DEVELOPING A CONCEPT OF POPULATION IN SURVEY SAMPLING CONSISTENT WITH THAT IN EXPERIMENTAL STATISTICS

C. H. Proctor, North Carolina State University
Department of Statistics, North Carolina State University, Raleigh, NC 27695-8203

The concept of a population in general statistics is almost the same as the concept of a probabilistic formulation, such as a model equation with distributional assumptions or as a stochastic process, in providing a distribution for data viewed as a sample. The notion of a population in survey sampling is somewhat more specialized. We will begin by defining population within the sampling context and then work around to experiments and observational studies.

Sampling textbooks agree that the population consists of objects, oftentimes people but also plants or animals or even organizations, on which variables are measured. An earlier viewpoint saw the population as the measurements themselves -- an "aggregate of values" (Yates, p. 20) on a large collection of objects. This concept of population is matched by that of the sample which also consists of (selected) measurements.

These two notions (objects versus numbers) lived peacefully side by side for some time but with the current interest in spatial and temporal sampling there is more opportunity now for confusion between them. Sampling units defined in space or in time may not be tied to any organisms or bounded objects. The measurements here are tangible enough, even for all units, but the objects may not be. For example, the "population" of interest to the wildlife biologist may still be the deer in Michigan's Upper Peninsula but the frame is of area segments and a sample of these is scanned for the variable of interest which is the number of deer droppings. Such a sample survey has a "population" but it's not the deer.

Theoretical formulations for finite population sampling must always begin with a frame. The frame consists of an indexed list of addresses. The index numbers allow for selection based on random numbers and the addresses consist of instructions on where and when to go to get a measurement. These instructions refer to materials such as directories or maps or meter readings. Because of the crucial importance and value of these materials they often are used in naming the frame, as a "list frame" or a "map frame," but in fact the addresses with their index numbers define the frame.

A sample survey can next be defined as a series of operations that start with the use of random numbers to select the addresses at which to measure the variable of interest, and end with the results of these measurements. The sample survey having simple random sample size equal to frame size is called a complete coverage survey. Quantities such as the mean, computed from the complete coverage survey, constitute population parameters.

Actual surveys, however, are carried out by fallible humans with fallible instruments. Sometimes the address is ambiguous, it is not followed correctly, the measuring instrument failed to work or erred, and so forth. Thus there has arisen the need to separate out the notion of an "equivalent" complete coverage su rvey (in Deming's (1960) words), from a complete coverage survey -- a "survey" population distinct from the "frame" population in Kish's (1987) terminology. Cochran (1977) contributed the term "sampled" population for Kish's "frame" population and both use "target" population as an ideal goal in which the list is of very high quality (really complete). I find some of the old confusions between values and units entering into these notions and thus I propose the following way of making the concept of population explicit.

A given finite survey population is defined by the distribution(s) of observations generated by drawing a simple random sample from the given frame and utilizing the given survey operations. The defining survey operations may include measurement error or not, locating problems or not, response problems or not, and so on. While it may be true that simple random sampling is seldom actually used, it is nonetheless a useful standard way to define a finite population. Under this viewpoint a single finite survey population can govern many variables' joint distributions. Although the definition generalizes the equivalent complete coverage survey to include all possible such surveys, it does not carry information on subdivisions or subpopulations. It gives essentially the marginal distributions of the variables of interest when all effects of measurement and procedural errors are included.

The purpose of such a definition is not to aid in deriving sampling distributions. In fact the theory of sampling is much better served just by definitions for frame and for the various sample designs. In particular, the very careful distinctions made by Tore Dalenius (1985), between "elements of the population" and "units of some other kind," allow all of sampling theory to proceed. Our definition's purpose is to establish continuity with the notions of population in experimental design, and in analyses of data from various kinds of observational studies.

Such data are invariably cast into a matrix format with variables as columns and cases as rows. When survey data are used for enumerative objectives (estimating a total or mean) then the cases of the dataset coincide invariably with frame addresses and one has a frame-addresses dataset. When survey data are used for analytical objectives (estimating relationships among variables) the cases must be experimental units. These may coincide with addresses but they may not. We then have, in addition to the address dataset, an experimental-units or an analytic-units dataset.

Even when there are several experimental units at some addresses or even when some experimental units extend over, or are associated with, several addresses, it can still be arranged that any sample draw makes a contribution to the experimental-units dataset. When one analytic unit belongs to several frame addresses it can be "prorated" or listed under each address with a proportional weight or it can, by some rule, be assigned only to one address. There is still a complete coverage or an equivalent complete coverage survey and data so produced can then be cast into the (complete) finite population experimental-units dataset from which one can compute measures of relationship among variables.

We use the word "can" advisedly since the finite population survey setting becomes somewhat too restrictive for studying relationships among variables. A survey is intended to furnish a description of the existing state of affairs - how things are, not how they got that way. When interest shifts to how experimental units develop or grow or exist through a time period, then the concept of finite population needs expanding.

In accord with the elaboration from the simple concept of "population" to those of "survey population," "frame population," and "target population," there has arisen the notion of "superpopulation." As with the other concepts it will be defined by distributions, but in this case of the values obtained at the addresses of the frame when all operations of locating and measurement are performed unerringly. Such results are sometimes called "true values" but are perhaps more realistically termed "preferred values."

The source of randomness here is thus neither sampling nor measurement, but is inherent in the process generating the units. For living organisms, both genetic and environmental forces are in play. Yield of corn from a 1/100*th* acre plot varies in accord with the plant material, the soil, the weather, and the grower's practices that impinge on the plot. While we observe just the one resulting (preferred value) yield on a given plot for one growing season, we visualize a range of yields that could have arisen under slightly changed circumstances, all within a minor alteration of the circumstances that actually took place.

This is the same viewpoint of statistical or probabilistic imagination that permits defining treatment effects. The same plots can be imagined to have been fertilized in different ways or treated for disease in different ways and then averages taken over hypothetically infinite numbers of trials under basically the same conditions at each treament. Differences among these means define the treatment effects.

The distributions resulting from these hypothetically infinite numbers of trials under basically the same conditions are revealed to some extent in the empirical distributions of residuals from statistical analyses. However, when one wishes to predict how well the estimated treatment differences will hold up in future years or under different conditions, then there may be no empirical evidence available and educated guess work must take over.

To return to the concept of superpopulation, it should be noted that there will be a given frame and certain information about the addresses will thus be available. There will generally be space and time coordinates and, with directories, there will be background information such as size or age. The nature of this auxiliary information will determine the form one will propose as the superpopulation model and getting some data will allow for checking and revision of this model.

Although the superpopulation concept is available and is used for both the enumerative survey as well as for the analytic survey, it plays different roles. For an enumerative survey the survey population is fixed and awaits discovery through sampling, even though it can also be recognized as just one realization of the superpopulation process. An ensemble of realizations can be visualized or simulated in order to answer questions of sample design, but su ch an ensemble would not ordinarily become the basis for judging sampling uncertainties.

For an analytic survey the superpopulation process, and the relationships among variables its cause-effect mechanisms create, become the objectives of the survey. Repeated realizations are visualized both in order to design the sample selections as well as to understand sampling uncertainties. Since the major interest is in the cause-effect mechanisms there is generally little initial attempt to estimate an overall average relationship. One is more concerned to characterize the variations in the relationship over sectors of the frame.

Roughly speaking, it is as inappropriate to apply sample design-based variance estimation when estimating an overall regression coefficient, as it is to estimate such a coefficient for just the survey population. In the same way it is inappropriate to apply model-based variance estimation when estimating a survey population mean. That is, the superpopulation process is not the proper survey objective when the objectives are enumerative. It also turns out, continuing to speak roughly, that either of these, so called, inappropriate calculations will not be very misleading. They will be very close numerically -- certainly within sampling error -- to their appropriate counterparts.

We have almost reached the concept of population in observational studies and in experiments which is essentially that of the superpopulation omitting the frame. That is, the experimental units available for use are ordinarily not all listed, but only implicitly visualized. In fact a vision of the population may not materialize until after the data are in and are being analyzed. It certainly undergoes adjustments as one examines residuals and checks assumptions. It continues under revision every time the results are applied by users.

References

Cochran, W. C. (1977). *Sampling Techniques*, Third Edition, John Wiley, NY.

Dalenius, Tore (1985). *Elements of Survey Sampling*, Swedish Agency for Research Cooperation with Developing Countries, distributed by Statistics Sweden, Stockholm.

Deming, W. E. (1960). *Sample Design in Business Research*, John Wiley, NY.

Yates, Frank (1981). *Sampling Methods for Censuses and Surveys*, Fourth Edition, MacMillan Publishing Company, NY.