

Sampling Frames and Coverage Error

Wayne Enanoria, PhD, MPH
Public Health Epidemiologist
Center for Infectious Disease Preparedness
UC Berkeley School of Public Health
Email: enanoria@berkeley.edu

*Slides created using free, open source software:
<http://www.openoffice.org>*

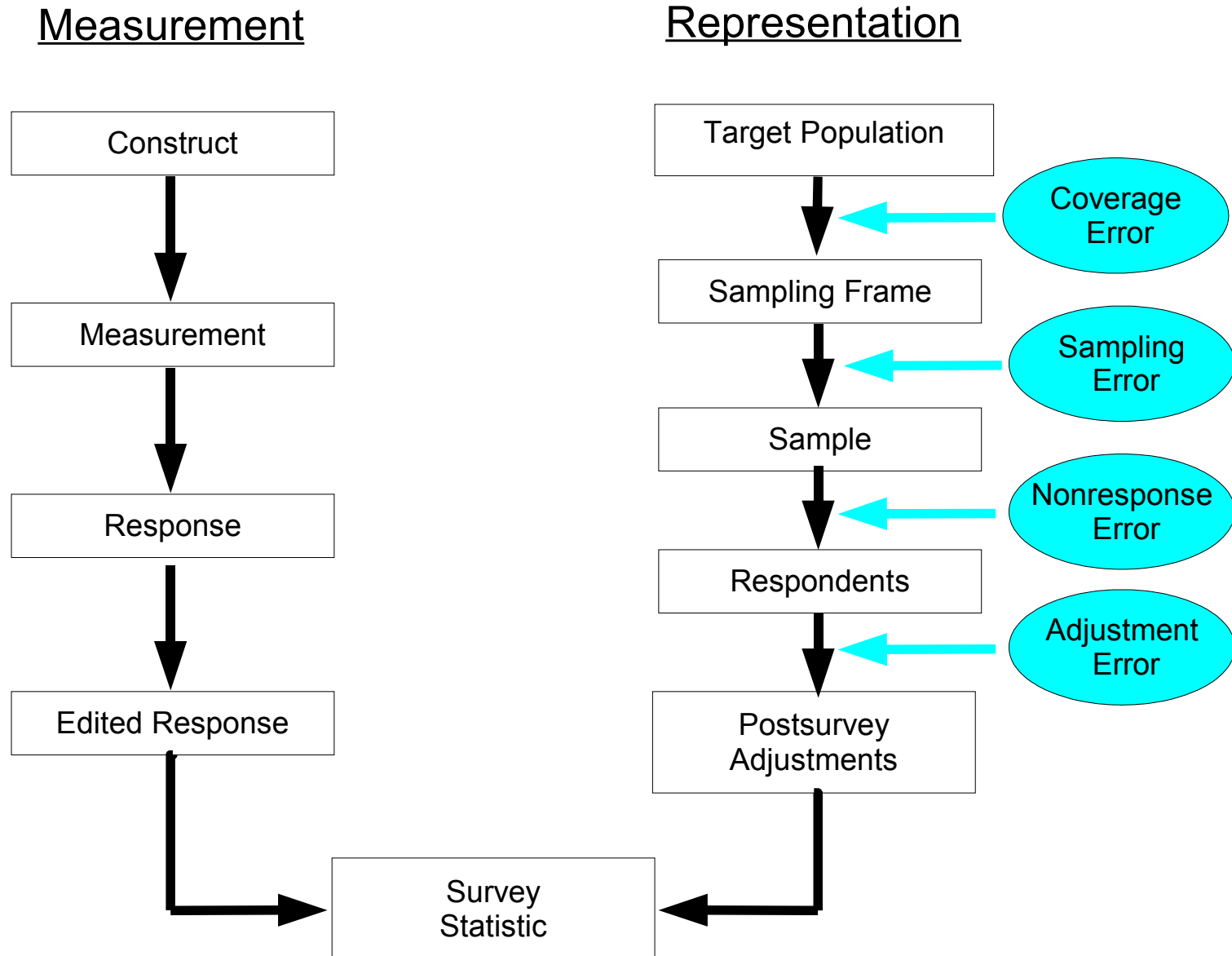


Overview

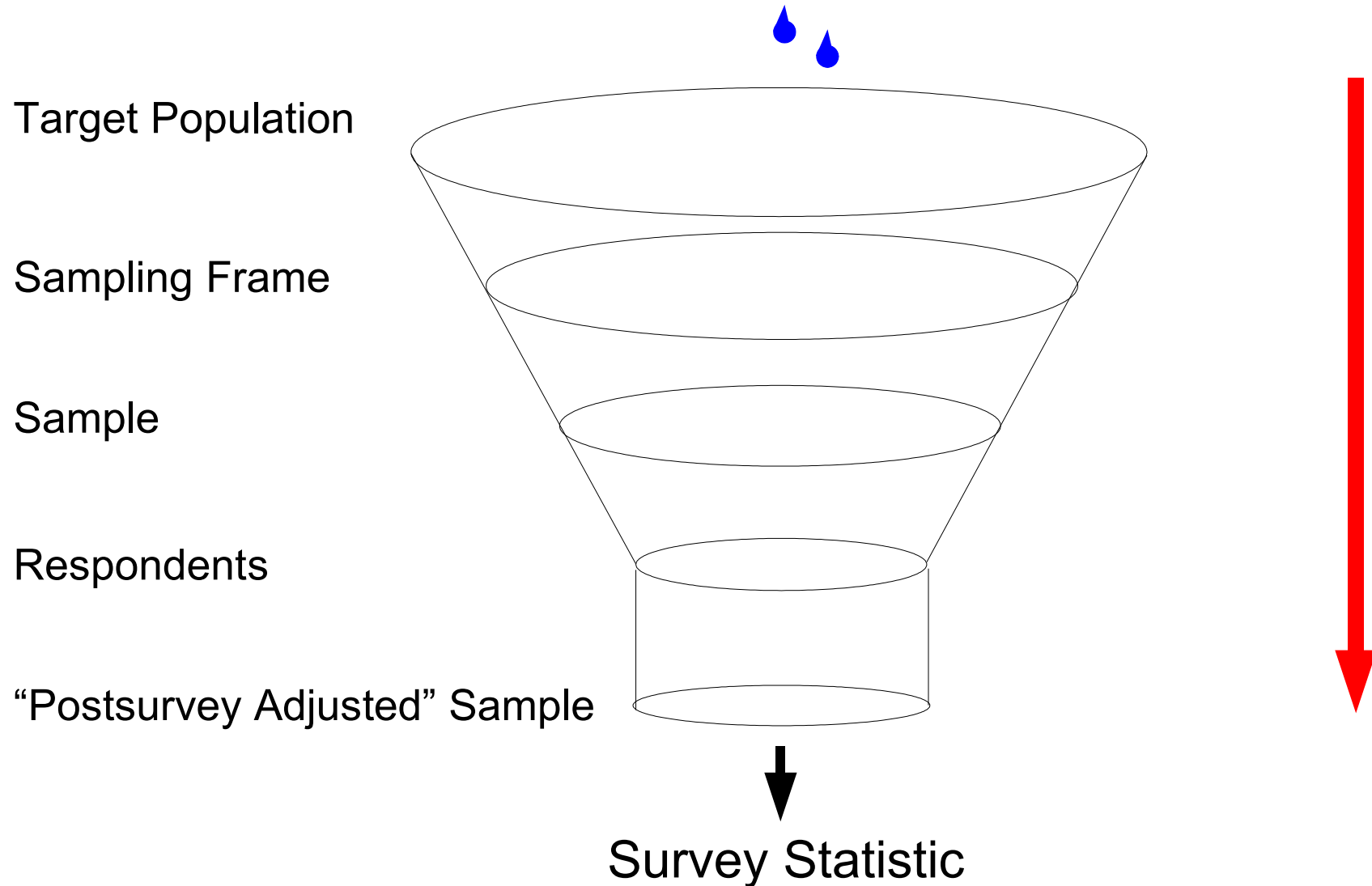
- ◆ Problems leading to coverage error:
 - ◆ Ineligible units
 - ◆ Clustering
 - ◆ Duplication
 - ◆ Clustering and duplication
- ◆ Additional methods for reducing undercoverage



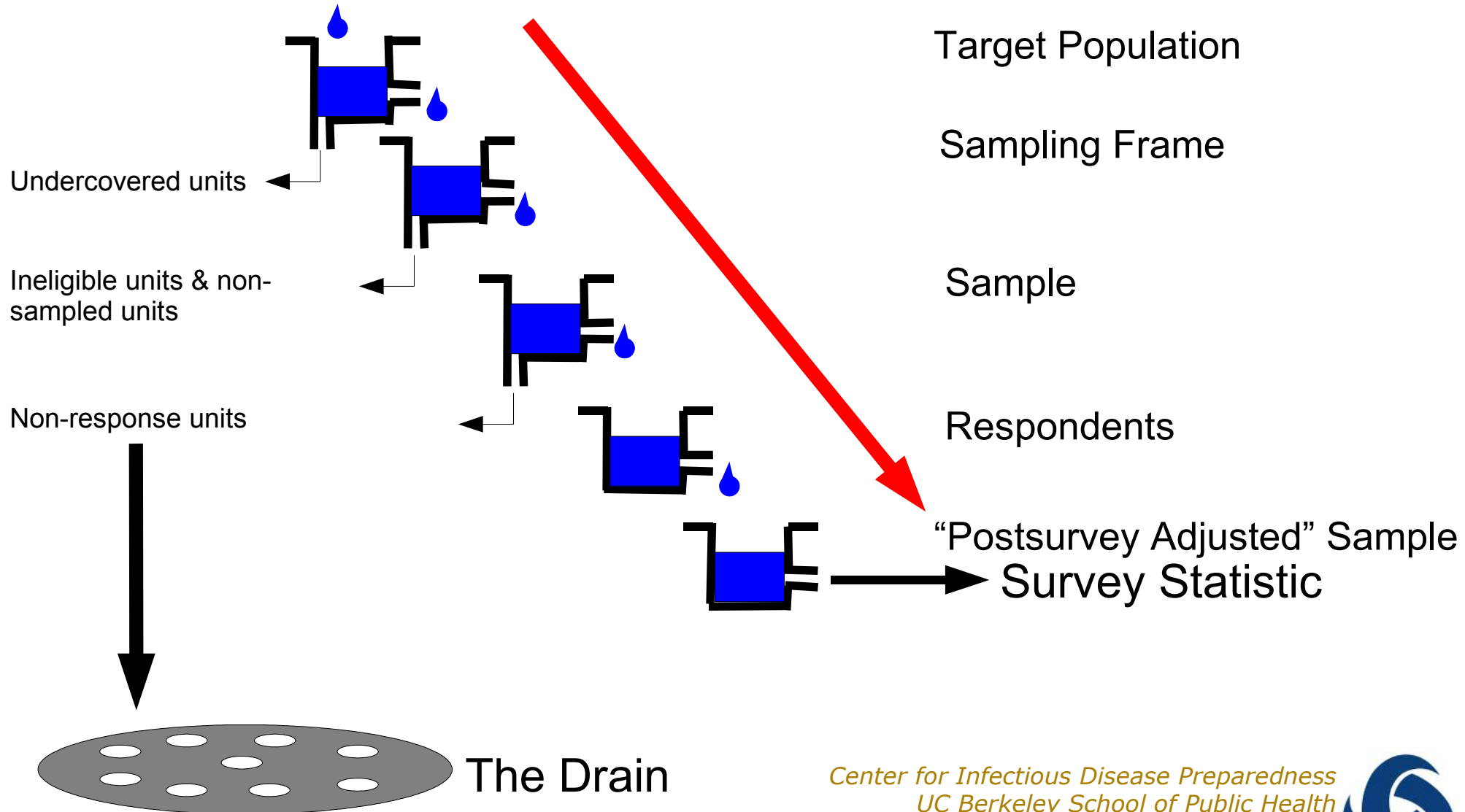
Survey Quality Perspective



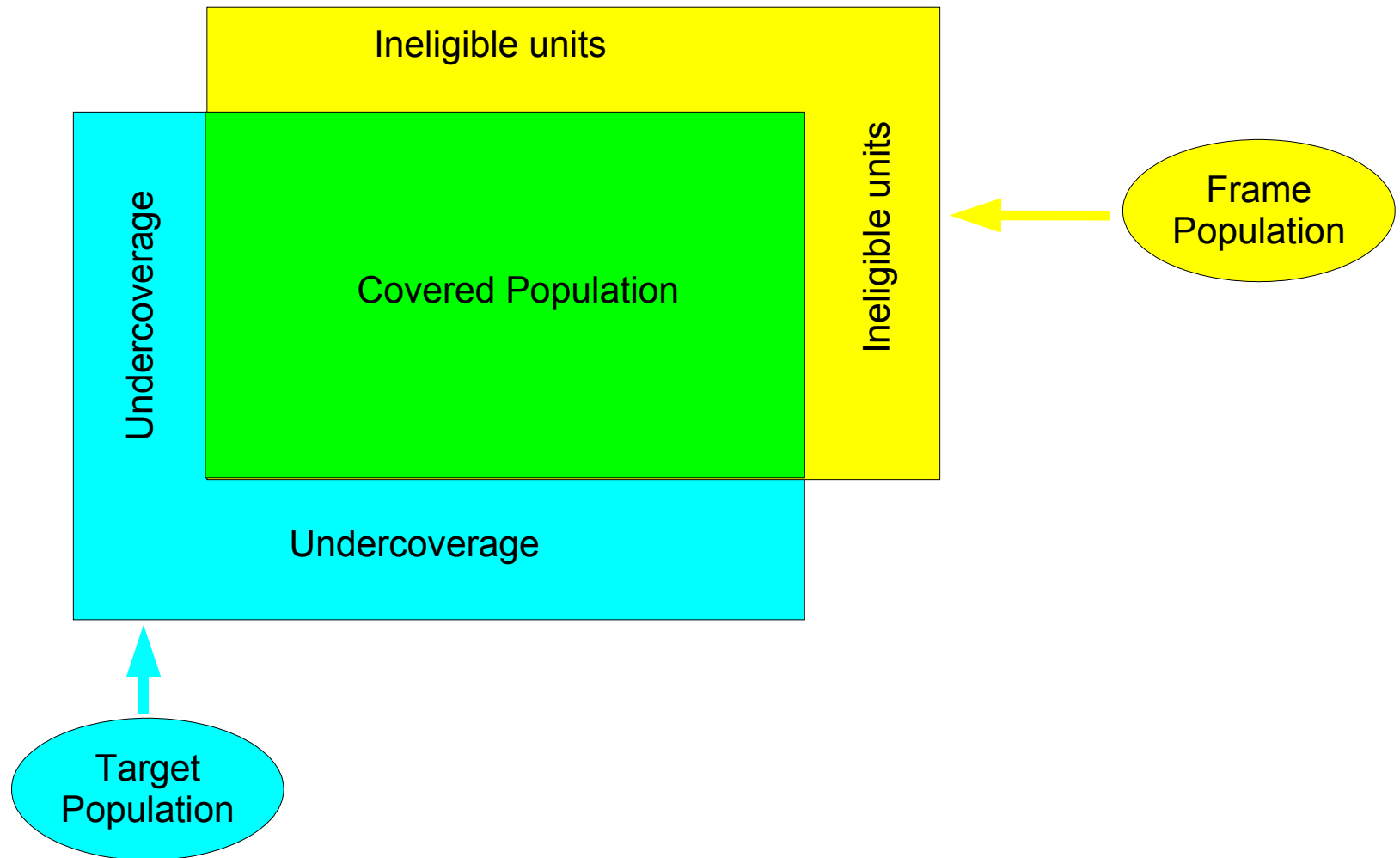
“Funnel” of Information



“Tubs” of Information



Coverage of a Target Population by a Frame



Options with Coverage Error

- ♦ Redefine the target population to fit the frame better.
 - ♦ By doing so, the survey may not be fully relevant to the population of interest.
- ♦ Admit the possibility of coverage error in statistics describing the original target population.
 - ♦ There will be criticisms of coverage error in survey statistics.

NOTE: the option you choose does not remedy the fact that the sampling frame does not fully cover the target population.



Problems leading to Coverage Error (Kish1965)

- (1) Missing elements: some population elements are not included in the frame;
- (2) Clustering: some listings refer to groups of elements, not individual elements;
- (3) Blank or foreign elements (ineligible units): when some listings do not relate to elements of the survey population; and
- (4) Duplicate listings: some population elements have more than one listing.



Ineligible Units

- ♦ Sampling frames can, at times, contain elements that are not part of the target population (“ineligible units”).
 - ♦ Many telephone numbers are not working or nonresidential numbers in a sampling frame of telephone numbers.
 - ♦ Unoccupied or business structures may appear to be housing units and can sometimes be included in area probability surveys.
- ♦ If ineligible or foreign units are identified on the sampling frame before selection begins, they can be purged with little cost.



Solution to Ineligible Units (1)

- ♦ Oftentimes, ineligible units cannot be identified until data collection begins.
- ♦ If few in number, after sampling they can be identified in a screening phase and dropped from the sample.
 - ♦ There will be a reduction in sample size.
 - ♦ Thus, select more than target sample to account for this.



Sample Calculation

- Let's assume that 15% of the entries in residential portions of national telephone directories are numbers that are no longer in service.
- We want 100 households in our survey.
- We do the following:

(number units sampled) - (number units "ineligible") = target number of units

X - $0.15 * X$ = 100

$(1 - 0.15) * X$ = 100

$X = 100 / (1 - 0.15) = 118$ entries



Solution to Ineligible Units (2)

- ♦ When the proportion of ineligible units is large, the sampling frame may not be cost-effective to use.
- ♦ Alternative sampling frames should be investigated.



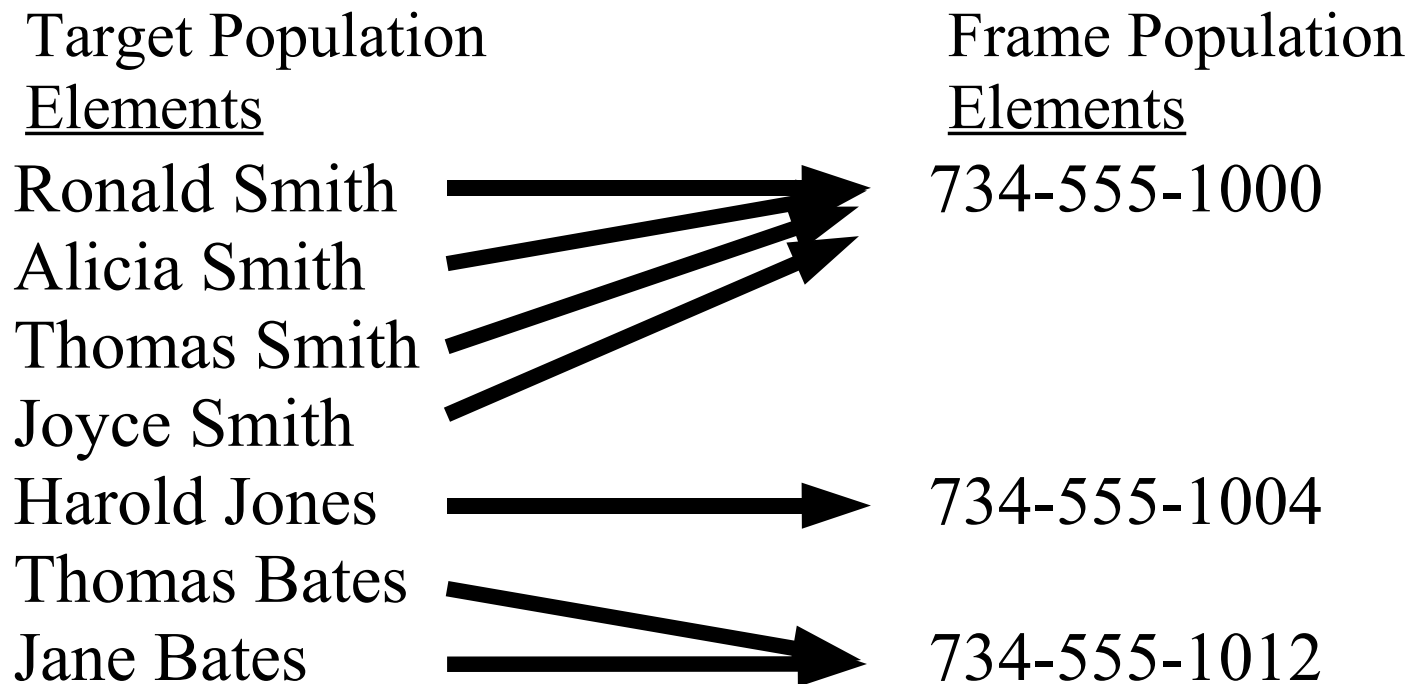
Clustering of Target Population Elements

- ♦ Multiple mappings of sampling frame to target population can also lead to problems in sample selection.
- ♦ “Clustering” means that multiple elements of the target population are represented by the same frame element.
 - ♦ A single sampling frame element represents a “cluster” of more than one target population element.
 - ♦ ratio target element to frame element = $n:1$



Clustering Example

(Figure 3.1. Groves 2004, page 75)



Solution to Clustering (1)

- ♦ Select *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.
- ♦ However, it may be difficult to collect information successfully from all elements in the cluster.
 - ♦ nonresponse can increase
 - ♦ members of the household can talk about the survey amongst themselves, thereby “tainting” future results
 - ♦ if the clusters are different in size, it may be difficult to control sample size



Solution to Clustering (2)

- ♦ Select only a sample of elements from each frame unit sampled.
 - ♦ In our telephone example, only select one adult from each sampled household.
- ♦ Advantages:
 - ♦ All efforts to obtain an interview are concentrated on one eligible person.
 - ♦ Contamination is eliminated within the household.
 - ♦ Sample size in persons is equal to the number of households sampled.



Difficulties with Solution (2)

- ♦ In order to sample one individual from the household, you need to know all of the eligible members in the household.
 - ♦ can lead to suspicion and to nonresponse
- ♦ Alternative approaches for sampling include:
 - ♦ “last birthday” method, if the survey is taken at one time point.



Difficulties with Solution (2) con't

- ♦ Overall probabilities of selection of individual household members vary by cluster size.
 - ♦ Elements in large clusters have lower overall probabilities of selection than elements in small clusters.
 - ♦ The sample will overrepresent persons from households with fewer eligible persons relative to the target population.
- ♦ Selection weights equal to the number of eligibles in the cluster can be used in survey estimation.



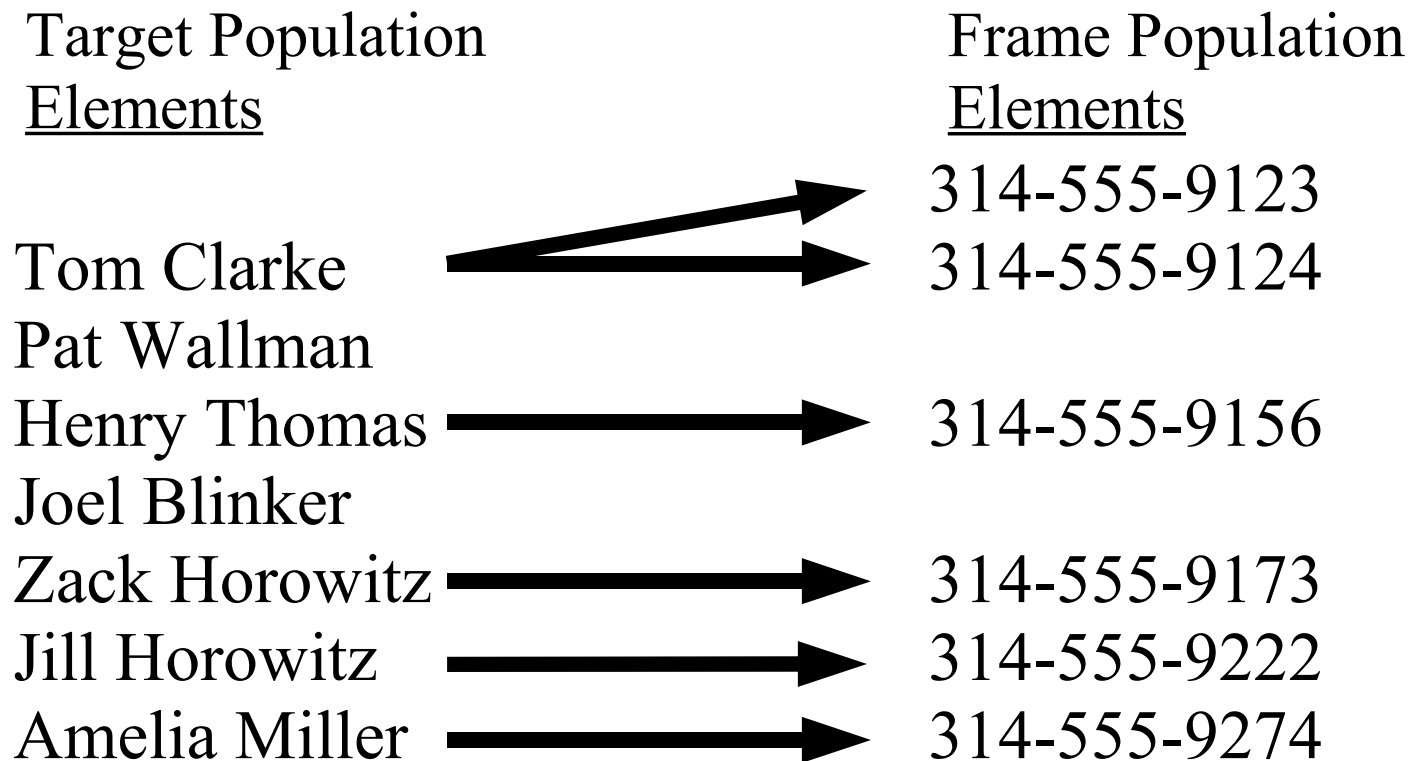
Duplication of Target Population Elements

- ♦ “Duplication” is another kind of multiple mapping problem between frame and target populations.
 - ♦ A single target population element is “duplicated” by more than one sampling frame element.
- ♦ Duplication is when a single target population element is associated with multiple frame elements.
 - ♦ ratio target element to frame element = 1:n



Duplication Example

(Figure 3.2. Groves 2004, page 77)



Problems with Duplication

- ♦ Target population elements with multiple frame units have higher probabilities of selection.
 - ♦ Such target elements will be overrepresented in the sample relative to the population.
 - ♦ If there is correlation between duplication and variables of interest, survey estimates will be biased.
 - ♦ The problem of duplication and the correlation between duplication and survey variables are often unknown.



Solution to Duplication (1)

- ♦ If the extent of duplication in the sampling frame is known, one can purge duplicates prior to sample selection.
 - ♦ may not be cost-effective, depending on the frame used



Solution to Duplication (2)

- ♦ Duplicate frame units may be detected at the time of sample selection or during data collection.
- ♦ A simple rule can be devised to eliminate duplicates, leaving only one of the frame entries available for sample selection.
 - ♦ For example, only the first entry in the directory is eligible; all other entries can be considered “foreign units” and ignored in selection.



Solution to Duplication (3)

- ♦ An alternative solution is to weighting the results to account for duplication.
- ♦ 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, a telephone household has two phone lines and three total entries in the directory.
 - ♦ The household receives a weight of $1/3$ in a sample using the directory frame (three entries in the directory).
 - ♦ The household receives a weight of $1/2$ in a sample using a Random Digit Dialing (RDD) frame (two valid phone numbers).



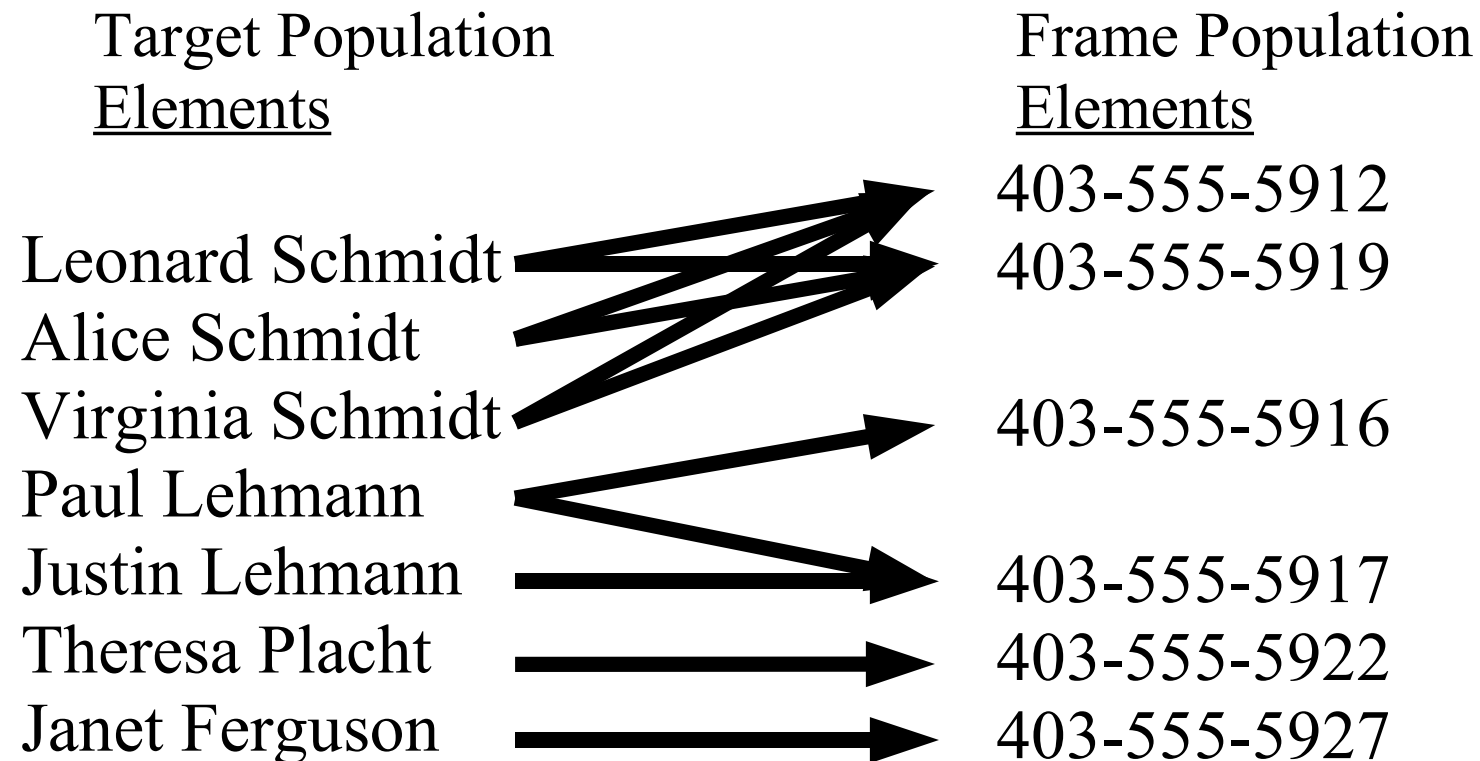
To Further Complicate Things...

- ♦ It is possible to have multiple frame units mapped to multiple population elements.
 - ♦ ratio target element to frame elements = n:m
(many-to-many relationship)



Clustering and Duplication Example

(Figure 3.3. Groves 2004, page 79)



Solution to Clustering and Duplication (1)

- ♦ A common solution to this problem is to weight survey results to handle both problems simultaneously.
- ♦ The compensatory weight for person-level statistics is the number of adults (or eligibles) divided by the number of frame entries for the household.
 - ♦ For the Schmidt household, the weight for the member selected would be:

$$\text{weight} = \frac{\# \text{eligible}}{\# \text{frame entries}} = \frac{3}{2}$$



In Addition...

There is a class of
coverage improvement
procedures!!!



Half-Open Interval

- ♦ Housing unit lists used in household surveys can become out of date and miss units quickly.
- ♦ They may have missed housing units that 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).
- ♦ 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.



Address List Example

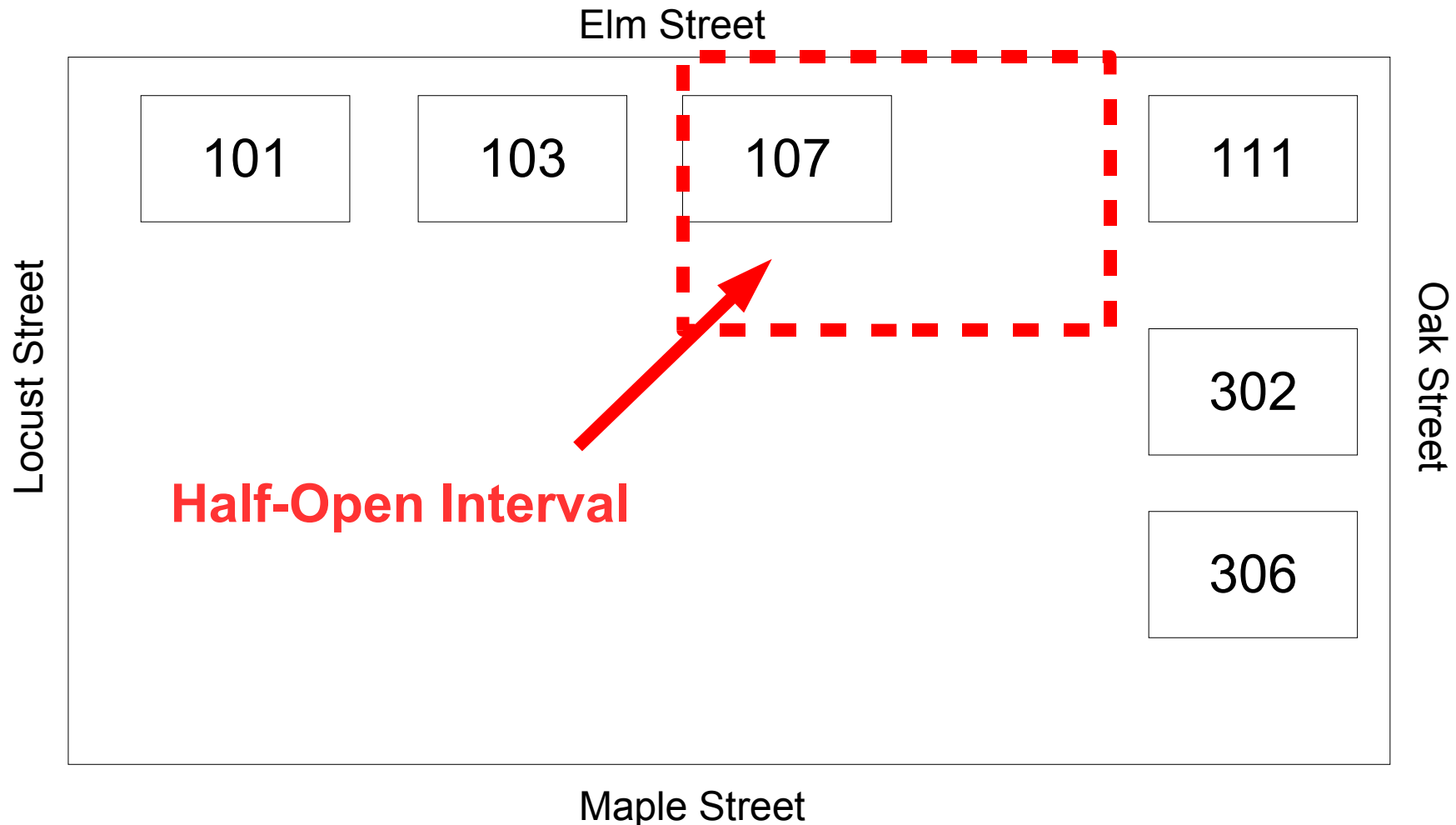
(Figure 3.4. Groves 2004, page 84)

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	
...



Sketch Map for Area Household Survey Block

(Figure 3.5. Groves 2004, page 85)



Open-Interval Example con't

- ♦ The address of 107 Elm Street is viewed as a geographic area bounded by “property lines” (not as a physical structure).
 - ♦ The area 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”).
- ♦ If a new or missed unit is discovered in this interval,
 - ♦ 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.



Multiplicity Sampling

- ♦ Some frame supplementation methods add elements to a population through network sampling (called “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 Example

- ♦ 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.
- ♦ Since network members have multiple chances of being selected, a weight can be applied that decreases the relative contribution of the data from the network to overall estimates.
 - ♦ For example, if a sampled adult identifies two living adult siblings, a weight of $1/3$ can be applied.



Comments on Multiplicity Sampling

- ♦ This “multiplicity” sampling has been used to collect data about networks to increase sample sizes for screening for rare conditions.
- ♦ The method has to be balanced against privacy concerns of individuals in the network.
- ♦ Response error (eg, failing to report 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 the characteristic.



Snowball Sampling

- ♦ “Snowball” sampling is when a sampled person is asked to identify others with the “condition” under study.
 - ♦ For example, we might use snowball sampling in a study of injection drug users.
- ♦ “Snowball” sampling cumulates sample persons by using a network information reported by sample persons.
- ♦ This method assumes that an individual with the “condition” will know others who also have the “condition”.



Limitations of Snowball Sampling

- ♦ Usually a non-probability method to supplement a sampling frame.
- ♦ Errors in reports, isolated individuals who are not connected to any network, and poorly defined networks make snowball sampling difficult in practice.



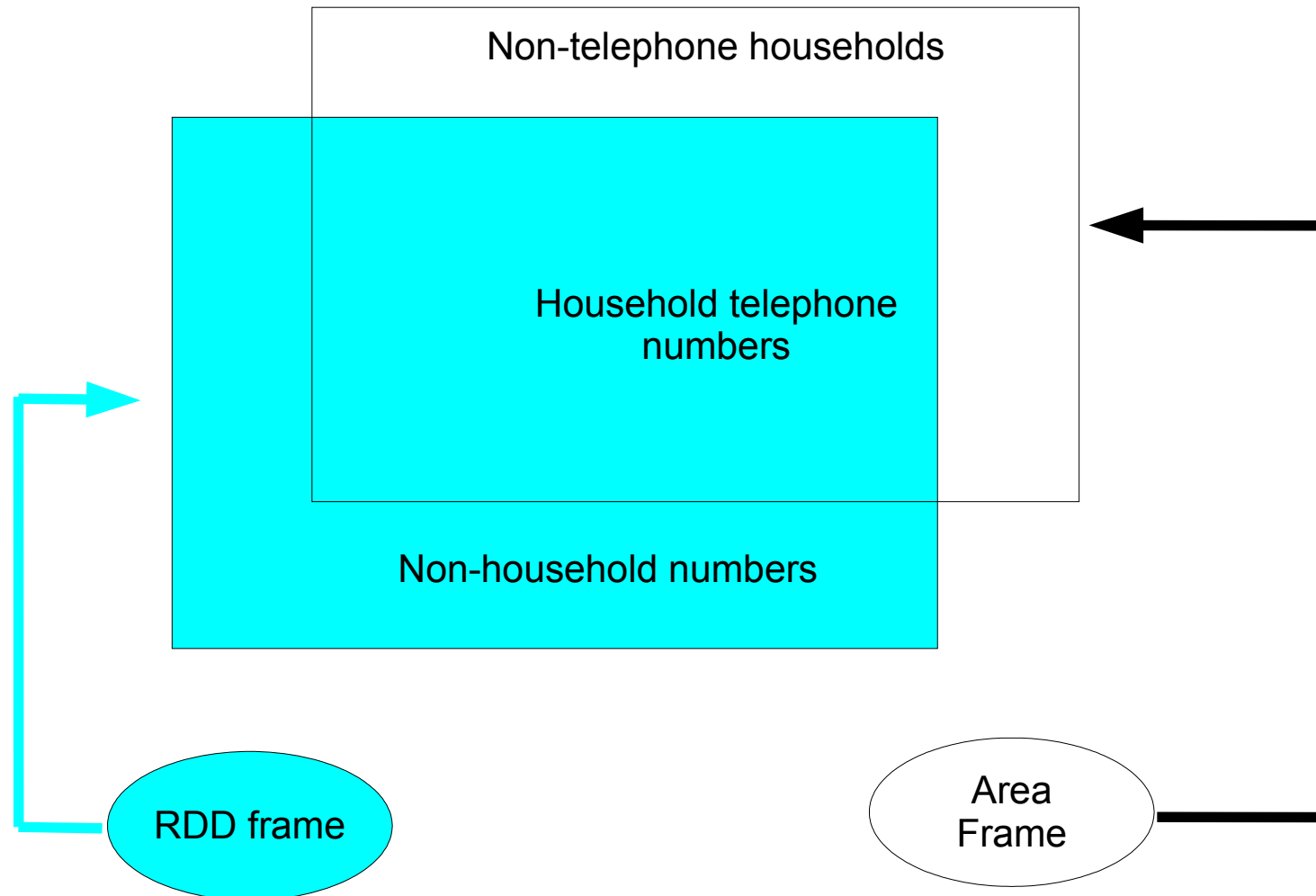
Multiple Frame Designs

- ♦ Coverage error can sometimes be reduced by the use of more than one sampling frame.
- ♦ The use of a second (or more) sampling frame can provide better or unique coverage for population elements absent or poorly covered in the principal frame.
- ♦ Alternatively, the supplemental frame may cover a completely separate portion of the population.
- ♦ Due to possible overlap between frames, multiple frame sampling and estimation procedures can be employed to correct for unequal probabilities of selection.



Dual Frame Sample Design

(Figure 3.6. Groves 2004, page 87)



Problems with Dual Frames

- ♦ The two frames overlap (both contain households with telephones).
- ♦ Thus, households with telephones will be overrepresented under such a design since they can be selected from both frames.
- ♦ Possible solutions:
 - ♦ screen out households with fixed-line telephones (screening)
 - ♦ attempt interviews at all sample households and determine the chance of selection for each household (weighting)
 - ♦ domain analysis (multiple frame estimation)



Summary

- ◆ Problems leading to coverage error:
 - ◆ Ineligible units
 - ◆ Clustering
 - ◆ Duplication
 - ◆ Clustering and duplication
- ◆ Additional methods for reducing undercoverage

