47210

UNITED NATIONS
ECONOMIC AND SOCIAL COUNCIL

Distr.
LIMITED
E/CN.14/SM/24
11 September 1979
Original: ENGLISH

ECONOMIC COMMISSION FOR AFRICA

Working Group on Organization, Content and Methodology of Household Surveys

Addis Ababa, 15-19 October, 1979

SOME COMMON SAMPLING SCHEMES, THEIR ADVANTAGES AND DISADVANTAGES

Content

Introduction

1. This paper examines only two aspects of household survey designs under African conditions but, in doing so, also takes into account some other related practical considerations. It is based on recent experience in two African countries and, although the information is in some respects incomplete, the paper should be able to highlight a few matters likely to be of general concern in organising surveys.

2. The first topic considered is the selection of units in the area stages of household samples with probabilities proportional to size (PPS) as compared with equal probabilities (EP). PPS arrangements have obvious attractions but a number of pitfalls are beginning to emerge and the paper argues that EP samples may be more satisfactory in the earlier stages of survey development if they can be properly organised.

3. Secondly there is the consideration that African countries have resources to investigate only fairly small samples, even at national level, and there is a need for the samples to perform as efficiently as possible. In the case of household economic surveys this can be achieved by introducing an income or similar stratification with variable sampling fractions at the ultimate sampling stage. The paper examines the prospects and problems so far indicated by African experience and then goes on to a consideration of the related estimation procedures and standard error calculations.

Selection of units at the area sampling stages

General considerations

4. When a developing country prepares a sample having national coverage, the principal factors which influence the work are probably convenience, durability and multi-subject utilisation. It is necessary to establish the sample without unduly expensive field operations and its area stages at least should serve satisfactorily for a few years in collecting data on a variety of topics.

5. Most African samples are based on frames derived from a previous population census or from administrative structures. Both of the countries considered here have taken a recent population census but neither has used the census enumeration areas for sampling because mapping was not adequate and in one country there was a significant re-organisation of local government after the census. In both cases the samples are based on administrative structures and new units have been established at subsequent sampling stages. The intention of the two countries is to extend their survey frames to provide a more satisfactory geographical basis for the next population census.

6. The two countries are referred to as A and B and their samples are described in the following notes. The discussion is limited to arrangements in the rural areas. The notation is a little different from that normally used and an explanation is given in a footnote to the description of the Country B sample.

Country A sample

7. The sample has four stages: location (a), chunk (k), cluster (c) and household (h).

8. There is no precise description of the arrangements but selection chances at the various stages (not taking into account first-stage ecological strata) generally appear to have been as follows:

(1) Any one of n locations (PPS): $n \times c_a/C$ (where c_a is the expected number of clusters in the selected location and C is the estimated total number of clusters in all locations).

(2) Any one of two chunks in selected location (PPS): $2c_k/c_a$ (where c_k is the expected number of clusters in selected chunk).

(3) Any one cluster in selected chunk (EP): $1/c'_k$ (where c'_k is the number of clusters in selected chunk established after an enumeration of the chunk).

(4) Any household in selected cluster (EP): f (where f is the ultimate-stage sampling fraction).

9. Then the overall selection chance for any household is (1) x (2) x (3) x (4) which cancels out to $2n/C$ x f and the sample is self-weighting provided c'_k is equal to c_k or the correction discussed below is applied.

10. There are two general points worth noting about this sample. The second stage comprising "chunks" was not envisaged in the original sample design but had to be introduced because it proved impracticable to delineate clusters throughout the whole of selected locations. Also the number of households selected in each sampled cluster is dependent on the cluster size if the sample is to be self-weighting; enumerators' workloads could have been held constant irrespective of variation in cluster sizes if the size estimates in the earlier stages of the sample had been expressed in numbers of households rather than clusters, which would have enabled a PPS selection at the third stage. No doubt there were good reasons for the arrangement actually used.

11. Estimates so far obtained from surveys using the sample have been lower than expected and, as a temporary measure, it has been necessary to introduce an ad hoc system of raising factors. There appear to be three kinds of problem which may be affecting the processing of results from the sample:

(1) Estimated numbers of clusters. C was estimated from population census data and, at the time locations were selected, no geographical units existed which corresponded with the concept of clusters. The figure C is therefore somewhat nebulous in the sense that it is not physically related to the sample frame. The same consideration applies to c_a and c_k. However there would be no problem from the estimates of numbers of clusters if arithmetical consistency was maintained throughout the first and second stage selection i.e. if C was equal to the sum of c_a and c_a was equal to the sum of c_k.

(2) Chunk enumeration. After selection, chunks were enumerated for the purpose of sub-dividing them into clusters identifiable on the ground. In doing this c'_k should have been kept equal to c_k i.e. the number of clusters actually formed in each chunk should have been the same as the number

originally estimated, which would not have been too difficult if the original census data were satisfactory. Apparently this was done, with the exception of a few cases which required special treatment. The point to be noted here is that, with a small number of clusters per chunk, any difference between c_k and c'_k would affect the self-weighting nature of the design. In such an event the two alternatives would be (a) to apply c_k/c_k as a weight to survey results at cluster level or (b) to use a variable sampling fraction (c'_k/c_k x f) in selecting households to put the sample back again on a self-weighting basis.

(3) Deviations from basic design. For practical reasons there were deviations from the basic design in a number of first-stage units. The position is being examined in detail and the current feeling is that the deviations may be responsible for the performance of the sample rather than the arithmetical aberrations suggested above. Perhaps the main points to be borne in mind are (a) the existence of such deviations makes the sample non-self-weighting, so weights at the cluster level have to be applied and (b) in working out a system of weights it is necessary to take into account not only the specific deviations but also any other adjustments of the more general kind noted above.

12. The position with respect to points (1) and (2) above could be determined by checking the internal consistency of the estimated numbers of clusters down to the chunk level and by comparing the numbers of clusters actually created with the estimated numbers of clusters used for the PPS selection of chunks. As indicated above, the position with respect to (2) seems to be fairly clear but records are not sufficiently detailed for an adequate check on (1). Additional investigations are concentrating on accuracy of the physical demarcation of chunks and clusters in the field and the accuracy of the household listing.

13. There is no doubt that country A has a good sample but there have been some shortcomings in recording the details of its construction and the arrangements seem unduly complex. It can be made to work effectively after detailed investigation of the selection procedures.

Country B sample

14. The sample design used for the rural sample of the survey involved three stages of selection: councils (c), segments (s), and households (h). It was first necessary to prepare a complete list of councils, with information on their locations and their estimated numbers of households (h_c). The work was carried out in collaboration with the local government authorities, and to some extent the size data could be checked against population census records, though often such a check was not possible because there had been changes in the local government boundaries since the last census.

15. The numbers of households shown in this council listing were used in the PPS selection of councils. The selected councils were then visited, and divided up into segments. The field work involved in this segmentation exercise provided new council estimates, based on the summation of the estimated number of households in individual segments within each council. One segment was then selected with PPS in each of the sample councils, using the new estimates. A listing exercise was then carried out to enumerate all the households in the selected segments, and the resulting figures were used in selecting the third stage sample of households.

16. With this sampling scheme the selection chances[1/] at each stage are as follows:

1/ Here it is necessary to explain the simplified notation of this paper in relation to the more usual presentation. The latter would define h_{ijk} as the k^{th} household in the j^{th} segment of the i^{th} council, where $i = 1, 2, \ldots\ldots N_c$

$j = 1, 2, \ldots\ldots s$

$k = 1, 2, \ldots\ldots h_s$

Then $\sum_{k=1}^{h_s} h_{ijk} = h_{ij.}$; $\sum_{j=1}^{s} h_{ij.} = h_{i..}$ and $\sum_{i=1}^{n} h_{i..} = h_{...}$.

In the notation of the present text $h_{ij.} = h_s$, $h_{i..} = h_c$ and $h_{...} = H$, while necessary explanations are incorporated in the text.

(1) Any one of n councils to be included in the sample (selected with PPS): nh_c/H (where h_c is the first estimate of the number of households in the selected council, and H is the estimated number of households in all rural councils, i.e. the sum of h_c).

(2) Any segment in selected council (PPS): h_s/h'_c (where h_s is the first estimate of the number of households in a selected segment, and h'_c is a revised estimate of the number of households in that council based on a summation of the size estimates for the different segments in that council).

(3) Any one of m households in selected segment (EP): m/h'_s (where m is a constant, and h'_s is the number of households in the selected segment obtained from an enumeration).

17. The overall selection probability for an individual household in the sample is therefore given by the product of the selection chances at the three stages: i.e. $nh_c/H \times h_s/h'_c \times m/h'_s$. This can more conveniently be written in the form $nm/H \times h_c/h'_c \times h_s/h'_s$.

18. To raise the sample figures to the population level, it would be necessary to multiply the sample data by the inverse of this probability. But to provide results comparable with those obtained from a sample of the same size as that actually used, one must multiply the population level figures by an additional factor, nm/H, i.e. sample size divided by number of households in total population. The overall reweighting factor thus becomes $H/nm \times h'_c/h_c \times h'_s/h_s \times nm/H$. The first and last fractions cancel out, as one would expect in a PPS sample, leaving a weighting factor of $h'_c/h_c \times h'_s/h_s$; this is the correction factor arising from the use of different size estimates at the different stages of sample selection.

19. If there is little difference between h_c and h'_c, and between h_s and h'_s, it may be reasonable to ignore the correction factor, and treat the sample as self-weighting. If however there are big differences in the size estimates, as was found to be the case in country B, then the correction factors must be applied.

Problems of PPS sampling

20. PPS designs are useful in providing a convenient means of accommodating variations in unit size at the preliminary area stages, while at the same time ensuring an equal distribution of work between enumerators and a self-weighting arrangement. However recent experience has brought to light a number of disadvantages which are listed below:

(1) As indicated in the descriptions of the two country examples above, it is possible for inconsistencies to occur in size estimates used in selecting the various stages of a sample. Such inconsistencies could invalidate the self-weighting nature of the sample and make it difficult to calculate remedial weights if they are not fully documented. It will be possible to evaluate the extent of the problem if the countries concerned can provide information on the size data actually used at each sampling stage.

(2) At the penultimate stage there is the disadvantage of an increased amount of work in enumerating households because the sample concentrates on the larger area units. This is probably not very serious because, if penultimate units are specially created, they need not vary too much in size.

(3) The household enumeration is used for the collection of general economic, social and demographic data, plus information for stratification of the ultimate household sample if needed. However selection probabilities for the penultimate units are not equal, so weights have to be applied to the data before summary tables can be prepared. Even by computer, the process is a little cumbersome. The weight to be

applied to the results for each segment in Country B would be $\bar{h}_s/h_s$ [1/] where $\bar{h}_s$ is the average number of households in sampled segments.

(4) Although a properly organised PPS sample leads to an unbiased selection of households, some statisticians still have misgivings because it concentrates on the more densely populated and often poorer area units.

21. In spite of these reservations there is no doubt that PPS samples are a very useful statistical tool. The main problem appears to be illustrated by the first of the points mentioned above: unless the work is very carefully planned and controlled, there are arithmetical pitfalls which can cause serious damage.

An alternative arrangement

22. In the situation described above it may be useful to examine the possibility of an alternative arrangement which attempts to avoid the problems already noted without introducing too many new ones. The main requirement is the acceptance of some additional work in preparing the units at the first area stage so that the size data do not have to be used directly in sample selection.

23. The position is examined below on the basis of the Country B rural illustration. Again there are three stages: council (c), segment (s) and household (h).

24. The first part of the work is to modify the council frame so as to reduce the variation in the number of households per unit as far as possible. It involves the same council listing exercise previously described, including the information on council locations and approximate numbers of households, using both census records and local sources. More equal sizes are then achieved by sub-dividing large councils and grouping small ones with neighbours. If care is taken to

---

1/ This is the inverse of (1) x (2) in paragraph 14 which is H/nh_s multiplied by $n\bar{h}_s/H$ for arithmetical convenience.

avoid bias, this can initially be done as a desk job and sub-division of larger councils in the field can be confined to selected units.

25. The selection of councils is made on an EP basis with a selection chance of n/C where n is the number of councils to be selected and C is the number in the modified frame.

26. Selected councils are divided into segments each containing around 100 households as before. One segment is drawn from each council on an EP basis with a chance of $1/s_c$ where s_c is the number of segments in the particular council.

27. In this arrangement the average selection chance for any segment is $1/\bar{s}_c$ but the chance for any segment in a particular council differs from the average by s_c/s_c. The difference calls for a correction at the third stage.

28. The third stage sampling fraction is then $f \times s_c/\bar{s}_c$ and the overall selection chance for any household is $n/C \times f/\bar{s}_c$ where f is the number of households required for the sample divided by the number enumerated at the second stage.

29. The following numerical example of a few councils serves to illustrate how enumerators' workloads can vary in the sample.

h_c	s_c	h_s	$f \times s_c/\bar{s}_c \times h_s$
600	6	100	24
600	10	60	24
600	5	120	24
1,000	6	166	40
1,000	10	100	40
500	5	100	20
500	10	50	20

overall sample data

$\bar{s}_c$: 6

$\sum h_s$: 12,000

f : 0.24

30. It will be seen that the number of segments into which a council is divided has no effect on workloads because $s_c/\bar{s}_c$ and h_s are compensatory. The important requirement is that all segments within a council should contain approximately

equal numbers of households. This should not be too difficult to achieve and then workloads are proportional to numbers of households in councils.

31. Success of this sampling arrangement is therefore dependent on the extent to which the frame of councils can be modified in the way described above. Possibilities in this respect can be seen at a very early stage in the work, i.e. as soon as the list of councils with approximate size data is available. If modification does not appear to be feasible, then there is little choice but to proceed with a PPS design.

32. The foregoing example is consistent with the aim of Country B to use field staff as economically as possible by selecting a sample of councils and investigating only one segment in each council. It has the advantage of ensuring that area unit size estimates have no effect on selection probabilities but it does not overcome the third problem mentioned in connection with PPS samples. The segments are selected with unequal probabilities and the results of the household enumeration still need to be weighted before summary tables are produced. In this case the weight for each segment would be $Cs_c/n \times n/C = s_c$.

33. Sampling arrangements would of course be much easier if it were possible to avoid the earlier area sampling stages and go straight to what is currently the the penultimate stage (segments/clusters/enumeration areas). With a few fairly important reservations relating to communications and logistics, this could be achieved in samples of national coverage as soon as African countries appreciate the need to use their population censuses in establishing satisfactory geographical frames and make the effort to up-date these frames on a continuous basis. One country is already organising a survey with a two-stage design and another is considering the possibility of doing so.

Strata at the ultimate sampling stage

General considerations

34. As indicated in the introduction, there is an interest in stratifying the ultimate stage sample of households in economic surveys in order to improve the precision of results. Strata based on income have already been used in a number of African surveys.

35. However it should be noted that income strata are not supported in the international recommendations on income, consumption and accumulation, Series M No. 61, which appear to envisage a more conservative approach based on fractile income groups and socio-economic class of head of household.

36. Of the two samples discussed in this paper, only that in Country B has an ultimate stage stratification. The method is described below and again relates to the rural sample.

Country B strata

37. It was decided to use three strata based on declared cash income and to have fixed strata boundaries for the whole of the rural area, determined so as to allocate an equal share of total cash income to each stratum. Information for this purpose was collected in the penultimate stage enumeration and consisted of income data for individual household members plus household income from agriculture and other entrepreneurial activities. The desirability of including subsistence consumption and other items of income in kind was appreciated but it was not considered practicable to request such data during a single-visit enumeration.

38. The enumeration records were sent to the central statistical office where they were sorted, for each penultimate unit (segment) separately, according to total household cash income. From a manual summary of the overall results it was then possible to determine the strata boundaries, assign households to the three strata and calculate the overall strata sampling fractions so as to allocate equal parts of the predetermined sample size to each stratum. A more detailed description is given in E/CN.14/CAS.10/18.

39. The strata sampling fractions could not be applied directly in selecting the final stage households because of the PPS design and the need to take the size variation of the penultimate units into account. With stratification, the selection chances at the three stages of the sample, as given in paragraph 16, can be re-written as follows:

(1) Any one of n councils (PPS): nh_c/H

(2) Any segment in selected council (PPS): h_s/h'_c

(3) Any household in each of the three strata in selected segment (EP):

$$\begin{aligned} \text{Stratum 1} &: f_1 \times \bar{h}'_s/h'_s \\ 2 &: f_2 \times \bar{h}'_s/h'_s \\ 3 &: f_3 \times \bar{h}'_s/h'_s \end{aligned}$$

(where $\bar{h}'_s$ is the unweighted average number of households per segment obtained from the enumeration of selected segments, and f_1, f_2 and f_3 are the overall strata sampling fractions calculated as described in paragraph 38).

40. The third stage selection chances multiplied by the numbers of households enumerated in the three strata in individual segments indicate the numbers of households drawn for the final stage of the sample. The overall selection chances for any household in each of the three strata was:

$$\begin{aligned} \text{Stratum 1} &: nh_c/H \times h_s/h'_c \times f_1 \times \bar{h}'_s/h'_s \\ 2 &: nh_c/H \times h_s/h'_c \times f_2 \times \bar{h}'_s/h'_s \\ 3 &: nh_c/H \times h_s/h'_s \times f_3 \times \bar{h}'_s/h'_s. \end{aligned}$$

The selection chance for any household in the first stratum can conveniently be re-written as $n \times f_1 \times \bar{h}'_s/H$, $\times\ h_c/h'_c \times h_s/h'_s$, and similarly for the second and third strata, with f_2 and f_3 replacing f_1.

41. If different size estimates for councils and segments had not been used at the different selection stages the sample would have been self-weighting at the stratum level, since the first fraction is a constant. In fact, if the difference between the size estimates is not too large, it may be reasonable to ignore the correction factor. However, this was not possible in the case of the Country B data, and the following approach was adopted.

42. Considering just the first income stratum, the third sampling fraction in the formula given in paragraph 39 was $f_1 \times \bar{h}'_s/h'_s$; but it can equally well be obtained by dividing the number of households in that segment actually selected in that income stratum (m_1) by the number enumerated in that income stratum at the listing stage (h_{s1}). The overall selection chance for any household in the first income stratum can then be written as $nh_c/H \times h_s/h'_c \times m_1/h_{s1}$.

43. Following the approach already described in paragraph 18, the sample figures could be raised to the population level by multiplying the sample data by the inverse of this probability. But to provide results comparable with those obtained from an unstratified sample of the same size as that actually used, one must multiply the population level figures by an additional factor, $n\bar{m}/H$, i.e. sample size divided by number of households in total population. The overall reweighting factor thus becomes $H/nh_c \times h'_c/h_s \times h_{s1}/m_1 \times n\bar{m}/H$, which simplifies to $\bar{m}/h_s \times h'_c/h_c \times h_{s1}/m_1$.

44. It should be noted that this reweighting factor contains allowances, not just for the correction factor required because of the use of different size estimates, but also for the weights to be applied because of the use of variable sampling fractions for the three income strata. The advantage of using the composite weighting factor is that the weighting figures for the different strata can be combined directly, without having to apply another weight. However there may be problems in using the composite factor, arising from the variations in m and the ommission of the factor H.

Possible problems

45. The stratification method described above could encounter a few problems of the kind indicated in the following notes:

(1) The use of fixed strata boundaries requires the centralisation of all penultimate stage enumeration records for the preparation of overall summary tables, calculation of strata boundaries and selection of the household sample. The records then have to be returned to the field

for confirmation of the sample and the insertion of substitutes if necessary. Such a procedure is clearly tedious and can be affected by transport and other problems. Against this, there is the advantage of a thorough centralised review of the household enumeration results and, if the manual processing is carried out by field supervisors who bring the records to the central office, these people have an opportunity of gaining a better understanding of the survey at one of its most crucial stages and can be given proper training on procedures for the ultimate-stage sample.

(2) It is generally believed that data on income are difficult to collect during a single-visit enumeration. This is certainly true of censuses and other inquiries which aim to make direct use of such information but the position is not borne out by experience in surveys where the data have been collected for stratification purposes. More research is needed into income and alternative indicators but there seems to be no reason for pessimism.

(3) The use of household cash income rather than total income as a means of stratification could lead to some problems in the rural areas, even though it apparently worked satisfactorily in urban areas. If subsistence consumption is mainly a characteristic of poorer households, its exclusion from the stratification data would result in a rather large number of households in the lowest income stratum, which is perhaps not too serious and could be dealt with through somewhat different strata sampling fractions from those in paragraph 38 or an arbitrary adjustment of strata boundaries. In the event of subsistence and other non-cash income being closely related to cash income, there would of course be no problem. Difficulties would occur in a situation where high income households receive a large proportion of their income in non-cash items, which seems rather unlikely. In the survey processing it is possible to check the effectiveness of the stratification by comparing it with the income data obtained during the detailed recording of transactions.

(4) The method involves some inequality in enumerators' workloads because of the variation in household income distribution between penultimate stage units. The problem does not appear to be too serious and could be overcome only by conducting the household enumeration at primary unit level, which would be prohibitively expensive. (An alternative method of adjusting the sample at the ultimate stage is too cumbersome to be considered here).

An alternative arrangement

46. The main alternative currently available is the use of percentile income groups as a basis for stratification, possibly involving a 2:3:5 allocation of households. In this case the selection chances given in paragraph 16 would be modified to read as follows:

(1) Any one of n councils (PPS): nh_c/H

(2) Any segment in selected council (PPS): h_s/h_c

(3) Any one of m/3 households in each of the three strata in selected segment (EP) (where m is the constant used in paragraph 14):

Stratum 1: $m/3 \times 1/0.2h'_s$

2: $m/3 \times 1/0.3h'_s$

3: $m/3 \times 1/0.5h'_s$

47. Without making any adjustment for possible differences between h_s and h'_s, the overall selection chance for any household in each of the three strata is:

Stratum 1: $n/H \times m/3 \times 1/0.2$

2: $n/H \times m/3 \times 1/0.3$

3: $n/H \times m/3 \times 1/0.5$

48. This is a neat self-weighting sample and various versions have been used in Africa. It avoids the need to centralise enumeration records for processing and sample selection and it ensures that all enumerators deal with the same number

of ultimate stage households. However the method no longer seems attractive because its overlapping strata are inefficient and selection of the household sample in the field leads to loss of control over the survey operation.

Application in non-PPS sample

49. To complete the review of ultimate-stage stratification within the limited context of this paper, it is necessary to consider how strata defined by fixed boundaries could be incorporated in the sample described in paragraphs 22-32.

50. Selection chances at the three stages would be as follows:

(1) Any one of n councils (EP): n/C (where n is the number of councils to be selected and C is the number in the mofified frame).

(2) Any one segment in selected council (EP): $1/s_c$ (where s_c is the number of segments in the selected council).

(3) Any household in each of the three strata in selected segment (EP):

$$\text{Stratum 1: } f_1 \times s_c/\bar{s}_c$$
$$\text{2: } f_2 \times s_c/\bar{s}_c$$
$$\text{3: } f_3 \times s_c/\bar{s}_c$$

51. The three strata sampling fractions would be determined by the procedure described in paragraph 38. Overall selection chances for any household in each of the three strata would be:

$$\text{Stratum 1: } n f_1/C \bar{s}_c$$
$$\text{2: } n f_2/C \bar{s}_c$$
$$\text{3: } n f_3/C \bar{s}_c$$

52. This is certainly the simplest of all the arrangements discussed in the present paper but, as emphasised in paragraph 31, it can be used only in situations where the frame of primary units can be modified for statistical purposes. There are of course many variations on the same theme.

Estimation of population values and standard errors

Country A sample

53. In paragraph 7-13, the country A sample is described as being selected in four stages:

(1) n locations with PPS

(2) two chunks with PPS from each location

(3) one cluster with EP from each selected chunk

(4) some households with EP from each selected cluster, with sampling fractions f

54. The probability of selecting a household in chunk (k) of location (a) is:

$$n\, c_a/C \times 2c_k/c_a \times 1/c'_k \times f$$

$$= 2nf/C \times c_k/c'_k \qquad - (1)$$

55. To find an estimate in that location (a), sum $[c'_k/c_k \times$ sample values in chunk $k]$ with similar products in the other chunk of the same location. Let such a value for the location (a) be denoted by:

$$y_a \qquad - (2)$$

56. As n locations have been selected, there will be n such values in the sample so that their sum will be:

$$\sum_{a}^{n} y_a \qquad - (3)$$

57. The estimate for the total population value will be: $\hat{Y}_A = C/2nf \times \sum_{a}^{n} y_a \qquad - (4)$

58. The standard error of $\hat{Y}_A$ is approximated by:

$$\sqrt{\hat{V}(\hat{Y}_A)} = C/2f \sqrt{1/n(n-1) \left[\sum_{a}^{n} y_a^2 - 1/n \left(\sum_{a}^{n} y_a\right)^2\right]} \qquad - (5)$$

59. To compute $\hat{Y}_A$

(i) in each location - find the sample values in chunk k and multiply it by c'_k/c_k i.e. c'_k/c_k x sample values in chunk k and then add this to similar product for the other chunk in the same location, and denote this sum by y_a

(ii) sum y_a overall selected n locations i.e. $\sum_a^n y_a$

(iii) multiply $\sum_a^n y_a$ by C/2nf to get

$$\hat{Y}_A = C/2nf \times \sum_a^n y_a$$

60. To compute standard error $V(\hat{Y}_A)$,

(i) find y_a^2 i.e. square of y_a in each location and sum them overall selected n locations to get $\sum_a^n y_a^2$

(ii) square $\sum_a^n y_a$ to get $(\sum_a^n y_a)^2$ and find $1/n \ (\sum_a^n y_a)^2$,

(iii) subtract $1/n \ (\sum_a^n y_a)^2$ from $\sum_a^n y_a^2$ i.e. $\sum_a^n y_a^2 - 1/n \ (\sum_a^n y_a)^2$ and multiply it by $1/n(n-1)$ i.e.

$$1/n(n-1) \left[\sum_a^n y_a^2 - 1/n \ (\sum_a^n y_a)^2 \right]$$

(iv) find the square root of this and on multiplying it by c/2f the standard error of $\hat{Y}_A$ will be obtained.

Country B sample

61. As described in paragraphs 14-19 and 37-44, the country B sample is selected in three stages:

(1) n councils with PPS

(2) one segment with PPS from each selected council

(3) stratified EP sampling having divided each selected segment into 3 income strata.

62. It is shown in paragraph 42 that the overall selection chance for any household in the first income stratum of a council (c) is:

$$nh_c/H \times h_s/h'_c \times m_1/h_{s1} = n/H \times h_c h_s/h'_c \times m_I/h_{s1} \quad - (6)$$

63. To find the estimate in that council, first calculate: h_{s1}/m_1 multiplied by sample values in stratum 1, and similar values for the other two strata, then sum them. Multiply this sum by $h'_c/h_c h_s$ and denote this product for council (c) by: y_c - (7)

64. As n councils have been selected there will be n such values in the sample and their sum will be: $\sum_c^n y_c$ - (8)

65. The estimate for the total population value will be:

$$\hat{Y}_B = H/n \sum_c^n y_c \qquad - (9)$$

66. The standard error of $\hat{Y}_B$ is approximately by:

$$\sqrt{\hat{V}(\hat{Y}_B)} = H\sqrt{\frac{1}{n(n-1)}\left[\sum_c^n y_c^2 - \frac{1}{n}\left(\sum y_c\right)^2\right]} \qquad - (10)$$

67. The procedure for computing $\hat{Y}_B$ and its standard error will be almost the same as that described for the country A sample, except that to get y_c for a council:

(i) find sample values in a stratum (i) and multiply it by h_{si}/m_i and sum these values over all three strata

(ii) multiply this sum by $h'_c/h_c h_s$ to get y_c.

Concluding comment

68. The paper summarises questions which have been discussed recently in respect of two African samples. At the present stage there is clearly not enough information available for any firm conclusions but there are a few points which survey statisticians ought to keep in mind.

69. Although PPS samples are convenient, the selection process requires great attention to detail and it is possible to make mistakes which are not readily apparent. When PPS samples are used it is clearly important that full details of the probability calculations for all units at every sampling stage should be kept on record.

70. To avoid some of these complications when organising a sample, it is worthwile first to look and see whether it would be possible to modify the first stage frame to make it suitable for EP sampling. Then unit size data would have no direct role in the selection procedure. Admittedly there are objections to this method on the grounds of additional preparatory work and the undesirability of sub-dividing or amalgamating units in the national administrative structure for statistical purposes. However the additional work may be worthwhile in producing an easier sampling arrangement and modification of administrative units is hardly important in a situation where surveys are not likely to provide valid results below the regional level.

71. A further simplification of sampling arrangements can be achieved if it is possible to use a two stage sample with the smallest area units (enumeration areas) as the first stage. The requirements are a comprehensive frame, normally established through a population census operation, and continuing means of up-dating the frame, plus a solution to transport and related problems. Some African countries already believe that two stage samples are practicable.

72. Stratification by income or some other indicator of household economic level is desirable to improve the accuracy of surveys dealing with household transactions, etc. but further development work is needed to identify the most reliable indicators and the feasibility of collecting the basic data for them during the preliminary household enumeration. The strata make the processing of survey results a little more complicated, but three strata are usually sufficient and the effort is rewarding in terms of improved accuracy.

73. It should be understood that the discussion in this paper is of a general nature and in no way provides a blueprint for any of the sample designs mentioned. Arrangements invariably need to be worked out in detail in the light of local conditions and data requirements; every survey is a tailor-made operation if it is to be of any use at all. This remark does not preclude the use of reasonably durable national samples which are utilised on a multi-subject basis.

74. As pointed out at the beginning, the paper has been limited to a discussion of two main sampling topics, with additional notes on estimation of population values and standard errors. A more complete description of the latter is given in Annex I.

Mathematical estimation of population values and standard errors

I. General estimation

(a) In multi-stage sampling suppose that n first-stage units out of N are selected with PPS such that π_i is the probability of selection of i^{th} unit. If $\hat{Y}_i$ is the estimate of total value of a variable for that unit based on the values observed in the successive stages in that unit then estimate for total population value is

$$\hat{Y} = \sum_{i}^{n} \frac{\hat{Y}_i}{\pi_i} \qquad - (11)$$

And variance of $\hat{Y}$ is $V(\hat{Y}) = \sum_{i}^{N} \sum_{j>i}^{N} (\pi_i \pi_j - \pi_{ij}) \left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j} \right)^2 + \sum_{i}^{N} \frac{1}{\pi_i} V(\hat{Y}_i) \qquad - (12)$

where $V(\hat{Y}_i)$ = variance of $\hat{Y}_i$ consisting of variation in the successive stages following the first-stage

and π_{ij} = joint probability of selection of i^{th} and j^{th} first-stage units

then estimate of this variance is

$$\hat{V}(\hat{Y}) = \sum_{i}^{n} \sum_{j>i}^{n} \frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \left(\frac{\hat{Y}_i}{\pi_i} - \frac{\hat{Y}_j}{\pi_j} \right)^2 + \sum_{i}^{n} \frac{\hat{V}(\hat{Y}_i)}{\pi_i} \qquad - (13)$$

where $\hat{V}(\hat{Y}_i)$ = estimate of variance of $\hat{Y}_i$ consisting of variation in the successive stages following the first stage.

Estimate of total value $\hat{Y}$ and standard error $\sqrt{\hat{V}(\hat{Y})}$ are required in sample surveys. However in equation (13), it is rather difficult to find π_{ij} when $n > 2$, although there are tedious estimates given by several statisticians. Therefore in practice, it might be more empedient to use a method which does not involve laborious calculation but at the risk of over-estimation. So by assuming that the first-stage units are selected with PPS with replacement, estimate of variance can be reduced to

$$\hat{V}(\hat{Y}) = \frac{n}{(n-1)} \sum_{i}^{n} \left(\frac{\hat{Y}_i}{\pi_i} - \frac{1}{n} \sum_{i}^{n} \frac{\hat{Y}_i}{\pi_i} \right)^2$$

$$= \frac{1}{n-1} \left(n \sum_{i}^{n} \frac{\hat{Y}_i^2}{\pi_i^2} - \hat{Y}^2 \right) \qquad - (14)$$

If n is fairly large, overestimation of this will not be very high.

These general estimates given in equations (11) and (14) can be applied in surveys of countries A and B by substituting π_i and $\hat{Y}_i$ with the appropriate values.

(b) If n first-stage units are selected by equal probabilities (EP) (assuming that units are of equal size) then

$$\pi_i = \frac{n}{N} \quad \text{and} \quad \pi_{ij} = \frac{n(n-1)}{N(N-1)}$$

Substituting these values in equations (11), (12) and (13), estimate of total and its variances become $\hat{Y} = \frac{N}{n} \sum_{i}^{n} \hat{Y}_i \qquad - (15)$

$$V(\hat{Y}) = N^2 \left(\frac{1}{n} - \frac{1}{N} \right) \frac{1}{N-1} \sum_{i=1}^{N} (Y_i - \bar{Y})^2 + \frac{N}{n} \sum_{i}^{N} V(\hat{Y}_i) \qquad - (16)$$

where $\bar{Y} = \frac{1}{N} \sum_{i}^{N} Y_i$ = mean of Y_i

$$\text{and } \hat{V}(\hat{Y}) = N^2 \left(\frac{1}{n} - \frac{1}{N} \right) \frac{1}{n-1} \sum_{i}^{n} (\hat{Y}_i - \hat{\bar{Y}})^2 + \frac{N}{n} \sum_{i}^{n} \hat{V}(\hat{Y}_i) \qquad - (17)$$

where $\hat{\bar{Y}} = \frac{1}{n} \sum_{i}^{n} \hat{Y}_i$ = sample mean of $\hat{Y}_i$

In practice if n is fairly large the first part in $\hat{V}(\hat{Y})$ i.e. variation due to the first-stage sampling is likely to dominate the total variation and hence equation (17) can be approximated by $\hat{V}(\hat{Y}) = N^2 \left(\frac{1}{n} - \frac{1}{N} \right) \frac{1}{n-1} \sum_{i}^{n} (\hat{Y}_i - \hat{\bar{Y}})^2$

$$= \frac{N(N-n)}{n(n-1)} \left(\sum_{i}^{n} \hat{Y}_i^2 - \frac{1}{n} \hat{Y}^2 \right) - (18)$$

Indeed, it will be an under-estimation if variation due to samplings in the successive stages following the first-stage is considerable. Normally this problem does not arise since samplings in the successive stages are confined to smaller areas in which units are more homogeneous.

II. Estimation for country A sample

As described in paragraphs 7-13 and 53, country A sample is selected in four stages:

(1) n locations by PPS

(2) 2 chunks by PPS from each selected location

(3) one cluster by EP from each selected chunk

(4) Some households by EP with sampling fraction f from each selected chunk.

If c_a is estimated number of clusters in location (a) and C is sum of c_a overall locations in the population, then probability that location (a) is selected is

$$\pi_a = nc_a/C \qquad (19)$$

In that location (a), if c_k and c'_k are the estimated and actually observed number of clusters in chunk (k), probability of selection of that chunk (k) is $2c_k/c_a$ and probability of selection of a cluster from that chunk (k) is $1/c'_k$

If some households are selected with EP and with sampling fraction f, the probability of selection of any household from that cluster is f.

Therefore the overall probability of selection of a household in the location (a) is

$$= nc_a/C \times 2c_k/c_a \times 1/c'_k \times f$$

$$= 2nf/C \times c_k/c'_k \qquad (20)$$

If y_k is the sample values of households observed in chunk (k) of location (a), then

$$\sum_{k}^{2} c'_k/c_k \times y_k = y_a \qquad (21)$$

is the same as equation (2).

(a) Estimate for total

Hence estimate for the population value based on a sample of n locations will be

$$\hat{Y}_A = C/_{2nf} \sum_{a}^{n} y_a \qquad (22)$$

which is the same as equation (4), It will become self-weighting when either $c_k = c'_k$ or ultimate stage sampling fraction is $f\, c'_k/c_k$.

If equation (22) is compared with equation (11),

$$\hat{Y}_a/\pi_a = C/_{2nf}\ y_a \qquad (23)$$

and so substituting this in equation (14), the variance of $\hat{Y}_A$ becomes

$$\hat{V}(\hat{Y}_A) = n/n-1 \times (C/2nf)^2 \sum_{a}^{n} (y_a - \frac{1}{n} \sum_{a}^{n} y_a)^2$$

$$= (C/2f)^2 \quad 1/n(n-1) \quad \left[\sum_{a}^{n} y_a^2 - \frac{1}{n} \left(\sum_{a}^{n} y_a\right)^2\right]$$

and standard error of $\hat{Y}_A$, being square root of this is

$$\sqrt{\hat{V}(\hat{Y}_A)} = C/2f \sqrt{1/n(n-1) \left[\sum_{a}^{n} y_a^2 - 1/n \left(\sum y_a\right)^2\right]} \qquad (24)$$

which is the same as that given in equation (5).

(b) Estimate for an average household

To find the estimate for an average household, it is required to up-date the total number of households in the light of sample data on the number of culsters c'_k and average number of households in a cluster.

If C' is the up-dated estimate of the number of clusters and $\bar{m}$ is the average number of households in a cluster then

$$C' = C/2n \sum_{a}^{n} \sum_{k}^{2} c'_k/c_k$$

$$= C/2n \sum_{a}^{n} x_a$$

where $x_a = \sum_{k}^{2} c'_k/c_k$

Hence estimate for an average household becomes

$$\hat{\bar{Y}}_A = \hat{Y}_A/\bar{m}C'$$

$$= \frac{1}{f\bar{m}} \quad \frac{\sum_{a}^{n} y_a}{\sum_{a}^{n} x_a} \qquad (25)$$

which is a biased ratio estimate whose standard error can be approximated by

$$\sqrt{\hat{V}(\hat{Y}_A)} = \sqrt{\frac{1}{n(n-1)f\bar{m}}} \quad \bar{y}/\bar{x} \sqrt{\sum_{a}^{n} (y_a/\bar{y} - x_a/\bar{x})^2}$$

$$= \sqrt{\frac{1}{n(n-1)f\bar{m}}} \quad \bar{y}/\bar{x} \sqrt{\sum_{a}^{n} y_a^2/y^2 + \sum_{a}^{n} x_a^2/\bar{x}^2 - 2\sum_{a}^{n} x_a y_a/\bar{x}\bar{y}} \qquad (26)$$

where $\bar{y} = 1/n \sum_{a}^{n} y_a$

and $\bar{x} = 1/n \sum_{a}^{n} x_a$

are the sample mean of y and x.

III. Estimation for country B sample

(a) Unstratified sample

As described in paragraphs 14-19, the rural sample of country B is selected in three stages:

(1) n councils by PPS,

(2) a segment by PPS from each selected council,

(3) n households by EP from each selected segment.

If h_c is the number of households in the council (c) and H is the sum of h_c over all councils then probability of selection of council (c) is

$$\pi_c = nh_c/H \qquad - (27)$$

If h_s is the first estimate of the number of households in segment (s) and h'_c is the sum of h_s in the council (c) then the probability of selection of segment (s) will be h_s/h'_c.

In the ultimate stage, if h'_s is the number of households in the selected segment (s) obtained from an enumeration then probability of selection of m households will be m/h'_s.

Therefore the overall probability of selection of a household in the council (c) is $nh_c/H \times h_s/h'_c \times m/h'_s = nm/H \times h_c h_s/h'_c h'_s$ - (28)

In that council (c), sum the sample values of households in the selected segment and multiply that sum by $h'_c h'_s/h_c h_s$. Let y_c be such a value which is the same as that denoted in equation (7).

Hence estimate of the population value based on a sample of n councils is

$$\hat{Y}_B = H/nm \sum_{n} y_c \qquad - (29)$$

which is the estimate given in equation (8). This estimate will be more efficient had at least two segments been selected from each council.

On comparing this estimate with the general estimate given in equation (11),

$$\hat{Y}_c/\pi_c = H/nm \; y_c \qquad - (30)$$

and substituting this value in equation (14), the variance of $\hat{Y}_B$ becomes

$$\hat{V}(\hat{Y}_B) = (H/m)^2 \; 1/n(n-1) \left[\sum_c^n y_c^2 - \frac{1}{n} \left(\sum_c^n y_c \right)^2 \right] \qquad - (31)$$

and standard error of $\hat{Y}_B$ is just the square root of $\hat{V}(\hat{Y}_B)$, giving the same value as that given in equation (10).

(b) Stratified sample

As described in paragraphs 37 - 44, three strata are formed using the data on income distribution of households in the selected segments and overall strata sampling fractions are also determined. In order to equalise the size of sample selected from each segment, a deflator is used such that probability of any household to be selected from

$$\begin{aligned} \text{stratum } 1 &: f_1 \; \bar{h}'_s/h'_s = m_1/h_{s1} \\ 2 &: f_2 \; \bar{h}'_s/h'_s = m_2/h_{s2} \\ 3 &: f_3 \; \bar{h}'_s/h'_s = m_3/h_{s3} \end{aligned} \qquad - (32)$$

(where $\bar{h}'_s$ is the unweighted average number of households per segment obtained from the enumeration of selected segments, and f_1, f_2, f_3 are the overall sampling fractions; m_1, m_2, m_3 are the number of households actually selected out of three strata of sizes h_{s1}, h_{s2} and h_{s3} respectively).

Then the probability of selection of a household from stratum 1 of council (c) is

$$\begin{aligned} & nh_c/H \times h_s/h'_c \times f_1 \; \bar{h}'_s/h'_s \\ &= nh_c/H \times h_s/h'_c \times m_1/h_{s1} \\ &= n/H \times h_c h_s/h'_c \times m_1/h_{s1} \end{aligned} \qquad - (33)$$

Such probability from strata 2 and 3 can be easily obtained by substituting m_1/h_{s1} by m_2/h_{s2} and m_3/h_{s3} correspondingly.

If y_k is the sum of values of sample households in stratum k then

$$h'_c/h_c h_s \sum_{k}^{3} h_{sk}/m_k \; y_k = y_c \qquad - (34)$$

is the value representing council (c) as denoted in equation (7).

As n councils have been selected, there will be n such values in the sample and therefore the estimate of the population value will be

$$\hat{Y}_B = H/n \sum_{c}^{n} y_c \qquad - (35)$$

It is almost the same as that given in equation (29) except m which has been taken into account in y_c.

The variance of $\hat{Y}_B$ is $\hat{V}(\hat{Y}_B)$ given in equation (31) from which $1/m^2$ will be dropped out. So the standard error of $\hat{Y}_B$ in this case is

$$\sqrt{\hat{V}(\hat{Y}_B)} = H \sqrt{\frac{1}{n(n-1)} \left[\sum_{c}^{n} y_c^2 - \frac{1}{n} \left(\sum_{c}^{n} y_c \right)^2 \right]} \qquad - (36)$$

(c) Estimate for an average household

To find the estimate for an average household, it is required to up-date H in the light of the number of households observed in the segments. If H' denotes the up-dated estimate of the number of households then

$$H' = H/n \sum_{c}^{n} h'_c h'_s / h_c h_s = H/n \sum_{c}^{n} x_c \qquad - (37)$$

where $x_c = h'_c h'_s / h_c h_s$

Hence estimate for a household using the values of $\hat{Y}_B$ and H' from equation (35) and (37) becomes

$$\hat{\bar{Y}}_B = \hat{Y}_B / H'$$

$$= \frac{H/n \sum_{c}^{n} y_c}{H/n \sum_{c}^{n} x_c} = \sum_{c}^{n} y_c / \sum_{c}^{n} x_c \qquad - (38)$$

It is a biased ratio estimate whose standard error is approximated by

$$\sqrt{\hat{V}\left(\hat{\bar{Y}}_B\right)} = \frac{1}{\sqrt{n(n-1)}} \times \bar{y}/\bar{x} \sqrt{\sum_c^n y_c^2/\bar{y}^2 + \sum_c^n x_c^2/\bar{x}^2 - 2\sum_c^n x_c y_c/\bar{x}\bar{y}} \quad - (39)$$

where $\bar{y} = 1/n \sum_c^n y_c$

and $\bar{x} = 1/n \sum_c^n x_c$

are the sample means of y and x.

(d) Estimation for stratum total

As in the previous section if y_k denotes the sum of sample households in stratum k of the council (c), write

$$y_{ck} = h'h_{sk}/h_c h_s m_k \times y_k$$

Then estimate for stratum (k) total will be

$$\hat{Y}_k = H/n \sum_c^n y_{ck} \quad - (40)$$

and its standard error is approximately compressed by

$$\sqrt{\hat{V}(\hat{Y}_k)} = H \sqrt{\frac{1}{n(n-1)} \left[\sum_c^n y_{ck}^2 - \frac{1}{n}\left(\sum_c^n y_{ck}\right)^2\right]} \quad - (41)$$

Equal probability sampling

As an alternative to the PPS sampling, an EP sampling is proposed in paragraph 22-33 for country B. At first councils are regrouped or divided into new units to reduce the variation in the number of households per unit as far as possible. A sample of n councils so formed is selected with EP so that chance of selecting a council = n/C (where C is the number of councils) in the first stage.

Selected councils are divided into segments each containing about 100 households and a segment is selected with EP = $1/s_c$ (where s_c is the number of segments in council (c). If s_c is different from an assigned number $\bar{s}$

then the third stage sampling fraction will become $f \times s_c/\bar{s}$ so that overall selection chance for any household is

$$n/C \times 1/s_c \times f \times s_c/\bar{s} = n/C \times f/\bar{s} \qquad - (42)$$

If y_c is the value of sample households in the selected segment of the council (c) then the estimate of the total population value will be

$$\hat{Y} = C\bar{s}/nf \sum_{c}^{n} y_c \qquad - (43)$$

comparing this equation with equation (15)

$$\hat{Y}_c = \bar{s}/f \; y_c$$

and on substituting this value in equation (18) the approximate standard error of $\hat{Y}$ is expressed by

$$\sqrt{\hat{V}(\hat{Y})} = \bar{s}/f \sqrt{(C/C - n)/n(n-1)\left[\sum_{c}^{n} y_c^2 - \frac{1}{n}(\sum y_c)^2\right]}$$

Regarding estimation for an average household, it will simply be $\hat{Y}$ divided by C x average size of a segment.

REFERENCE

(1) Cochran, W.G. (1963) "Sampling Techniques" Second Edition, Wiley, New York

(2) Hartley, H.O. and Rao J.N.K. (1962) "Sampling with unequal probabilities and without replacement" Ann Math. Stat. 33, 350-374.

(3) African Household Survey Capability Programme "Some Aspects of Household Survey Methodology" E/CN.14/CAS.10/18.