J. R. Statist. Soc. A (1986)
149, *Part* 1, *pp.* 65–82.

Sampling Rare Populations

By GRAHAM KALTON† and DALLAS W. ANDERSON‡
University of Michigan, USA
National Institute of Neurological and Communicative Disorders and Stroke, USA

SUMMARY

The design of an efficient procedure for sampling a rare population can present a challenging task. This paper reviews a variety of methods for sampling rare populations, including screening methods, the use of disproportionate sampling, multiplicity sampling, multiple frames and snowballing. The practical feasibility of the methods is discussed, and examples of applications are given.

Keywords: RARE POPULATIONS; SCREENING; DISPROPORTIONATE SAMPLING; MULTIPLICITY OR NETWORK SAMPLING; MULTIPLE FRAMES; SNOWBALLING

1. INTRODUCTION

The design of an efficient sample for surveying a rare population is one of the most challenging tasks confronting the sampling statistician. This paper reviews a variety of methods available for sampling rare populations. Although emphasis here is on health surveys, the methods are applicable more generally. For present purposes a rare population is taken to be a small subset of the total population. "Small" may be as large as one tenth or as small as one hundredth, one thousandth or even less.

The initial consideration in designing a sample for a rare population is whether there exists a separate frame for that population. If a separate frame exists, is available for sample selection, and is deemed adequate, the sample may be selected from it using standard methods and no special problems arise. That situation will not be discussed further. Frequently no single separate frame of adequate coverage is available; the methods discussed below have been developed for this situation.

Three purposes in surveying a rare population need to be distinguished:

(i) The estimation of the number of persons in the rare population, M, the prevalence of the rare population in the total population, $P = M/N$, where N is the size of the total population, and the prevalence in various subgroups of the population. (Following Elandt-Johnson, 1975, prevalence is defined here as the ratio of the number of cases at a given time to the size of the population at that time.) An example is a survey conducted to estimate the number of persons with a hearing defect, the prevalence of the defect in the total population, and the prevalence in age, sex and other subgroups.

(ii) The estimation of means of certain variables for persons in the rare population, or proportions of the rare population with certain characteristics. These population parameters may be denoted by

$$\bar{Y} = \sum_{i}^{M} Y_i/M;$$

in the case of a proportion, $Y_i = 1$ if the ith person has the characteristic, and $Y_i = 0$ if not. An example is a survey to estimate the mean annual cost of treatment for hearing

† *Present address:* Survey Research Center, University of Michigan, Ann Arbor, Michigan, 48106-1248, USA
‡ *Present address:* National Institutes of Health, Federal Building, Bethesda, Maryland, 20892, USA

0035—9238/86/149065 $2.00

defects among persons with such defects, and the proportion of these persons with hearing aids.

(iii) As Kish (1965 a, b) points out, sometimes the mean or total of the Y-variable may also be needed for the entire population, with $Y_i = 0$ for all those who are not members of the rare population. The population parameters are then

$$\bar{Y}_t = \sum_i^N Y_i/N = P\bar{Y} \quad \text{and} \quad Y_t = N\bar{Y}_t = M\bar{Y}.$$

An example is a survey to estimate the mean cost of treatment for hearing defects in the entire population, where the rare population comprises all persons with hearing defects.

In practice, of course, surveys are often required to serve more than one of these purposes.

In many cases, the sample design for a rare population includes the selection of a large sample from a more general population and then the identification of members of the rare population within that sample. The ease and costs of screening a large sample depend on the prevalence of the rare population in the more general population, the ease with which members of the rare population can be identified, and the methods of data collection used for screening and for the main survey. Issues of screening are taken up in Section 2. Sometimes information is available on the concentration of the rare population in different sectors of the more general population. In this case, disproportionate sampling may be used to reduce screening costs, with the sectors as strata and sampling the sectors in which the rare population is highly concentrated at higher rates; this approach is discussed in Section 3. Screening costs may also be reduced by allowing more than one member of the more general population to report on a rare event. This fact has led to the development of multiplicity (or network) sampling methods, covered in Section 4. Even though no single complete sampling frame is available for a rare population, there are sometimes incomplete lists that can be used to sample parts of the population. In this case, a combination of incomplete lists and a "catch-all" frame may be used; this approach is discussed in Section 5. Another possibility is to compile a frame for the rare population. Snowballing, discussed in Section 6, is one method for constructing such a frame. Section 7 then briefly describes the use of sequential sampling and some other methods for sampling rare populations. The paper concludes with a comment on the usefulness of combinations of the methods previously discussed.

2. SCREENING

A common feature of most samples of rare populations is the need to screen a large sample in order to identify members of the rare population. Sometimes most or all of the listings that do not represent members of the rare population can be identified as such from the information on the sampling frame. In this case, the sampling issues are reasonably straightforward: either the frame can be cleaned of those listings before the sample is selected, or a larger sample can be selected initially and those selections that are clearly not members of the rare population based on the information given in the sampling frame can be eliminated prior to the fieldwork. Any remaining non-members of the rare population found in the sample are then eliminated at the data collection stage. Usually, however, the sampling frame does not contain the information necessary to distinguish between members and non-members of the rare population. Then, all the non-members have to be identified and eliminated at the data collection stage. If the rare population constitutes a sizeable proportion of the frame population, the screening costs may not be too severe, but screening costs increase rapidly as the degree of rarity rises.

A number of techniques may be employed to reduce screening costs. The most important is the use of an economical data collection method to obtain the screening information. Telephone interviewing is widely used in the U.S. for this purpose. If screening is conducted by telephone, it is convenient if the main survey is conducted by the same mode. If, say, the

telephone is used to identify members of the rare population, and the main survey data are then collected by face-to-face interviews, the telephone sample generally needs to be clustered geographically to control the travel costs of the interviewers working in the main survey. A serious concern in using telephone sampling for a rare population is the coverage of the telephone frame for that population. Although only about 7 per cent of U.S. households are inaccessible by telephone, there is the risk that this rate may be much higher for the rare population. Telephone surveys underrepresent low-income households, households in the South, households with non-white heads, households with heads under 35, and households with single, divorced or separated heads (Thornberry and Massey, 1983). If the rare population is concentrated in such groups, telephone sampling will be open to the possibility of significant bias. A telephone sample obtained by random digit dialling has, for instance, been found to underrepresent the deaf population and those limited in activity because of chronic conditions (Freeman *et al.*, 1982).

An alternative data collection method for screening is the mail questionnaire. This method is widely used in the U.K. where telephone coverage is not as high as in the U.S. and where the Register of Electors provides a sampling frame of names and addresses. Harris (1971), for instance, describes a survey of the handicapped and impaired for which 250,000 households were screened by mail questionnaires at the initial phase; a response rate of 85.6 per cent was obtained for this phase. Cartwright (1964) sent a mail questionnaire to 29,400 persons to identify those who had been in hospitals in the last six months, and obtained a response rate of 87 per cent. Hunt (1978) sent a screening questionnaire seeking basic demographic details about the members of 11,500 households to identify a sample of the elderly, and achieved a response rate of 80 per cent. Although these response rates are high for mail surveys, there is the risk that the nonrespondents include a sizeable proportion of members of the rare population.

Most sample designs employ clustering for reasons of economy. In face-to-face interview surveys, clustering is used to reduce both sample selection costs and travel costs of interviewers, while in telephone surveys clustering is used to increase efficiency in locating working household numbers. As a rule, clustering leads to a loss of precision in the survey estimates as compared with an unclustered design of the same sample size. The design effect for the mean of a variable from a cluster sample can be represented by $[1 + (\bar{b} - 1)roh]$, where *roh* is a synthetic measure of the homogeneity of the variable in the clusters (Kish, 1965a). In surveys of the general population $\bar{b}$ is the average sample size per selected cluster; when dealing with a subclass that is fairly evenly spread across the clusters, however, $\bar{b}$ is replaced by $\bar{b}_0$, the average subclass sample size per cluster. For a rare population (i.e. subclass) evenly spread across the clusters, a sample with a large $\bar{b}$ value for the general population can be tolerated since it will result in a small $\bar{b}_0$, and hence small design effects for estimated means of the rare population. Thus an efficient sampling procedure for selecting a sample of a rare population is to screen large samples in selected clusters. In face-to-face interview surveys, complete screening of blocks or other units can be an effective design for some rare characteristics, giving good coverage as well as being efficient.

When a rare population is geographically clustered, it can be efficient to sample the clusters in which the rare population is more heavily concentrated at higher rates. The widely-used method for improving the efficiency of random digit dialling sampling in telephone surveys described by Waksberg (1978) can be adapted for this purpose. The first step is to select a cluster with probability proportional to the size of its total population, and to select one element in that cluster. If the selected element is a member of the rare population, it is included in the sample and further screening interviews are taken in the cluster until a set number of additional members of the rare population (K) is selected. If the initially selected element is not a member of the rare population, the cluster is rejected. The process is repeated until the required number of clusters is accepted for the sample. This procedure in effect samples the clusters with probabilities proportional to their numbers of members of the rare population. It

may be noted that clusters with no members of the rare population are always rejected at the first screening interviews. Sudman (1985) gives a formula for the optimum choice of K and describes a variety of possible applications of this approach, and Blair and Czaja (1982) describe its use with random digit dialling sampling for sampling black and affluent households. A problem that can occur with the procedure is that in some of the clusters accepted for the sample the number of members of the rare population may be less than $(K + 1)$, the number required for the sample; this creates the need for weighting which causes a loss in efficiency (see Waksberg, 1983).

The screening information needed to identify members of the rare population is not always straightforward to collect. Moreover, even when the screening questions are fairly simple, due attention needs to be paid to response errors. As Sudman (1972) observes, if the rare population is defined by some combination of characteristics (e.g. black, adult, employed males under 30 years of age), response errors can occur to each of the several screening questions, thus possibly leading to serious levels of misclassification in the combination.

Identification of members of the rare population is sometimes a costly operation, as for instance when a medical diagnosis requires a trained physician or expensive, perhaps not transportable, equipment. When this applies, a two-phase approach can sometimes be profitable: at the first phase a relatively cheap but imperfect screening is used to divide a large sample into two or more strata according to the likelihood of being in the rare population, and then a subsample is selected for the expensive measurement, sampling the strata with higher likelihoods at greater rates. In many applications, the first phase sample is divided into just two strata, those who may be members of the rare population and those who are not; then often all of those in the first stratum and none of those in the second stratum are included in the second phase of the survey. Anderson *et al.* (1982), for example, conducted a survey of the prevalence of major neurological disorders in a rural county in the U.S. using a two-phase design. At the first phase, a screening questionnaire was administered in inhabited dwellings to collect certain demographic and medical information on the occupants. The medical information was used to identify persons who required the second-phase clinical examination by a neurologist because of their likelihood of having one or more of the disorders under study.

A two-phase approach is beneficial only when the first phase screening costs are much lower than the second phase costs. Deming (1977) suggests that the ratio of the second to first phase screening costs needs to be at least 6:1 and it should be preferably much larger, e.g. 40:1 or more. With an initial screening into two strata, those classified as members and nonmembers of the rare population, the stratum of those classified as nonmembers should contain very few members of the rare population if the first phase screening is to be effective. In other words the screening needs to keep the number of "false negatives" to a very small proportion: the proportion of "false positives" is not so critical. Deming (1977) gives tables from which he concludes that for the first phase screening to be economical, the proportion of false negatives needs to be less than one quarter of the overall proportion that members of the rare population represent of the total population.

One way that can sometimes be used to keep the proportion of false negatives small, at the cost of an increase in the proportion of false positives, is to use less stringent criteria for defining the rare population at the first phase. Thus, for instance, Fry *et al.* (1969) used a lower level of hearing loss for defining significant deafness in a sample of children at the first phase, when their hearing was measured at home, than was used at the second phase which was conducted in a controlled laboratory setting. Similarly, in a fertility study of women aged 15–45, the initial screening may specify a wider age range to deal with minor misreporting of ages at the boundaries. When the screening criteria are set such that the proportion of false negatives is negligibly small, there is no need to subsample from the negative stratum at the second phase: however, the assumption that the proportion of false negatives is negligible needs to be made with due caution, for if the negative stratum is large even a small proportion of false negatives can constitute a sizeable proportion of the rare population.

3. DISPROPORTIONATE SAMPLING

It is sometimes possible to identify strata with higher concentrations of the rare population. Thus, for instance, the rare population may be concentrated in certain geographical areas, or it may be concentrated in a stratum comprising a list frame. In such situations, it may be efficient to sample the strata in which the rare population is concentrated at higher sampling fractions. We will consider first the use of disproportionate sampling for estimating the prevalence of the rare population and afterwards its use for estimating a mean for members of the rare population.

3.1. Estimating Prevalence

Suppose that the purpose of the survey is to estimate the prevalence of the rare population, $P = M/N$. For simplicity, suppose that there are two strata, with stratum 1 having a high prevalence of the rare population (P_1) and stratum 2 a low prevalence (P_2), and assume that the selections are to be made by simple random sampling in each stratum. Then, by standard theory for optimum allocation, the sample sizes (n_h) in the two strata should be made proportional to $W_h S_h/\sqrt{c_h}$, where W_h is the proportion of the overall population, $S_h \cong \sqrt{(P_h Q_h)}$ is the element standard deviation, $Q_h = 1 - P_h$, and c_h is the cost per unit of sampling, all in stratum h ($h = 1, 2$). Often the costs c_h in the two strata are approximately equal, in which case the optimal allocation reduces to the Neymann allocation, $n_h \propto W_h\sqrt{(P_h Q_h)}$. Ignoring fpc's, the ratio of the variance of the sample estimate of prevalence with Neyman allocation to that with a proportionate allocation based on the same sample size is (Cochran, 1977, p. 110)

$$R_0 = \{\sum W_h\sqrt{(P_h Q_h)}\}^2/\sum W_h P_h Q_h.$$

To illustrate the magnitude of the gains from using Neyman allocation, consider a situation where the prevalence of the rare characteristic is $P = 0.05$ or 5 per cent. Table 1 gives the values of R_0 for various combinations of W_1 and P_1. Also given in each cell of the table are the values of: $k_0 = \sqrt{(P_1 Q_1/P_2 Q_2)}$, the value of the ratio of the optimum sampling rate in stratum 1 to that in stratum 2; $P_2 = (P - W_1 P_1)/(1 - W_1)$, the proportion of the second stratum with the rare trait; and $A = W_1 P_1/P$, the proportion of the rare population in stratum 1. For convenience in Table 1 and subsequent tables W_1, P_1, P_2 and A are expressed as percentages.

The results in the table show that sizeable gains in precision in the estimation of P (i.e. sizeable reductions in R_0 below 1) accrue only when P_1 is much larger than P and when W_1 is large. For a given value of P_1, the minimum value of R_0 occurs when $P = W_1 P_1$ so that $P_2 = 0$; in this case, no sampling is needed in stratum 2, and $R_0 = W_1 = P/P_1$. The minimum value of R_0 with $P_1/P = 2$ is thus $R_0 = 0.5$; with $P_1/P = 4$ it is $R_0 = 0.25$, etc. The values of R_0 increase rapidly above their minimum values when W_1 is less than its maximum value of P/P_1. Thus, for instance, with $P/P_1 = 2$, the minimum value of $R_0 = 0.5$ occurs when $W_1 = 50$ per cent; as can be seen from the table, with $P = 5$ per cent and $P_1 = 10$ per cent, when $W_1 = 45$ per cent, R_0 is substantially higher than 0.5 at 0.77; similarly it can be shown that with $P = 10$ per cent and $P_1 = 20$ per cent, when $W_1 = 45$ per cent, $R_0 = 0.79$.

The above findings can be expressed in a variety of alternative ways. Large values of P_1 and W_1 imply small values of P_2 and large values of A, the proportion of the rare population in stratum 1. Thus substantial reductions in R_0 require a large value of P_1 and a small value of P_2, or large values of P_1 and A. Note also that a large value of P_1 is not sufficient to ensure a substantial reduction in R_0; a large value of A, or a small value of P_2, is also needed.

3.2. Estimating a Mean for the Rare Population

Consider now the estimation of a mean $\bar{Y}$ for members of the rare population. As before, suppose that there are two strata, with stratum 1 having a high prevalence and stratum 2 a low prevalence of the rare trait. Suppose that simple random samples are selected from the two strata, with the sampling fraction in stratum 1 being k times that in stratum 2, and assume that

TABLE 1

Estimating prevalence: values of R_0, k_0, P_2 and A for combinations of P_1 and W_1, with $P = 5\%$

Values of P_1		Values of W_1						
		5%	10%	15%	20%	25%	35%	45%
10%	R_0	0.99	0.98	0.97	0.96	0.94	0.89	0.77
	k_0	1.4	1.5	1.5	1.6	1.7	2.0	3.2
	P_2	4.7	4.4	4.1	3.8	3.3	2.3	0.9
	A	10	20	30	40	50	70	90
20%	R_0	0.96	0.90	0.82	0.68	0.25	—	—
	k_0	2.0	2.2	2.6	3.6	∞	—	—
	P_2	4.2	3.3	2.4	1.3	0	—	—
	A	20	40	60	80	100	—	—
30%	R_0	0.92	0.79	0.49	—	—	—	—
	k_0	2.4	3.1	6.0	—	—	—	—
	P_2	3.7	2.2	0.6	—	—	—	—
	A	30	60	90	—	—	—	—
40%	R_0	0.89	0.61	—	—	—	—	—
	k_0	2.8	4.7	—	—	—	—	—
	P_2	3.2	1.1	—	—	—	—	—
	A	40	80	—	—	—	—	—
50%	R_0	0.85	0.1	—	—	—	—	—
	k_0	3.1	∞	—	—	—	—	—
	P_2	2.6	0	—	—	—	—	—
	A	50	100	—	—	—	—	—

the strata are large so that fpc terms may be ignored. Let A be the proportion of the rare population in stratum 1 and $(1 - A)$ be the proportion in stratum 2. We assume initially that A is known, although in practice this will seldom be the case. With this assumption $\bar{Y} = A\bar{Y}_1 + (1 - A)\bar{Y}_2$ may be estimated by $\bar{y} = A\bar{y}_1 + (1 - A)\bar{y}_2$, where $\bar{y}_1$ and $\bar{y}_2$ are the means of sample members in the rare population in the two strata. For simplicity we assume that the element variances of the y variable for members of the rare population are the same in the two strata. We also assume that the costs of data collection are the same for the two strata. However, we allow for a difference in the costs of sampling members of the rare and nonrare populations. This difference is needed to reflect the fact that when a member of the rare population is sampled a full interview is conducted, whereas when a member of the nonrare population is sampled the interview can be terminated at the end of the screening questions.

Letting c be the ratio of the cost of sampling a member of the rare population to that of sampling a member of the nonrare population, the ratio of the variance of $\bar{y}$ under a disproportionate allocation to that under a proportionate allocation is approximately

$$R \cong \frac{[kP - (k-1)W_1P_1][(c-1)\{P + (k-1)W_1P_1\} + (k-1)W_1 + 1]}{kP[(c-1)P + 1]} \tag{1}$$

and the optimum choice for k, the ratio of the sampling fractions in the two strata, is given by

$$k_0^2 = \frac{P_1[(c-1)(P - W_1P_1) + (1 - W_1)]}{(P - W_1P_1)[(c-1)P_1 + 1]}. \tag{2}$$

The derivations of these results are given in the Appendix.

If the costs of sampling members of the rare and nonrare populations are assumed to be the same, i.e. $c = 1$, the above formulae simplify considerably, with k_0 reducing to $\sqrt{(P_1/P_2)}$. For this case, equivalent results — but expressed in terms of different parameters — are derived by Waksberg (1973), who also gives a discussion of them. Table 2 presents values of R_0 (the value of R when $k = k_0$), k_0, P_2 and A for various combinations of P_1 and W_1 when the overall population percentage with the rare trait is 5 per cent and when $c = 1$. As with Table 1, Table 2 shows that the reduction in variance from using the optimum disproportionate allocation over proportionate allocation is small unless P_1 and A are large. For small values of A, even large values of P_1 are not adequate to cause R_0 to be small; for instance, when $P_1 = 50$ per cent and $W_1 = 5$ per cent (hence $P_2 = 2.6$ per cent), R_0 is as high as 0.72. For a given value of P_1, the value of R_0 declines at a relatively slow rate as A increases to about 80 per cent, and then

TABLE 2

Estimating a mean for the rare population: values of R_0, k_0, P_2 and A for combinations of P_1 and W_1, with $P = 5\%$ and $c = 1$

Values of P_1		*Values of* W_1						
		5%	10%	15%	20%	25%	35%	45%
10%	R_0	0.99	0.98	0.97	0.95	0.93	0.88	0.76
	k_0	1.5	1.5	1.6	1.6	1.7	2.1	3.3
	P_2	4.7	4.4	4.1	3.8	3.3	2.3	0.9
	A	10	20	30	40	50	70	90
20%	R_0	0.94	0.87	0.78	0.64	0.25	—	—
	k_0	2.2	2.4	2.9	4.0	∞	—	—
	P_2	4.2	3.3	2.4	1.3	0	—	—
	A	20	40	60	80	100	—	—
30%	R_0	0.88	0.71	0.43	—	—	—	—
	k_0	2.9	3.7	7.1	—	—	—	—
	P_2	3.7	2.2	0.6	—	—	—	—
	A	30	60	90	—	—	—	—
40%	R_0	0.80	0.50	—	—	—	—	—
	k_0	3.6	6.0	—	—	—	—	—
	P_2	3.2	1.1	—	—	—	—	—
	A	40	80	—	—	—	—	—
50%	R_0	0.72	0.10	—	—	—	—	—
	k_0	4.4	∞	—	—	—	—	—
	P_2	2.6	0	—	—	—	—	—
	A	50	100	—	—	—	—	—

	Values of P_1				
$W_1 = 5\%$	60%	70%	80%	90%	100%
R_0	0.62	0.52	0.40	0.27	0.05
k_0	5.3	6.7	8.7	13.1	∞
P_2	2.1	1.6	1.1	0.5	0
A	60	70	80	90	100

declines rapidly as A increases from 80 per cent to 100 per cent. The maximum reduction occurs for a given value of P_1 when $A = 100$ per cent in which case $R_0 = W_1$.

When $A = 100$ per cent, $P_2 = 0$ and $k_0 = \infty$: no sampling is needed in stratum 2 since it contains no members of the rare population. When A is close to 100 per cent, P_2 is close to 0, so that a sample taken in stratum 2 will yield only a small number of members of the rare population. For this reason, it is common practice to confine the sample to stratum 1 when $(1 - A)$ is known to be sufficiently small. This practice leads to biased estimators, but provided $(1 - A)$ is sufficiently small and the members of the rare population in stratum 2 are not too different from those in stratum 1, the bias will be negligible compared with the standard error of the estimator. One application of this approach can arise when using area sampling to sample members of a minority population. If the minority is highly concentrated geographically, it may be possible to identify a small proportion of areas that contains a high proportion of the minority, and then restrict the sample to those areas (Hedges, 1979). A sufficiently high degree of concentration occurs seldom in practice, however. Moreover, considerable caution is needed in applying this cut-off method, since the data on which the exclusions are made are generally out-of-date: if the distribution of the minority has changed markedly in the interim, a serious bias can result.

Table 3 illustrates the effect of differential costs for including members of the rare and nonrare populations in the survey. The table gives values for R_0 and k_0 for varying values of P_1 and c, holding P and W_1 fixed at 5 per cent and 10 per cent respectively. The results show that for a given value of P_1, the value of R_0 increases and the value of k_0 declines as c increases. With $P_1 = 40\%$, for instance, $R_0 = 0.50$ if $c = 1$ but $R_0 = 0.76$ if $c = 10$. When c is much greater than 1, as would often be the case in practice, the gains from optimum allocation are appreciably less than those given in Table 2.

TABLE 3

Estimating a mean for the rare population: values of R_0 and k_0 for various values of P_1 and c, with $P = 5\%$ and $W_1 = 10\%$

c	P_1									
	10%		20%		30%		40%		50%	
	R_0	k_0	R_0	k_0	R_0	k_0	R_0	k_0	R_0	k_0
1	0.98	1.5	0.87	2.4	0.71	3.7	0.50	6.0	0.10	∞
2	0.98	1.5	0.89	2.3	0.75	3.3	0.55	5.1	0.14	∞
5	0.99	1.4	0.92	1.9	0.82	2.6	0.66	3.8	0.25	∞
10	0.99	1.3	0.95	1.7	0.88	2.1	0.76	2.9	0.38	∞

A limitation to the preceding results is that they are based on the assumption that A, the proportion of the rare population that is in stratum 1, is known, whereas in practice this will rarely be the case. When A is unknown, $\bar{Y}$ may be estimated by $\bar{y}' = a\bar{y}_1 + (1 - a)\bar{y}_2$, where $a = m_1 f_2/(m_1 f_2 + m_2 f_1)$, with m_1 and m_2 being the sample sizes of members of the rare population and f_1 and f_2 being the sampling fractions in the two strata. When $\bar{y}'$ is used as the estimator, formulae (1) and (2) for R_0 and k_0 need modification to include terms involving the difference between the stratum means in the survey variable (see the Appendix). However, if $\bar{Y}_1$ and $\bar{Y}_2$ are assumed equal, these additional terms disappear, and the preceding formulae hold as approximations for this case also. Moreover, even if $\bar{Y}_1$ and $\bar{Y}_2$ are not equal, formulae (1) and (2) will still serve as good guides provided that the between stratum variance is small relatively to the within stratum variance, as will often be the case in practice.

Finally, it should be noted that the values of R_0 reported in the above tables are the

minimum values obtained with the optimum allocation. The value of the optimum ratio of the sampling fractions in the two strata, k_0, depends on the parameters P_1, W_1 and c. In practice, none of these will be known precisely, so that the optimum allocation will not be achieved exactly. Provided a value of k close to k_0 is used, however, the associated value of R will not be much greater than R_0.

3.3. General Comments

In line with Kish (1965a, p. 406) and Waksberg (1973), the general conclusion from the above results is that the gains from disproportionate stratification will be sizeable only when two conditions apply: first, the strata to be oversampled need high concentrations of the rare population and, second, the strata need to contain a substantial proportion of the rare population. This conclusion applies both for estimating the prevalence of the rare population and for estimating characteristics of the rare population.

As Waksberg (1973) points out, the computations of optimum sampling rates for disproportionate stratification are usually based on past data such as the most recent Census. Changes in the distribution of the rare population across the strata since the time those data were collected commonly lead to lower concentrations of the rare population in the strata to be oversampled. When this is so, the gains in precision will be smaller than indicated above. The sample size for the rare population will also be smaller than predicted based on the past data.

Ericksen (1976) provides an interesting illustration and analysis of the use of disproportionate stratification with area sampling in the selection of a sample of females aged 15 to 19 years old, with a higher sampling fraction for blacks.

4. MULTIPLICITY SAMPLING

The essential problem with sampling a rare population is that many contacts are required to identify the sample members with the rare trait. Multiplicity, or network, sampling can sometimes be used to reduce the number of contacts needed.

It is generally desirable that a sampling frame provides a single listing for each population element; otherwise the frame problem of duplicate listings arises (see, for instance, Kish, 1965a, Section 11.2). However, instead of seeking to avoid duplicate listings, multiplicity sampling creates such duplicates and capitalizes on them for data collection.

In a conventional household survey, information on persons with a rare trait is collected in respect of members of the sampled households only. Thus each member of the population is included in the sample only if his or her household is selected. The information may be collected from each person individually, or one member of the household may report the information for all eligible household members. The latter procedure has benefits in terms of economy of data collection, but it requires that the informant can provide the requisite information accurately for other household members. With a multiplicity sample design, information is provided by a selected household not only about its own household members but also about other persons who are linked to that household in clearly defined ways. One common form of linkage, for instance, is to include all close relatives of household members —e.g., children, parents and siblings; linkages of this type have been used in methodological and pilot studies of diabetes (Sirken *et al.*, 1978), cancer (Sirken *et al.*, 1980; Czaja *et al.*, 1984), births and deaths (Nathan, 1976) and Vietnam era veterans (Rothbart *et al.*, 1982). Another form of linkage is to include addresses adjacent to the selected households. Such linkages have been used in surveys of ethnic minorities (Brown and Ritchie, 1981), and in a pilot survey of home vegetable gardeners using sewage sludge (Bergsten and Pierson, 1982).

Linkages need to be clearly specified so that the selection probabilities of sample members can be determined. Thus, for instance, in an equal probability sample of households with all members of selected households being included and with a multiplicity linkage to siblings, the probability that a given individual is included in the sample is proportional to the number of different households in which he and his siblings are living. Weights inversely proportional to

these selection probabilities are needed in the analysis. Note that an individual living in an institution has no chance of being selected in a conventional household survey, but he or she may have a chance of being selected for a multiplicity sample: this is another advantage of multiplicity sampling.

The critical consideration in the choice of linkages is the ability of informants to provide the necessary information. At the least, they need to be able to provide details on the linkages and to give accurate information on whether those linked to them have the rare trait. These data are the minimum required for estimating the prevalence of the rare trait, but even for prevalence surveys more details are needed: for instance, a prevalence survey of diabetics would also need information on the age and sex of each sampled person so that age- and sex-specific prevalences could be computed. In a survey of the characteristics of those with the rare trait, information needs to be collected on those characteristics (for instance, on the cost of medical care for cancer patients). It is frequently impossible for a sampled person to provide information on these characteristics for others linked to him. In this case, he has to be able to provide a means of locating persons with the rare trait who are linked to him so that direct contact can be made with them. When direct contact is needed, this factor needs to be taken into account in defining the linkages. With a face-to-face interview survey, it is economically advantageous to restrict the linkages to persons living in the same geographical area as the initial informant. With telephone surveys, the initial informants have to provide information to make contacting the linked persons possible. Bergsten and Pierson (1982), for instance, found that the information they collected from initial informants was sufficient to enable only 70% of neighbours to be contacted in a telephone survey.

The preceding illustrations of linkages are ones that have been purposively introduced in household surveys to extend the range of data collected from each informant. Multiplicity also occurs naturally when rare populations are sampled by means of their contacts with certain establishments, as for instance when patients with a rare illness are sampled from hospital records or from pharmacy records of prescriptions for drugs that treat that illness (Sirken, 1984). The establishments may have several records for each patient (e.g. several prescriptions) and the patient may have records with several establishments. The establishment records rarely contain information on the multiplicity and they seldom contain the substantive information to be collected in the survey. Thus, follow-up surveys with the patients are generally needed. Often a major obstacle to the conduct of a follow-up survey is that many establishments will refuse to grant permission for contact to be made with the patients. Even when they grant permission, a problem can arise with obtaining accurate information from the sampled patients on their multiplicity. Lessler (1981) describes a method for reducing this problem. Bryan *et al.* (1984) describe an application of this use of multiplicity sampling in a pilot study on epilepsy.

Since the ideas of multiplicity sampling were first propounded by Birnbaum and Sirken (1965), there have been a number of theoretical and conceptual developments of the technique (e.g., Sirken, 1970, 1972a, 1972b; Sirken and Levy, 1974; Levy, 1977; Nathan, 1976). There has also been a variety of applications. While multiplicity sampling is a useful addition to the battery of techniques for sampling rare populations, it is no panacea. A number of factors need to be taken into account when choosing between a conventional sample and a multiplicity sample. The clear advantage of a multiplicity sample is that it needs a smaller sample to yield the required sample size of members of the rare population. However, the use of informants in multiplicity sampling is frequently likely to increase significantly the level of response error (although their use may on occasion lead to a reduction of certain types of response bias—Sirken and Royston, 1970), and the need for weights in the analysis of a multiplicity sample causes an increase in sampling error. Problems can arise in determining the linkages with a multiplicity design, and in any case additional questions need to be asked to identify the linkages; item nonresponse to these additional questions also creates difficulties (Sirken *et al.*, 1978). When sampled members of the rare population need to be contacted in person, the costs

of tracing and contacting them in a multiplicity design may be considerable. An ethical question with multiplicity sampling is whether it is appropriate to collect the survey data from an informant who is not even a member of the linked person's household. In some cases this question may not raise much concern. However, with private information (which could include health matters) there may be serious difficulties, and especially so if the survey design calls for a subsequent interview with persons identified as having a rare trait. Thus, for instance, in a survey of cancer patients, there could be serious ethical concerns about interviewing a person identified as having cancer by a sibling living in a different household. For these various reasons the choice between a multiplicity and a conventional sample is a complex one that requires a careful assessment of relative costs and relative quality of the resultant estimators.

5. MULTIPLE FRAMES

Even though a complete list of the rare population is not available, there may exist one or more partial lists. Thus, for instance, hospital records may provide a means of identifying a sizeable proportion of persons with a particular illness, but they may fail to cover an important minority who have different characteristics from hospital attenders. In such a case, the sample may be made up of two components: first, a sample from the partial list and second a sample from the total population to screen for persons with the illness. Since the screening sample is expensive, an economic design will sample the partial list at a higher sampling fraction than is used for the screening sample. Sometimes several partial lists are available, but a screening sample from the total population may still be needed to give representation to those on none of the lists.

When multiple frames are used, it is likely that some members of the rare population will be included on more than one frame: for instance, a person with a particular illness may attend one or more hospitals for that illness and should also be covered by the area frame. The representation of some population members on more than one frame gives rise to the frame problem of duplicates (see Kish, 1965a, Section 11.2). There are two basic approaches for handling duplicates: one is to redefine the frames so that they are non-overlapping and the other is to make compensations in the analysis.

5.1. Eliminating Overlaps

One way to eliminate overlaps is to collate the several list frames into a single list without duplicates, and to define the screening sample from the total population to be a sample of rare population members not on the combined list. The collation of a single list requires the matching of listings across frames, a process that is notoriously error-prone. Problems such as misspelt names and alternative versions of addresses can give rise to failures to match and mismatches, and the resultant biases need to be taken into account. In practice, it is often wiser to define the frames in a way that makes matching simpler, even if this is achieved at the cost of some loss of statistical efficiency in the sampling methods.

One example of this general approach is the sample of retail stores described in Hansen, Hurwitz, and Madow (1953, pp. 516-558). Within selected primary sampling units, all retail stores on a combined list were included in the sample and an area sample was taken to give representation to stores not on the list. Another example is a study of the deaf population in Washington in the 1960's. For this study, as complete a list of deaf persons as possible was compiled from organizations for the deaf, schools for the deaf, deaf informants, social agencies, etc. All those on this list found to be eligible were then included in the survey. In addition, an area sample with an overall sampling fraction of 1 in 120 was taken to check on the completeness of this list and, if needed, to give representation to unlisted deaf persons (Schein, 1968). It may be noted that when a complete enumeration is taken from the combined list, failures to match listings across the separate frames are not so serious a problem since they will be resolved during fieldwork.

An alternative procedure for eliminating overlaps is to use a unique identification to specify one of the listings as the real listing, with the other listings being treated as blanks. This may be illustrated by the earlier example of persons with a particular illness. The hospital lists could be ordered, with a person being identified with the first list on which he or she appears, and then his or her inclusions on the lists of later hospitals are treated as blank entries; the area frame could be defined to include only persons with the illness who were on none of the hospital lists. The procedure involves selecting a sample of listings throughout the several frames, and then for each sampled listing searching earlier frames to determine whether there is a prior listing. If there is a prior listing, the element is rejected for the sample; if not, the element is accepted. The procedure avoids the need to collate the several frames; the searches are made only for sampled listings and only in prior frames. Even in this case the searches for matching listings are not always easy to conduct. Sometimes the searching task can be simplified by choosing a suitable order for the frames. For instance, a frame not in alphabetical order may be best placed last and a well-organized long frame placed first in the frame order (Kish, 1965a, Section 11.2D).

The choice between these two procedures depends on the ease and accuracy with which a combined list can be generated and on the purpose of the survey. When the lists are not computerized and when they are long and not ordered in a systematic fashion (e.g. not in alphabetical order), merging can be a major undertaking. Merging is much more feasible when the lists are computerized, but even in this case failures to match and mismatches can pose severe problems.

The creation of a combined frame without duplicates can be especially useful when the purpose of the survey is to estimate the size of the rare population, M, and rare population totals like $Y = M\bar{Y}$. Indeed, when the combined frame is comprehensive and contains no blanks, M is known exactly and Y may be estimated by $M\bar{y}$. Without combining the frames, M has to be estimated by Fm and Y by Fy, where F is the inverse of the sampling fraction (assumed the same for all lists), m is the rare population sample size and y is the sample total for members of the rare population. For a given sample size, $M\bar{y}$ has smaller variance than Fy (see Kish, 1965a, Section 11.8).

5.2. Compensating for Overlaps

When a sample is drawn from two or more overlapping frames, the chance of an element being selected depends on the number of frames on which it appears. Compensation for the varying inclusion probabilities of different population elements may be made by means of a weighting adjustment in the analysis. One widely-used adjustment method is to assign sample elements weights made inversely proportional either to their inclusion probabilities or to their expected number of selections. Assuming that no population element can appear in the sample more than once, the weights should be made inversely proportional to inclusion probabilities. With independent sampling between lists, the inclusion probability for the ith sample member is

$$\sum_j p_{ij} - \sum\sum_{j<k} p_{ij}p_{ik} + \sum\sum\sum_{j<k<l} p_{ij}p_{ik}p_{il} \cdots,$$

where p_{ij} is the probability that the ith sample member is selected from the jth list. If a sample element selected from more than one list is included in the sample once for each selection, the weight should be made inversely proportional to the expected number of selections, i.e. inversely proportional to $\sum_j p_{ij}$. This latter weighting scheme is easier to compute. When the inclusion probabilities are small for all lists, the overall inclusion probability may also be approximated by $\sum_j p_{ij}$. Application of these weighting schemes requires knowledge of the lists on which each sample element is to be found. Where possible, this information is best obtained

from the lists; otherwise, it may be obtained from the sampled elements, but in this case the reports may be subject to response errors.

A general approach for dealing with overlapping frames is obtained from the multiple frame estimator introduced by Hartley (1962, 1974). Since a full treatment is not possible here, we will discuss only some aspects of the simple and common case with just two frames, labelled A and B. The members of the population of interest then fall into one of three mutually exclusive subsets: members on frame A only, members on frame B only, and members on both frames. One or both of the frames may also include listings that do not refer to members of the population of interest. The essence of the procedure is to divide those population members on both frames between their two listings. Thus the survey variables Y_i are divided into two parts, pY_i being associated with the listing on frame A and qY_i being associated with the listing on frame B ($p + q = 1$). In the same way, the count variable $X_i = 1$ may be divided into parts, p for frame A and q for frame B. When the numbers of members of the population of interest are known for the three subsets (N_a, N_b and N_{ab}), then a population total Y may be estimated by the poststratified estimator

$$\hat{Y}_1 = N_a\bar{y}_a + N_b\bar{y}_b + N_{ab}(p\bar{y}'_{ab} + q\bar{y}''_{ab}),$$

where $\bar{y}_a$ is the sample mean for those who are only on frame, A, $\bar{y}_b$ is the corresponding quantity for frame B, $\bar{y}'_{ab}$ is the mean of those on both frames sampled from frame A and $\bar{y}''_{ab}$ is the mean for those on both frames sampled from frame B. When N_a, N_b and N_{ab} are unknown, then Y may be estimated by

$$\hat{Y}_2 = F_A(y_a + py'_{ab}) + F_B(y_b + qy''_{ab}),$$

where F_A and F_B are the inverses of the sampling fractions for the two frames and y_a, y_b, y'_{ab} and y''_{ab} are sample totals. An important special case occurs when frame A, say, provides complete coverage (e.g. a frame of the total population) while frame B provides only partial coverage. In this case $N_b = 0$ and the terms $N_b\bar{y}_b$ in $\hat{Y}_1$ and F_By_b in $\hat{Y}_2$ drop out.

This approach provides a general framework for handling overlapping frames. The use of unique identification discussed in the previous subsection, for instance, can be viewed as a special case with $p = 1$. If p is set equal to $F_B/(F_A + F_B)$, the estimator $\hat{Y}_2$ is equivalent to the one obtained by weighting sample members inversely to their expected numbers of selections. The approach has the attraction that it provides a means for determining the optimum sampling fractions and the optimum value of p to be used. Thus, the values of the sampling fractions and of p can be determined to minimize the variance of the sample estimator subject to some cost function for sampling from each of the frames. For further details, the reader is referred to the papers by Hartley (1962, 1974), Cochran (1964), Fuller and Burmeister (1972), and the sizeable recent research on the use of dual frame estimation techniques to augment telephone surveys by face-to-face interviews (Lund, 1968; Cassady *et al.*, 1981; Groves and Lepkowski, 1982; Lepkowski and Groves, 1984).

6. SNOWBALLING

The techniques discussed so far can be valuable tools for sampling rare populations, but even with their use a survey of an extremely rare population can still remain prohibitively expensive. For extremely rare populations, researchers sometimes resort to what is generally known as snowball or reputational sampling.

A necessary condition for successful applications of snowballing is that members of a rare population know each other. This condition does not hold for all rare populations, but it may well hold for certain rare ethnic minorities, religious groups, persons with disabilities (e.g. deaf people), etc. One application of snowballing is to create a frame of members of the rare population. The approach is to identify a few members of that population, to ask each of them to identify other members, to contact those so identified and ask them to identify others, and so on. When the frame has been compiled, a probability sample can then be drawn from it. The

critical issue with this use of snowballing is simply the completeness of the frame. Those missing from the frame are likely to be those socially isolated from other members of the rare population; the survey estimates will be biased if this factor is associated with the survey variables.

A more common application of snowballing avoids the construction of the frame by simply continuing the snowballing process until a sufficient number of members of the rare population has been found for the survey. Survey interviews can be conducted with the members of the rare population as they are identified, thus avoiding the recontacts that are needed with the frame construction approach. With this approach, those with many contacts with other members of the rare population are more likely to be included in the survey than those with few contacts (unless, as sometimes applies, the survey aims to take all the members of the rare population). However, since the sample is not a probability sample, objective weighting adjustments cannot be employed in the analysis to compensate for this factor. Steps may be taken to make the sample conform to known or hypothesized distributions for certain background variables, as in quota sampling, but this cannot ensure that the sample produces unbiased estimates for other variables. Moreover, distributions of important background variables are seldom known for a rare population; the use of hypothesized distributions in place of known distributions introduces its own potential biases. Given the likelihood of substantial bias with this use of snowball sampling, the results from a snowball sample need to be assessed with considerable caution. The technique seems more suited for exploratory and qualitative investigations, as with case finding in the initial stages of an epidemiological study, rather than for statistical surveys.

Biernacki and Waldorf (1981) review problems and techniques of snowball sampling, and Welch (1975) and Snow *et al.* (1981) describe applications of its use. A theoretical paper entitled "Snowball sampling" by Goodman (1961) is often improperly cited in discussions of this topic; that paper does not deal with the type of snowball sampling discussed here.

7. SEQUENTIAL SAMPLING AND OTHER TECHNIQUES

For surveys estimating characteristics of a rare population, a reasonably accurate estimate of the prevalence of the rare population is required so that the sampling fraction needed to yield the desired sample size can be determined. In practice, however, often no good estimate of prevalence is available. This difficulty can be handled by some form of sequential sampling. One approach is to select an initial sample of size sufficient to give the desired sample size of members of the rare population (n) based on the highest estimate of prevalence. This sample will yield, say, n' members of the rare population, and also an estimate of prevalence. If $n' < n$, a second sample is selected to produce the remaining $(n - n')$ members of the rare population based on the prevalence figure obtained from the initial sample. Another approach is to construct a large sample as a set of replicate samples, with the size of the large sample being sufficient to generate the desired number of members of the rare population based on the lowest estimate of prevalence. The replicates are then assigned for fieldwork in turn until the desired sample size is attained. These sequential approaches cause some inefficiencies in fieldwork for face-to-face interview surveys, but they can be fairly easily implemented with telephone surveys.

In addition to the techniques discussed here, three others should be mentioned (Kish, 1965a). One is the use of multipurpose surveys, where several investigators pool resources to conduct a single survey to study several rare traits simultaneously. Market researchers frequently use multipurpose surveys to study reactions of users to infrequently purchased products. The second technique is the cumulation of cases with the rare trait over several rounds of a continuous survey. The continuous survey may be used either simply to screen for members of the rare population or to collect the full details required for the survey of the rare population. Another possible way of cumulating cases is by secondary analysis of a set of surveys that have collected the necessary information to identify members of the rare

population and that provide the required measures for them. However, problems of inconsistent definitions, classification errors and different populations sampled can present severe difficulties in pooling cases from diverse surveys (Reed, 1975). The third additional technique for finding rare elements has a different purpose from the others. The method is batch testing, which is applicable when the rare trait is detected by means of material that is expensive to test. Samples of the material from several units may then be pooled and tested together. Thus, for instance, samples of drinking water from several households may be pooled and tested for contaminants.

8. CONCLUDING REMARKS

When a sample of a rare population is required, the first consideration is whether there is a special list (or a combination of lists) that gives complete, or almost complete, coverage of the population and that does not contain many foreign elements. If an adequate list is found, the sampling problems reduce to the usual ones encountered in surveying any population. If no such list is available, the techniques described above may often be used effectively to select a sample of a rare population. However, when the population is extremely rare or when the identification of members of the rare population is expensive, there may be no satisfactory solution to the sampling problem.

For convenience of exposition, the techniques discussed above have been treated separately, although in practice they are commonly used in combination. Thus, for instance, disproportionate sampling is frequently used with screening and can be used with multiplicity sampling; multiplicity sampling involves screening, so that the issues discussed under screening apply with multiplicity sampling also.

ACKNOWLEDGEMENTS

We wish to thank Joseph Waksberg for valuable comments on an earlier version of this paper, and also a referee who made a number of helpful suggestions.

REFERENCES

Anderson, D. W., Schoenberg, B. S. and Haerer, A. F. (1982). Racial differentials in the prevalence of major neurological disorders: background and methods of the Copiah County Study. *Neuroepidemiology*, **1**, 17-30.

Bergsten, J. W. and Pierson, S. A. (1982). Telephone screening for rare characteristics using multiplicity counting rules. *Proc. of the Section on Survey Research Methods, Amer. Statist. Ass.*, 145-150.

Biernacki, P. and Waldorf, D. (1981). Snowball sampling. Problems and techniques of chain referral sampling. *Sociol. Methods and Res.*, **10**, 141-163.

Birnbaum, Z. W. and Sirken, M. G. (1965). *Design of Sample Surveys to Estimate the Prevalence of Rare Diseases: Three Unbiased Estimates.* National Center for Health Statistics, Series 2, No. 11, U.S. Washington, D.C.: Government Printing Office.

Blair, J. and Czaja, R. (1982). Locating a special population using random digit dialing. *Publ. Opin. Quart.*, **46**, 585-590.

Brown, C. and Ritchie, J. (1981). *Focussed Enumeration. The Development of a Method for Sampling Ethnic Minority Groups.* London: Policy Studies Institute and Social and Community Planning Research.

Bryan, F. A., Lessler, J. T., Weeks, M. F. and Woodbury, N. N. (1984). Pilot study for a national survey of epilepsy. In *Health Survey Research Methods, 1982* (C. F. Cannell and R. M. Groves, eds), pp. 329-334. Publication No. (PHS) 84-3346. Washington, D.C.: U.S. Department of Health and Human Services.

Cartwright, A. (1964). *Human Relations and Hospital Care.* London: Routledge and Kegan Paul.

Casady, R. J., Snowden, C. B. and Sirken, M. G. (1981). A study of dual frame estimators for the National Health Interview Survey. *Proc. of the Section on Survey Research Methods, Amer. Statist. Ass.*, 444-447.

Cochran, R. S. (1964). Multiple frame sample surveys. *Proc. of the Social Statistics Section, Amer. Statist. Ass.*, 16-19.

Cochran, W. G. (1977). *Sampling Techniques.* 3rd ed. New York: Wiley.

Czaja, R., Warnecke, R. B., Eastman, E., Royston, P., Sirken, M. and Tutuer, D. (1984). Locating patients with rare diseases using network sampling: frequency and quality of reporting. In *Health Survey Research Methods, 1982* (C. F. Cannell and R. M. Groves, eds), pp. 311-324. Publication No. (PHS) 84-3346. Washington, D.C.: U.S. Department of Health and Human Services.

Deming, W. E. (1977). An essay on screening, or on two-phase sampling, applied to surveys of a community. *Int. Statist. Rev.*, **45**, 29-37.

Elandt-Johnson, R. C. (1975). Definitions of rates: some remarks on their use and misuse. *Amer. J. Epidemiol.*, **102**, 267-271.

Ericksen, E. P. (1976). Sampling a rare population: a case study. *J. Amer. Statist. Ass.*, **71**, 816-822.

Freeman, H. E., Kiecolt, K. J., Nicholls, W. L. and Shanks, J. M. (1982). Telephone sampling bias in surveying disability. *Publ. Opin. Quart.*, **46**, 392-407.

Fry, J., Dillane, J. B., McNab Jones, R. F., Kalton, G. and Andrew, E. (1969). The outcome of acute otitis media (a report to the Medical Research Council). *Brit. J. Prev. Soc. Med.*, **23**, 205-209.

Fuller, W. A. and Burmeister, L. F. (1972). Estimators for samples selected from two overlapping frames. *Proc. of the Social Statistics Section, Amer. Statist. Ass.*, 245-249.

Goodman, L. A. (1961). Snowball sampling. *Ann. Math. Statist.*, **32**, 148-170.

Groves, R. M. and Lepkowski, J. M. (1982). Alternative dual frame mixed mode survey designs. *Proc. of the Section on Survey Research Methods, Amer. Statist. Ass.*, 154-159.

Hansen, M. H., Hurwitz, W. N. and Madow, W. G. (1953). *Sample Survey Methods and Theory, Vol. I.* New York: Wiley.

Harris, A. (1971). *Handicapped and Impaired in Great Britain.* London: H.M.S.O.

Hartley, H. O. (1962). Multiple frame surveys. *Proc. of the Social Statistics Section, Amer. Statist. Ass.*, 203-206.

——(1974). Multiple frame methodology and selected applications. *Sankhya C*, **36**, 99-118.

Hedges, B. M. (1979). Sampling minority populations. In *Social and Educational Research in Action* (M. J. Wilson, ed.), pp. 244-261. London: Longman.

Hunt, A. (1978). *The Elderly at Home.* London: H.M.S.O.

Kish, L. (1965a). *Survey Sampling.* New York: Wiley.

——(1965b). Selection techniques for rare traits. In *Genetics and the Epidemiology of Chronic Diseases.* Public Health Service Publication No. 1163. Washington, D.C.: U.S. Department of Health, Education, and Welfare.

Lepkowski, J. M. and Groves, R. M. (1984). The impact of bias on dual-frame survey designs. *Proc. of the Section on Survey Research Methods, Amer. Statist. Ass.*, 265-270.

Lessler, J. (1981). Multiplicity estimators with multiple counting rules for multistage sample surveys. *Proc. of the Social Statistics Section, Amer. Statist. Ass.*, 12-16.

Levy, P. S. (1977). Optimum allocation in stratified random network sampling for estimating the prevalence of attributes in rare populations. *J. Amer. Statist. Ass.*, **72**, 758-763.

Lund, R. E. (1968). Estimators in multiple frame surveys. *Proc. of the Social Statistics Section, Amer. Statist. Ass.*, 282-288.

Nathan, G. (1976). An empirical study of response and sampling errors for multiplicity estimates with different counting rules. *J. Amer. Statist. Ass.*, **71**, 808-815.

Reed, J. S. (1975). Needles in haystacks: studying "rare" populations by secondary analysis of national sample surveys. *Pub. Opin. Quart.*, **39**, 514-522.

Rothbart, G. S., Fine, M. and Sudman, S. (1982). On finding and interviewing the needles in the haystack: the use of multiplicity sampling. *Pub. Opin. Quart.*, **46**, 408-421.

Schein, J. D. (1968). *The Deaf Community: Studies in the Social Psychology of Deafness.* Washington, D.C.: Gallaudet College Press.

Sirken, M. G. (1970). Household surveys with multiplicity. *J. Amer. Statist. Ass.*, **65**, 257-266.

——(1972a). Stratified sample surveys with multiplicity. *J. Amer. Statist. Ass.*, **67**, 224-227.

——(1972b). Variance components of multiplicity estimators. *Biometrics*, **28**, 869-873.

——(1984). Discussion: survey methods for rare populations. In *Health Survey Research Methods, 1982.* (C. F. Cannell and R. M. Groves, eds), pp. 347-349. Publication No. (PHS) 84-3346. Washington, D.C.: U.S. Department of Health and Human Services.

Sirken, M. G. and Levy, P. S. (1974). Multiplicity estimation of proportions based on ratios of random variables. *J. Amer. Statist. Ass.*, **69**, 68-73.

Sirken, M. G., Graubard, B. I. and McDaniel, M. J. (1978). National network surveys of diabetes. *Proc. of the Section on Survey Research Methods, Amer. Statist. Ass.*, 631-635.

Sirken, M. G. and Royston, P. N. (1970). Reasons deaths are missed in household surveys of population change. *Proc. of the Social Statistics Section, Amer. Statist. Ass.*, 361-364.

Sirken, M. G., Royston, P., Warnecke, R., Eastman, E., Czaja, R. and Monsees, D. (1980). Pilot of the national cost of cancer care survey. *Proc. of the Section on Survey Research Methods, Amer. Statist. Ass.*, 579-584.

Snow, R. E., Hutcheson, J. D. and Prather, J. E. (1981). Using reputational sampling to identify residential clusters of minorities dispersed in a large urban region: Hispanics in Atlanta, Georgia. *Proc. of the Section on Survey Research Methods, Amer. Statist. Ass.*, 101-106.

Sudman, S. (1972). On sampling very rare human populations. *J. Amer. Statist. Ass.*, **67**, 335-339.

——(1985). Efficient screening methods for the sampling of geographically clustered special populations. *J. Marketing Res.*, **22**, 20-29.

Thornberry, O. T. and Massey, J. T. (1983). Coverage and response in random digit dialed national surveys. *Proc. of the Section on Survey Research Methods, Amer. Statist. Ass.*, 654-659.

Waksberg, J. (1973). The effect of stratification with differential sampling rates on attributes of subsets of the population. *Proc. of the Social Statistics Section, Amer. Statist. Ass.*, 429-434.
——(1978). Sampling methods for random digit dialing. *J. Amer. Statist. Ass.*, **73**, 40-46.
——(1983). A note on "Locating a special population using random digit dialing". *Publ. Opin. Quart.*, **47**, 576-579.
Welch, S. (1975). Sampling by referral in a dispersed population. *Pub. Opin. Quart.*, **39**, 237-245.

APPENDIX

This appendix presents a derivation of the formulae given in Section 3.2.

First, consider the case of the sample estimator $\bar{y} = A\bar{y}_1 + (1 - A)\bar{y}_2$ where A, the proportion of the rare population in stratum 1, is known. Suppose that simple random samples of sizes n_1 and n_2 are selected from the two strata, and that of these m_1 and m_2 are members of the rare population. Provided $E(m_1)$ and $E(m_2)$ are sufficiently large, the variance of $\bar{y}$ is approximately, ignoring the fpc terms,

$$V(\bar{y}) \cong \frac{A^2 S_1^2}{E(m_1)} + \frac{(1 - A)^2 S_2^2}{E(m_2)}, \tag{3}$$

where S_1^2 and S_2^2 are the element variances of the y variable for members of the rare population in the two strata. Let N be the size of the total population, W_1 be the proportion of the total population in stratum 1, P and P_1 be the prevalences of the rare population in the total population and in stratum 1 respectively, and kf_2 and f_2 be the sampling fractions in strata 1 and 2 respectively. Then $A = W_1 P_1 / P$, $E(m_1) = kf_2 P_1 W_1 N$ and $E(m_2) = f_2(P - W_1 P_1)N$. Letting $S_1^2 = \Delta S_2^2$, $V(\bar{y})$ may then be expressed as

$$V(\bar{y}) = S_2^2[kP - (k - \Delta)W_1 P_1]/kf_2 P^2 N.$$

In order to compare proportionate and disproportionate stratification for the estimation of $\bar{Y}$, the economics of data collection need to be considered. A simple cost model to represent the expected cost of a sample of size n is

$$E(C) = c_0 + c_1^* E(m_1) + c_2^* E(m_2) + c_1' E(n_1 - m_1) + c_2' E(n_2 - m_2),$$

where c_0 is an overhead cost, c_1^* and c_2^* are the costs of data collection for each selected member of the rare population in the two strata, and c_1' and c_2' are the costs of including each selected member of the nonrare population in the sample in the two strata. We assume for simplicity that $c_1^* = c_2^* = c^*$ and that $c_1' = c_2' = c'$. Letting $c = c^*/c'$, the expected cost reduces to

$$E(C) = c_0 + c'[(c - 1)E(m_1 + m_2) + n_1 + n_2] = c_0 + c'D,$$

where D denotes the term in brackets. We now compare disproportionate stratified designs with a proportionate stratified design of the same expected cost, or equivalently the same value of D. With $n_1 = kf_2 W_1 N$ and $n_2 = f_2(1 - W_1)N$, the quantity D may be expressed as

$$D = f_2 N[(c - 1)\{P + (k - 1)W_1 P_1\} + (k - 1)W_1 + 1].$$

The constant D determines the value of f_2, which can then be substituted in $V(\bar{y})$ to give

$$V(\bar{y}) \cong (S_2^2/DkP^2)[kP - (k - \Delta)W_1 P_1][(c - 1)\{P + (k - 1)W_1 P_1\} + (k - 1)W_1 + 1]. \tag{4}$$

This general formula includes proportionate stratification as the special case with $k = 1$.

The ratio of the variance of $\bar{y}$ under a disproportionate allocation to that under a proportionate allocation is then

$$R \cong \frac{[kP - (k - \Delta)W_1 P_1][(c - 1)\{P + (k - 1)W_1 P_1\} + (k - 1)W_1 + 1]}{k[P - (1 - \Delta)W_1 P_1][(c - 1)P + 1]}. \tag{5}$$

Formula (1) in Section 3.2 is the special case of R under the assumption that $S_1^2 = S_2^2$, i.e. $\Delta = 1$.

The optimum choice of k to minimize $V(\bar{y})$ can be obtained by solving $\partial V(\bar{y})/\partial k = 0$ from (4). The solution is

$$k_0^2 = \frac{\Delta P_1[(c-1)(P - W_1P_1) + (1 - W_1)]}{(P - W_1P_1)[(c-1)P_1 + 1]}. \tag{6}$$

Formula (2) in Section 3.2 is the special case with $\Delta = 1$.

Consider now the case where A is unknown and is estimated by $a = m_1f_2/(m_1f_2 + m_2f_1)$. The variance of the estimator $\bar{y}' = a\bar{y}_1 + (1 - a)\bar{y}_2$ is approximately (from Kish, 1965a, p. 137)

$$V(\bar{y}') \cong \frac{A^2[S_1^2 + (1 - P_1)(\bar{Y}_1 - \bar{Y})^2]}{\mathrm{E}(m_1)} + \frac{(1 - A)^2[S_2^2 + (1 - P_2)(\bar{Y}_2 - \bar{Y})^2]}{E(m_2)}.$$

This formula is of the same form as $V(\bar{y})$ in (3), for which A is assumed known, but it has S_h^2 replaced by $S_h^2 + (1 - P_h)(\bar{Y}_h - \bar{Y})^2$. If the assumption that $\bar{Y}_1 = \bar{Y}_2$ is made, the two formulae are equivalent. Without making this assumption, formulae (5) and (6) still hold if Δ is redefined to be

$$\Delta = [S_1^2 + (1 - P_1)(\bar{Y}_1 - \bar{Y})^2]/[S_2^2 + (1 - P_2)(\bar{Y}_2 - \bar{Y})^2].$$

Under the assumptions that $S_1^2 = S_2^2$ and that $(\bar{Y}_h - \bar{Y})^2/S_h^2$ is small, $\Delta \cong 1$, and formulae (1) and (2) in Section 3.2 will hold approximately for this case also.