A TAXONOMY OF ELUSIVE POPULATIONS

Leslie Kish
The University of Michigan

KEY WORDS: rare items, mobile populations, multiple events, multiple frames, changing units

Introduction

Since agreeing in December '87 to talk on "elusive populations" I queried dozens of samplers about this term and they seemed to be no more confident than I about the need and place for this new term in sampling. Do we need this new term to cover the many selection problems, over thirty altogether, that I organized below into ten classes? These problems have already been presented in separate articles in the sampling literature- three of them today - and I also have written about some of them. See the papers on "Estimating the size of elusive populations" in the Proceedings of the Section on Survey Research Methods [1986]. What is common to all of these problems to deserve a new name in order to embrace them with a collective name?

All those several problems have in common the failure of compact sampling frames to cover adequately the elusive populations they describe. Those failures are not only occasional and accidental (as occurs often in practice), but massive and inherent in the elusive nature of the populations. However, adequate population coverage by sampling frames is assumed as basic to standard, classic techniques of probability selections. Thus the operations of probability sampling, i.e., assigning known probabilities of selection to all of the N elements in the frame, are beset with widespread difficulties. The nature of these difficulties appears to vary a great deal between the many problems described below, but they exist for all of them.

We may still ask whether the term "frame problems," which has been used for a narrower set of problems, should be stretched to also cover the broader, more diverse problems presented below. (See, for example, Kish 1965, 2.7, 11.1-11.6 and Wright and Tsao, 1983.) And if not, is "elusive populations" the collective name we need for the problems described below?

I counted over thirty kinds of problems, and I sorted them for convenience into the ten classes listed below by finding meaningful similarities for the several problems within each of the classes. This classification is offered in the hope that such organization of the many problems will have heuristic value. Some of you have read or thought about most of these problems; nevertheless some of you may find some of them novel and interesting. This must be a rapid overview, from a bird's eye--or from a satellite, to be modern. Please consider this as only a first attempt, and feel free to construct a more consistent, a more heuristic classification. The much too brief list of references also needs an apology: each of the thirty problems has its own list, some of them long, and assembling them would be a task above my present ambition.

1. Rare items, small domains and small area estimation refer to related problems, which have been treated separately. "Rare" may refer to estimating a small proportion $M = M/N$ (and $M_i = 0$ or 1) of the population of N elements; or to a mean $\bar{Y}_m = Y/M$ of a variable Y based on the rare elements M; or, less often, to the mean of the rare variable based on the entire population $\bar{Y}_t = Y/N = \bar{M}\bar{Y}_m + (1-\bar{M})0$. There is no agreement on how rare is "rare," but let me use it for proportions smaller than, say, 1 in 10,000 of the population, so that $\bar{M} < 1/10,000$. Think of searching for centenarians, or for homeless people, or for incomes over $1,000,000 in the entire population, and these are nearly hopeless tasks for ordinary survey sampling.

On the other hand, the problems and tasks posed by small domains need not be hopeless, though they are often difficult. Just how difficult depends partly on how small the domains are, and to just speak or write vaguely about "subnational" populations is not specific enough and confusing. To facilitate discussions I proposed the trichotomy of major, minor and mini domains, each of which pose different practical problems and with different feasible solutions for the three sizes of subclasses, which represent in samples the three magnitudes of domains. Furthermore, for all three size classes we should also distinguish crossclasses and domains from designed classes and domains, because the two types have different statistical properties. (Kish 1987, 2.3)

Supplements for small areas, in particular, can be readily and separately designed for one or two areas; however estimation for all small areas poses problems and conflict for modest sized samples. [Kish 1987, 7.3D] About eight methods can be used for sampling rare cases and small domains: screening and multiphase selection, disproportionate stratified selection, cumulated and multipurpose samples, special lists, large clusters, multiplicity selection, batch sampling in clinical trials; these are described in the literature. [Platek et al. 1987; Kalton and Anderson 1986; Kish 1965, 11.4]

2. Noncoverage, nonresponses, missing items and imputations pose separate but related problems, as do the undercounts of censuses. The relations are close, and the separate problems are sometimes confused with each other. Methods of dual frames and multiple frames are designed chiefly to deal with them. It would be difficult to add here to the frequent and thorough treatments of this vast topic (Madow, Olkin, and Rubin 1983).

3. Mobility, mobile and nomad populations, diurnal and seasonal mobility, de facto/de jure and "daytime" populations, random timers for changing activities. These terms all refer to methods for and problems due to mobility of many populations, whereas the uses of sampling frames assume some stability. The mobility of populations of humans and of other animals exhibits great variety and great differences. The greatest mobility is of fishes, birds and insects, who live in three dimensions, many (not all) without clearly identifiable and permanent homes. Methods for sampling them are called capture/recapture or capture/tag/recapture; these are related to dual or multiple frame methods [El-Khorazaty 1977]. Many nomads and migrant farm workers migrate seasonally and so do the millions of owners of vacation homes. On the other hand, millions of suburban dwellers move daily into (and out of) offices in city centers; this is of great concern for disaster planning, among others. De facto vs. de jure residence affects statistics of births, deaths, hospitalization and university education. Furthermore, the location of a crime may be pointed to the scene, to the victim or to the criminal. Residents of mobile homes pose problems; also the "homeless" and street people, who are rare as well as mobile. Random timing devices have been used to track the instant time use of hospital personnel and other mobile people. The literature here is disparate, but comparisons and technology transfer between them could be heuristically and mutually beneficial.

4. Multiple events, waiting times, and large observational units: these terms cover an extreme variety of problems and methods. Their appearances vary so greatly that connections between them are seldom recognized, and their treatments appear as separate, disconnected novelties. The common feature comes from single individuals and units being connected to several replicate events, and these events serve as selector probabilities. Some examples are: 1) Medical visits to clinics and doctors, also visits to stores, libraries, theaters; 2) Sizes of families or households as selection factors; 3) waiting times as selection factors.

These replicated events differ in numbers between the units and they can thus lead to unrecognized selection biases, or to increased variances when recognized with weights in the analysis. However, when they are recognized the biases can usually be avoided. Furthermore, their recognition in the selection process can also avoid often the increased variances. {Kish 1987, 7.4}

5. Network sampling has denoted the selection and analysis of the network of the $N_i(N_i-1)/2$ possible relations between the N_i members of the group, when those relations themselves are the chief objects of study. For example, the relations between the N_i members of workgroups, or between N_i siblings; etc. However, network sampling has also been used as a synonym for multiplicity sampling, a needless and confusing redundancy, perhaps. Both of these refer to procedures for defining several selector events (in the sense of the 4 above) in order to increase selection probabilities for rare events. Snowball sampling is the colorful term for procedures of building up lists of rare populations by using initial sets of selected members as informants. Naive optimism about its possibilities should be tempered with caution, because in practice the probabilities of the identifications remain grossly unknown [Kalton and Anderson 1986; Sudman et al. 1988].

6. Multiple occupancy of areas, of periods, and of causes. Each of these problems has been recognized here and there, but they have not been subjected to consistent methodological treatments. In crop surveys, for example, the same plots of land may be growing two crops simultaneously, as well as two or three crops in rotation during the entire year. Multiple use of areas, addresses, or buildings may also occur in other surveys and must be discovered in the collection or the analysis and presentation of data. Also multiple uses of time periods (e.g., working while travelling or while eating) must be recognized and reported in surveys of time use. Multiple reasons and motives for behavior and attitudes are often reported, and they can thus sum to over 100 percent.

7. Changing units in panel studies: families, firms, communities. Problems of "elusive populations" come into sharp focus in panel studies. Persons can be identified over their lifetimes, regardless of chemical, physiological and psychological changes of the individual. However, families are subject to frequent changes of composition of its members, as these move or die out of the families or as they move into or are born into them. Business firms may be even more dynamic than families in membership; and they can split into several units, or establishments, or contrariwise several may coalesce into one firm (or more). Longitudinal studies of communities must overlook (or explain) the processes whereby they maintain their identity in spite of constant changes due to migration and to vital processes, as well as possible changes in their boundaries (Kish 1987, 3.1, 6.3B).

8. Samples cumulated over time from a changing, dynamic, hence "elusive," population, pose daunting problems of definitions, methods and inferences. They become even more relevant as periodic or repeated surveys become ever more widely available. Cumulations offer greater precisions from larger sample bases, especially valuable for smaller domains. These problems of changing populations are not the same as for studies of change from panels noted in number 7. These problems become important in methods for combining data, now called meta-analysis (Kish, 1987, 6.2B, 6.6).

9. Changing variables: population censuses, economic indicators, labor statistics, health indexes. Here we are

concerned not with changes of identifiable elements of the population, but with changes and response errors in the survey variables measured on the elements. Perhaps this topic of measurements is too large to belong to the set of problems defined as elusive populations, and can serve merely to define the boundaries of that set. There are interesting problems of design for periodic surveys, where the changes in variables, in populations, and in samples, may all intersect [Kish 1987, 6.3].

Furthermore, elusive populations can also result from response errors in finding the elements: e.g., illegal aliens, AIDS victims, young unemployed males, illegal practitioners. This problem was mentioned under point

10. *Trace sampling and unobtrusive observations* may be used for sampling the elusive traces, prints and remains left behind by populations that have vanished or eluded us. Examples are: bones, fossils, wastes, footprints, objects studied by anthropologists, archaeologists, paleontologists; tombstones, diaries, letters; footprints and worn floors in museums. (Webb et al. 1966).

Conclusion. After seeing the ten classes above and the examples within each we may better ponder questions of necessity for the new term: "elusive populations." In a broad sense all populations are elusive because they are constantly changing and moving. While you read that last sentence there have been changes in the number of persons alive in the USA, in the number of red blood cells in your body, and in the size of the ozone hole over Antarctica. Nevertheless the concept of frames covering (reasonably well) definite and finite populations still remains useful. The dozens of exceptions I noted, which massively strain that concept, deserve our attention. And that attention may be helped with the common name "elusive populations" and by a classification similar to the above, but perhaps with future improvements.

References

El-Khorazaty et al. [1977], Estimating the total number of events with data from multiple record systems, *International Statistical Review*, 45, 129-57.

Kalton G and Anderson DW [1986], Sampling rare populations, *JRSS*(A), 149, 65-82.

Kish L [1965], *Survey Sampling*, New York: John Wiley and Sons.

Kish L [1987], *Statistical Design for Research*, New York: John Wiley and Sons.

Madow WG, Olkin I and Rubin DB [1983], *Incomplete Data in Sample Surveys*, 3 Volumes, New York: Academic Press.

Platek R, Rao JNK, Sarndal CE, and Singh MP, ed., [1987], *Small Area Statistics*, New York: John Wiley and Sons.

Sirken MG et al. [1986] *Proceedings of the Section on SSRM, American Statistical Association*, 159-186.

Sudman S, Sirken MG, and Cowan CD [1988], "Sampling rare and elusive populations," *Science* 240, 991-995.

Webb EJ, Campbell DT, Schwartz RD and Sechrest L [1966], *Unobtrusive Measures*, Skokie, IL: Rand McNally.

Wright T and Tsao H [1983], A frame on frames, in Wright ed *Statistical Methods and the Improvement of Data Quality*, Orlando Fl: The Academic Press.