BEBR

**FACULTY WORKING
PAPER NO. 781**

**Efficient Screening Methods for the
Sampling of Special Populations**

Seymour Sudman

College of Commerce and Business Administration
Bureau of Economic and Business Research
University of Illinois, Urbana-Champaign

FAC

College of C

University

Effic
the Sa

S
Departm

ABSTRACT

Rare special populations for which no list exists require costly screening. Efficient procedures are discussed in this paper for reducing these screening costs, if the populations are geographically clustered. These procedures involve telephone, face-to-face screening and mixed modes. If the special population is not geographically clustered, multiplicity procedures may be useful.

EFFICIENT SCREENING METHODS FOR
THE SAMPLING OF SPECIAL POPULATIONS

1. Introduction

A special population is defined as a subgroup of a general population
for which no complete list is available. Such populations require screen-
ing of the general population before data collection. If the population
is rare or very rare, screening costs may become very large and account for
the major share of costs of data collection. Procedures are available,
however, for reducing these screening costs.

There are many special populations that are geographically clustered.
A few examples are:

a. Racial and ethnic groups such as Blacks, Hispanics, Cuban,
Vietnamese and recent immigrants from the Soviet Union.

b. Special types of housing such as very large apartments, trailers,
substandard housing, etc.

c. Employees in specified industries such as asbestos workers.

d. Purchasers of new products with limited distribution.

Other special populations are not geographically clustered. Examples
of these are:

e. Cancer patients

f. Alcholics

g. Accident victims

In this paper some alternative procedures for reducing screening
costs are discussed. Two dimensions are considered:

a. Geographic clustering - If the special population is geographically

clustered, costs may be reduced substantially by rapidly identifying zero segments in which no members of the special population can be found, or by undersampling clusters with very few members of the special population. If the special population is not geographically clustered, multiplicity procedures may be useful.

b. <u>Methods of screening and data collection</u> - Screening and data collection may be conducted using mail, telephone or face-to-face procedures and mixed methods are possible with screening being done in a less expensive way than interviewing.

The standard procedure for screening such special populations is to increase the size of the cluster. See, for example, Kish (1965, pp. 904-10). Thus, if a special population is p percent of the total population and a cluster size of $\bar{n}$ completed cases is optimum, initial clusters of $\bar{n}/p$ would be required.

2. <u>Sampling When There Are Many Zero Geographic Segments</u>

Many geographically clustered special populations are located in a limited number of geographic areas. Conversely, there are a large fraction of total geographic segments in which <u>no</u> members of the special population are located. The standard procedure in this case often leads to hundreds of screening calls in these zero segments and no eligible respondents.

If the zero segments are known in advance from Census data or some other source, substantial cost savings are possible by eliminating the screening of these zero segments. In this case, the optimum procedures for screening and data collection using mail or telephone procedures involve no clustering while the optimum procedures for face-to-face interviewing would utilize the standard optimum cluster procedures

developed by Hansen, Hurwitz and Madow (1953). Frequently, however, zero segments are not known in advance.

Zero segments unknown in advance - The rapid elimination of zero segments through use of a one (or a few) screening contacts can sub-stantially reduce screening costs, particularly if the proportion of zero segments is high. The widely used method for improving the efficiency of random digit dialing procedures described by Waksberg (1978) may be adapted for special populations with even greater cost savings than for general populations.

The procedure, as used in either mail or telephone screening, requires that initially a single unit be sampled within a geographic segment. If that unit is a member of the special population, additional screenings are conducted in the segment until the desired cluster size is reached. This procedure produces a sample in which each selected unit has an equal probability of selection.

Generally following Waksberg's notation, let:

n = the total sample size of the special population

m = the number of geographic segments

$k+1$ = the cluster size in the sample

ρ = homogeneity (intraclass correlation) within the geographic segments

t = proportion of segments with no members of population (zero segments)

π = proportion of special population to total population

C_S = cost of screening a contact

C_I = cost of data collection from an eligible unit

C_p = Total costs of this method

Waksberg shows that the expected screening costs for this procedure are:

$$\frac{m}{\pi} [1+(1-t)k]C_S \qquad (2.1)$$

Total costs for screening and interviewing are:

$$C_P = m(k+1)C_I + \frac{m}{\pi}[1 + (1-t)k]C_S$$

$$= m(k+1)\left[C_I + C_S\frac{(1-t)}{\pi}\right] + \frac{mt}{\pi} C_S \qquad (2.2)$$

It therefore follows immediately from Hansen, Hurwitz and Madow (1953, Ch. 6) that:

$$(k+1)_{optimum} = \left[\frac{tC_S}{\pi[C_I + \frac{(1-t)}{\pi}C_S}\left(\frac{1-\rho}{\rho}\right)\right]^{\frac{1}{2}} \qquad (2.3)$$

The cost for a sample of size n is found by substituting the value of k+1 from formula (2.3) into formula (2.2). The cost for a nonclustered sample with the equivalent variance is

$$C_U = \frac{m(k+1)}{1+k\rho}\left[C_I + \frac{C_S}{\pi}\right] \qquad (2.4)$$

The ratio of costs from formulas (2.2) and (2.4) is:

$$\frac{C_P}{C_U} = (1+k\rho)\frac{[C_I\pi + C_S(1-t) + tC_S/k+1]}{[C_I\pi + C_S]} \qquad (2.5)$$

Table 2.1 presents optimum values of the function $k+1 / (\frac{1-\rho}{\rho})^{\frac{1}{2}}$ for values of t, π, and C_I / C_S. Table 2.2 presents relative data collection costs for this optimum clustering as compared to unclustered samples. It may be seen that very substantial cost savings of about 70 percent or more are possible in the upper left hand corner of the table when t is around .9, π is correspondingly low, and ρ is around .01. On the other hand, there

is no advantage to these clustered screening methods in the lower right hand corner where t is less than .5 or .6, π is greater than .2 and ρ is about .10.

Two examples illustrate the effectiveness of geographic screening. For both examples, it is assumed that telephone sampling and interviewing are used and that virtually all eligible households have telephones.

Example 2.1:   Phone Screening of Black Households Using Random Digit Dialing

Suppose one wishes to select a national phone sample of 1,200 Black households. (Such a sample is currently being used to obtain Black attitudes on public policy and marketing issues.) The proportion of Black households in the United States is about .12 so that the estimate of π is (.12) (.30) = .036. (The .30 estimate of working phone numbers is based on Groves and Kahn (1979) estimates that 65 percent of banks of 100 numbers are nonworking and on the additional estimate that about 20 percent of numbers in working banks are not working household numbers.

Based on an analysis of some Census Block Statistics, it is estimated that about 70 percent of working banks have no Black households. Thus, t = 1-(.35)(.30) = .9. Assume that ρ = .05, C_I =10 and C_S = \$2. Then:

$$(k+1)_{optimum} = \left(\left(\frac{.9}{(.036)5 + .1}\right)\frac{.95}{.05}\right)^{\frac{1}{2}} = 8$$

or using Table 2.1 and interpolating, $\frac{k+1}{(\frac{1-\rho}{\rho})^{\frac{1}{2}}} = 1.82$ and k + 1 = (1.82)(4.36) = 8.

The actual cost for an unclustered sample with an equivalent variance is:

$$C_U = \frac{1200}{1+7(.05)} \left[10 + \frac{2}{.036}\right] = \$58,272$$

2.1 Optimum Values of $\dfrac{k+1}{[(1-\rho)/\rho]^{1/2}}$ for Values of t, π, C_I/C_S *

| | | | | | π | | | |
t	C_I/C_S	.025	.05	.10	.20	.30	.40	.50
.95	2	3.08	2.52	-	-	-	-	-
	5	2.33	1.79	-	-	-	-	-
.90	2	2.45	2.12	1.73	-	-	-	-
	5	2.00	1.60	1.22	-	-	-	-
.80	2	1.79	1.63	1.41	1.15	-	-	-
	5	1.57	1.33	1.07	.82	-	-	-
.70	2	1.41	1.32	1.18	1.00	.88	-	-
	5	1.28	1.13	.94	.73	.62	-	-
.60	2	1.15	1.10	1.00	.87	.77	.71	-
	5	1.14	.96	.82	.65	.56	.50	-
.50	2	.95	.91	.85	.75	.67	.62	.58
	5	.89	.82	.71	.58	.50	.45	.41
.25	2	.56	.54	.51	.47	.43	.40	.38
	5	.53	.50	.45	.38	.33	.30	.28

*To find optimum k+1; if ρ=.1, multiply by 3;
if ρ = .05, multiply by 4.36; if ρ=.02 multiply by 7;
if ρ = .01, multiply by 9.95; etc.

2.2 Relative Data Collection Costs for Optimum Clustering Compared to Unclustered Samples

ρ / t	c_T/c_S	π=.025		π=.05		π=.10		π=.20		π=.30		π=.40	
		.01	1.0	.01	.10	.01	.10	.01	.10	.01	.10	.01	.10
.95	4/2	.16	.35	.21	.42	—	—	—	—	—	—	—	—
	10/2	.23	.44	.33	.55	.34	—	—	—	—	—	—	—
.90	4/2	.22	.42	.26	.48	.44	.56	—	—	—	—	—	—
	10/2	.29	.50	.37	.59	.50	.72	—	—	—	—	—	—
.80	4/2	.33	.55	.37	.59	.52	.65	.53	.74	—	—	—	—
	10/2	.38	.60	.46	.68	.57	.77	.70	.88	—	—	—	—
.70	4/2	.43	.65	.46	.68	.60	.73	.60	.80	.66	.86	—	—
	10/2	.48	.69	.54	.75	.63	.83	.73	.91	.80	.95	—	—
.60	4/2	.53	.74	.55	.76	.67	.80	.67	.86	.72	.89	.76	—
	10/2	.57	.77	.62	.82	.70	.83	.79	.94	.80	.97	.88	—
.50	4/2	.62	.82	.64	.85	.73	.87	.74	.90	.78	.93	.81	.92
	10/2	.65	.86	.70	.88	.76	.88	.83	.96	.87	—	.90	.95
.25	4/2	.84	—	.85	—	.87	—	.89	—	.91	—	.92	—
	10/2	.86	—	.87	—	.90	—	.93	—	.95	—	.96	—

For the optimum cluster design, the cost is:

$$C_\rho = 1200 \left[10 + \frac{2(.1)}{.036} \right] + \frac{150}{.036} (.9)(2) = 18667 + 7500$$

$$= \$26167$$

and the ratio $\dfrac{C_p}{C_U} = .44$ which could also be obtained from Table 2.2

Example 2.2 <u>Phone Survey of American Jews</u>

It is estimated that Jews account for about three percent of the total U.S. population and are heavily concentrated in the largest cities. For this example, the following estimates are made:

$$\pi = (.03)(.35) = .01$$

$$t = 1 - (.35)(.4) = .86$$

$$C_I = \$10, \quad C_S = \$2, \quad \rho = .02$$

Then $(k+1)_{optimum} = 15$ and from formula (2.5), for any sample size

$$\frac{C_p}{C_I} = 1.28 \left[\frac{.1 + 2(.14) + .86(2)/15}{.1+2} \right] = .30$$

<u>Clusters with too few eligible households</u> - As illustrated by the last example, optimum cluster sizes may be fairly large for relatively rare population screening. The procedure becomes biased if a selected cluster contains fewer than k+1 members of the special population. Two procedures are possible to prevent this bias:

a. Increasing the size of the cluster from banks of 100 to banks of several hundred or more.

b. Weighting the data in a cluster by the ratio k+1/r where r is the number of eligible households within the cluster. Optimum procedures would usually involve weighting. This is discussed in Section 7.

3. Face-to-Face Screening Required

For some purposes, such as determining the condition of housing
units and interviewing in those that meet, or fail to meet, prespecified
criteria, or in ethnic groups with low phone coverage, it may be necessary
to conduct face-to-face screening. The cost function for this process is
strongly affected by the cost of listing and travel to and from the segment.
Thus, the procedure used for mail and phone surveys of selecting non-zero
segments on the basis of the characteristics of a single unit may no
longer be optimum. Rather, it may be more efficient to conduct multiple
screening calls before deciding whether to include or exclude a cluster.

Consider the following design. An interviewer makes j screening
calls at points within a cluster that are relatively close geographically.
(For specificity, we can assume that these might be housing units on
different blocks within the same Census Tract.) If the screening call
yields an eligible household, then additional calls are made sequentially
until k additional eligible households are located. If the screening
call does not yield an eligible household, no additional screening calls
are made. This is an unbiased sampling procedure which is a direct
extension of the method of the previous section.

Note that this procedure, unlike the one discussed for phone and mail
screening, produces variability in the total number of completed cases in
a cluster. The number of completed cases will range from $k+1$ to $j(k+1)$.
This variability increases the sampling variance as does the clustering.
Even so, it may be shown that in some cases, this procedure is optimum.
Using the same notation as in the previous section:

Let C_T = travel costs for one trip to an average segment.

It is assumed that the segments have been listed previously so that no additional listing costs are required.

C_F = total costs of face-to-face screening and interviewing procedures.

Assume that j screening calls can be made on a single trip so that no additional travel costs are required. These j calls will yield 0-j eligible units. For each eligible unit found, continue screening until k additional eligible units are found. It is assumed that one additional trip is required for interviewing in a segment once an eligible unit has been found. More complex cost functions describing travel may be used, but the general result would still follow.

It can be demonstrated that for most applications this is more efficient than either conducting only a single screening call/segment or conducting a very large number of screening calls in every segment. We use the same cost function as previously, adding, however, a term C_T for travel to screen the cluster. Let C_F be the total costs for this face-to-face procedure.

$$\text{Then } C_F = m(k+1)\ [C_I = C_S\frac{(1-t)}{\pi}] + m\left(\frac{t}{\pi}\,C_S + C_T\left(\frac{\pi+1}{\pi}\right)\right) \quad (3.1)$$

$$(k+1)_{opt} = \left[\frac{t\,C_S + C_T\,(\pi+1)}{C_I\,\pi + (1-t)\,C_S}\ \left(\frac{1-\rho}{\rho}\right)\right]^{\frac{1}{2}} \quad (3.2)$$

The value of multiple starts depends on the fact that the homogeneity between elements in a cluster typically declines as the cluster increases in geographic size. If this is not the case, then the new cost function merely means an increase in optimum cluster size. This is immediately evident since the numerator in formula (2.7) has a C_T term, but is otherwise identical to the phone optimum in formula (2.3).

Example 3.1

As an illustration, suppose the sample of Black households in Example 2.1 required a face-to-face interview because of the topic. Let $\pi = .12$, $t = .7$ (now working banks are irrelevant) $C_I = \$10$, $C_S = \$4$ and $C_T = \$16$, $\rho = .05$. Note that C_S is increased since face-to-face screening is more costly than telephone screening. From formula 3.2,

$$(k+1)_{opt.} = \left[\frac{.7(4) + 16(1.12)}{10(.12) + .3(4)} \left(\frac{.95}{.05}\right)\right]^{\frac{1}{2}} = 12.8$$

Now suppose the homogeneity ρ' between blocks within a tract is low, say .01 while ρ within the block is .05. Assuming no additional cost for screening, it is possible to compute alternative values for j and compare to a single screening call. For specificity assume that m = 100 tracts. Then

$$C_F = 100(12.8) \left[10 + 4 \left(\frac{.3}{.12}\right)\right] + 100 \left[\frac{.7}{.12}(4) + 16 \left(\frac{1.12}{.12}\right)\right]$$
$$= \$40,767$$

$$n_{equiv} = \frac{n}{[1 + \rho'(j-1)][1 + \rho(k)]} \tag{3.3}$$

$$C_F = mj(k+1)[C_I + C_S \left(\frac{1-t}{\pi}\right)] + m[\frac{jt}{\pi} C_S + C_T(\frac{\pi+1}{\pi})] \tag{3.4}$$

$$(k+1)_{opt} = \left[\frac{tC_S + \frac{C_T}{j}(1+\pi)}{C_I\pi + (1-t)C_S} \left(\frac{1-\rho}{\rho}\right)\right]^{\frac{1}{2}} \tag{3.5}$$

Consider value of j = 2,3,... For alternate values of j and ρ', optimum cluster sizes and costs may be determined and the equivalent sample sizes compared to the equivalent sample size for j = 1. This is done in Table 3.1. As an illustration, consider j = 3, $\rho = .05$, $\rho' = .1$.

$$\text{Then } (k+1)_{opt} = \left[\frac{.7(4) + \frac{16}{3}(1.12)}{10(.12) + .3(4)} \left(\frac{.95}{.05}\right)\right]^{\frac{1}{2}} \approx 8.3$$

$$\text{From (3.4) } m = \frac{40,767}{3(8.3)[10 + 4\left(\frac{.3}{.12}\right)] + \frac{3(.7)(4)}{.12} + 16\left(\frac{1.12}{.12}\right)} = 56.7$$

so $n = mj(k' + 1) = 1411$

$$n_{equiv} = \frac{1411}{[1 + .1(2)][1 + .05(7)]} = 871$$

3.1 Equivalent Sample Sizes for Values of j and ρ' in Example 3.1

Equivalent n

ρ'

	Actual n	.01	.05	.10	.20	.30
j = 1	1280	805	805	805	805	805
2	1350	990	952	909	833	769
3	1411	1025	950	871	747	
4	1458	1049				
5	1483	1056				
6	1505	1062				
7	1515	1059				

For the general population, ρ' is usually in the range of .05 - .2. For rare populations, ρ' could be smaller. Unfortunately, usually one would not know this until after the study. A conservative approach would be to use general population values. For this example, j = 2 would be optimum over a broad range of ρ'. While this result depends on the specific cost function and value of ρ, additional calculations suggest that two or three starting points per segment would be better than one for most face-to-face screenings, unless ρ' is large. As ρ increases relative to ρ', or for ρ' very small, even more starting points are optimum. It is not clear that many such situations exist.

4. The Use of Incomplete Lists

Even incomplete lists may be very useful in identifying areas where the special population is located. In the simplest case, assume that a random (or systematic) sample of starting points is chosen from the list and screening continues at each starting point until k additional eligible respondents are located. It is evident that this procedure is almost identical to those just discussed.

A cluster will have an initial probability of selection proportional to the number of members of the special population in it. Then, sampling within the cluster is inversely proportional to this probability so that the ultimate sample is self-weighting.

The sample is not unbiased, however, if there exist non-zero clusters that have no eligible respondents on the incomplete list. These clusters have no chance of selection. It is possible to measure the sample bias from such a procedure if one has an estimate of the total size of the special population. One would also estimate from the list the number of nonzero clusters and from the screening the average number eligible per cluster. The product of these last two is an estimate of the number of persons in the special population who have a nonzero probability of selection. The difference between the total estimate and the estimate of those with nonzero probabilities would indicate potential bias.

The cost function for using lists is very similar to those already discussed.

Let C_L be the cost for a unit on the list. Then the total cost

$$C_U = m(k)[C_I + \frac{(1-t)}{\pi} C_S] + m [C_I + C_L] \qquad (4.1)$$

$$k_{opt} = \left[\frac{C_I + C_L}{C_I + (\frac{1-t}{\pi}) C_S} \; (\frac{1-\rho}{\rho}) \right]^{\frac{1}{2}} \qquad (4.2)$$

The cost per case using optimum clustering is:

$$C_I + \frac{(1-t)}{\pi} C_S + \frac{C_L}{(k+1)_{opt}} \qquad (4.3)$$

and the cost per an equivalent unclustered case is:

$$C_{UE} = [1+\rho k] \left[C_I + \frac{(1-t)}{\pi} C_S + \frac{C_L}{(k+1)_{opt}} \right] \qquad (4.4)$$

Example 4.1

Suppose, following Example 2.2 that a list were available of Hadassah members and one wished to sample American Jews in a telephone survey. Assume that the list was estimated to include clusters in which 95 percent of American Jews live. The cost of the list C_L is 10¢ for each member.

As before, let, $C_I = 10$, $C_S = 2$, $\rho = .02$, $\pi = .01$, $t = .86$.

$$\text{Then } k_{opt} = \left[\frac{10 + .1}{10 + \frac{.14}{.01} \, (2)} \; (\frac{.98}{.02}) \right]^{\frac{1}{2}} = 3.6$$

Note that since screening costs are small, much smaller clusters are optimum. The cost per case is $10 + \frac{.14}{.01}(2) + \frac{.10}{4.6} = \$38.02$ and per equivalent case is $[1 + .02 \, (3.6)][38.02] = \$40.76$, about 2/3 of the costs of phone screening.

5. Use of Combined Methods

The use of combined procedures is common in survey sampling. The standard optimization methods allocate sampling rates to the procedures inversely to the square root of the ratios of costs. Bayesian allocation procedures result in the elimination of very costly procedures when the

reduction in variance would be small and resources are limited.

In the screening of special populations, one would be very unlikely to use combined methods if, as in Example 4.1, the list is 95 percent complete, since the marginal reduction in total survey error would be negligible. Suppose, however, a list includes clusters where 50-90 percent of a special population live. Then, combined methods become optimum, especially if the study uses lists and face-to-face screening. Note that if one uses lists, the costs of the other screening methods increase since one is now attempting to locate only those clusters with a zero probability of selection from the list.

Example 5.1

Suppose, following Example 3.1, we want to use a list of subscribers to a Black magazine to reduce screening costs. As in Example 3.1, let $\pi = .12$, $t = .7$, $C_I = 10$, $C_S = 4$, $C_I = 16$, $C_L = .1$, $\rho = .05$. (We assume $\rho' > .2$ so only one start is optimum). The cost function for this method is:

$$C_K = nk \left[C_I + \frac{(1-t)}{\pi} C_S \right] + m \left[C_I + C_T + C_L \right] \qquad (5.1)$$

$$k_{opt} = \left(\frac{C_I + C_T + C_L}{C_I + (\frac{1-t}{\pi})C_S} \cdot \frac{1-\rho}{\rho} \right)^{\frac{1}{2}} \qquad (5.2)$$

and $C_{KE} = (1+\rho k) \left[C_I + (\frac{1-t}{\pi}) C_S + \frac{C_T + C_L}{(k+1)_{opt}} \right] \qquad (5.3)$

So $k_{opt} = \left[\left(\frac{26.10}{10 + \frac{.3}{.12} (4)} \right) \left(\frac{.95}{.05} \right) \right]^{\frac{1}{2}} = 5.0$

$$\text{and } C_{KE} = 1.25 \left[10 + \frac{.3}{.12}(4) + \frac{16.10}{6}\right] = \$28.35$$

Assume that the list is estimated to cover two-thirds of the clusters in which the population lives so that a combined procedure using initial face-to-face screening is required. Then the new values for π and t for this screening are π = .04, t = .9. The optimum k+1 is 21.9 and the cost per equivalent case is $88.15. Therefore, the sampling rate for intitial face-to-face screening should be .57 that for list screening. A total sample of 1,000 would contain 776 cases from 155 clusters selected using the list and 224 cases from 10 clusters using face-to-face initial screenings. As with all disproportionate samples, weighting would be required for unbiased estimates.

6. Phone Screening and Face-to-Face Interviewing

In some situations, it is possible to screen by telephone, but inter-viewing must be face-to-face. If this procedure is followed, screening costs are obviously reduced, but interviewer travel costs to locate the respondent are added. In an earlier paper, Sudman showed that for most situations, the joint use of phone screening and face-to-face interviewing was more efficient than face-to-face screening and interviewing. (Sudman, 1978).

In that paper, however, the procedures for screening zero segments that are discussed above in Sections 2 and 3 were not used. When these are considered, both face-to-face and phone screening are more efficient. It remains true, however, that for relatively rare populations, the joint procedure is still much more efficient than face-to-face screening.

To see this, consider the following cost function that describes the joint procedure:

$$C_{PF}^{\bullet} = m(k+1)[C_I + C_L + C_{SP} \frac{(1-t)}{\pi}] + m\frac{-t}{\pi}C_{SP} + 2\ C_T] \quad (6.1)$$

$$(k+1)_{opt} = \frac{t\ C_{SP} + 2\ \pi\ C_I}{(C_I + C_L)\ \pi\ C_{SP}\ (k-t)}\ (\frac{1-\rho}{\rho})^{\frac{1}{2}} \quad (6.2)$$

The only new term is C_L which is the cost of locating an eligible respondent after the phone screening; C_{SP} is the cost of screening a case on the telephone. The subscript is added as a reminder that this is not identical to the cost of face-to-face screening. It is assumed that two trips to the segment are necessary to locate and interview all eligible respondents. This is comparable to the assumptions made in section 3.

Comparing formulas (3.1) and (6.1), the major tradeoffs are between travel costs to zero sites in face-to-face screening and location costs in telephone screening.

Example 6.1

Assume that the face-to-face screening procedure in Example 3.1 is to be compared to a joint procedure. The cost per equivalent case in that example is C_{FE} = $50.64.

For the joint procedure, use the same estimates as in earlier examples: $t = .9$, $\pi = .12$, $\rho = .05$, $C_I = \$10$, $C_{SP} = \$2$, $C_T = \$16$, and let $C_L = \$2$.

$$\text{Then } (k+1)_{opt} = [\frac{(.9)(2) + .24(16)}{12\ (.12) + 2(.1)}\ \frac{.95}{.05}]^{\frac{1}{2}} = 8.1$$

$$C_{PFE} = 1.355\ [10 + 2 + 2(\frac{.1}{.12}) + \frac{.9}{.12}\ (2) + 32] = 26.38$$
$$\frac{}{8.1}$$

which is roughly half of the face-to-face screening cost.

This advantage for joint procedures disappears for values of π equal to or greater than about .3. Thus, the results are similar to those of section 2. Joint procedures are optimum for the same populations for which phone screening is optimum.

7. <u>Variations in Density of Special Populations in Non-Zero Clusters</u>

We now consider the situation where the special population is unevenly distributed among the non-zero clusters. This would be likely to occur with ethnic groups where most members would live in a few clusters with high proportions of the population, but others would be thinly spread among the general population. While it is possible to have identified these clusters from earlier screening or Census data, it is also possible to estimate the proportion of the special population by asking the first contacted household(s) to estimate π_j.

<u>Phone Screening</u> - Assume first that the non-zero clusters have been identified and catagorized into strata where π_j is the proportion of the special population to the total population in that stratum. In this case, no clustering is required and the cost per case in the jth stratum is $C_{Pj} = C_I + \dfrac{C_S}{\pi_j}$.

An optimum allocation procedure would be to sample from the strata with rates inversely proportional to the square roots of costs. Thus, the relative rates in strata A and B would be:

$$\frac{r_A}{r_B} = \left[\frac{C_I + \dfrac{C_S}{\pi_B}}{C_I + \dfrac{C_S}{\pi_A}} \right]^{\frac{1}{2}} \tag{7.1}$$

If there is a screening or list cost, the procedures of Section 2 apply. Then, from formula 2.5

$$\frac{r_A}{r_B} = \left\{ \frac{[1 + \rho(k_B-1)] \; [C_I + (1-t)\dfrac{C_T}{\pi_B} + tC_S/k_B-1]}{[1 + \rho(k_A-1)][C_I + C_S\dfrac{(1-t)}{\pi_A} + t_{CS}/k_A-1]} \right\}^{\frac{1}{2}} \quad (7.2)$$

Example 7.1

Suppose one wishes to select a sample of Hispanics in a community using phone screening. Ignoring zero clusters, assume most members live in clusters where $\pi_1 = .5$, but a few live in areas where $\pi_2 = .01$. As before, let $C_I = 10$, $C_S = 2$

$$\text{Then, } \frac{r_1}{r_2} = \left[\frac{10 + 2/.01}{10 + 2/.5}\right]^{\frac{1}{2}} = 3.9$$

Face-to-Face Screening

Face-to-face screening would require clustering. Again assuming the non-zero clusters have been identified, the total cost in the jth stratum would be:

$$C_{Fj} = mk \; [C_I + \frac{C_S}{\pi_j}] + mC_T \quad (7.3)$$

$$\text{and } k_{opt} = \left[\left(\frac{C_T}{C_I + \dfrac{C_S}{\pi_j}}\right)\left(\frac{1-\rho}{\rho}\right)\right]^{\frac{1}{2}} \quad (7.4)$$

The cost per equivalent case in the jth stratum is:

$$C_{Ej} = [1 + \rho \; (k-1)] \; [C_I + \frac{C_S}{\pi} + \frac{C_T}{k}] \quad (7.5)$$

$$\text{and } \frac{r_A}{r_B} = (C_{EB}/C_{EA})^{\frac{1}{2}}$$

If the non-zero clusters had not been previously identified, ratios of the cost functions in Section 3 could be computed.

Example 7.2

Assume one wishes a sample of Hispanics, but phone ownership in the community is too low so that face-to-face screening is required. As in Example 7.1, assume zero clusters have already been located, let $\pi_1 = .5$, $\pi_2 = .01$, $C_I = 10$, $C_S = 2$, $\rho = .05$ and let $C_T = 32$ to cover the costs of two trips to the cluster. Then:

$$k_{1opt} = 6.6 \qquad\qquad k_{2opt} = 1.7$$
$$C_{E_1} = \$24.13 \qquad\qquad C_{E_2} = \$236.83$$
$$\text{and } \frac{r_1}{r_2} = (9.8)^{\frac{1}{2}} = 3.1$$

Sudman (1972) discussed the problem earlier for very rare populations where zero clusters had not been identified. In that situation, it was sometimes optimum either from a Bayesian or a total survey error perspective to omit strata with very few members of the special population. The same would be the case here if π_j is much less than .01 and the jth stratum contains a small fraction of the total special population.

Procedures for variable sampling rates require weighting and introduce administrative complexities. They are worthwhile, however, to reduce the very high screening costs associated with locating isolated members of special populations.

8. Non-Clustered Populations

For special populations that are not geographically clustered the

network procedures that have been developed by Sirken and Nathan can be generalized (Nathan, 1976; Sirken, 1970, 81). As initially used, these procedures improved estimates of births and deaths by obtaining information from a respondent, not only about members in the household, but also about close relatives (sons and daughters, brothers and sisters) who lived outside the house. The procedure permitted the computation of multiplicity weights to account for the fact that there were known probabilities that the same birth or death could be reported in several households.

The direct extension is simply to ask for the name and address of members of a special population using a fixed inclusion rule. Thus, in two separate recent examples, relatives have been asked to identify Viet Nam War veterans and cancer patients. The probability of a person being identified is proportional to the number of households which contain a person who can identify the member of the special population.

While the theory of this procedure is well developed, it is limited by possible response errors. The empirical question is whether respondents can report with reasonable accuracy about the characteristics or behavior of prominated persons. The empirical data reported by Nathan (1976) and by Rothbart, Fine and Sudman (1981) does indicate that individuals can report well about close relatives such as children and siblings, but with lower levels of accuracy about nephews.

In theory, there is no reason this procedure could not be expanded to larger networks such as neighbors, friends, co-workers or members of an organization. It would be necessary, however, to develop and test procedures for specifying who is to be included or excluded, and to be able to estimate response errors.

<u>Example 8.1</u>

Suppose one wished to identify a sample of persons who are missing one or more limbs so that interviews could be conducted with them. The following series of location question might be considered:

<u>Conventional</u>: Is there anyone in this household who is missing one or more limbs?

<u>Close relatives</u>: How many sons or daughters do you and your spouse have living away from home? How many brothers or sisters do you and your spouse have living away from home?

Do any of your children have any missing limbs?

Do any of your or your spouses brothers or sisters have any missing limbs?

<u>Distant relatives</u>: How many nieces and nephews do you have, whom you keep in touch with, at least once in a while?

How many aunts or uncles do you have, whom you keep in touch with, at least once in a while?

How many cousins do you have, whom you keep in touch with, at least once in a while?

Do any of your nieces of nephews have a missing limb?

Do any of your aunts or uncles have a missing limb?

Do any of your cousins have a missing limb?

<u>Neighbors</u>: About how many neighbors living _____*_____ would you recognize if you met them?

* Multilisting buildings : in the building

  Single family - urban:on this block

              - rural : around here.

Do any of your neighbors have any missing limbs?

Co-Workers: About how many other people work with you in your department (group, unit, etc.)? Do any of your co-workers have a missing limb?

It is evident from the example that different response errors, of greater or lesser seriousness, are possible. The error of misclassifying a person as falling into the special population when he or she did not would be corrected when that person was contacted directly. The reverse error of not including a person in the special population when the person belongs there cannot be corrected and leads to an overstatement of the probability of selection, and thus to a sample bias. This undercoverage bias depends both on the topic and the closeness of the acquaintance between the nominator and the nominees.

One potential problem, locating a nominated respondent, appears to require only modest effort. Even if the nominator does not have an exact address (or even name) it is usually possible to enlist intermediaries who are also in the network. Thus, when an aunt did not know the address of a nephew who was a Viet Nam war veteran, she did know the address or phone of her sister who did know her son's address. (Rothbart, Fine and Sudman, 1980).

Potentially the most serious and least understood response error, except for close relatives, is in the estimate of the size of the network. This estimate can be obtained from either the nominator or the nominee. Almost certainly the absolute error is a function of network size, which would limit the use of very large networks. Further research is needed to determine the size and directions of response errors in estimating network

size, as well as the possible demographic and social psychological correlates
of these response errors.

Nevertheless, as the Viet Nam veteran study demonstrated, it is
possible to reduce screenings in half by using relatives, and even larger
reduction in cost are possible for larger networks. For rare special
populations, total survey error may be minimized by using fairly large
networks, even in presence of response errors.

REFERENCES

1. Groves, Robert and Kahn, Robert. Surveys by Telephone: A National
    Comparison with Personal Interviews. New York: Academic, 1979.

2. Hansen, Morris, Hurwitz, William and Madow, William. Sample Survey
    Methods and Theory, Volume 1. New York: Wiley, 1953.

3. Kish, Leslie. Survey Sampling. New York: Wiley, 1965.

4. Nathan, Gad. "An Empirical Study of Response and Sampling Error for
    Multiplicity Estimates with Different Counting Rules." Journal of the
    American Statistical Association, 71, (December, 1976) 808-15.

5. Rothbart, George, Fine, Michelle and Sudman, Seymour. "On Finding and
    Interviewing the Needles in the Haystack: The Use of Multiplicity
    Sampling." Working Paper, Center for Policy Research, New York, 1981.

6. Sirken, Monroe, "Household Surveys with Multiplicity," Journal of the
    American Statistical Association, 63, (March, 1970) 257-66.

7. Sirken, Monroe. "Network Sampling in Health Surveys." Proceedings,
    Third Biennial Conference on Health Survey Research Methods, Hyattsville,
    Maryland: National Center for Health Service Research, 1981.

8. Sudman, Seymour. "Optimum Cluster Designs Within a Primary Unit Using
    Combined Telephone Screenings and Face-to-Face Interviewing." Journal
    of the American Statistical Association, 73, (June, 1978) 300-4.

9. Sudman, Seymour. "On Sampling of Very Rare Human Populations." Journal of
    the American Statistical Association, 67 (June, 1972) 335-9.

10. Waksberg, Joseph. "Sampling Methods for Random Digit Dialing." Journal
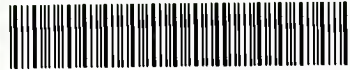   of the American Statistical Association, 73 (March, 1978) 40-46.