CHAPTER THREE

TARGET POPULATIONS, SAMPLING FRAMES, AND COVERAGE ERROR

3.1 INTRODUCTION

elements

Sample surveys describe or make inferences to well-defined populations. This chapter presents the conceptual and practical issues in defining and identifying these populations. The fundamental units of populations are referred to as "elements." The totality of the elements forms the full population. In most household populations, elements are persons who live in households; in NAEP and other school samples, the elements are often students within the population of schools; in business surveys like CES, the element is an establishment. Elements can be many different kinds of units, even in the same survey. For example, in a household survey, in addition to persons, the inference might also be made to the housing units where the persons live, or to the blocks on which they live, or to the churches that they attend. In short, statistics describing different populations can be collected in a single survey when the populations are linked to units from which measurements are taken.

Surveys are unique among the common research tools in their concern about a well-specified population. For example, when conducting randomized biomedical experiments, the researcher often pays much more attention to the experimental stimulus and the conditions of the measurement than to the identification of the population under study. The implicit assumption in such research is that the chief purpose is identifying the conditions under which the stimulus produces the hypothesized effect. The demonstration that it does so for a variety of types of subjects is secondary. Because surveys evolved as tools to describe fixed, finite populations, survey researchers are specific and explicit about definitions of populations under study.

3.2 POPULATIONS AND FRAMES

target population

The "target population" is the group of elements for which the survey investigator wants to make inferences using the sample statistics. Target populations are finite in size (i.e., at least theoretically, they can be counted). They have some time restrictions (i.e., they exist within a specified time frame). They are observable (i.e., they can be accessed). These aspects of target populations are desirable for achieving clear understanding of the meaning of the survey statistics and for permitting replication of the survey.

household

housing unit

The target population definition has to specify the kind of units that are elements in the population and the time extents of the group. For example, the target population of many US household surveys is persons 18 years of age or older, – "adults" who reside in housing units within the United States. A "household" includes all the persons who occupy a housing unit. A "housing unit" is a house, an apartment, a mobile home, a group of rooms, or a single room that is occupied (or if vacant, is intended for occupancy) as separate living quarters. Separate living quarters are those in which the occupants live and eat separately from any other persons in the building and which have direct access from the outside of the building or through a common hall. The occupants may be a single family, one person living alone, two or more families living together, or any other group of related or unrelated persons who share living arrangements. Not all persons in the United States at any moment are adults; not all adults reside in housing units (some live in prisons, long-term care medical facilities, or military barracks).

Not all US national household surveys choose this target population. Some limit the target population to those living in the 48 coterminous states and the District of Columbia (excluding Alaska and Hawaii). Others add to the target population members of the military living on military bases. Still others limit the target population to citizens or English speakers.

Since the population changes over time, the time of the survey also defines the target population. Since many household surveys are conducted over a period of several days, weeks, or even months, and since the population is changing daily as persons move in and out of the US households, the target population of many household surveys is the set of persons in the household population during the survey period. In practice, the members of households are "fixed" at the time of first contact in many surveys.

survey population

There are often restrictions placed on a survey data collection operation that limit the target population further. For example, in some countries it may not be possible to collect data in a district or region due to civil disturbances. These districts or regions may be small in size, and dropped from the population before sample selection begins. The restricted population, sometimes called a "survey population" is not the intended target population, and yet it is realistically the actual population from which the survey data are collected. For example, the CES target population consists of all work organizations with employees in a specific month. Its survey population, however, consists of employers who have been in business for several months (long enough to get on the frame). A survey organization may note the restriction in technical documentation, but users of available public use data may not make a clear distinction between the target population (e.g., persons living in the country) and the survey population (e.g., persons living in the country, except for districts or regions with civil disturbances).

sampling frame

A set of materials, or "sampling frame," is used to identify the elements of the target population. Sampling frames are lists or procedures intended to identify all elements of a target population. The frames may be maps of areas in which elements can be found, time periods during which target events would occur, or records in filing cabinets, among others. Sampling frames, at their simplest, consist of a simple list of population elements. There are populations for which lists are readily available, such as members of a professional organization, business establishments located in a particular city or county, or hospitals, schools, and

other kinds of
number of cour
There are ı
are not readily ɛ
able in one plac
in prisons, or e
bers in a single
across institutic
have to be crea
units are often ı
geographic area
tion, and staff a
ating a list of h
a large area.
When ava
entirely, the sur

1) Redef
2) Admit
nal tar

A common ex
household surv
Although the c
holds, the attra
alter the target
the researcher
approximately
phones. Using
tion is not the
full household
that there is co
ferences for su
still be an imp
A more dr
vey of student
children in the
in public scho
wealthier fami
those of public
fit the frame (a
would be subj
population – ir
est to US polic
ing the survey
error for privat
focuses on issu
focuses on sta

other kinds of institutions. There are registries of addresses or of persons in a number of countries that also serve as sampling frames of persons.

There are many populations, though, for which lists of individual elements are not readily available. For example, in the United States lists are seldom available in one place for all students attending school in a province or state, inmates in prisons, or even adults living in a specific county. There may be lists of members in a single institution or cluster of elements, but the lists are seldom collected across institutions or combined into a single master list. In other cases, lists may have to be created during survey data collection. For example, lists of housing units are often unavailable for household surveys. In area sampling, well-defined geographic areas, such as city blocks, are chosen in one or more stages of selection, and staff are sent to selected blocks to list all housing units. The cost of creating a list of housing is thus limited to a sample of geographic areas, rather than a large area.

When available sampling frames miss the target population partially or entirely, the survey researcher faces two options:

1) Redefine the target population to fit the frame better.
2) Admit the possibility of coverage error in statistics describing the original target population.

A common example of redefining the target population is found in telephone household surveys, where the sample is based on a frame of telephone numbers. Although the desired target population might be all adults living in US households, the attraction of using the telephone frame may persuade the researcher to alter the target population to adults living in telephone households. Alternatively, the researcher can keep the full household target population and document that approximately 6% of the US households are missed because they have no telephones. Using a new target population is subject to the criticism that the population is not the one of interest to the user of the survey statistics. Maintaining the full household target population means that the survey is subject to the criticism that there is coverage error in its statistics. Clearly, these are mostly labeling differences for survey weaknesses that are equivalent – the telephone survey will still be an imperfect tool to study the full adult household population.

A more dramatic example of the options above could affect NAEP, the survey of students in US schools. Imagine that the target population was all school children in the United States, but the sampling frame consisted only of children in public schools. Because children in private schools on average come from wealthier families, their mathematical and verbal assessment scores often exceed those of public school children. The choice of redefining the target population to fit the frame (and reporting the survey as describing public school children only) would be subject to the criticism that the survey fails to measure the full student population – in essence, the survey is not fully relevant to the population of interest to US policy makers. Using the target population of all students (and reporting the survey as describing all students), but noting that there may be coverage error for private school students, leads to coverage errors. In short, the first option focuses on issues of relevance to different users of the survey; the second option focuses on statistical weaknesses of the survey operations.

3.3 Coverage Properties of Sampling Frames

Although target and survey populations can be distinguished, a central statistical concern for the survey researcher is how well the sampling frame (the available materials for sample selection) actually covers the target population. In Figure 2.6 in Chapter 2, the match of sampling frame to target population created three potential outcomes: coverage, undercoverage, and ineligible units.

coverage

undercoverage

ineligible units

When a target population element is in the sampling frame, it is labeled as "covered" by the frame. There can be elements in the target population that do not, or cannot, appear in the sampling frame. This is called "undercoverage," and such eligible members of the population cannot appear in any sample drawn for the survey. A third alternative, "ineligible units," occurs when there are units on the frame that are not in the target population (e.g., business numbers in a frame of telephone numbers when studying the telephone household target population).

A sampling frame is perfect when there is a one-to-one mapping of frame elements to target population elements. In practice, perfect frames do not exist; there are always problems that disrupt the desired one-to-one mapping.

duplication

clustering

It is common to examine a frame to measure the extent to which each of four problems arises. Two of these have already been discussed briefly above: undercoverage and ineligible or foreign units. The other two concern cases in which a unit is present in the frame, and it maps to an element in the target population, but the mapping is not unique, not one to one. "Duplication" is the term used when several frame units are mapped onto the single element in the target population. In sample surveys using the frame, the duplicated elements may be overrepresented. "Clustering" is the term used when multiple elements of the target population are linked to the same single frame element. In sample surveys using the frame, the sample size (in counts of elements) may be smaller or larger depending on the clusters selected. There are also cases in which multiple frame units map to multiple target population elements, many-to-many mappings. We consider this more complicated problem only briefly in this section, viewing the problem as a generalization of a combination of duplication and clustering.

3.3.1 Undercoverage

area frame

area probability sample

Undercoverage is the weakness of sampling frames prompting the greatest fears of coverage error. It threatens to produce errors of omission in survey statistics from failure to include parts of the target population in any survey using the frame. For example, in telephone household surveys where the target population is defined as persons in all households, undercoverage occurs because no telephone sampling frame includes persons in households without telephones. This is true for the BRFSS and SOC. In many countries of the world, because telephone subscription requires ongoing costs, poor persons are disproportionately not covered. In countries in which mobile telephones are replacing fixed-line service, younger persons are likely to be uncovered by frames limited to line telephone numbers because they are adopting the new technology more quickly. As we will discuss in Section 3.5, the impact on survey statistics (whether based on censuses or surveys) of noncoverage depends on how those on the frame differ from those not on the frame.

The
the proc
frame. Tl
of the su
the surv
outside s
surveys,
list of ar
tion area
of housi
meration
within tl
"area fi
samples.
three lev

In a
area inc
which su
housing
them. St
block or
housing
more di
such as
or other
readily i
unit is ir
not be
Bounda
lines of
top of a
tation, a
conditio
in the ar
pretatio
rated fr
because
will be

Ho
task in
defined
dwellin
other u
meals n
or detac
fied. H
tion ma
barriers
locked
gether.

The causes of coverage problems depend on the processes used to construct the sampling frame. Those processes may be under the control of the survey design, or they may be external to the survey (when a frame is obtained from an outside source). For example, in some household surveys, the survey sample is initially based on a list of areas, such as counties, blocks, enumeration areas, or other geographic units; then on lists of housing units within selected blocks or enumeration areas; and finally, on lists of persons within the households. These samples are called "area frame samples" or "area probability samples." Coverage problems can arise at all three levels.

In area probability designs, each selected area incurs a second frame development, in which survey staffs develop sampling frames of housing units, usually using addresses to identify them. Staffs are sent to sample areas, such as a block or group of blocks, and instructed to list all housing units in them. The task is considerably more difficult than it may appear. A boundary such as a street or road, railroad tracks, or river or other body of water are relatively fixed and readily identified. Whether a particular housing unit is in or out of the area, and should or should not be listed, is relatively easily determined. Boundaries based on "imaginary lines" based on lines of sight between natural features such as the top of a mountain or ridge, are open to interpretation, and more difficult to identify under field conditions. Whether a particular housing unit is in the area or not also becomes a matter of interpretation. Housing units that are widely separated from others may be left out of the listing because of boundary interpretation errors. These will be part of the noncovered population.

Housing unit identification is not a simple task in all cases. A housing unit is typically defined to be a physical structure intended as a dwelling that has its own entrance separate from other units in the structure and an area where meals may be prepared and served. Single family or detached housing units may be readily identified. However, additional units at a given location may not be easily seen when walls or other barriers are present. Gated communities or locked buildings may prevent inspection altogether. It is also possible to miss a unit in rural

Robinson, Ahmed, das Gupta, and Woodrow (1993) on US Decennial Census Coverage

In 1993, Robinson et al. assessed the completeness of coverage in the census of 1990 using demographic analysis.

Study design: Demographic analysis uses administrative data on population change: births, deaths, immigration, and emigration. This yields counts of persons by age, race, and gender, independent of the census, which enumerates people through mail questionnaires and enumerator visits. Robinson et al. compared these independent population estimates with the census population counts and calculated the net census undercount as well as the net undercount rate.

Findings: The overall estimated net undercount in the 1990 Census was 4.68 million, or 1.85%. The estimate for males exceeded the one for females (2.79% vs. 0.94%) and the net undercount of Blacks was higher than that of non-Blacks (5.68% vs. 1.29%). Black men age 25–64 as well as Black males and females age 0–9 had higher than average net undercount.

Limitations of the study: The undercount estimates assume no error in demographic analysis. Net undercount measures are not able to reflect their different components in detail, the omission of persons from the census (undercount), and the duplication of persons of the census (overcount). In addition, definitional differences in demographic categories complicate the comparison.

Impact of the study: This study provided empirical support for ethnographic studies identifying transient membership, unrelated individuals in large households, and young children of divorced parents as underreported in listings of the household members.

areas or under crowded urban conditions because it is not visible from public areas. Housing units located along alleyways or narrow lanes with an entrance not clearly visible from a public street can be easily missed during listing. Each missed unit may contribute to undercoverage of the target population.

Housing units in multiunit structures are also difficult to identify. External inspection may not reveal multiple units in a particular structure. The presence of mailboxes, utility meters (water, gas, or electricity), and multiple entrances are used as observational clues about the presence of multiple units. Hidden entrances, particularly those that cannot be seen from a public street, may be missed.

There are also living arrangements that require special rules to determine whether a unit is indeed a housing unit. For example, communal living arrangements are not uncommon in some cultures. A structure may have a single entrance, a single large communal cooking area, and separate sleeping rooms for families related by birth, adoption, or marriage. Procedures must be established for such group quarters, including whether the unit is to be considered a household, and whether the single structure or each sleeping room is to be listed.

Institutions must also be identified, and listing rules established. Some institutions are easily identified, such as prisons or hospitals. Caretaker housing units on the institutional property must be identified, however, even if the institution itself is to be excluded. Other institutions may not be as easily identified. For example, prison systems may have transition housing in which prisoners still under the custody of the system live in a housing unit with other inmates. Procedures must be established for whether such units are to be listed as housing units, or group quarters, or excluded because of institutional affiliation. Similarly, hospitals or other health care systems may use detached housing units for care of the disabled or those requiring extended nursing care. To the extent that housing units are left off a list because staff is uncertain about whether to include them, coverage error in survey statistics might arise.

Another common concern about undercoverage in household surveys stems from the fact that sampling frames for households generally provide identification of the housing unit (through an address or telephone number) but not identifiers for persons within the household. (Countries with population registers often use the registry as a sampling frame, skipping the household frame step.) In a census or a survey using a frame of addresses, but with a target population of persons, a small sampling frame of persons in each household must be developed. Interviewers list persons living in the household, but if the listings are not accurate reflections of who lives in the household, coverage problems arise.

de facto residence rule

The frames of persons in household surveys generally list "residents" of the household. Residency rules must be established so that an interviewer can determine, based on informant reports, whether to include persons in the household listing. Two basic residency rules are used in practice. In the *de facto* rule used in census and some survey operations, persons who slept in the housing unit the previous night are included. This rule is typically reserved for shorter-term data collection activities to avoid overcoverage of individuals who may have frequent residence changes, could appear in more than one housing unit across a short time period, and be overrepresented in the sample. It is easy to apply, because the definition is relatively clear. Undercoverage may arise for individuals traveling and staying in institutions (such as a hotel) the previous evening, even though the person usually sleeps in the household in the evening.

A more common residency rule in surveys is the *de jure* rule, based on "usual residence," who usually lives in the housing unit. This rule can be straightforward to apply for many individuals, but there are also many circumstances where the application of the rule is difficult. Usual residency for individuals whose employment requires travel, such as sales representatives, truck drivers, or airline pilots, may be unclear. If the informant says that the housing unit is their usual residence when not traveling, the rule uses the residence for the majority of some time period (such as the previous year or month). If the individual intends to use the housing unit as their residence (for those who have just moved into the unit) the *de jure* rule will include them as usual residents.

de jure residence rule

usual residence

Most US household censuses and surveys use such procedures, and their coverage properties are well documented. Younger males (18–29 years old), especially those in minority racial groups, appear to have looser ties with households. They may live with their parents some days of the week, with friends on other days. Similarly, young children in poorer households, especially those without two parents, may live with their mother some time, their grandparents sometimes, their father or other relative other times. In such housing units, when the interviewer asks the question, "Who lives here?" it appears that such persons are disproportionately omitted (see Robinson, Ahmed, das Gupta, and Woodrow, 1993; box on page 71) and are a source of undercoverage.

Sometimes, the set of persons residing in a housing unit are not approved by legal authority. For example, a rental agreement for an apartment may specify that only one family of at most five persons can occupy the unit. However, poor persons may share rental expenses among several families in violation of the agreement. They may be reluctant to report the additional family as residents of the unit. If social welfare rules limit eligibility to married couples, an unmarried woman may fail to report a male resident in the unit. This leads to systematic omissions of certain types of persons (de la Puente, 1993).

In some cultures, certain individuals are not considered to be part of the household, even though they fit the usual resident requirements of the *de jure* rule. Infants, for example, may not be considered residents, and left off the list. Indeed, the fit between the traditionally defined "household" and the population is a ripe area for research in survey methodology. So central has been the use of the household as a convenient sampling unit that most survey research employs it for person-level surveys. When people are only ambiguously related to housing units, however, the practice needs scrutiny.

In an establishment survey, the creation, merger, and death of establishments are important factors in undercoverage. The definition of an establishment, particularly with very large and very small firms, is difficult to apply in practice. Firms with many locations, such as franchised establishments, may have to be separated into multiple establishments based on geographic location. Firms with several offices or factories, warehouses, or shipping locations may also have to be listed separately. The distinction between a survey-defined establishment and a business unit in a firm may be difficult to determine.

Establishments may be in existence for very short periods of time, or may be so small that they are not included in available frames. For example, the CES misses newly established employers for a period of months. Establishment frames may be in part based on administrative registries, but these can be out of date or incomplete, particularly for newly created establishments. Mergers or subdivision of firms complicate administrative record keeping, and may lead to overcoverage as well as undercoverage. Keeping establishment frames up to date is a sizable and ongoing task.

Undercoverage is a difficult problem to identify and to solve. If population elements do not appear in the frame, additional frames might be used to try to identify them (see multiple frame surveys in Section 3.6.3). In telephone household surveys, nontelephone households may be covered through the use of an area sampling frame that in principle covers all households, regardless of telephone subscription. There are techniques for expanding the coverage of the frame through respondent reports about other population elements (see multiplicity techniques in Section 3.6.2), but in US household surveys, proxy reports about other households or persons are increasingly restricted due to human subjects' concerns about privacy.

3.3.2 Ineligible Units

Sometimes, sampling frames contain elements that are not part of the target population. For example, in telephone number frames, many of the numbers are nonworking or nonresidential numbers, complicating the use of the frame for the target population of households. In area probability surveys, sometimes the map materials contain units outside the target geographical area. When the survey staff visits sample areas to list housing units, they sometimes include unoccupied or business structures that appear to be housing units.

When interviewers develop frames of household members within a unit, they often use residence definitions that do not match the meaning of "household" held by the informant. Parents of students living away from home often think of them as members of the household, yet many survey protocols would place them at college. The informants might tend to exclude persons unrelated to them who rent a room in the housing unit. Studies show that children in shared custody between their father and mother disproportionately are omitted from household listings.

foreign element

ineligible element

Although undercoverage is a difficult problem, "ineligible" or "foreign" units in the frame can be a less difficult problem to deal with, if the problem is not extensive. When foreign units are identified on the frame before selection begins, they can be purged with little cost. More often, foreign or ineligible units cannot be identified until data collection begins. If few in number, after sampling they can be identified in a screening step and dropped from the sample, with a reduction in sample size. If the prevalence of foreign units is known, even approximately, in advance, additional units can be selected, anticipating that some will be screened out. For example, it is known that approximately 15% of entries in residential portions of national telephone directories are numbers that are no longer in service. To achieve a sample of 100 telephone households, one could select a sample of 100/(1 – 0.15) = 118 entries from the directory, expecting that 18 are going to be out-of-service numbers.

When the proportion of foreign entries is very large, the sampling frame may not be cost-effective to use. In telephone household surveys in the United States, for example, one frame contains all the known area code–prefix combinations (the first six digits of a US 10 digit phone number). Surveys based on the frame are often called "random digit dialed surveys." Of all the possible 10 digit phone numbers in the frame, more than 85% of the numbers are not in service (foreign units). It is time-consuming to screen numbers with that many foreign units.

random digit dialing (RDD)

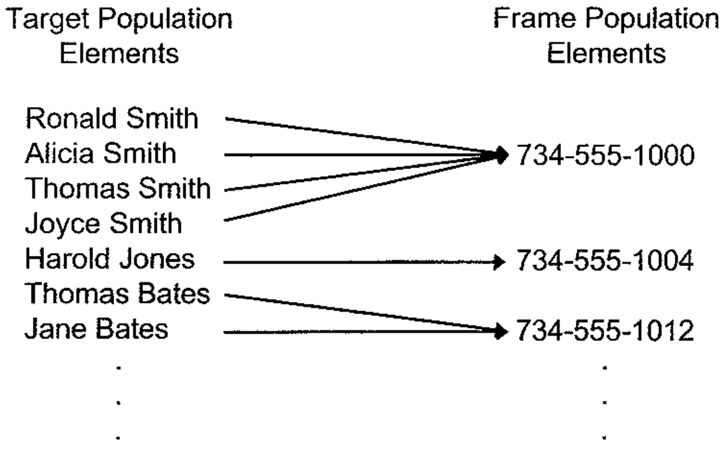

Figure 3.1 Cluster of target population elements associated with one sampling frame element.

Other sampling frames and sampling techniques have been developed that are more cost-effective for selecting telephone households (some of these are described in Section 4.8).

3.3.3 Clustering of Target Population Elements Within Frame Elements

multiple mappings

As mentioned previously, multiple mappings of frame to population (clustering) or population to frame (duplication) are problems in sample selection. A sample of adults living in telephone households (the target population) using a telephone directory (the sampling frame) illustrates each of these problems.

clustering

A telephone directory lists telephone households in order by surname, given name, and address. When sampling adults from this frame, an immediately obvious problem is the clustering of eligible persons that occurs. "Clustering" means that multiple elements of the target population are represented by the same frame element. A telephone listing in the directory may have a single or two or more adults living there.

Figure 3.1 illustrates clustering. The left side of the figure shows seven different target population elements, persons who live in telephone households. The Smith family (Ronald, Alicia, Thomas, and Joyce) lives in the same household, which has the telephone number 734-555-1000, the sampling frame element. All the Smith's are associated with only one frame element, even though together they may form four elements of the target population.

One way to react to clustering of target population elements is by simply selecting all eligible units in the selected telephone households (or all eligible units in a cluster). With this design, the probability of selection of the cluster applies to all elements in the cluster.

Clustering poses important issues that often lead to subsampling the cluster. First, in some instances it may be difficult to collect information successfully from all elements in the cluster. In telephone surveys, nonresponse sometimes increases when more than one interview is attempted in a household by telephone.

Second, when interviews must be conducted at more than one point in time, initial respondents may discuss the survey with later respondents, affecting their answers. In opinion surveys, an answer to a question may be different if the respondent is hearing the question for the first time, or has already heard the question from another person who has already completed the interview. Even answers to factual questions may be changed if respondents have talked among themselves. Third, if the clusters are different in size (as in the case of clusters of adults at the same telephone number), control of sample size may become difficult. The sample size of elements is the sum of the cluster sizes, and that is not under the direct control of the survey operation unless cluster size is known in advance.

To avoid or reduce these problems, a sample of elements may be selected from each frame unit sampled (the cluster of target population members). In the case of telephone household surveys of adults, one adult may be chosen at random from each sampled household. All efforts to obtain an interview are concentrated on one eligible person, contamination is eliminated within the household, and the sample size in persons is equal to the number of households sampled.

In the case of telephone and other household surveys in which a single eligible element is chosen, the development of the within-household frame of eligible persons, the selection of one of them, and the interview request is done in real time, at the time of data collection. Since the primary responsibility of the interviewer is data collection, and since interviewers are seldom statistically trained to draw good samples, simple procedures have been designed to allow selection during data collection of a single element rapidly, objectively, and with little or no departures from randomization. One widely used procedure (Kish, 1949) designates in advance a particular element on a newly created frame of eligible elements in the household. In order to maintain objectivity in selection, the elements are listed in a well-defined and easily checked order (say, by gender, and then by age within gender). When data collection is implemented by computer-assisted methods, the selection can be randomly made by computer at the time of listing, avoiding the need for a designated element to be selected in advance.

One difficulty that arises with these procedures is that early in the first contact with the household, the interviewer must ask for a list of eligible persons. Such a request may arouse suspicion about the intent of the interviewer, and lead to nonresponse, especially in telephone surveys. Alternative selection procedures have been devised, such as the "last birthday" method. The informant is asked to identify the eligible person in the household whose birthday occurred most recently. Given a specific time for the data collection, this assigns zero chance of selection to all but one person in the household (that person whose birthday was most recent). If, on the other hand, the survey data collection is continuous over time, probabilities of selection are equal across persons. Although the procedure does not yield a probability sample, for time-limited surveys, it has no apparent biases if correct responses were obtained. In practice, the choice of the eligible person may be influenced by subjective criteria used by the respondent. Repeated studies show a tendency for the person who answers the telephone to self-identify as that person who had the most recent birthday (suggesting response error). In the United States, there is a tendency for females to identify as having had the most recent birthday. Hence, the procedure is, in practice, typically biased.

After sample selection, there is one other issue that needs to be addressed in this form of cluster sampling – unequal probabilities of selection. If all frame elements are given equal chances, but one eligible selection is made from each, then elements in large clusters have lower overall probabilities of selection than elements in small clusters. For example, an eligible person chosen in a telephone household containing two eligibles has a chance of one half of being selected, given that the household was sampled, whereas those in a household with four eligibles have a one in four chance.

The consequence of this kind of sampling is that the sample ends up overrepresenting persons from households with fewer eligibles, at least relative to the target population. In other words, more eligibles are in the sample from smaller households than one would find in the target population. If, for some variables collected in the survey instrument, there is a relationship between cluster size and the variable, the sample results will not be unbiased estimates of corresponding target population results. For example, persons living in households with more persons tend to be victims of crime more often than persons living in smaller households.

Some compensation must be made during analysis of the survey data in order to eliminate this potential source of bias. Selection weights equal to the number of eligibles in the cluster can be used in survey estimation. Weighting and weighted estimates are described in detail in Chapter 10.

3.3.4 Duplication of Target Population Elements in Sampling Frames

duplication

The other kind of multiple mapping between frame and target populations that arises is duplication. "Duplication" means that a single target population element is associated with multiple frame elements. In the telephone survey example, this may arise when a single telephone household has more than one listing in a telephone directory. In Figure 3.2, Tom Clark, a target population member, has two frame elements associated with him: the telephone numbers 314-555-9123 and 314-555-9124. Multiple listings of the target population of households may occur because the household has more than one telephone number assigned to it, or

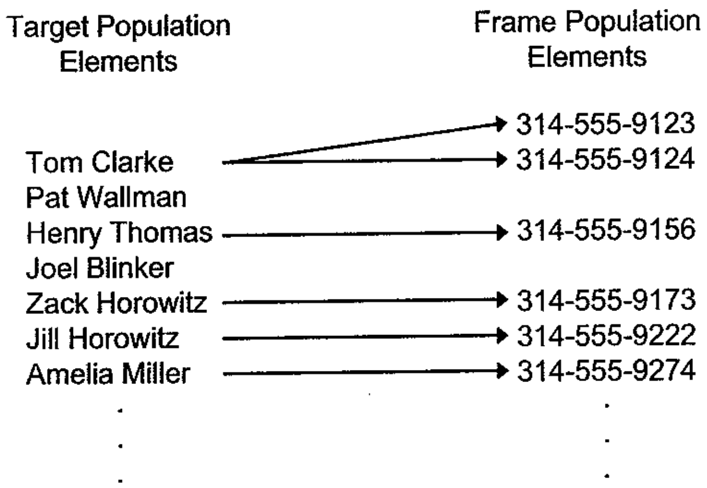

Figure 3.2 Duplication of target population elements by more than one sampling frame element.

because individuals within the household request and pay for additional listings in the directory. For example, in towns with universities and colleges, unrelated students often rent housing together, and acquire a single telephone number for the household. One listing in the directory is provided with the telephone subscription, and one person is given as the listed resident. Other residents may add listings for the same phone number under different names. This gives multiple frame listings for one household.

The problem that arises with this kind of frame problem is similar to that encountered with clustering. Target population elements with multiple frame units have higher chances of selection and will be overrepresented in the sample, relative to the population. If there is a correlation between duplication and variables of interest, survey estimates will be biased. In survey estimation, the problem is that both the presence of duplication and the correlation between duplication and survey variables are often unknown.

The potential for bias from duplication can be addressed in several ways. The sampling frame can be purged of duplicates prior to sample selection. For example, an electronic version of a telephone directory can be sorted by telephone number, and duplicate entries for the same number eliminated. When the frame cannot be easily manipulated, though, purging of duplicates may not be cost-effective.

Duplicate frame units may also be detected at the time of selection, or during data collection. A simple rule may suffice to eliminate the problem, designating only one of the frame entries for sample selection. Any other duplicate entries would be treated as foreign units, and ignored in selection. For example, one selection technique is that only the first entry in the directory is eligible. At the time of contact with the telephone household, the household informant can be asked if there is more than one entry in the directory for the household. If so, the entry with the surname that would appear first is identified. If the selection is for another entry, the interview would be terminated because the selection was by definition a foreign unit.

Another solution, as in the case of clustering, is weighting. If the number of duplicate entries for a given population element is determined, the compensatory weight is equal to the inverse of the number of frame elements associated with the sampled target element. For example, if a telephone household has two phone lines and three total entries in the directory (identified during data collection by informant report), the household receives a weight of ⅓ in a sample using the directory frame and a weight of ½ in a sample using an RDD frame.

3.3.5 Complicated Mappings between Frame and Target Population Elements

It is also possible to have multiple frame units mapped to multiple population elements. For example, in telephone household surveys of adults, one may encounter a household with several adults who have multiple entries in the directory. This many-to-many mapping problem is a combination of clustering and duplication. For example, in Figure 3.3 the three member Schmidt household (Leonard, Alice, and Virginia) has two telephone number frame elements (403-555-5912 and 403-555-5919). They might represent three target population elements mapped onto two sampling frame elements. A common solution to this problem is to weight survey results to handle both problems simultaneously. The compensatory weight

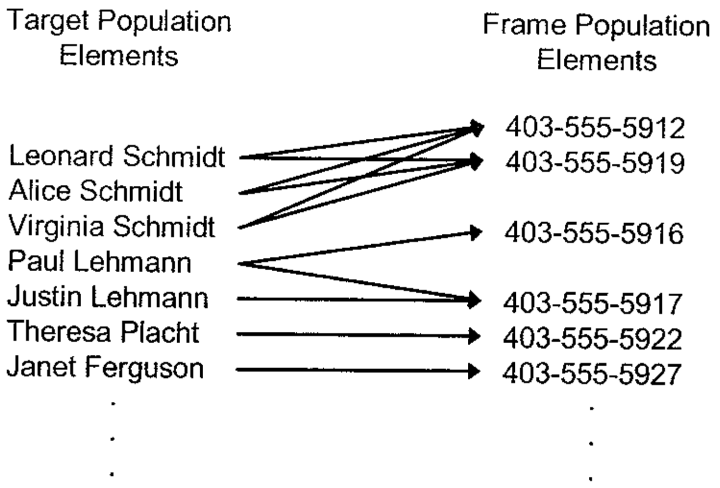

Figure 3.3 Clustering and duplication of target population elements relative to sampling frame elements.

for person-level statistics is the number of adults (or eligibles) divided by the number of frame entries for the household. In the example, the weight for the Schmidt household member selected would be $3/2$. More complicated weighting methods may be required for more complicated many-to-many mappings occurring in other types of surveys.

3.4 Common Target Populations and Their Frame Issues

Given the description of different frame problems above, we are now ready to describe some common target populations and frame issues they present. Surveys most commonly target the household population of a geographical area; employees, customers, or members of an organization; or organizations and groups. Some surveys are designed to select events such as surgical procedures or trips taken by car.

3.4.1 Households and Persons

In the United States, the common sampling frames for households are area frames (lists of area units like census tracts or counties), telephone numbers, telephone listings, and mailing lists. The area frame, because it is based on geographical units, requires an association of persons to areas, accomplished through a residency linking rule (*de facto* or *de jure*). Such a frame requires multiple stages when used to sample persons. First, a subset of area units is selected; then listings of addresses are made. If good maps or aerial photographs are available, the frame offers theoretically complete coverage of residences. The frame suffers undercoverage if the listing of residences within selected area units misses some units. The frame suffers duplication when one person has more than one residence. The frame suffers clustering when it is used to sample persons because multiple persons live in the final frame. These are the frame issues for the NCVS and NSDUH.

Another household population frame is that associated with telephone numbers for line telephones in housing units. This fails to cover about 6% of US households. A minority of households has more than one line number and are "over" covered. There are many nonresidential numbers on the frame that have to be screened out when it is used for person-level samples. These are the frame issues for BRFSS and SOC, both of which are random digit dialed surveys, using the telephone number frame.

The frame of listed residential telephone numbers in the United States is smaller than the telephone number frame. Commercial firms derive the frame from electronic and printed telephone directories. They sell samples from the frame to mass mailers and survey researchers. For uses in household surveys, it is much more efficient because most of the nonworking and nonresidential numbers are absent. However, a large portion of residential numbers, disproportionately urban residents and transient persons, is not listed in the directory. There are duplication problems because the same number can be listed under two different names, typically names of different members of the same household.

With the growing interest in Web surveys, there is much attention paid to the possibility of developing a frame of e-mail addresses for the household population. The e-mail frame, however, fails to cover large portions of the household population (see Section 5.3.3). It has duplication problems because one person can have many different e-mail addresses, and it has clustering problems because more than one person can share an e-mail (e.g., smithfamily@aol.com).

Mobile or cell phones are rapidly replacing fixed-line service in many countries. As early as the mid-1990s in Finland, for example, fixed-line telephone subscriptions began to decline while mobile phone subscribers rapidly increased (Kuusela, 1996). This shift represented a loss of fixed-line telephone coverage because cell phone numbers were not included in the existing frames. The coverage loss was greatest among younger persons and those just forming households independent of their parents.

In addition, mobile phones differ from line phones in that they are often associated with one person, not an entire household (as with line phones). Eventually, telephone surveys will sample mobile phone numbers, and this will require movement away from the household as a frame and sampling unit. At the present time, though, there are a number of frame problems associated with a mix of clustering and duplication occurring in fixed-line and cell service telephone numbers that are unsolved. There is much methodological research to be conducted in this area.

3.4.2 Customers, Employees, or Members of an Organization

Most surveys that study populations of customers, employers, or members of organizations use a list frame. Sometimes, the list frame is an electronic file of person records; other times it can be a physical set of records. Such record systems have predictable coverage issues. Undercoverage issues stem from the out-of-date files. New employees or customers tend to be missed if the files require several administrative steps before they are updated.

Similarly, the lists can contain ineligible elements, especially if persons leaving the organization are not purged from the list quickly. For example, in a file of customers, some of the customers may have experienced their last transaction so long ago that in their minds they may not perceive themselves as customers. In addition, there may be some ambiguity about whether a person should be counted as a member of the organization. In surveys of employees of a business, how should "contract" employees be treated? Although they may work day to day at the business, they are employees of another company that has a contract to supply services to the business.

Duplication of elements within a frame of employees or members can occur, but is rarer than with the household frames. In a frame of customers, duplication is inherent if the record is at the level of a transaction between the customer and the organization. Every person who is a customer has as many records as transactions they had with the organization. Hence, the survey researcher needs to think carefully about whether the target population is one of people (i.e., the customers) or transactions or both.

Survey researchers attempt to learn how and why the list was developed as part of the evaluation of alternative frames. For example, payroll lists of employees of a hospital may fail to cover volunteer and contract staff, but records from a security system producing identification cards may do so. Learning how and why the frame is updated and corrected is important. For example, payroll records for employees on monthly pay periods may be updated less frequently than those for daily or weekly pay periods. Updating procedures for temporary absences may complicate coverage issues of the frame. If an employee is on extended medical leave, is his/her record still in the frame? Should the target population include such persons? All of these issues require special examination for each survey conducted.

3.4.3 Organizations

Organizational populations are diverse. They include churches, businesses, farms, hospitals, medical clinics, prisons, schools, charities, governmental units, and civic groups. Sampling frames for these populations are often lists of units. Of all the types of organizational populations, perhaps businesses are the most frequent target population for surveys.

Business populations have distinct frame problems. First, a very prominent feature of business populations is their quite large variation in size. If the target population of software vendors is chosen, both Microsoft (with revenues over $20 billion per year) and a corner retail outlet that may sell $5,000 in software per year should be in the frame. Many business surveys measure variables where size is related to the variables of interest (e.g., estimates of total employment in the industry). Hence, coverage issues of business frames often place more emphasis on including the largest businesses than the smallest businesses.

Second, the business population is highly dynamic. Small businesses are born and die very rapidly. Larger businesses purchase others, merging two units into one (e.g., Hewlett-Packard buys Compaq and becomes a single corporation). A single business splits into multiple businesses (e.g., Ford Motor Company splits off its parts division, creating Visteon, an independent company). This means that frame populations need to be constantly updated to maintain good coverage of new businesses and to purge former businesses from the lists.

Third, the business population demonstrates a distinction between a legally defined entity and physical locations. Multiunit, multilocation companies are common (e.g., McDonald's has over 30,000 locations in the world, but only one corporate headquarters). Hence, surveys of businesses can study "enterprises," the legally defined entities, or "establishments," the physical locations. Some legally defined businesses may not have a physical location (e.g., a consulting business with each employee working on his/her own). Some locations are the site of many businesses owned by the same person.

Besides the business population, other organizational populations exhibit similar features, to larger or smaller extents. These characteristics demand that the survey researcher think carefully through the issues of variation in size of organization, the dynamic nature of the population, and legal and physical definitions of elements.

3.4.4 Events

Sometimes, a survey targets a population of events. There are many types of events sampled in surveys: a purchase of a service or product, marriages, pregnancies, births, periods of unemployment, episodes of depression, an automobile passing over a road segment, or a criminal victimization (like those measured in the NCVS).

Often, surveys of events begin with a frame of persons. Each person has either experienced the event or has not. Some have experienced multiple events (e.g., made many purchases) and are in essence clusters of event "elements." This is how the NCVS studies victimizations as events. It first assembles a frame of persons, each of whom is potentially a cluster of victimization events. NCVS measures each event occurring during the prior six months and then produces statistics on victimization characteristics.

Another logical frame population for event sampling is the frame of time units. For example, imagine wanting to sample the visits to a zoo over a one-year period. The purpose of the survey might be to ask about the purpose of the visit, how long it lasted, what were the most enjoyable parts of the visit, and what were the least enjoyable parts. One way to develop a frame of visits is to first conceptually assign each visit to a time point, say, the time of the exit from the zoo. With this frame, all visits are assigned to one and only one point in time. If the study involves a sample, then the research can select a subset of time points (say, 5-minute blocks) and attempt to question people about their visit as they leave during those 5-minute sample blocks.

Some time-use surveys (which attempt to learn what population members are doing over time) use electronic beepers that emit a tone at randomly chosen moments. When the tone occurs, the protocol specifies that the respondent report what they were doing at the moment (e.g., working at the office, watching television, shopping) (see Csikszentmihalyi and Csikszentmihalyi, 1988; Larson and Richards, 1994).

Surveys that study events may involve multiple populations simultaneously. They are interested in statistics about the event population, but also statistics about the persons experiencing the event. Whenever these dual purposes are involved, various clustering and duplication issues come to the fore. In a study of car purchases, for the event element of a purchase by a family, which persons experienced the event – the legal owner(s), all family members, or just the potential drivers of the car? The NCVS produces statistics like the percentage of households experiencing a crime (based on the household units) and the percentage of

household
incident p
different s

3.4.5 F

"Rare pop
est to rese
size but it:
the popula
persons re
would be
When cho
for identif
The th
ulations. F
example, c
these migh
Sometime:
the discus
commonly
be screene
families th
members c
is possible

3.5 Cc

There are r
3.3, but th
difficult p
It is impor
tics and es
large cove
same cove
duplicatio
Coverage
The n
presented
bias was g

where $\overline{Y}$ d
population
covered), r

household break-ins occurring while the residents were at home (based on the incident population). Careful thinking about the most informative population for different statistics is important in choosing the target and frame populations.

3.4.5 Rare Populations

rare population

"Rare populations" is a term that is used to describe small target groups of interest to researchers. Sometimes, what makes a population rare is not its absolute size but its size relative to available frames that cover it. For example, consider the population of welfare recipients in the United States. If there were 7.5 million persons receiving welfare benefits in a population of 280 million, the population would be rare principally because it forms less than 3% of the total population. When chosen as target populations, rare populations pose considerable problems for identifying suitable sampling frames.

The there are two basic approaches to building sampling frames for rare populations. First, lists of rare population elements themselves can be made. For example, one can attempt to acquire lists of welfare recipients directly (although these might be kept confidential) through records in welfare disbursement offices. Sometimes, no single list has good coverage and multiple lists are assembled (see the discussion of multiple frame designs in Section 3.6.3). Second, and more commonly, a frame that includes the rare population as a subset of elements can be screened. For example, the household population can be screened to locate families that receive welfare payments. If all elements of the rare population are members of the larger frame population, complete coverage of the rare population is possible (albeit at the expense of screening to locate the rare population).

3.5 Coverage Error

There are remedies for many of the sampling frame problems discussed in Section 3.3, but the remedies do not always eliminate coverage error. Undercoverage is a difficult problem, and may be an important source of coverage error in surveys. It is important to note, though, that coverage error is a property of sample statistics and estimates made from surveys. One statistic in a survey may be subject to large coverage errors; another from the same survey can be unaffected by the same coverage issues. In the jargon of survey methodology, undercoverage, duplication, clustering, and other issues are problems of a sampling frame. Coverage error is the effect of those problems on a survey statistic.

The nature of coverage error in a simple statistic like the sample mean was presented in Section 2.3.4. Recall that if a mean is being estimated, the coverage bias was given as

$$\bar{Y}_C - \bar{Y} = \frac{U}{N}\left(\bar{Y}_C - \bar{Y}_U\right),$$

where $\bar{Y}$ denotes the mean for the total population, $\bar{Y}_C$ and $\bar{Y}_U$ are the means in the population of the eligible units on the frame (covered) and not in the frame (not covered), respectively, U is the total number of target population elements off the frame and N is the full target population. Thus, the error due to not covering the $N–C$ units left out of the frame is a function of the proportion "not covered" and the difference between means for the covered and the not covered.

The survey (regardless of its sample size) can only estimate the mean of the covered, $\bar{Y}_C$. The extent to which the population of U noncovered units is large, or there is a substantial difference between covered and undercovered, determines the size of the bias, or coverage error. The proportion not covered will vary across subclasses of the eligible persons. That is, undercoverage could be higher for the total sample than for a particular subgroup. In addition, since coverage error depends on the difference of estimates between covered and undercovered, coverage error can vary from one statistic to another, even if they are based on the same subclass of eligible units.

3.6 Reducing Undercoverage

Remedies for common frame problems such as duplication, clustering, many-to-many mappings, undercoverage, and foreign units in frames have been examined in Section 3.3. However, specific remedies for undercoverage, and consequent coverage error, were not addressed in detail in that section. There is a general class of coverage improvement procedures that involve frame supplementation designed to reduce coverage error more specifically.

3.6.1 The Half-Open Interval

Frames that are slightly out of date, or that provide reasonably good coverage except for some kinds of units, may be brought up to date through additions in update listing operations during or shortly before data collection. If there is a logical order to the list, it may be possible to repair the frame by finding missing units between two listed units.

No.	Address	Selection?
1	101 Elm Street	
2	103 Elm Street, Apt. 1	
3	103 Elm Street, Apt. 2	
4	107 Elm Street	Yes
5	111 Elm Street	
6	302 Oak Street	
7	306 Oak Street	
...	...	...

Figure 3.4 Address list for area household survey block.

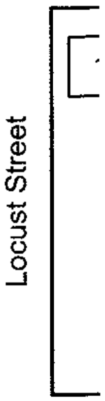

Figure 3.

Consider, fo:
veys (see Figure
They may also h
added to the list.
it is possible to a
than updating the
during data colle
One such too
block with the a
addresses is avail
that an address fi
frame perspective
structure but a ge
up to but not incl
what mathematic
each address app
closed end of the
(the open end of
When an int
open interval" as
constructed units
covered, the inter
to conduct an int
open interval thu
address), and mi
frame during data
On occasion
too large for the
instance, if a nev

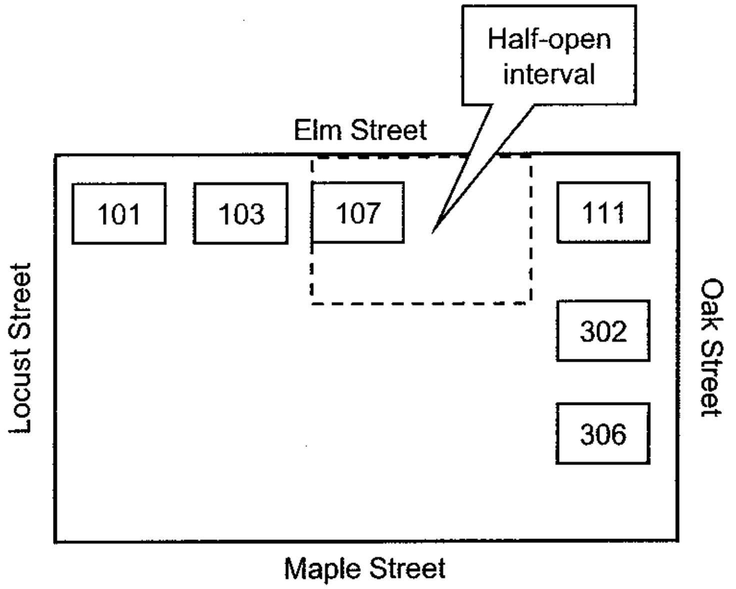

Figure 3.5 Sketch map for area household survey block.

Consider, for example, address or housing unit lists used in household surveys (see Figure 3.4). These lists may become out of date and miss units quickly. They may also have missed housing units that upon closer inspection could be added to the list. Since address lists are typically in a particular geographic order, it is possible to add units to the frame only for selected frame elements, rather than updating the entire list. That is, frames can be updated after selection and during data collection.

half-open interval

One such tool is called the "half-open interval." Consider the example of one block with the address list shown in Figure 3.4. The geographic distribution of addresses is available for the block in a sketch map shown in Figure 3.5. Suppose that an address from this frame has been selected: 107 Elm Street. From an area frame perspective, the address of 107 Elm Street will be viewed not as a physical structure but a geographic area bounded by "property lines" from 107 Elm Street up to but not including the next listed address, 111 Elm Street. List order defines what mathematicians would define from set theory as a half-open interval for each address appearing on the list. The interval begins with 107 Elm Street (the closed end of the interval) and extends up to but does not include 111 Elm Street (the open end of the interval).

When an interviewer arrives at the selected address, he inspects the "half-open interval" associated with 107 Elm Street to determine if there are any newly constructed units or any missed units in the interval. If a new or missed unit is discovered, the interviewer adds it to the list, selects it as a sample unit, and attempts to conduct an interview at all addresses in the interval. All addresses in the half-open interval thus have the same probability of selection (that of the selected address), and missed or newly constructed units are automatically added to the frame during data collection.

On occasion, the number of new addresses found in a half-open interval is too large for the interviewer to be able to sustain the added workload. For instance, if a new apartment building with 12 apartments had been constructed between 107 and 111 Elm Street, the interviewer would be faced with conducting 13 interviews instead of an expected single household interview. In such instances, the additional addresses may be subsampled to reduce the effect of this clustering on sample size and interviewer workload. Subsampling, as for clustering in sample frames, introduces unequal probabilities of selection that must be compensated for by weighting (see Chapter 10 for discussion of weighting and weighted estimates). In continuing survey operations, it is also possible to set aside such added units into a separate "surprise" stratum (see Kish and Hess, 1959) from which a sample of missed or newly constructed units are drawn with varying probabilities across additional samples selected from the frame.

Similar kinds of linking rules can be created for coverage checks of other logically ordered lists that serve as frames. For example, a list of children attending public schools ordered by address and age within address could be updated when the household of a selected child is visited and additional missed or recently born children are discovered in the household.

3.6.2 Multiplicity Sampling

The half-open interval concept supplements an existing frame through information collected during the selection process. Some frame supplementation methods add elements to a population through network sampling. This is commonly termed "multiplicity sampling." A sample of units can be selected, and then all members of a well-defined network of units identified for the selected units.

multiplicity sampling

For instance, a sample of adults may be selected through a household survey, and asked about all of their living adult siblings. The list of living adult siblings defines a network of units for which information may be collected. Of course, the network members have multiple chances of being selected, through a duplication in the frame, since they each could be selected into the sample as a sample adult. The size of the network determines the number of "duplicate chances" of selection. If an adult sample person reports two living adult siblings, the network is of size three, and a weight of ⅓ can be applied that decreases the relative contribution of the data from the network to the overall estimates. This "multiplicity" sampling and weighting method (Sirken, 1970) has been used to collect data about networks to increase sample sizes for screening for rare conditions, such as a disease. The method does have to be balanced against privacy concerns of individuals in the network. In addition, response error (see Chapter 7) such as failing to report a sibling, including someone who is not a sibling, or incorrectly reporting a characteristic for a member of the network, may contribute to errors in the network definition and coverage as well as in the reported levels of a characteristic.

"Snowball sampling" describes a closely related, although generally nonprobability, method to supplement a frame. Suppose an individual has been found in survey data collection who has a rare condition, say blindness, and the condition is such that persons who have the condition will know others who also have the condition. The sample person is asked to identify others with the condition, and they are added to the sample. Snowball sampling cumulates sample persons by using network information reported by sample persons. Errors in reports, isolated individuals who are not connected to any network, and poorly defined networks make snowball sampling difficult to apply in practice. It generally does not yield a probability sample.

Although multiplicity sampling offers theoretical attraction to solve frame problems, there is much research left to be conducted on how to implement practical designs. These include problems of measurement error in reports about networks, nonresponse error arising from incomplete measurement of networks, and variance inflation of multiplicity estimators.

3.6.3 Multiple Frame Designs

Coverage error can sometimes be reduced by the use of multiple frames, in several ways. A principal frame that provides nearly complete coverage of the target population may be supplemented by a frame that provides better or unique coverage for population elements absent or poorly covered in the principal frame. For example, an out-of-date set of listings of housing units can be supplemented by a frame of newly constructed housing units obtained from planning departments in governmental units responsible for zoning where sample addresses are located. Another example concerns mobile homes that may be present on an address list but poorly covered. A supplemental frame of mobile home parks may be added to the principal address list to provide better coverage of the population residing in mobile homes.

At times, the supplemental frame may cover a completely separate portion of the population. In most cases, though, supplemental frames overlap with the principal frame. In such cases, multiple frame sampling and estimation procedures are employed to correct for unequal probabilities of selection and possibly to yield improved precision for survey estimates.

multiple frame sampling

Suppose, for example, that in a household survey random digit dialing (RDD) is used to reach US telephone households. RDD will, in principle, cover all telephone households in the country, but it fails to cover approximately 6% of the households that do not have telephones. Figure 3.6 shows the telephone frame as a shaded subset of the area frame of housing units. A remedy to undercoverage

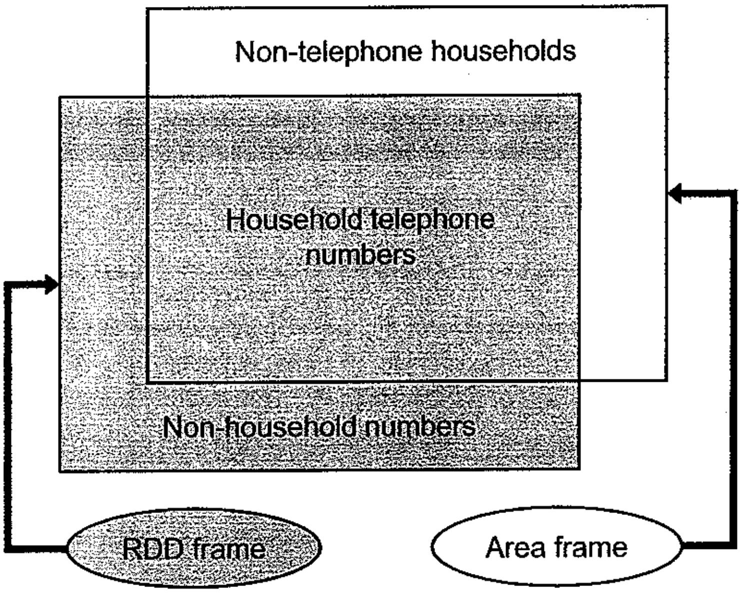

Figure 3.6 Dual frame sample design.

is a supplementary area frame of households. Under such a dual frame design, a sample of households is drawn from an area frame, but it will require visits to households, considerably more expensive than contact by telephone. Together, the two frames provide complete coverage of households. For the NCVS, Lepkowski and Groves (1986) studied the costs and error differences of an RDD and an area frame. They found that for a fixed budget most statistics achieve a lower mean square error when the majority of the sample cases are drawn from the telephone frame (based on a simulation including estimates of coverage, nonresponse, and some measurement error differences).

These two frames overlap, each containing telephone households. The data set from this dual frame survey combines data from both frames. Clearly, telephone households are overrepresented under such a design since they can be selected from both frames.

There are several solutions to the overlap and overrepresentation problem. One is to screen the area household frame. At the doorstep, before an interview is conducted, the interviewer determines whether the household has a fixed-line telephone that would allow it to be reached by telephone. If so, the unit is not selected and no interview is attempted. With this procedure, the overlap is eliminated, and the dual frame sample design has complete coverage of households.

A second solution is to attempt interviews at all sample households in both frames, but to determine the chance of selection for each household. Households from the nonoverlap portion of the sample, the nontelephone households, can only be selected from the area frame, and thus have one chance of selection. Telephone households have two chances, one from the telephone frame and the other from the area household frame. Thus, their chance of selection is $p_{RDD} + p_{area} - p_{RDD} * p_{area}$, where p_{RDD} and p_{area} denote the chances of selection for the RDD and area sample households. A compensatory weight can be computed as the inverse of the probability of selection: $1/p_{area}$ for nontelephone households and $1/(p_{RDD} + p_{area} - p_{RDD} * p_{area})$ for telephone households, regardless of which frame was used.

A third solution was proposed by Hartley (1962) and others. They suggested that the overlap of the frames be used in estimation to obtain a more efficient estimator. They proposed that a dual frame (in their case, multiple frame) design be examined as a set of nonoverlapping domains, and results from each domain combined to obtain a target population estimate. In the illustration in Figure 3.6, there would be three domains: nontelephone households (*Non-tel*), RDD telephone households (*RDD-tel*), and area sample telephone households (*area-tel*). The *RDD-tel* and *area-tel* households are combined with a mixing parameter chosen to maximize mathematically the precision of an estimate (say a mean). The telephone and nontelephone domains are combined using a weight that is the proportion of the telephone households in the target population, say W_{tel}. The dual frame estimator for this particular example is

$$\bar{y} = \left(1 - W_{tel}\right) p_{non-tel} + W_{tel}\left[\theta\, p_{RDD-tel} + \left(1-\theta\right) p_{area-tel}\right],$$

where θ is the mixing parameter chosen to maximize precision.

The dual frame illustration in Figure 3.6 is a special case of the multiple frame estimation approach. The method can be applied to more complex situations involving three or more frames, where the overlap creates more domains.

Even in dual frame sampling, there are at least four domains: frame 1 only, frame 2 only, frame 1 sample overlapping with frame 2, and frame 2 sample overlapping with frame 1. (Although the last two are intersections of the two frames, which frame they are actually sampled from may affect the survey statistics; hence they are kept separate.) These kinds of designs are found in agriculture surveys. Suppose that a sample is to be drawn in a state of farm holdings that have a particular kind of livestock, say dairy cows. Suppose also that there is a list of dairy farmers available from the state department of agriculture, but it is known to be out of date, having some dairy farms listed that no longer have dairy cows and not listing small farms with dairy cows. A second area frame is used to draw a sample of all farms. There will be four domains: list frame only, area frame only, list frame also found on area frame, and area frame also found on list frame. Again, screening, weighting, or multiple frame estimation may be used to address the overlap problem.

One last example of more recent interest concerns Web survey design. Suppose that a list of e-mail addresses is available from a commercial firm. It is inexpensive to use for self-administered surveys, but it has foreign elements (addresses that are no longer being used, persons who are not eligible) and lacks complete coverage of the eligible population. A second supplementary RDD frame can be used to provide more expensive and complete coverage of eligible persons in all telephone households. Samples would be drawn from each frame, interviews conducted with sample eligible persons in both, and a dual frame estimation procedure used to combine results from these overlapping

Tourangeau, Shapiro, Kearney, and Ernst (1997) and Martin (1999) on Household Rosters

Two studies on why persons are omitted from listings of household members inform survey practice.

Study designs: The Tourangeau et al. study mounted a randomized experimental design of three different rostering procedures: asking for names of all living at the unit, asking for names of all who spent the prior night at the unit, and asking for initials or nicknames of all those who spent the prior night. Follow-up questions asked whether all persons listed fulfilled the definition of living in the unit. Face-to-face interviews were conducted in 644 units on 49 blocks in three urban areas. The Martin study used an area probability sample of 999 units. Roster questions asked for all persons with any attachment to the household during a two-month period. Reinterviews were conducted on a subsample with further questions about residency. In both studies, follow-up questions after the roster asked whether all persons listed fulfilled the definition of living in the unit.

Findings: The Tourangeau et al. study found that after probes to identify usual residence, only the technique of asking for initials produced more than the standard procedure. They conclude that concealment of the identity of some residents contributes to undercoverage. The Martin study found inconsistent reporting for unrelated persons, persons away from the home for more than a week, and persons not contributing to the financial arrangements of the unit. Martin concluded that informant's definitions of households do not match the survey's definition, causing underreporting.

Limitations of the studies: There was no way to know the true household composition in either study. Both assumed that follow-up questions producing larger counts of persons were more accurate reports.

Impact of the studies: They helped to document the size of household listing errors. They demonstrated that both comprehension of the household definition and reluctance to report unusual composition produce the undercoverage of household listings.

frames. There is much methodological research to be conducted in these mixes of frames and modes (see more in Section 5.4 regarding these issues).

3.6.4 Increasing Coverage While Including More Ineligible Elements

The final coverage repair arises most clearly in coverage of persons within sample households (first described in Section 3.3.3), as part of a survey to produce person-level statistics. When a housing unit is identified for measurement, the interviewer attempts to make a list of all persons who live in the household. There appear to be consistent tendencies of underreporting of persons who are inconsistent residents of the household or who fail to fit the household informant's native definition of a "household."

Both Tourangeau, Shapiro, Kearney, and Ernst (1997) and Martin (1999) investigated what happens when questions about who lives in the unit are altered to be more inclusive. The typical question that generates the frame of persons within a household is "Who lives here?" Additional questions included who slept or ate there the previous evening, who has a room in the unit, who has a key to the unit, who receives mail at the unit, who usually is at the home but was away temporarily, etc. (see box on page 89). The questions appeared to increase the number of different people mentioned as attached to the household.

The next step in the process is the asking of questions that determine whether each person mentioned did indeed fit the definition of a household member according to the survey protocol. After such questions, those mentioned who have households elsewhere are deleted.

In essence this repair strategy "widens the net" of the frame and then trims out those in the net who were erroneously included. The burden of the strategy is that it requires more time and questions to assemble the frame of eligible persons in the household. Many times, this questioning is one of the first acts of the interviewer, at which point continued cooperation of the household informant is most tenuous. Hence, at this writing, adoption of the new approach is limited.

3.7 Summary

Target populations, sampling frames, and coverage are important topics in survey design because they affect the nature of the inference that can be made directly from survey data. The problems that arise when comparing frame to target population have remedies, many of which are standard approaches in survey research. They are also not necessarily complete corrections to the coverage error that may arise.

Coverage errors exist independent of the sampling steps in surveys. The sample selection begins with the frame materials. Samples can be no better than the frames from which they are drawn. We examine in the next chapter how samples for surveys are drawn. The discussion assumes that the kind of coverage errors and frame problems examined here are considered separately from the issues of how to draw a sample that will yield precise estimates for population parameters.

KEY

area f
area p
cluste
cover
duplic
de fac
de jur
eleme
foreig
half-o
housel
housir

FOR

Kish,

Lessle
3-

EXER

1) N
w
to

2) U
hc
sa
tic
th

3) N
fa
tic

4) Y
ar
an
m
sq
fr
of

Keywords

area frame
area probability sample
clustering
coverage
duplication
de facto residence rule
de jure residence rule
elements
foreign element
half-open interval
household
housing unit
ineligible element
ineligible unit
multiple mappings
multiplicity sampling
multiple frame sampling
random digit dialing (RDD)
rare population
sampling frame
survey population
target population
undercoverage
usual residence

For More In-Depth Reading

Kish, L. (1965), *Survey Sampling*, Chapter 11, Section 13.3, New York: Wiley.

Lessler, J. and Kalsbeek, W. (1992), *Nonsampling Error in Surveys*, Chapters 3–5, New York: Wiley.

Exercises

1) Name two conditions (whether or not they are realistic) under which there would be no coverage error in a statistic from a telephone survey attempting to describe the target population of adults in US households.

2) Using one of the six example surveys (see Chapter 1), give an example of how altering the definition of the target population or the definition of the sampling frame could eliminate a coverage error in a resulting survey statistic. (Mention the survey statistic you have in mind, the target population, and the sampling frame.)

3) Name three concerns you would have in transforming an area probability face-to-face survey into a Web survey attempting to estimate the same statistics.

4) You are interested in the target population of farm operators in a three-county area encompassing 360 square miles. You lack a list of the farm operations and instead plan on using a grid placed on a map, with 360 square-mile segments. You plan to draw a sample of farm operations by drawing a sample of square-mile segments from the grid. Identify three problems with using the frame of 360 square-mile grids as a sampling frame for the target population of farm operations in the three-county area.

5) Five years after the last census, you mount a household survey using a telephone number frame. If a selected telephone number is a household number, interviewers ask to speak to the person most knowledgeable about the health of the household members. After the survey is over, someone suggests evaluating your survey by comparing the demographic distributions (i.e., age, sex, race/ethnicity, gender) of your "most knowledgeable" health informants to the demographic distributions of adults from the last census. Comment on the wisdom of this suggestion.

CHAP

SAM
SAM

4.1 INT

The selecti
critical par
naire, highl
managemei
collected, a
haphazardl
making inf
In the
"Survey 2C
the frequen
ing, etc. In
ple to com
page and c
Howard, 2(
A little
the Survey
the Arts, 1
holds selec
was approx
It measurec
The re
however, v
survey. For
ater in the
for nonmu
dents but o
gallery.
How t
selected na
Geographi
and active
Contra
of Consum
at random
Alaska and
ence, both