enable one to use tax records as a frame for deciding which establishments belong to the listed category, even if the tax laws are different. Within each state, each metropolitan area, groups of towns by population size groups and groups of villages in compact geographical areas within the state can be the secondary state. Even for a town, some sort of geographical stratification can be considered instead of by size group if there is reason to believe that such a procedure will ensure less heterogeneity. Such stratification will also ensure a wider spread of the total sample. If some new industrial undertakings have been established in certain parts of rural areas, it may be desirable to treat each of these new industrial areas as separate strata, as the income pattern prevalent in these areas may give rise to a different distributive trade pattern in these areas compared to the rest.

Type of sampling

41. From each secondary area stratum, large listed establishments can be completely enumerated. For medium establishments, in strata other than a metropolitan area, a certain number of towns or villages can be selected according to a pre-determined selection procedure with known probability of selection. Every metropolitan area will no doubt be included. Having selected thesearea units, lists of establishments for these area units can be employed for further sub-stratification within places. These lists are, it may be re-emphasized, the ones

chosen with or without a cut-off point for the purposes of defining listed units. To the list can be added some large establishments (defined in a convenient manner) which may not have been included in the list or which may have newly come up. A check of the locality for those establishments in the list that have disappeared will also help in bringing the list up to date. The listed medium establishments within a selected place, excluding the largest ones which are separately treated, can be first classified according to size groups, the measure of size being turnover or employment in a past period, as the case may be, and by kind of activity and kind of business. If the size characteristic used is highly correlated with the actual sales during the current period under consideration, the stratification of establishments by size will be highly effective. It may be possible to cover one hundred per cent of the establishments some of the large size-groups. The balance of the given number of establishments on the list can be covered through a sample, where the given desired sample size can be distributed to the various sub-strata in proportion to their respective total turnover or employment. The allocation of a given sample size between rural and urban areas can first be decided on the basis of the proportion of turnover or employment (in preference to proportion of number of establishments) in these two divisions. The urban component can be treated as illustrated above. The rural areas can also be similarly treated. If there is geographical stratification of villages, the sample size for each geographical stratum can be allocated on the basis of the proportionate share of size , total turnover or employment, claimed by each region.

42. As the distribution of sales turnover is very likely to be skew, this procedure will help to select a substantial percentage of big establishments which account for a large share of sales turnover. This would not be the case if the allocation to strata were proportional to the number of establishments. If the cut-off point is fairly high, it may not be necessary to have more than two or three size groups and this will simplify the sampling and estimation procedures to that extent.

43. The case where no lists are available has been already mentioned. The efficiency of the design will be improved to the extent that use can be made of census data in not only charting out area segments but also in allocating the size of the sample. If the census data on total of size measure and number of establishments are available, it will be possible to allocate the total sample size to the strata according to the size of the stratum. Then for a given stratum, an overall sampling fraction can be worked out on the basis of the known size of the stratum and the size of the allocated sample. Thereafter the places within stratum and area segments within selected places can be worked out in detail.

44. If no lists and no census data are available, the procedure for sampling has to be a crude one. No doubt, the metropolitan areas, the towns and villages can be selected as already outlined. But the allocation of the sample to various strata, various selected places within the strata present difficulties. It may, however, be a wise course, in view of the known skewness of the distribution of trading establishments, to over-represent the larger places. Thus the metropolitan areas and larger towns may be allocated, for example, twice or thrice as many as would be allocable on the basis of the population proportion. The balance could be distributed to the other groups in the proportion of their population to the balance of total population. Since this would again mean building up a frame for selection of establishments for each selected place, the work can be further limited. It is possible to divide each place into blocks or area segments, either through detailed mapping or on the basis of census of population records, and to select a few of them. Here, local knowledge about principal concentrations of business and trade localities can be fruitfully utilized; for, with this knowledge, the sampling of establishments within such blocks or area segments can be made more intensive. Having selected a number of blocks or area segments, a complete listing has to be made with a view to building up a frame for selection of distributive trades establishments. These listed establishments can be separated out by size, kind of activity and kind of business if this information is also collected in the basic lists. The total number of establishments allocated to a place can then be distributed among the selected blocks suitably. Alternatively, after the complete listing of selected blocks or area segments, all establishments can also be covered in the survey.

45. Sampling within each ultimate stratum may preferably be simple random or systematic (with random start). The complexity of the design can be reduced to the extent that stratification is reduced to a minimum and the design made self-weighting within each stratum at least. This will mean that only stratum-level weighting will be involved.

46. It is, as already mentioned, not possible to lay down a particular design which will serve all purposes most efficiently. When a multi-purpose survey is undertaken involving estimation of several characteristics, very often not quite related in the sense that a design for one would not subserve the purpose of the other, the sampling design may be a form of compromise for several of these characteristics. Such a design is not efficient for a particular characteristic, but it is a balanced design to give a tolerably good estimate for all characteristics. Whether this should be done or whether the design to be evolved will be such as to be efficient for one at the expense of others will depend on the circumstances of each case.

47. In the foregoing, simple random sampling or systematic sampling with random start has been mentioned. This may go well with deep stratification and more complicated sampling procedures may not, in fact, be necessary. However, it is important to state that other procedures for sampling can be employed. For instance, if turnover or employment figures are available for each establishment and therefore for each place, it is possible to have limited space stratification (e.g. by states) and then select places according to probability proportional to size (and with replacement), where size may be measured according to number of establishments, total turnover or total employment. Then sub-sampling of a constant number of establishments within the selected places could be done. If activity-business stratification is possible within the selected places, the constant number fixed for a selected place can be distributed over the sub-strata so as to make the design self-weighting. It is also possible to cover the largest, say first 5 to 10 per cent of establishments on a census basis, sampling being applied to the remaining establishments. Thus, variants of designs can be applied. The important thing to remember is that all available data should be fully pressed into service in evolving a sampling design, and that the design need not be unduly complicated.

48. In collecting information from the selected establishments, it needs to be considered how best the questionnaire can be distributed and the best possible response secured. We have earlier pointed out that the mail survey method, though the easiest and the cheapest one, is of doubtful accuracy as the non-response is generally very high. The largest 5 or 10 per cent of the establishments could be tried out for mail survey methods. The remaining sector consisting as it does of small establishments, an on-the-spot visit by a field investigator may in any case become necessary for a large proportion of establishments that return the questionnaires, as some clarification may be required in respect of the data furnished. Ultimately the purpose of using a mail survey, namely reducing the survey costs, may not be fully achieved. The mail survey method is not the one which can be recommended for this sector in the developing countries.

Joint use of mail and personal follow-up

49. In this method, the questionnaire is sent by mail to the selected establishment with a letter specifying the objectives of the survey and stating that a field investigator will call and collect the questionnaire during a stated period. By that time, the completed schedule or questionnaire may be ready. A further advantage of the personal visit by a field staff member is that the completed schedule can be checked for completeness and for internal consistency of certain key figures.

50. Since each of the very large establishments is to be included in the sample, no matter where located, questionnaires would be mailed to this small group and followed up with field interviews as necessary to ensure substantially complete coverage. The follow-up work will be facilitated by the fact that most of the very large establishments are concentrated in the larger cities and that a prompt mail response can be expected from most establishments in this group and a visit of the field investigator may not be necessary. However, this method will be very important and perhaps will constitute the only choice to obtain response from the medium large establishments.

Use of cluster sampling to facilitate field work

51. Field follow-up of non-respondents to a mail survey is more expensive in the case of the medium large establishments than with the very large establishments, since the former are more widely scattered than the latter. If it were not for this expense, the optimum procedure described earlier, would call for designating the establishments to be included in the sample with no clustering of the sample establishments whatever. To deal with the problem of cost, however, it is highly desirable to use the device of clustering the sample. A sample of areas would be designated and mailing to the medium large establishments would be limited to those located within the selected areas. A cross-section of such areas should be designated throughout the country such that, if the establishments within these areas are adequately represented in the sample, sufficiently reliable information for all establishments should become available. This does not mean that the sample should be adequate to represent well each area separately. It does mean that the sample should be adequate to represent all areas combined. Questionnaires would be mailed to all the establishments on medium large establishment list; then an enumerative follow-up would be made of a sample of those not responding to the mail canvass. The plan should include rotation in the follow-up so that over a period of a relatively few months all of those not responding to the mail canvass would be visited in an effort to build up co-operation to the mail survey.

Selection of areas

52. With the adoption of this cluster sampling approach the question that now remains is: What kind of a sample of areas ought to be selected? The solution to this problem may be fairly easy in some countries where there is a permanent and operating field organization conducting continuing inquiries such as monthly population or labour force surveys. For this type of surveys, a sample of several areas is designated within each of which multi-stage sampling operations are carried out, and a full-time supervisor and full-time and/or part-time enumerators are available.

53. An obvious question to be answered before the adoption of these sample areas for business sampling is: What reason is there to assume that a sample of areas designated for a population or labour force survey will prove to be a satisfactory set of areas in which to take a distribution trade sample? The answer is not too difficult. In the first place, if the design of the population or labour force sample is unbiased, then it follows that unbiased estimates of both population and business as well as of other characteristics can be obtained through a sample survey covering these areas. The real question then is: Will this sample produce results having a sufficiently small sampling error? Since distributive trade exists to serve the population, it is not surprising to find that it is distributed among areas much as the population is distributed and, therefore that a sample designed to estimate labour force characteristics will be comparatively efficient also for estimating distributive trade, if appropriate methods of estimation are used.

54. Labour force and general population samples may be expected to produce distributive trade estimates of reasonable precision, but it should be clear that at least some gains in precision, whether small or large, could be achieved if use were made of a sample of areas designed explicitly for distributive trade rather than for some other problems. The question arises then: Why not design a new sample? The answer to this question has two aspects. The first is that, for the same sample size, some gains undoubtedly could be achieved by a more specialized sample. Nevertheless, it appears that, on the average, for the many different statistics in distributive trade that need to be estimated, the resulting gains would not be large. The second and the principal point is that, with field facilities already existing in the sample areas, the cost of obtaining returns would be substantially less if use were made of the available sample areas than if an independent sample of other areas were selected. Therefore, <u>per monetary unit</u>, the most reliable results are obtained in the manner specified above. If more reliable results are needed, one needs merely to expand that sample.

<u>Sample from the sub-population of the remaining establishments</u>

55. The discussion up to this point concerned sampling only the sub-population of listed large establishments. For those establishments not on the list, which include the great bulk of them, it may be desirable to confine the sample to the same areas that are being used for sampling the listed medium large establishments. In addition to the advantages of minimizing cost of travel, the use of these areas for the remaining establishments makes it feasible to develop and maintain an up to date sample of births and deaths of stores as they occur.

56. Some of the areas may include only a small number of establishments, and in these areas it is feasible to prepare locally a complete

list of establishments in the area and to be currently informed about births and deaths of establishments as they occur. For most of the areas, however, the establishments involved are too numerous to make it feasible to list all of them and to follow all births and deaths of establishments within the areas as they occur, and for these areas a method of multi-stage sampling from the selected areas has to be used involving a further application of an area sampling procedure.

57. This area sampling procedure consists in designating small areas within the selected first-stage areas. The maintenance of up to date listings of the smaller establishment (not included on the large store list) in these areas should reflect the turnover in establishments due to births and deaths. The problem in designating small areas to be sampled within the selected first-stage areas is to locate and utilize the best available information and facilities which would aid in defining efficient higher stage sampling units.

58. For cities of size, say, 25,000 inhabitants or over, detailed maps may provide an excellent source for defining higher stage sampling units. It is desirable that the maps be brought up to date at frequent intervals about once a year, and show the locations of most establishments within blocks in all such cities. Using such maps, it is possible to approximate the number of establishments in each block and draw a sample of blocks stratified by size. An alternative possibility which may be followed sometimes is to designate small areas, which have, according to the map, one, two, or as many establishments as one wishes, for use as higher stage sampling units. First, blocks can be selected by using probability proportional to size in terms of number of establishments; and then each block so selected can be segmented into areas containing an approximately equal number of establishments. Second, one of such areas is chosen at random from each sample block for inclusion in the sample of higher stage sampling units.

59. The methods used in sampling these small areas should be such that any errors in the number of establishments shown on the map would not bias the sample, although the poorer the counts made from the maps, the larger would be the sampling variance.

60. In smaller towns of, say, between 2,500 and 25,000 inhabitants, a different technique may be used. The particular towns of this size to be included in the sample may be obtained by a multi-stage sampling process within the selected first-stage areas and a suitable sampling frame (such as a previous distribution census, social insurance or tax records) may be used in designating the towns to be included in the sample. It is desirable that each town has a chance of selection proportional to the number of establishments in that town. In many of the first-stage sample areas, however, there may be only a few, if any, towns belonging to this group, and all of them may be included in the sample in the first-stage sample area. A random sample of town blocks may be drawn for enumeration outside the main business area

of such towns, and a complete listing of establishment addresses may be made for the main business areas and a sub-sample of establishments or of small areas may be designated for coverage based on information obtained from such a listing.

61. In the case of very small towns and rural areas, minor civil divisions may be used assampling units. For these, it is highly unlikely that previous information on the number of establishments in each area would be available, and therefore, rough estimates of number of establishments in each minor civil division may have to be made. These estimates may be based on the previous population of the areas and on the proportion of that population living inside of towns and villages. Then the actual areas to be included in the sample would be selected with probability proportional to this measure of size, or would be stratified on the basis of this measure of size as has been indicated above in the case of the selection of blocks in the larger cities.

62. Under the above scheme, for cities and towns of all sizes, once a set of small areas is drawn into the sample, all establishments in those areas are listed. The enumerator should be given a map showing the boundaries of the areas and should be instructed to canvas the areas and list every establishment included therein. The advantage of the area approach is that it gives every establishment not on the large establishment list a chance of being drawn into the sample, including both new establishments and those inadvertently omitted from the large establishment list.

63. Although in the above scheme, which is largely illustrative, urban-rural stratification is stressed, there needs to be a great deal of flexibility in the stratification scheme.

The use of rotation samples and ratio estimation

64. The point concerning rotation was briefly touched upon in connexion with the field follow-up of the medium large establishments not responding to a mail inquiry. Rotating samples may be used in monthly and other repetitive surveys of distribution because of one or more of the following advantages:

(i) rotation spreads the burden of reporting among more respondents;

(ii) rotation permits the use of data from past samples to improve current estimates; and

(iii) rotation may make possible an unbiased solution of the problem of large observations which occur in the sample.

65. The advantage of rotation may be so great, that the possibility of rotation should be considered for every repetitive survey. This is especially true if the data being surveyed are such that they are expected to have high period-to-period correlation.

66. The first advantage, that of spreading the burden of reporting during a sample survey among more respondents, could be very important from the standpoint of maintaining a high rate of response.

67. The sample as described earlier can be divided into two main categories , the list sample and the area sample. The list sample consists of very large establishments which have been identified from previous censuses or other administrative records and which are large enough to justify their inclusion in a non-rotating sample to be surveyed periodically. All remaining establishments are represented by the area sample. Each of the small sample areas is divided into say, 12 divisions of equal numbers of establishments and then these divisions are rotated during the field work by enumerating only one set of such divisions every month. Suppose, for the sake of illustration, two months of data (current and preceding ) are obtained from each respondent during each enumeration which is made by personal visit of the enumerator each year.

68. The fundamental principle behind the use of such rotating samples is that, in order to develop the most efficient estimate possible, a search for correlated data should always be made. In order to be useful, these correlated data must be either data from the entire population (universe) or data based on a sample different from that used for the estimate. Then a means of linking these correlated data to the desired estimate must be found. This linking can be usefully done through a sample survey where data on both the items to be estimated and the correlated items are obtained for an identical sample. There may be many ways of using correlated data which may already be available or developing such data when they are not available.

69. The use of rotating samples is one such way of using correlated data and is a direct application of the principle mentioned in the preceding paragraph. If one considers the estimates which can be made from a rotating sample for, say, the month of November, one can of course obtain the simple estimate for the month of November from the November set of divisions. However, one can also obtain an estimate for the month of November from the October division by applying the ratio of the November-October results from the November set of divisions. Progressively less reliable estimates for the month of November can be produced from the September, August, etc., sets of divisions by using products of the month-to-month ratios which can be developed from the sample. Now instead of a single estimate for the month of November, one has a number of estimates for November at practically no additional cost and by a proper weighting of these estimates one can produce a much more reliable composite estimate than

the single simple unbiased estimate.

70. The estimates are usually of the ratio form. Regression or difference estimates can also be used. The regression estimate would result in a gain in reliability if regression coefficients were properly computed. However, computation of the regression coefficients involves considerable labour and where correlations are very high, regression and ratio estimates yield very similar results.

71. The occurrence of large observations is one of the principal problems in sampling applied to statistics of distribution. If the sample is non-rotating, one is usually confronted with the unhappy choice of accepting the considerable increase in variance they create or of taking a bias by arbitrarily reducing their weight. In the rotating system these large observations can be placed in a special set of divisions which can be sampled at heavier than normal rates, thus permitting the weights to be reduced without biasing the results.

72. The principle of this procedure is simple. One identifies in $n-1$ previous sets of divisions these large observations and surveys them in the last set of divisions. The weight of these observations is then divided by $n$ which may drastically reduce their effect on the estimate and the variance of the estimates.

## VI. HOUSEHOLD RETAIL TRADE

73. In addition to easily recognizable establishments, distributive trade sector also includes households which are engaged in commercial activities on their own account and which have no place of business outside the home. In developing countries, household retail trade constitutes a significant segment of distributive trade and also of the economy. In view of the general tendency (already briefly mentioned) to omit small independent traders and rural areas in distributive inquiries, it is particularly important that this vast segment of household retail trade is covered adequately. Sampling inquiries have been found to be the best means of achieving this objective.

74. In the context of sampling inquiries of household distributive trade in developing countries, household trading is defined as trading by individual households or jointly by two or more households as distinct from operations by non-household corporate organizations, co-operative societies and other public bodies. By trading activity is meant here the purchasing of goods from producers and intermediaries and selling them to intermediaries or consumers. In this distributive service no element of transformation of the commodity should be considered to be involved. If any transformation is reported, the activity should be excluded from the scope of trading and included under small-scale manufacturing. In households combining both trading and manufacturing, where the commodities traded are different from the products manufactured, only the trading activity is considered and the common costs are allocated in proportion to respective gross earnings. Intermediaries, who do not actually purchase or sell goods but only arrange for purchases and sales and earn remunerations by way of brokerage or commission, should be excluded from such inquiries.

75. The approach, in such inquiries, to household trade should be through individual households having trade as a principal activity or as one of the subsidiary activities. Generally speaking, a sample of villages and blocks is taken from rural and urban areas, as the case may be, and the investigation should be conducted in a manner such that the sample of villages and blocks is uniformly distributed over the entire survey period. During such inquiries, several types of information may be collected such as:

(i) general particulars;

(ii) fixed capital;

(iii) purchases and sales of merchandise;

(iv) trading costs and other expenses; and

(v) man-days utilized and wages paid to hired labour. The first category of information relates to particulars about type of household trading -- retail or wholesale, perennial, seasonal or casual, etc.

76. It has been found that wholesale trading households form a very minor and insignificant part in the total sample of trading households and hence are not deemed adequate to furnish reliable estimates regarding household wholesale trading. In this connexion, it is convenient to treat retail trading as that part of trading where sales of merchandise are confined to consumers only.

77. The character and the nature of the household trade obviously has a great bearing on the design of the sample survey. In considering the question of a total fixed sample size for this sector of the distributive trade, it is necessary to know how much of the total sample size agreed to for the entire distributive trade sector has been set apart for the non-listed establishments.

78. A generally applicable sampling design for a household retail trade inquiry may be two-stage stratified, with villages and households as the first and second stage sampling units respectively. The rural part of the country may be divided into a number of strata. The total number of sample villages may be allocated to the strata on the basis of population. The allocated number of sample villages in the respective stratum may be selected from the total number of villages in the stratum with probability proportional to population. In the selected village a list may be prepared of households which are usually self-employed in at least one activity of.

(i) manufacture and handicrafts;

(ii) transport; and

(iii) trade, either as principal or subsidiary means of livelihood.

From such a list in each village, sample households may be selected by, say, systematic procedure with a random start.

79. The total sample of villages may finally be split into a number of independent sub-samples, each of which would serve to furnish independent estimates of the items investigated. It may be noted here that, since the sample households may be selected from a frame that may also include households other than those engaged in retail trading, the effective sample for the retail trade inquiry would form a fraction of the total sample.

80. For the inquiry in urban areas also, a two-stage stratified sampling design may be adopted. Blocks and households may constitute the first-and-second stage sampling units respectively.

81. All towns in a country may be grouped into a number of strata. The sample blocks may then be selected from each stratum by using simple random sampling procedure. The selection of households within each selected block may be made on the same line as the selection of households in villages.

82. As in rural areas, in urban areas as well, a number of independent sub-samples may be formed from the total sample of blocks.

83. One difficulty likely to be encountered is the lack of adequate and proper accounts in the household distributive trade. The method of exhaustive field survey may be employed to obtain information from this sector. It is often necessary not only to check the details but also to fill up the questionnaire at the household premises. Owners who do not keep accounts suitable for the purposes of the questionnaire would not fill it up if it were just left with them. It is advisable, therefore, that the field staff member goes out to the selected household, explains the purpose, establishes the necessary rapport and obtains data on the spot or in two or three visits, either on the basis of oral discussions and questions and cross-questions, or where possible through an inspection of such records as may have been kept by the household. This method is of particular use in regard to the household retail trade for the obvious reasons that there is no extra staff in such establishments to devote much attention to filling up the questionnaires, nor are there detailed records on the basis of which data could be filled up.

84. The problem of sampling homeless persons trading on their own account without a fixed place of business is quite complex. One way of dealing with the problem is to carry out the survey of homeless persons engaged in trading activity during the night, when they can be expected to sleep on the pavements. The same sampling procedure adopted for household retail trade can be adopted for these also. It may not be feasible nor perhaps necessary to get very accurate data from these persons as they contribute a negligible portion of the distribution activities. In view of this, a decision to exclude them altogether from the scope of the survey can also be taken.

## VII. NON-SAMPLING ERRORS

85. Whatever refinements are incorporated in a sampling design, the ultimate results could be vitiated if proper care is not taken to control and reduce the non-sampling errors as much as possible. As a general rule, bias entering into an estimate because of non-sampling errors is not easily measurable, and if these are sizable, the results could be misleading. Furthermore, in the presence of such a bias, a meaningful measure of variation cannot be worked out.

86. In order to obtain some insight into non-sampling errors, a small-scale recheck project may be undertaken. If the sample size for such a recheck is small, the results cannot be used as estimates of non-sampling errors, but may be viewed as rough qualitative indications of their nature.

### Response errors in sales

87. In a study of response errors, a field reinterviewer may be asked to secure a previously answered sales figure again. The reinterviewer should be furnished with the figure originally reported and should be instructed to probe into the causes of difference when the figure obtained is different from the original sales figure. The probe should consist in determining whether the recheck figure is a "book" or estimated figure, obtaining ranges where estimates are given, and any comments offered by the respondent in explanation of the differences in sales figures. For those establishments which do not respond originally, estimates have to be made by the collecting personnel. In the recheck, however, some of these establishments would furnish their sales figures.

### Listing errors

88. This type of error occurs when an interviewer fails to record an establishment which he should record or when he records an establishment which he should not. Small areas should be checked by giving the reinterviewer the original listing, instructing him on how to find errors (with appropriate safeguards being taken so that no interviewer checks his own work), and budgeting twice as much cost per establishment to ensure a thorough job. In a large number of cases the errors result from wrongly interpreting small area boundaries.

### Scope differences

89. After an interviewer lists an establishment, he obtains information on the nature of its activity, and on the proportion of its receipts which result from sales at retail, from sales at wholesale, and from receipts for personal services. He then makes a judgement on whether the establishment is within the scope of the inquiry. Later his work is reviewed at the office of the collecting agency, and his judgement may be reversed. A final determination of "scope" is then a function

of the inquiries made of the establishment, the responses received, the judgement of the interviewer obtaining the information, and the judgement of the particular reviewer of the interviewer's decision. Furthermore, the same interviewer or the same reviewer may arrive at different decisions (in a certain proportion of the cases) at different times. It is clear that, when two decisions are different, it is frequently possible that neither is clearly in error and consequently reference is made here to "scope differences" rather than to "scope errors".

90. In the recheck, responses to questions pertaining to scope should be obtained independently of the original enumeration, and this information should be reviewed and coded independently.

Kind-of-business differences

91. After a judgement is made that an establishment is within the scope of a given inquiry, a further judgement is made as to the kind of business (e.g., grocery store, chemist's shop etc.) to which it belongs. The kind-of-business code assigned to an establishment is, in general, an attempt to characterize the pattern of the money volume sales of different kinds of commodities. Few establishments, however, keep records of their sales of different commodity classes, and consequently their estimates must be accepted. Evidence obtained from surveys of distributive trade shows that the differences in responses are not the only cause of kind-of-business differences. Even with the same information on hand, different coders are likely to assign different codes and even the same coder may assign different codes at different times. It is not surprising, therefore, that two independent surveys, in which a kind-of-business code is assigned to the same establishment, will show substantial differences.

Additional remarks on the non-sampling errors

92. The effect of response errors in sales is small in relation to the effect of differences arising from the other elements discussed, particularly scope and kind-of-business determinations, which are a complex result of many variable elements, such as the type of information requested, the availability of information, differences in the ability of interviewers, differences in coders' judgements.

93. If careful work is not performed, under-enumeration of the area sample may be a serious type of error. The area sample enumeration is carried out by enumerators who have in their possession the enumeration carried out on a previous occasion. The old enumeration should serve as a starting point, which is to be brought up to date and the differences should be explained. Examination of the resulting enumeration should furnish evidence that by this means boundary line errors are reduced to the level of the probability of two different interviewers making the same mistake.