Journal of Official Statistics
Vol. 7. No. 3, 1991, pp. 339-347
© Statistics Sweden

Taxonomy of Elusive Populations

Leslie Kish[1]

Abstract: Ten types of problems, comprising over thirty subtypes, suffer in common the failure of compact sampling frames to cover adequately the "elusive populations" they comprehend. Those failures are structural to the entire population, rather than mere imperfections and exceptions, such as are treated commonly as "frame problems." The widespread failures are inherent in the elusive nature of the populations, which lack the structure of sampling frames for all N elements that are basic to finite population theory. The many specific problems are sorted into ten meaningful types, which are referenced to applications with specific solutions. This classification or taxonomy has heuristic value, I hope, over the *ad hoc* treatments these problems have often received.

Key words: rare events; mobile populations; multiple events; multiple uses; changing units.

0. Introduction

The term "elusive populations" has been used recently to refer to rare and mobile populations, also sometimes to other similar difficulties in survey sampling (Kish 1988; Sirken 1986; Sudman, Sirken, and Cowan 1988).

Such difficulties and problems, over thirty of them, are organized below into ten types (or classes). They have generally appeared before in the sampling literature separately, under their own names. For example, the literatures on mobile populations and nomads, on rare items and subclasses, are quite extensive. What is common to all these ten types, with over thirty subtypes, to justify a new collective name, "elusive populations" to embrace all of them?

All these separate problems have in common the failure of compact sampling frames to cover adequately the elusive populations they describe. That failure is common, although the nature and source of elusiveness differs for each type. A sampling frame for all N elements denoted by i ($i = 1, 2, \dots N$) is basic to all finite population sampling theory. The operations of probability sampling assume assigning known probabilities of selection to all the N elements in the population frame. The need and justification for population frames and probability sampling have been presented often. I cannot repeat here those justifications, but I reaffirm the need for them, despite the difficulties posed by the frequent occurrence of elusive populations.

With these ten types we refer not only to occasional and accidental failures, which can occur as exceptions to any rule in practice, but to widespread failures that are inherent in the elusive nature of many populations. Furthermore these failures are more structural to the whole population than the kinds that can be simply comprehended as "frame problems" (Kish 1965, sec. 2.7, 11.1–11.6; Wright and Tsao 1983). Frame problems refer to the imperfections of lists or frames of well defined populations. They have been separated into four types: missing units and their opposites, blanks in the

[1] University of Michigan, Ann Arbor, MI 48106, U.S.A.

frame, replicate listings and their opposites; and clusters of units. But elusiveness refers to the very definitions and delimitations of the populations themselves.

I counted over thirty kinds of problems and sorted them for convenience into the ten types listed below, by finding meaningful similarities for the several problems within each type. This organization of the many problems is offered with the hope that it will have heuristic value. The several problems within each type have a great deal in common, so that solution(s) to one can be used for others of the same type. Practical experience in designing samples facilitates such creative imitation of solutions from one problem to another.

Some of the problems treated here have their own literature, which may be followed separately, and it would take too long for us to follow them all here. Types 1 and 2, rare cases and mobile populations, are well known and documented. Types 3, 4, and 5, multiple events and multiple uses and network sampling, have not been generally recognized as such, but only as individual problems. Nonresponse and noncoverage (type 6) are too important and well known for more than perfunctory mention. Trace sampling (type 7) has been unrecognized. Types 8 and 9 deal with two distinct sets of problems that occur in sampling over time. The inclusion of type 10 may be somewhat questionable.

1. Rare Items, Small Domains, and Small Area Estimation

These have been treated separately and widely. They refer to related problems, but they also have practical differences. Exploring both the differences and the connections seems worthwhile.

Sample designs for rare items have been conceived as directed toward estimating any of three kinds of parameters.

a. A small proportion $\bar{M}_n = M/N$ of rare elements, where $M = \Sigma M_i$ and $M_i = 1$ for M elements and 0 for the other $(N–M)$.
b. Or, to a mean $\bar{Y}_m = Y_m/M$ of a variable Y_i with $Y_m = \Sigma Y_i$, summed and based on the M rare elements only; and $Y_i = 0$ for the other $(N–M)$. This should be generalized to other statistics based on the rare domain.
c. Or, less often, to the mean of the rare variable, but based on the entire population $\bar{Y}_n = Y_m/N = \bar{M}\bar{Y}_m + (1 - \bar{M})0$.

For example, we may study the proportion of an ethnic group; or their mean income or education; or, less often, the mean of their income in the entire population.

There is no agreement on how rare "rare items" are, but their problems and treatments should be distinguished from those of subclasses, domains, and subpopulations, three terms that are often confused. We shall use *subclasses* as subdivisions of the sample representing the *domains M* of the population. To facilitate discussions I also proposed that we distinguish a trichotomy of domains and subclasses: major, minor, and mini (Kish 1980; Kish 1987, sec. 2.3). Most national surveys, even small samples, are designed to yield adequate data also for 5, 10, or 20 major domains. These will have standard errors only 2, 3, 4, or 5 times greater than the overall means, often a tolerable margin. This is true for "designed" subclasses like provinces, as well as "crossclasses" like major age or occupation classes.

However, for minor subclasses, like counties, which may number 100, 1000, or more, most sample surveys cannot yield adequate data. For these "small domains" the sizes of sample surveys are usually inadequate. The usual resources for data have been censuses and administrative registers; and recently combined methods of small area estimates have been developed (Platek, Rao, Särndal, and Singh 1987). Furthermore, cumulation of periodic samples has also been practiced and sometimes proposed (as "rolling samples") especially for administrative areas (e.g., states and counties in the USA) (Kish 1987, sec. 6.6; Kish 1990). An excellent example of cumulation comes from weekly samples of diseases, each of which is a small domain or a rare item. However, the weekly samples of 1000 households are cumulated into yearly samples of 52,000 households (100,000+ persons) for small domains (and 1.6 millions decennially for rare items). Note also that the households cumulate different diseases and persons hence multipurpose data (National Center for Health Statistics 1958). These cumulated estimates are usually sought for *all* the designated small domains. When estimates are needed for only one or a few domains distinct samples are usually designed and targeted.

For rare items, however, we are concerned with designing separate samples for single domains that may comprise only 1/100, or 1/1000, or 1/10,000 of the whole population. Think of searching for centenarians, or for homeless people, or for millionaires, or for a rare disease or trait. Designing such a narrowly targeted sample that is both valid and practical can be a difficult or impossible task. But it is often sought and a large literature of various methods has been developed and reviewed (Kalton and Anderson 1986; Kish 1965, sec. 11.8)

1. When a good, recent list of the rare population (RP) is available it can be used if the uncovered portion is acceptably small or similar.

2. When large portions of (RP) are highly concentrated (on lists or in sampling units) then the principles of "optimal allocation" and of "dual frames" may be applied to add samples of the missing portions (RP).

3. Screening, double sampling, and multiphase sampling all refer to methods for larger and cheaper samples for obtaining concentrations of the RP, then treated as in (2).

4. Cumulated and multipurpose samples were noted above.

5. Multiplicity sampling, snowball sampling, batch sampling and others are noted in the references above.

2. MOBILITY: Mobile and Nomad Populations; Diurnal and Seasonal Mobility; De Facto and De Jure Populations; Daytime Populations; Travellers and Traffic; Wild Animals; Random Times for Activities

Humans are mobile animals, hence a survey or census of the population of a nation or of a city involves careful definitions and operations, also some artificiality. Using areal unity for surveys depends on identifying people and dwellings with those units; and area surveys of people, families, households, dwellings, etc, are being done worldwide fairly well, even if not without exceptions. All area sampling frames assume considerable spatial stability, whereas humans and other animals exhibit a great variety and great differences of mobility.

Domesticated animals can be identified with homes, barns, and other confined areas, but wild (undomesticated) animals are mobile. The greatest mobility is of fishes, birds, and insects who live in three dimensions, many without clearly identifiable and permanent homes (nests). Methods for sampling them are called *capture-tag* (or mark)-*recapture* (or catch-tag-recatch) *sampling*: two successive *independent epsem* selections from the same population, n_1 are captured first and tagged, then among the n_2 captured later, n_{12} are tagged as recaptured; $\hat{N} = n_1 n_2 / n_{12}$ estimates the population total, and this can be extended from *srs* to other selection methods. These could be applied also to sampling human populations, but they seldom are; however *dual frame* methods for mobile and rare populations also use estimators $n_1 n_2 / n_{12}$. The literature on dual and *multiple frames* is extensive (El-Khorazaty, Imrey and Koch 1977).

Many nomads and migrant farm workers migrate *seasonally* in cycles that tend to recur annually. There are similar patterns for others: people with vacation homes, owned or rented, also workers in vacation hotels; university students on campuses. Defining "*de jure*" (legal; regular) residence versus *de facto* location at the time of the survey causes operational difficulties. Hospital patients should also be assigned to their de jure residence to avoid overestimates of birth and death rates in small cities with large hospitals.

Mobile homes and houseboats can be localized to their usual, legal bases. Travelling people (salesmen, perfomers) also have home bases. But the "homeless" and street people are the most difficult, because they are rare and also mobile, and without ascertainable home bases. We have focused here on locating people and units to their unique de jure bases, so that they may be counted uniquely, once and only once, for representations in censuses and surveys.

However, sometimes we should also be interested in their actual de facto location of people and units while they move. An important problem is posed by the "daytime" populations of central cities, ports, and other workplaces, which are crowded by employees whose homes are elsewhere, but whose working locations concern city planners and officials. The activities as well as the location of mobile people (such as hospital personnel, policemen, drivers) can be monitored either with diaries, which are tiresome, or with random tracking devices, which "beep" at random times.

Surveys of traffic are often conducted and pose interesting problems, whose solutions depend greatly on specific situations. There are surveys of vehicles, of pedestrians, of airplane travellers, of international tourists, and of visitors to public facilities. A recent review of references (Kalton 1990) lists the following kinds of surveys: library use, museum visits, exit polls for voting, ambulatory medical care, international passengers by land, sea, and air, customers at shopping centers, road traffic. See also papers on game kills by Eberhardt and Murray (1960), and on highway traffic by Kish, Lovejoy, and Rackow (1961); also in (Kish, 1965). Usually these count the events, such as vehicle or personal passages, rather than the separate individuals effecting replicate passages and events. But the next section deals with the problem of counting the populations of separate individuals who are located by multiple events.

3. Multiple Events, Family Members, Waiting Times, Variable Observational Units

The common feature for these problems, which seem so distinct, is their similar origin: a population of individuals (elements) are connected to replicate events that serve as unequal selector probabilities for the elements, which are of chief concern. Thus they lead to biased estimates often, when those unequal probabilities pass unnoticed or uncorrected.

Let us begin with a simple and common problem: sampling the visits to public facilities, such as hospitals and doctors, libraries, theaters, museums, stores, or airports. Any individual visitor may make a different number ($x_i = 1, 2, 3 \ldots$) of visits in a year, or some other fixed period. Thus the total number $N = \Sigma N_i$ of *visitors* makes a total of $X = \Sigma x_i N_i$ visits during the year, and an average of $\bar{X} = X/N = \Sigma x_i N_i / \Sigma N_i = \Sigma w_i x_i$ visits per visitor during the year, where $w_i = N_i / \Sigma N_i$ is the proportion of visitors (people) who make x_i visits each.

When visits are selected with equal probability f, a visitor who makes x_i visits has $x_i f$ probabilities of selection, so that the expected proportion among the sample visits with x_i visits will be $w_i x_i / \Sigma w_i x_i = w_i x_i / \bar{X}$; whereas the proportion of such visitors (people) is $w_i = N_i / N$. The mean of these observations then is the weighted mean of these proportions times the variable x_i

$$\begin{aligned} \bar{C} &= \Sigma w_i x_i^2 / \Sigma w_i x_i = \Sigma w_i x_i^2 / \bar{X} \\ &= [\Sigma w_i (x_i - \bar{X})^2 + \Sigma w_i \bar{X}^2] / \bar{X} \\ &= \bar{X}[\sigma_x^2 / \bar{X}^2 + 1]. \end{aligned}$$

Thus $\bar{C} = \bar{X}(1 + CV_x^2)$, with the relative bias of the weighted mean $(\bar{C} - \bar{X})/\bar{X} = CV_x^2$ is the *relvariance* of the number of visits x_i per visitor. The numbers of visits (per year etc.) for visitors, customers, clients, patients to diverse facilities, can vary a great deal; hence large relvariances, hence large biases can occur readily if the bias is not detected and corrected.

Detection and correction can often (not always) be readily made. The sample can be reweighted with weights $k_i \propto 1/x_i$. Or the inverse weights can be introduced into the selection by retaining the selection with probability $p_i = 1/x_i$, that is, discarding selections with $(x_i - 1)/x_i$. Or a unique selector can be designated, e.g., only *first* visits during the year. The most feasible and efficient method depends on circumstances (Kish 1965, sec. 11.2).

Family size as selection factor is a frequent problem with similar sampling origins. For example, samples drawn from school records will give x_i chances to families with x_i children in school, and thus exaggerate by $(1 + CV_x^2)$ the number of children per family. Not only the size variable x_i but any variable y_i that is positively correlated with $R_{xy} > 0$ will also have a positive biased mean of $\bar{Y}(1 + RC_x C_y) = \bar{Y}(1 + \sigma_{xy}/\bar{X}\bar{Y})$. This is a more general rule, of which $y_i = x_i$ and $R = 1$ above was a special case.

The problems of waiting times and queues arise in different contexts but also have similar structures. A sample of prisoners will show a relative bias of the length of their sentences x_i also depending on CV_x. A survey of people waiting for buses will have a mean that differs similarly from the average minutes $\bar{X}$ between buses; but the $\bar{C}$ represents better the average time of riders spent in queues. And so on (Kish 1987, sec. 7.4).

A problem that frequently needs the attention of statistical samples concerns large units with highly variable sizes: factories, stores, hospitals, schools, etc. Suppose that a single observation is made on each unit so that an unweighted mean of epsem selection estimates the unit mean $\bar{X}$. But very often the weighted mean $\bar{C}$ would be more meaningful. The relative bias can be very large because CV^2 is large. Weighting the observations by x_i can produce the desired $\bar{C}$, but with much increased variance. However, selecting the units with probabilities proportional to sizes x_i and then self-weighting will estimate the desired $\bar{C}$ (Kish 1965, sec. 11.6; Kish 1987, sec. 7.5).

4. Multiple Use (Occupancy) of Areas and Periods

These problems occur often and they are sometimes recognized and described in survey reports dealing with specific situations. However, they have not received recognition as a general problem and a general theory may not be either feasible or needed.

We need not dwell on such common problems as two or multiple households or families at a single dwelling, or address, or telephone number. Or a single address or telephone may be shared by a household and a business or enterprise. Or two firms may share an address and they may be either of the same kind or different. The field workers must be alerted to these problems of multiple uses. Also a single respondent may harbor two or more diseases, occupations, or characteristics. Also multiple reasons and motives for behavior and attitudes are often held and sometimes reported, and these should often sum to over 100%.

An important question is whether the multiple uses are both (all) covered in a survey or only one of them. In crop surveys, the same plot of land may be growing two crops simultaneously, as well as two or more crops in rotation during the year; for example, trees (olive or fruit) and grains (wheat or grass), or a crop and grazing animals. Only one or some of them may be sought for some surveys, whereas surveys of total agricultural production need them all.

Multiple uses of time periods must be recognized and reported in surveys of time use. Some people may be working while travelling, or while eating. People may have two or more occupations during the same day, week, or year, also working students or housewives. Time use studies should also sum to more than 100%. Surveys that cover only some uses may overstate the total time for those uses.

5. Network Sampling

This term has been used recently and fairly frequently, but for two distinct problems and methods. First, it has denoted the selection and analysis of the $(C_2^{N_i}) = N_i(N_i - 1)/2$ possible, pairwise, reciprocal relations between the N_i members of the group, when those relations themselves are the chief objects of the study, rather than the N_i members. Usually this has been done for small groups, so that the numbers of relations and their inequalities are not overwhelming; for example, the relations between the N_i members of a work team, or between N_i siblings. The relations can be regarded as a special case of events generated by individuals, as in Section 3.

Second, however, this term has also been used as a synonym for "*multiplicity*" sampling, perhaps a needless and confusing redundancy. Both of these terms can refer to finding, defining and operationalizing several (N_i) selector events, in the sense of Section 4, in order to increase selection probabilities for rare events. *Snowball sampling* is the colorful term for procedures of constructing lists of rare populations (e.g., a rare ethnic or language group) by using initial sets of selected members as informants for names and addresses of unknown members. Naive optimism about its possibilities should be tempered with caution, because in practice the probabilities of the identifications remain grossly unknown, hence probability sampling cannot be approximated (Kalton and Anderson 1986; Sudman, Sirken, and Cowan 1988).

6. Noncoverage, Nonresponses, Missing Items

These sources must be mentioned in any list of types of elusiveness. It would be difficult to make meaningful additions to the hundreds of treatments devoted to various aspects of this vast topic (Kish 1965, sec. 13.3–13.6; Madow, Olkin, and Rubin 1983).

7. Trace Sampling and Unobtrusive Measures

These may perhaps be treated together. "Unobtrusive measures" (Webb, Campbell, Schwartz, and Sechrest 1966) was the term invented to describe observations based on footprints, worn floors and similar traces of past activities left by visitors and audiences in public areas. The statistical research for quantitative measures differs from a detective's search for unique individuals. But it has much in common with other kinds of research based on elusive traces, prints, and remains left behind by populations that have vanished or eluded us. Anthropology, paleontology, archeology, and history are based on bones, fossils, wastes, remains, letters, diaries, books, records, artifacts, buildings, tools, and weapons left by people and other beings. Stone and metal survive better than paper and wood, and these last better than speech; but linguistics can build theories by tracing phonemes through the ages.

Great biases may occur in survival, favoring the nobler metal, the better homes, the bigger tombs, etc. Scientific sampling would face great challenges, beyond those of the brilliant detective.

Finally, trace sampling of past events differs, I believe, from sampling and monitoring current behavior, whether obtrusively or unobtrusively. Hence I tend to favor my new term *trace sampling* for such research, over the better known "unobtrusive," which refers to the method of observation rather than of sampling.

8. Changing Units in Panel Studies: Families, Firms, and Communities

Panels denote longitudinal studies on the same elementary units of observation, and there is a growing literature concerned with them (Kasprzyk, Duncan, Kalton, and Singh 1989; Kish 1987, sec. 6.4). All living organisms are constantly changing and dying, and this vital and mortal characteristic raises problems for all panel studies that increase with the length of the study. Panels of persons suffer from the problems of deaths and missed births in the population. But at least a person can be identified over his/her lifetime regardless of physiological and psychological changes of the individual. Families, however, are subject to frequent changes of composition, as family members move, die, or give birth to new members (Survey Research Center 1984). 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 single firms.

Longitudinal studies of communities are common and useful, but they generally ignore a basic problem. They must overlook (or explain) the processes whereby they maintain their identity in spite of constant changes due to migrations and vital processes, as well as possible changes of their boundaries (Kish 1987, sec. 3.1].

9. Cumulating Periodic Samples

Cumulating samples is becoming more common with increases in numbers and kinds of periodic surveys. The changing, dynamic, hence elusive populations pose daunting problems of definitions, methods, and inference. These problems differ from those of Section 8, which dealt with the same units, whereas here we combine everchanging samples from the same population. Cumulations offer greater precision from larger sample bases than single surveys, and those are especially needed for smaller domains. "Rolling samples" is a new term for designs for cumulating samples (Kish 1990). The problems of cumulations are also salient for methods of combining data sets, now called meta-analysis (Kish 1987, sec. 6.2B, 6.6).

10. Fluid Population Definitions

To clarify matters we may begin with some examples of the kind of problems this type may encompass. Suppose we want to sample one of the following populations:

-the unemployed or the labor force

-the handicapped, or the blind, deaf, or lame
-families with low or with high income
-illiterates
-elites
-members of an ethnic or religious group.

Research on these populations has been done, and on others like them. These populations share a great vagueness in size and nature, which differs from the usual problems of limits in space and time. Most of all I wish to separate this problem from the operational problems of measurement errors. We are not dealing here with measuring the percentage of unemployment or illiteracy in the entire population, but with taking a sample from the population of unemployed or illiterate, etc. Second, we focus here on the drastic effects that different definitions can have, aside from the operational problems of measurement. There are many variables which may present measurement errors, without serious vagueness of definitions (age, some diseases, and many behaviors etc). For this reason I would exclude the difficult problems of finding elusive cases (rapes, incest, AIDS, murderers), because they don't fit the definition of elusive populations I propose.

It is difficult to give common advice for the different problems of this type. Nevertheless, where various valid definitions can lead to very different populations, it may be wise to use two or more, either for the entire sample or for subsamples, the broader including the narrower. These several bases can facilitate comparisons with other studies, past or future, or in other countries. Furthermore they permit studies of the sensitivity of the definitions.

11. References

Eberhardt, L. and Murray, R.M. (1960). Estimating the Kill of Game by Licenced Hunters. American Statistical Association, Proceedings of the Section on Social Statistics, 182–188.

El-Khorazaty, M.N., Imrey, P.B., and Koch, G.G. (1977). Estimating the Total Number of Events with Data from Multiple Record Systems. International Statistical Review, 45, 129–157.

Kalton, G. and Anderson, D.W. (1986). Sampling Rare Populations. Journal of the Royal Statistical Society, ser. A, 149, 65–82.

Kalton, G. (1990). Sampling Flows of Mobile Human Populations. Proceedings of Symposium 90, Ottawa: Statistics Canada (to be published).

Kasprzyk, D., Duncan, G.J., Kalton, G., and Singh, M.P. (eds.) (1989). Panel Surveys. New York: John Wiley & Sons.

Kish, L. (1965). Survey Sampling. New York: John Wiley & Sons.

Kish, L. (1980). Design and Estimation for Domains. The Statistician, London, 29, 209–222.

Kish, L. (1987). Statistical Design for Research. New York: John Wiley & Sons.

Kish, L. (1988). A Taxonomy of Elusive Populations. American Statistical Association, Proceedings of the Section on Survey Research, pp. 44–46.

Kish, L. (1990). Rolling Samples and Censuses. Survey Methodology, 16, 63–94.

Kish, L., Lovejoy, W., and Rackow, P. (1961). A Multi-Stage Probability Sample of Continuous Traffic Surveys. American Statistical Association, Proceedings of the Section on Social Statistics, 227–230.

Madow, W.G., Olkin, I., and Rubin, D.B. (eds.) (1983). Incomplete Data in Sample Surveys, 3 Volumes. New York: Academic Press.

National Center for Health Statistics (1958). Statistical Designs of the Health Household Interview Survey. Public Health Series 584–A2, 15–18.

Platek, R., Rao, J.N.K., Särndal, C.E., and Singh, M.P. (eds.) (1987). Small Area Statistics. New York: John Wiley & Sons.

Sirken, M.G. (1986) Estimating the Size of Elusive Populations. American Statistical Association, Session of five papers in Proceedings of the Section on Survey Research Methods, 159–186.

Sudman, S., Sirken, M.G., and Cowan, C.D. (1988). Sampling Rare and Elusive Populations. Science, 240, 991–995.

Survey Research Center (1984). User's Guide to PSID. Ann Arbor, MI, ISR: ICPSR.

Webb, E.J., Campbell, D.T., Schwartz, R.D., and Sechrest, L. (1966). Unobtrusive Measures. Skokie, IL: Rand McNally.

Wright, T. and Tsao, H.J. (1983). A Frame on Frames. In T. Wright, ed., Statistical Methods and the Improvement of Data Quality, Orlando FL: The Academic Press.

Received November 1988
Revised February 1991