A COMPARISON ON THE BASIS OF THE NASHVILLE MORBIDITY SURVEY OF TWO FRAMES WITH EQUAL AND UNEQUAL SIZED FIRST-STAGE UNITS

*T. Khosla, University of Aberdeen
A. L. Finkner, Research Triangle Institute,
and
J. C. Koop, N. C. State College

1. Introduction

In any statistical investigation where it may be costly to employ unrestricted random sampling and where units that are close together exhibit little variability, multi-stage sampling is appropriate. The advantages of multi-stage sampling are reduced cost both in travelling, because the ultimate units are less dispersed, and in developing the frame since it is constructed in successive stages.

However, with multi-stage sampling the question of whether to select the first-stage units with equal or unequal probability arises. Now the selection of the first-stage units with equal probability is usually less efficient than the corresponding selection with unequal probability if the first-stage units are large and vary greatly in their sizes. Difficulties of unequal probability sampling without replacement are found in (1) involved variance formulas with complicated functions of the probabilities used in the selection of the units and (2) estimates of variance which may be negative.

These reasons have prompted us to make a comparative study of a frame in which the first-stage units are constructed of equal size (in terms of ultimate sampling units) and selected with equal probabilities and without replacement, with another frame in which the first-stage units are unequal and selected with probability proportional to the size of the ultimate sampling units and also without replacement. Clearly when first stage units are equal in size equal probability sampling and sampling with probability proportional to size are equivalent.

The two methods will be compared on the basis of relative precision and also with an eye on relative costs. Precision will be judged from the closeness with which estimates centre round their own mean, in repeated application of the same sampling scheme.

2. The Nashville Morbidity Survey

The data used in the present study were obtained from the Nashville Morbidity Survey (N.M.S.), a detailed report of which is given by Finkner et al. (1960). In this section a brief description of this survey will be given with particular reference to (1) the construction of its frame and (2) the method of sampling. These aspects of the original study have an important bearing on the reconstruction of the frame for the purpose of the present study.

The universe consisted of eligible households found in the 51 city planning units of the city of Nashville, Tennessee, and its environs. Each of these planning units were divided up into a certain number of strata. Each stratum was further subdivided into 45 sampling units. The sampling unit was defined as an area segment, or as a strip along a street, with an expectation of about one household per sampling unit. The total number of dwelling units included in the universe at the time the frame was constructed was 69,244.

The structure of this frame which is based on maps is shown in the following table.

Table 1. The Structure of the Frame of the Nashville Morbidity Survey

Planning unit no.	Number of dwelling units	Number of strata	Number of sampling units	"Expected" size of sampling unit
I	II	III	IV	V
12	886	19	855	1.036
13	999	22	990	1.009
14	2287	50	2250	1.016
15	1331	29	1305	1.020
16	1731	38	1710	1.012
17	2268	50	2250	1.008
18	1605	35	1575	1.019
19	1171	26	1170	1.001
20	1003	22	990	1.013
21	242	5	225	1.076
22	1800	40	1800	1.000
23	1208	26	1170	1.032
24	1636	36	1620	1.010
25	1223	27	1215	1.007
26	1130	25	1125	1.004
27,28,29	939	20	900	1.043
30	2945	65	2925	1.007
31	2203	48	2160	1.020
32	1863	41	1845	1.010
33	1771	39	1775	1.009
35	1861	41	1845	1.009
34,36	1940	43	1935	1.003
37	1730	38	1710	1.012
38	684	15	675	1.013
39	152	3	135	1.126
40	487	10	450	1.082
41	340	7	315	1.079
42	2313	51	2295	1.008
43	2921	64	2880	1.014

* Formerly at North Carolina State College

Table 1. (continued)

44	938	20	900	1.042
45,46	1423	31	1395	1.020
47	939	20	900	1.043
48	2791	62	2790	1.000
49	2529	56	2520	1.004
50	3782	84	3780	1.001
51	1931	42	1890	1.022
52	463	10	450	1.029
53	210	4	180	1.167
54	1080	24	1080	1.000
55	934	20	900	1.038
56	1049	23	1035	1.014
57	575	12	540	1.065
65	1467	32	1440	1.019
66	1614	35	1575	1.025
67	1768	39	1755	1.007
73	1914	42	1890	1.013
74	1148	25	1125	1.020
Total	69,224	1516	68,220	1.015

In column I of Table 1 the serial number of the planning unit is given. Not all of the planning units were included in the survey. Thus planning unit number 12, containing 886 dwelling units (existing about the time the frame was constructed), was divided up into 19 strata each of which was further subdivided into 45 sampling units (each an area segment) giving in all 855 sampling units shown in column IV. It will be noted that all the entries in this column are multiples of 45. The "expected" size of the sampling unit, shown in column V, is 886/855, i.e. a little over one household. In all there were 1516 strata.

In regard to the method of sampling, two sampling units (area segments) were selected with equal probability and without replacement from the forty-five units constituting each of the 1516 strata. This observation has an important bearing on the statements on the structure of the sample design of the present study made in Section 4. Also it will be noted that the sample was self-weighting.

Out of 2 x 1516 = 3032 sampling units selected 282 were found to have no dwelling units. In the remaining sampling units, 2649 completed interviews from households were obtained. This number included 85 out of 264 households selected at random which did not respond at the first, but which responded at the second or the third interview*. The details of all these complexities in the resulting data can be found in the report referred to above. At any rate, the present study was not noticeably affected by these complexities in the data.

*Vaivanijkul (1961) studied the differences between respondents and these initial non-respondents for 34 morbidity characteristics of the survey. She found significant differences in only six characteristics.

3. Reconstruction of Frames for Present Study

In this section, the procedure of forming two new frames from existing materials shown in Table 1 will be given. Consider the universe as defined by the 51 planning units of the N.M.S. as shown in column I of Table 1. We desire only 5 large strata each composed of an equal number of large first-stage units. There are 51 planning units. Deleting no. 17 and no. 23 (for reasons which will appear in the subsequent discussion) and compounding the contiguous units 27, 28 and 29 (all of which contain only 939 dwelling units) as one unit, also 34 and 36 as one, and finally 45 and 46 as one, leaves us with the original 42 planning units and the 3 new compounded units whose respective constituents are contiguous. Thus in all we have 45 units and these we redefine as first-stage units (FSU's) and the original strata of the N.M.S. which make up each of these new units as second-stage units (SSU's), and lastly the sampling units (area segments) which the original N.M.S.-strata contained as the third-stage units.

Frame A. Dividing up the 45 FSU's into contiguous sets of 9, each to constitute a stratum, we have Frame A as shown in Table 2.

It will be noted that the nine FSU's in each stratum are quite unequal in size as measured by the number of SSU's contained by each of them. The FSU's are selected with probability proportional to size, e.g. the selection probability for FSU no. 1 stratum I is
19/(19+22+50+29+39+35+26+22+5).

Table 2. Frame A[a/]

Strata	Planning unit no. of N.M.S.	FSU no. of Frame A	Number of SSU's in Frame A
	12	1	19
	13	2	22
	14	3	50
	15	4	29
I	16	5	38
	18	6	35
	19	7	26
	20	8	22
	21	9	5
	22	1	40
	24	2	36
	25	3	27
	26	4	25
II	27,28,29	5	20
	30	6	65
	31	7	48
	32	8	41
	33	9	32
	35	1	41
	34,36	2	43
	37	3	38
	38	4	15
III	39	5	3
	40	6	10
	41	7	7

	42	8	51
	43	9	64
	44	1	20
	45,46	2	31
	47	3	20
	48	4	62
IV	49	5	56
	50	6	84
	51	7	42
	52	8	10
	53	9	4
	54	1	24
	55	2	20
	56	3	23
	57	4	12
V	65	5	32
	66	6	35
	67	7	39
	73	8	42
	74	9	25

[a] Each SSU contains 45 sampling units

We shall now construct another frame (on the basis of the same material) consisting of 5 strata in which the FSU's are of equal size as measured by the number of SSU's.

Frame B. In this frame, our object is to construct FSU's of equal size. With the deletion of two planning units (no. 17 and no. 23) from the original N.M.S. frame, we are left with 1440 SSU's. These are equally allotted to the 45 FSU's, so that each FSU contains exactly 32 SSU's under this scheme.

Consider Frame A. The first FSU (no. 12) has 19 SSU's and with 13 consecutive SSU's taken from the second FSU (no. 13) and added to it the size of the first FSU comes to 32. This leaves the second FSU with 9 SSU's, which in turn is compensated by taking 23 consecutive SSU's from the third FSU (no. 14) and the process is repeated in like manner for the remaining units. This procedure is diagrammatically represented for Stratum I in Table 3. After redefining the new first-stage units as above, 5 strata are formed, each stratum consisting of 9 first-stage units taken in order.

Here again, it may be mentioned that a number of different combinations can be made to form the first-stage units of equal size. For example, 13 second-stage units are required to be added to the size of the first FSU. Theoretically, this can be done in a number of ways. We may take 13 SSU's in any arbitrary manner from the remaining FSU's. Geographical proximity and administrative conveniences were again the guiding factors.

All these FSU's are selected with equal probabilities.

Table 3. Frame B (Stratum I)[a]

Planning unit no. of N.M.S.	Number of strata in N.M.S.	Number of SSU's in Frame B		FSU no. of Frame B
12	19	19	32	1
		13		
13	22	9	32	2
		23		
14	50	27	32	3
		5		
15	29	24	32	4
		8		
16	38	30	32	5
		2		
18	35	32	32	6
		1		
19	26	26	32	7
		5		
		17		
20	22	5	32	8
		10		
21	5			
22	40	30	32	9
		2		
24	36	32		

[a] Here also each SSU contains 45 sampling units

4. The Samples for Present Study

In the present study, two first-stage units are selected without replacement from each stratum, but with probability proportional to size of first-stage unit in sampling from Frame A, and with equal probability in sampling from Frame B. Now as the FSU's are composed of the 1440 SSU's, which were formerly strata of the N.M.S., all information is already available and are recorded on cards. Thus in any given stratum of Frame A or B, all the SSU's constituting each selected FSU play the role of second-stage units which are completely "sampled" and the two sampling units selected out of 45 from the original strata now play the role of the third-stage units under the present system of dual reconstruction of the original frame for the purpose of comparison.

The situation now is as if two different sample surveys relating to the same universe had been carried out with two different types of frames described in Section 3.

5. Theoretical Basis for Comparisons

We now introduce the formulas for estimates and their variances and also discuss methods for comparing the precision of the estimates obtained

on the basis of Frames A and B. The formulas will be given for a single stratum to avoid making the notation more involved than it is now.

Let Y_{iks} be the measure of a character of interest in the sth third-stage unit of the kth second-stage unit of the ith first-stage unit ($i=1,2,\ldots N$; $k=1,2,\ldots K_i$; $s=1,2,\ldots L$). We shall be concerned with the estimation of totals. The total T for a given stratum is given by

$$T = \sum_{i=1}^{N} \sum_{k=1}^{K_i} \sum_{s=1}^{L} Y_{iks} \qquad (1)$$

The linear unbiased estimate of T on the basis of a sample of n first-stage units, selected with probability proportional to the size of the units and without replacement from Frame A, and ℓ third-stage units selected with equal probability and without replacement from each of the second-stage units is

$$\hat{T}_A = \sum_{i=1}^{n} \frac{\sum_{k=1}^{K_i} \frac{L}{\ell} \sum_{s=1}^{\ell} Y_{iks}}{P_i} \qquad (2)$$

in which P_i is the probability of the first-stage unit i being included in a first-stage sample of n. In terms of the selection probabilities (which are chosen proportional to size of the FSU's), that is

$$p_i = K_i / \sum_{i=1}^{N} K_i, \quad i=1,2,\ldots N,$$

this inclusion probability when $n = 2$ is given by

$$P_i = p_i + \sum_{j(\neq i)=1}^{N} \frac{p_i p_j}{1-p_j} . \qquad (3)$$

The variance of $\hat{T}_A$ is given by

$$V(\hat{T}_A) = \sum_{i=1}^{N} Y_i^2 \left(\frac{1-P_i}{P_i}\right) + \sum_{i \neq j} Y_i Y_j \left(\frac{P_{ij}-P_i P_j}{P_i P_j}\right) + \sum_{i=1}^{N} \frac{1}{P_i} \sum_{k=1}^{K_i} L^2 \frac{S_{ik}^2}{\ell}\left(1-\frac{\ell}{L}\right) \qquad (4)$$

in which

P_{ij} = the probability of FSU's i and j being included in a first-stage sample of size n,

$$Y_i = \sum_{k=1}^{K_i} \sum_{s=1}^{L} y_{iks},$$

and

$$S_{ik}^2 = \sum_{s=1}^{L} (y_{iks} - y_{ik.})^2/(L-1)$$

where

$$y_{ik.} = \frac{1}{L} \sum_{s=1}^{L} y_{iks}$$

In the present study $N = 9$, $n = 2$, $L = 45$, $\ell = 2$. The essential theory underlying formulas (2) and (4) is due to Horvitz and Thompson (1951, 1952) and Narain (1951). When $n = 2$,

$$P_{ij} = p_i p_j \left(\frac{1}{1-p_j} + \frac{1}{1-p_i}\right).$$

The best linear unbiased estimate of T on the basis of a first-stage sample of size n from Frame B and third-stage samples of size ℓ selected from each of the second-stage units found in the n first-stage units, all selected at each stage with equal probability and without replacement is given by

$$\hat{T}_B = \frac{N}{n} \sum_{K=1}^{K} \frac{L}{\ell} \sum_{s=1}^{\ell} y_{iks} \qquad (5)$$

noting that $K_i = K = 32$ for all i in Frame B. The variance of $\hat{T}_B$, which can also be derived from (3) by putting $P_i = \frac{n}{N}$, $P_{ij} = \frac{n(n-1)}{N(N-1)}$ and $K_i = K$ and simplifying, is found to be

$$V(\hat{T}_B) = N^2 \left[\frac{S^2}{n}\left(1-\frac{n}{N}\right) + \frac{1}{nN} \sum_{i=1}^{N} \sum_{K=1}^{K} L^2 \frac{S_{ik}^2}{\ell}\left(1-\frac{\ell}{L}\right)\right] \qquad (6)$$

where $S^2 = \sum_{i=1}^{N} (Y_i - Y.)^2/(N-1)$ in which $Y. = \sum_{i=1}^{N} Y_i/N$.

Now the variance of the estimate having its basis in Frame A is $\sum_A V(T)$ and the corresponding variance for the estimate having its basis in Frame B is $\sum V(T_B)$ where the summation sign relates to summation over the five strata. Each of these expressions is a quadratic form in the underlying variates y_{iks}. The sign of the expression $\sum_A V(T) - \sum_B V(T)$ determines which of the two sampling procedures based on A or B is more efficient. Even for the case $n = 2$ it is not possible to determine the sign of the expression in the context of this study where $p_i = K_i/\sum K_i$. Generally the problem appears to be intractable.

The next approach is to make use of classical formulas for unbiased estimates of variance. This, too, has been avoided because the variability of the variance estimates might make the comparisons uncertain.

The approach which remains in such a situation is that of independent interpenetrating or replicated samples in the sense defined by Lahiri (1954). The technique consists of drawing two or more sets of samples from the same population using the same procedure of sampling for each set of samples. Sets of samples drawn in this manner are independent if and only if the sets of first-stage units selected are replaced after each drawing of a set is completed. This sometimes results in the same first-stage unit appearing in one or more sets.

For our study, we have drawn one hundred independent samples for each of the two sampling procedures under study.

The selection of first-stage units with

probability proportional to size of first-stage units and without replacement is very laborious if the usual method of cumulative totals is used. A simplified procedure of selecting first-stage units with unequal probability introduced by Lahiri was used.

One of us (Koop, 1960) has proved that the unbiased estimate of $V(\hat{T})$, the variance of an estimate T, from a set of m unbiased estimates

$$_1\hat{T},\ _2\hat{T},\ \cdots\ _m\hat{T}$$

each derived from independent replicated samples, whatever the underlying probability system and sample design, is given by

$$V(T) = \frac{\sum_{\alpha=1}^{m} (_\alpha T - \hat{T})^2}{m-1}$$

where $\hat{T} = \sum_\alpha \hat{T}/m.$ (7)

This result is used to compute the variances $\sum\hat{V}(\hat{T}_A)$ and $\sum\hat{V}(\hat{T}_B)$ on the basis of one hundred sets of independent estimates for each of the two sampling procedures. The variances of course are computed stratum by stratum and then added as indicated by the respective formulas.

Regarding the estimates for strata it may be noted that $\hat{T}_B$ is self weighting and is given by $\frac{9 \times 45}{4} \sum_i \sum_k \sum_s y_{iks}$, whereas $\hat{T}_A$ is not so and is given by $22.5 \sum_{i=1}^{2} \frac{1}{P_i} \sum_k \sum_s y_{iks}$. Each P_i is computed using formula (3).

6. Data Used for Investigation

The following four characteristics studied in the Nashville Morbidity Survey are selected for the purpose of comparing the frames:

(1) Number of people in households
(2) Number of deaths reported
(3) Number of employed individuals
(4) Number of households bothered by smog.

In the N.M.S. the selected sampling units were designated by five digit numbers. The first two digits were for the planning unit, the next two referred to the stratum and the last represented the sampling unit. The data is available on punched cards and is shown in skeleton form in Table 4.

Table 4. Data for sampling units selected in N.M.S.

Sampling Unit	Characteristics (1)	(2)	(3)	(4)
12 01 1	5	0	1	0
12 01 2	4	0	3	0
==				
12 19 1	2	0	2	0
12 19 2	4	1	0	0
13 01 1	2	0	2	0
13 01 2	3	0	1	0
==				
13 22 1	4	0	1	0
13 22 2	5	0	1	0
--				

For the present study, the above data is used as follows:

Frame A: The cards are first sorted and then grouped for each of the 45 first-stage units. The totals for each of the four characteristics are found by running the cards in I.B.M. 407 Tabulator. These totals for each of the 45 FSU's are used in the sampling design based on Frame A. The card numbers running from 12 01 1 to 12 19 2 fall in the first FSU. The allocation of the card numbers to the different FSU's and their totals for each of the four characteristics are given elsewhere (Khosla, 1961).

Frame B: We first make the necessary divisions of the cards to correspond to the new FSU's. For example cards running from 12 01 1 to 13 13 2 are put in the first FSU and from 13 14 1 to 14 23 2 in the second FSU. These totals are also shown in the above reference.

The totals from Frame B can be used as they are for the computation of stratum variances but the totals for each FSU from Frame A have to be multiplied by the relevant factor $\frac{1}{P_i}$ as indicated in the last paragraph of the previous section.

7. Comparison of Precision and Costs

The estimates for 4 characteristics studied are given in Table 5.

In three of the four characteristics under study the estimate of variance based on Frame A is greater than the corresponding estimate of variance based on Frame B. The relative precision of the sampling procedure based on Frame B compared to the sampling procedure based on Frame A is given by the ratio of the estimate of variance for Frame A to the corresponding estimate of variance for Frame B. The increase in the relative precision of Frame B expressed in percentage is found to be 27.65, 32.90, and 20.97 for the characteristics (1) number of people in households, (3) number of employed individuals and (4) the number of households bothered by smog, respectively. These results are shown in Table 6. In characteristic (2), number of deaths, there is a decrease of 0.62 per cent in the relative precision of Frame B.

There are 36^5 distinct samples possible for each of the two sampling procedures used. We have selected only a set of a hundred samples for each. It should be noted that any inference based on a sample study is subject to the usual uncertainties of sampling.

The costs involved in the two schemes are considered under the following headings:

(1) cost of constructing frames,
(2) interviewing costs in a sample under the two procedures,
(3) computing costs.

Table 5. Estimates of totals for entire population*

	From Frame A	From Frame B
(1) Number of people in households	200299	201673
(2) Number of deaths reported	7725	7805
(3) Number of employed individuals	70210	70878
(4) Number of households bothered by smog	14285	14457

*Each of these estimates are the means of the hundred independent estimates based on the relevant formula (2) or (4).

Table 6. Estimates of variances for each procedure

	$\sum_{h=1}^{5} \hat{V}(\hat{T})_{h\,A}$ Frame A	$\sum_{h=1}^{5} \hat{V}(\hat{T})_{h\,B}$ Frame B	Relative precision of Frame B (per cent)
(1) Number of people in households	89857314.5	70394481.1	127.65
(2) Number of deaths reported	776405.25	781252.06	99.38
(3) Number of employed individuals	10721893.6	8067576.8	132.90
(4) Number of households bothered by smog	4224824.8	3492453.1	120.97

a/ A multiplying factor of 1/100 applies to variances of estimates given in Table 5 since each of these estimates is based on 100 sets of independent estimates.

It was difficult to evaluate the comparative costs of constructing the two frames as the materials used for our study were already available. It can, however, be stated that it is possible to construct an area frame of equal sized first-stage units from the different materials available for area sampling work in this country. The additional material expense for constructing Frame B would be in the form of extra maps and attendant materials necessary for redefining areas of equal sized first-stage units.

Although we have ensured that the total sets of samples under each scheme is 100, there is no possible way to make the number of sampling units equal for each sample. The total number of sampling units in a sample under the two sampling schemes are

$$2\sum_{h=1}^{5}\sum_{i=1}^{2} K_{hi}$$

for Frame A and 640 for Frame B. The expected number of sampling units based on a sample from Frame A is 804.8. The number of sampling units in our set of a hundred samples based on Frame A was found to vary from 662 to 948 with a mean of 801.5. For the sampling system considered, the cost of enumeration using Frame A can be expected to be higher.

The computing time with Frame B was roughly thirty man-hours. This included all operations of computations manually performed with the Monromatic machine and without any recourse to I.B.M. Compared to this, seventy man-hours were necessary for the initial set up alone for Frame A and in addition I.B.M. was used to calculate estimates and their variances. On the basis of this experience, it seems that the computing time involved with Frame A (unequal probability sampling) with always be higher than with Frame B (equal probability sampling), whenever n the first-stage sample size from each stratum is greater than one.

With Frame B the selection of more than two FSU's from each stratum can be made with a proportionate increase in the computing cost. Compared to this, the cost of computation with Frame A will increase at a much faster rate as we progressively increase the number of first-stage units in the sample.

Our study at least shows that sampling with equal sized first-stage units deserves more attention.

References

Finkner, A. L., Monroe, J., and Fleischer, J. (1960) The Nashville Morbidity Survey, Institute of Statistics Mimeo. Series No. 252. N. C. State College, Raleigh, N. C.

Horvitz, D. G. and Thompson, D. J. (1951) A generalization of sampling without replacement from a finite universe. Abstract. Ann. Math. Stat., 22 (2):315

_____ (1952) Ibid. J. Amer. Statist. Assoc. 47:663-685.

Khosla, T. (1961) Comparison of two frames with equal and unequal sized first-stage units. Unpublished M. S. Thesis, N. C. State College Library.

Koop, J. C. (1960) On theoretical questions underlying the technique of replicated or interpenetrating samples. 1960 Proceedings of the Social Statistics Section. Amer. Statist. Assoc., 196-205.

Lahiri, D. B. (1954) National Sample Survey: Some Aspects of Sample Design. Sankhya, 14: 268-316.

Narain, R. D. (1951) On sampling without replacement with varying probabilities. J. Ind. Soc. Agric. Statist., 3 (2):169-174.

Vaivanijkul, N. (1961) A comparison of respondents and non-respondents in a sample survey. Unpublished M. S. Thesis, N. C. State College Library.