

Inference and Error in Surveys

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>*



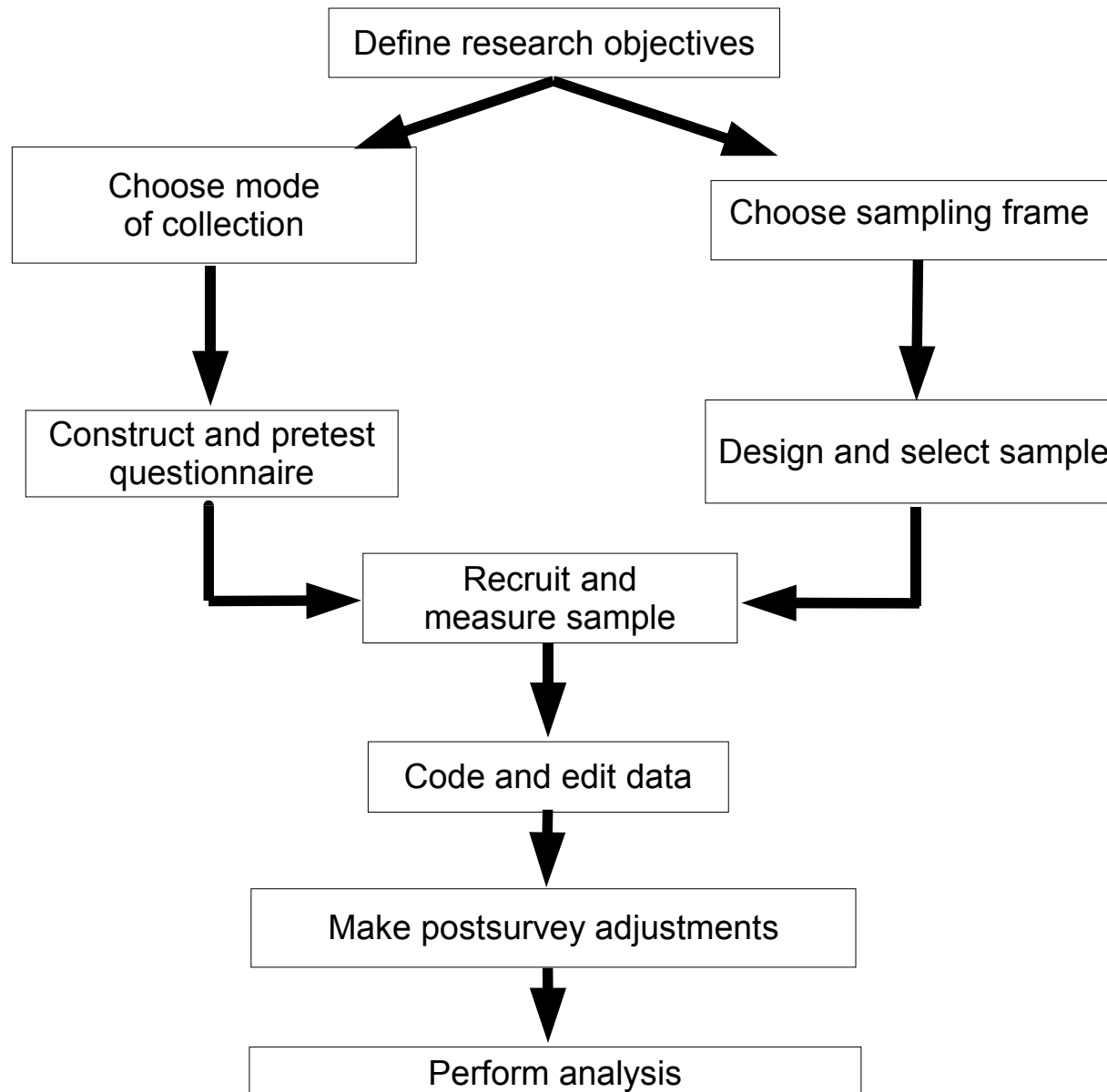
Overview

- ♦ Defining Research Objectives
- ♦ Target Populations and Elements
- ♦ Sampling Frames
- ♦ Coverage Bias



Survey Process Perspective

Groves et al. *Survey Methodology* 2004



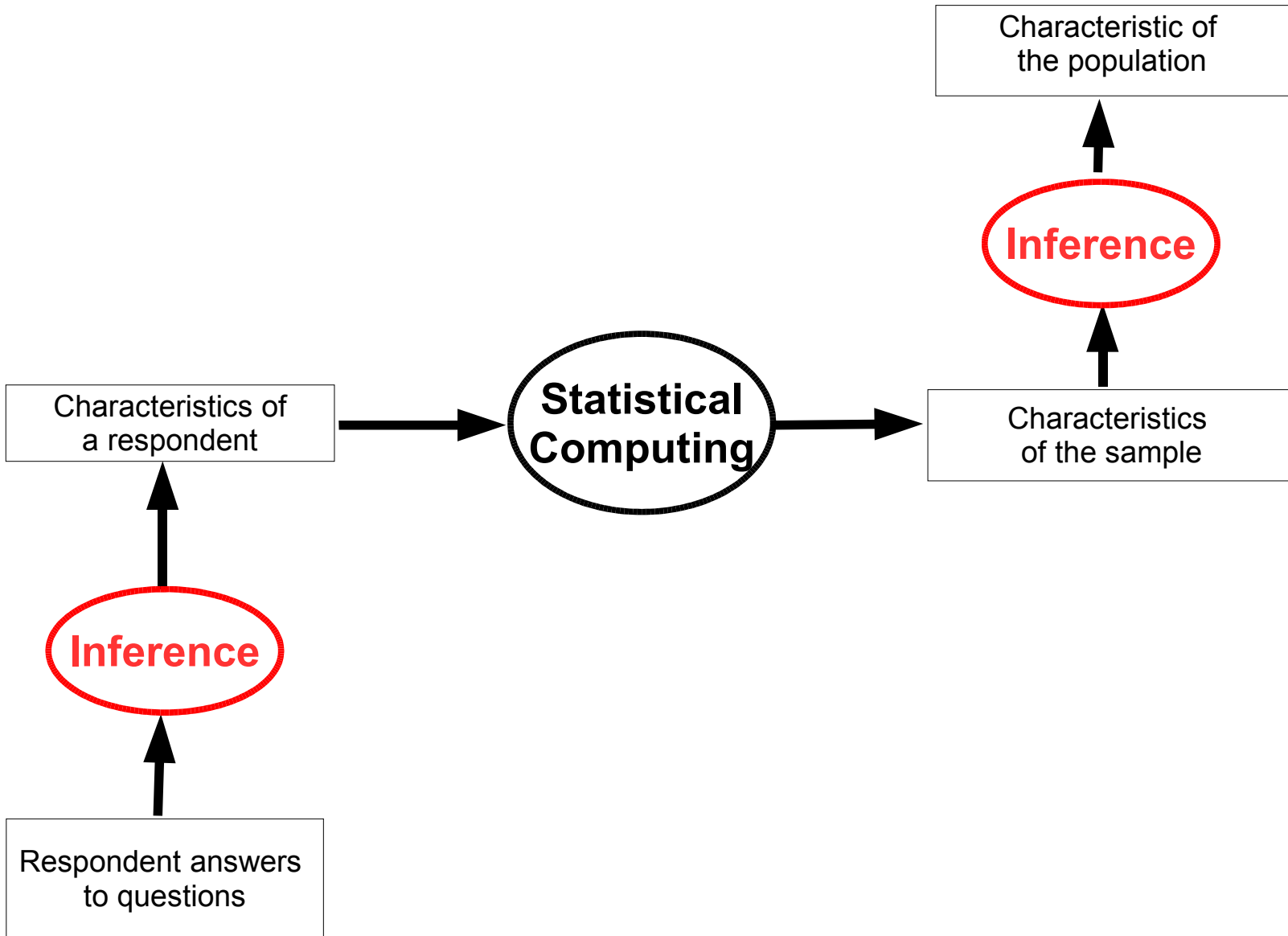
Defining Research Objectives

- ♦ What's the research question?
 - ♦ What are you interested in finding out?
 - ♦ Whom do you want to study?
 - ♦ Where are these people located?
 - ♦ When do you want to do the survey?
 - ♦ What do you expect to learn and why?
- ♦ Deciding who will be the focus of the study is critical for determining the ultimate sampling plan for the survey.



Survey Inference

Groves et al. *Survey Methodology* 2004



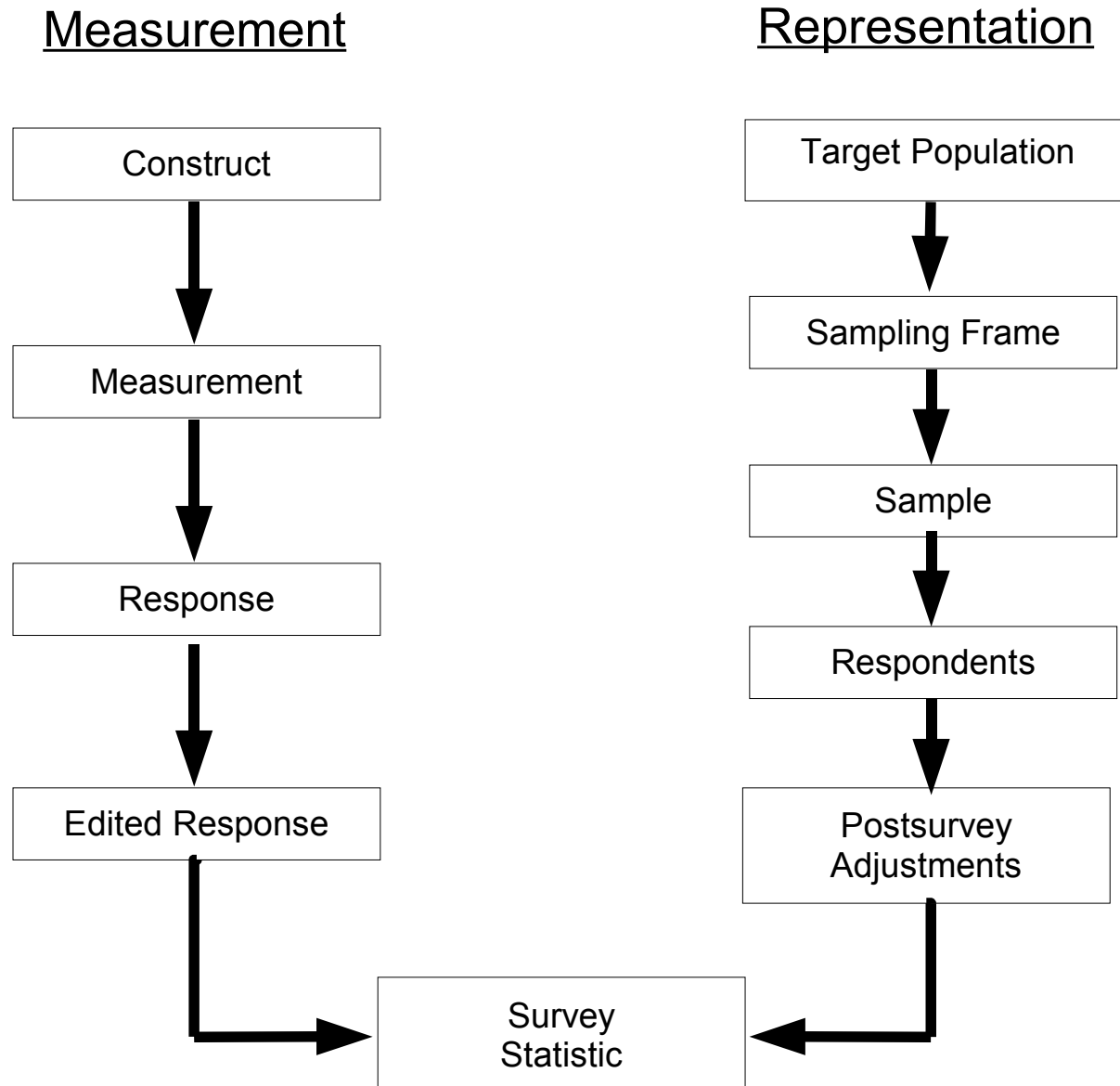
Two Types of Inference

- ♦ Inference in surveys
 - ♦ The formal logic that permits description about unobserved phenomena based on observed phenomena.
- ♦ We use an answer to a question from an individual respondent to infer something about the characteristics of that person.
- ♦ We use characteristics of the study sample to infer something about the characteristics of the larger population from which they are a member.

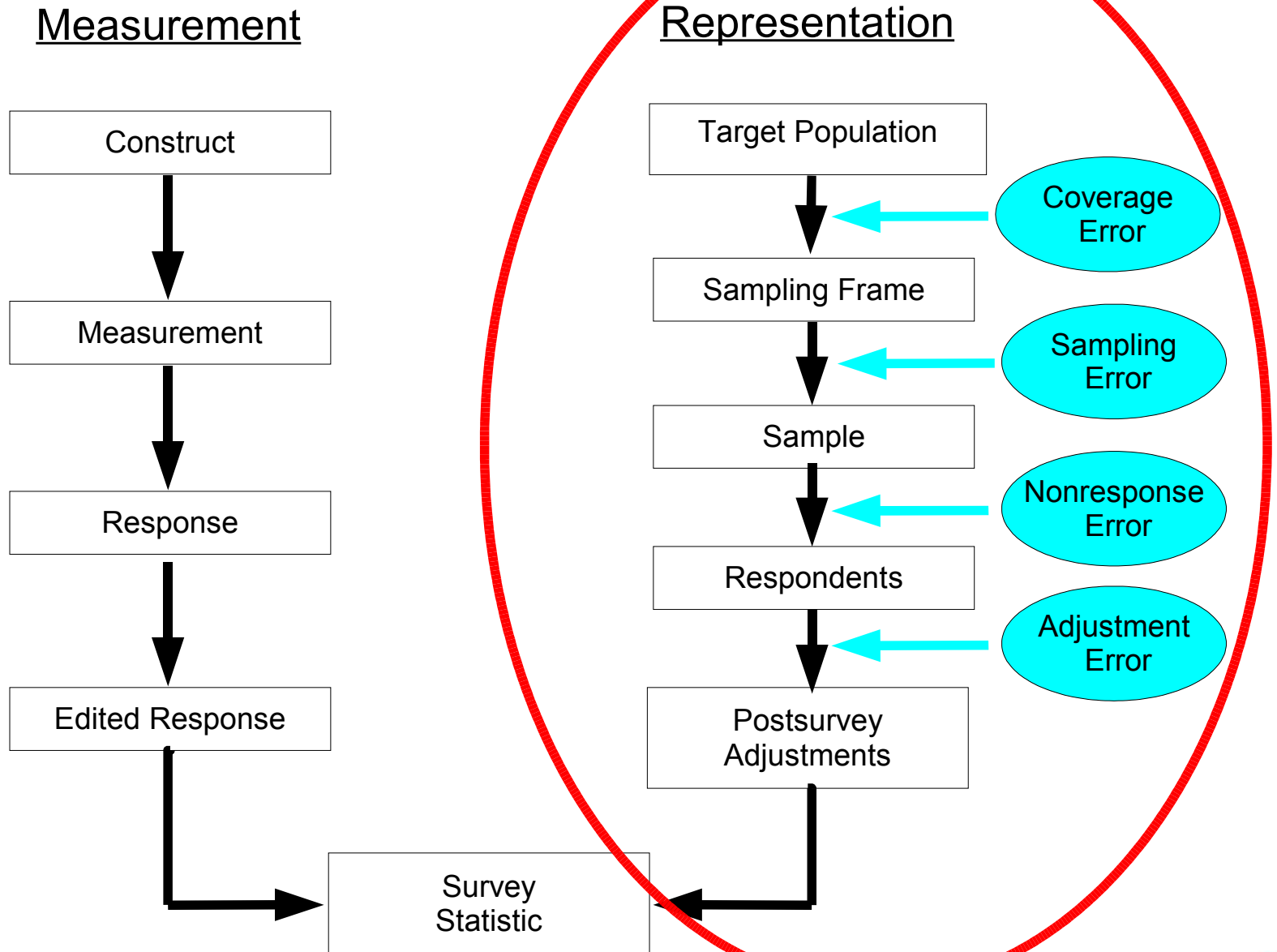


Survey Design Perspective

Groves et al. *Survey Methodology* 2004



Survey Quality Perspective



Target Population and Elements

- ♦ The *target population* for the survey is the entire set of individuals to which findings of the survey are to be extrapolated.
 - ♦ Group to which one wishes to generalize (“make inferences”) from the survey sample.
- ♦ Sample inclusion criteria or exclusion criteria are reflective of the target population for the study.
- ♦ The individual members of the population whose characteristics are to be measured are called *elementary units* or *elements* of the population.



Example

- ♦ Sample survey of hospital medical records to estimate the number of hospital discharges in one year having a specific diagnosis
 - ♦ hospital discharge occurring during the year is an *element*
 - ♦ the totality of such discharges constitutes the *target population*



Classes of Sample Surveys

- ♦ In sample surveys, we would like to make inferences to the target population by obtaining information on a sample from this population.
- ♦ Two broad classes of sample surveys (based on how sample is selected):
 - ♦ non-probability samples
 - ♦ probability samples



Probability Sampling (PS)

- ♦ PS allows one to make inferences about large populations without observing every member.
- ♦ PS avoids selection bias by giving each element a known, nonzero probability of selection.
- ♦ Knowing these probabilities, it is possible to select a subset of the population from which to make estimates about the entire population (with some uncertainty).
- ♦ This course will focus on sampling designs that utilize probability sampling.



Nonprobability Sampling

- ♦ A nonprobability sample is one based on a sampling plan that does not have this feature.
- ♦ There is no firm method of evaluating either the reliability or the validity of the resulting estimates.
 - ♦ Example: quota survey in which interviewers are told to contact and interview a certain number of individuals from certain demographic subgroups.
- ♦ This course will not focus on nonprobability sampling designs.



Sampling Frame

- ♦ To draw a probability sample from a population, it is necessary to have a list or other selection process called a sampling frame.
- ♦ The sampling frame is the list of “elements” of the target population from which the sample will actually be drawn.
 - ♦ “Elements” are the units that you will take measurements on (eg, households, individuals, hospitals, etc.).
 - ♦ The sampling frame is the “operational definition” of the target population.
- ♦ Thus, the quality, completeness, and availability of possible sample frames are major considerations when selecting a study population.

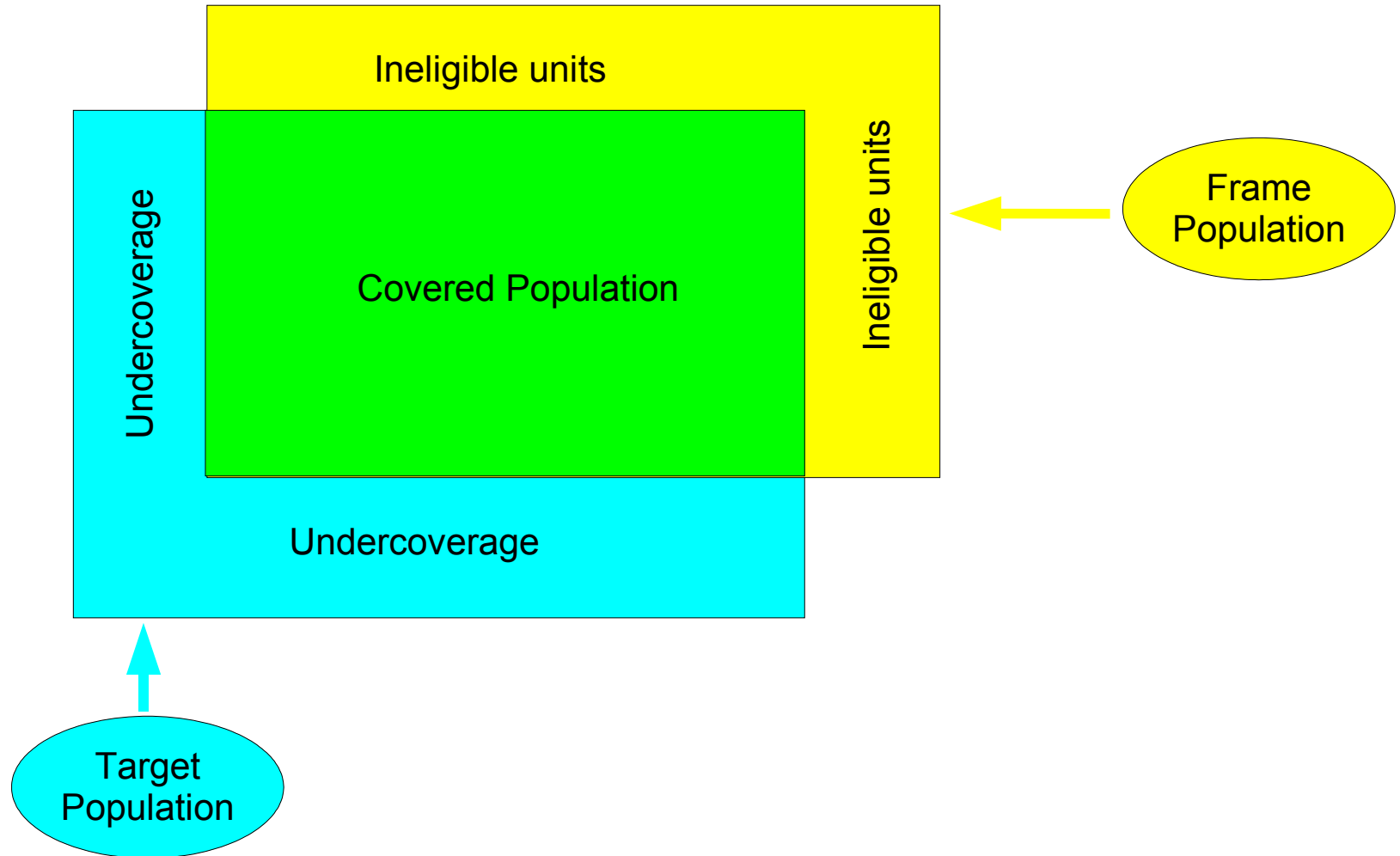


What is the “right” sampling frame?

- ♦ The “ideal” sampling frame is one that provides a one-to-one correspondence between elements in the frame and the target population.



Coverage of a Target Population by a Frame



Examples of Sampling Frames

- ♦ Frames are sometimes defined by:
 - ♦ geographic listings of blocks
 - ♦ maps
 - ♦ directories (telephone or other electronic formats)
 - ♦ membership
 - ♦ etc.



Enumeration Units

- ♦ In sample surveys, it may not be feasible to sample the elementary units directly.
- ♦ Lists of elementary units from which the sample can be taken are often not readily available.
 - ♦ For example, a study of homeless individuals
- ♦ However, elementary units can be associated with other types of units called enumeration units for which lists are either available or can be readily constructed for sampling purposes.
 - ♦ For example, homeless shelters



Example

- ♦ Suppose that a sample survey is being planned to estimate the number of persons living in San Francisco who received the influenza vaccine this year.
 - ♦ What is the target population?
 - ♦ What are the elementary units?
 - ♦ What is a possible enumeration unit?



Specify a Sampling Frame

- ♦ Target population: incarcerated population in California
- ♦ How would you develop a sampling frame for this target population?



Specify a Sampling Frame

- ♦ Target population: injection drug users living in Alameda County
- ♦ How would you develop a sampling frame for this target population?



Specify a Sampling Frame

- ♦ Target population: individuals who have been exposed to recreational water in California
- ♦ How would you develop a sampling frame for this target population?



Specify a Sampling Frame

- ♦ Target population: general US population
- ♦ How would you develop a sampling frame for this target population?



Issues to Consider

- ♦ In conducting a survey of the US population, there may be:
 - ♦ subgroups that are difficult (or impossible) to interview
 - ♦ subgroups that differ from general in ways that make them “nonrepresentative” of the population
 - ♦ subgroups that are housed in ways that make them unavailable for interview

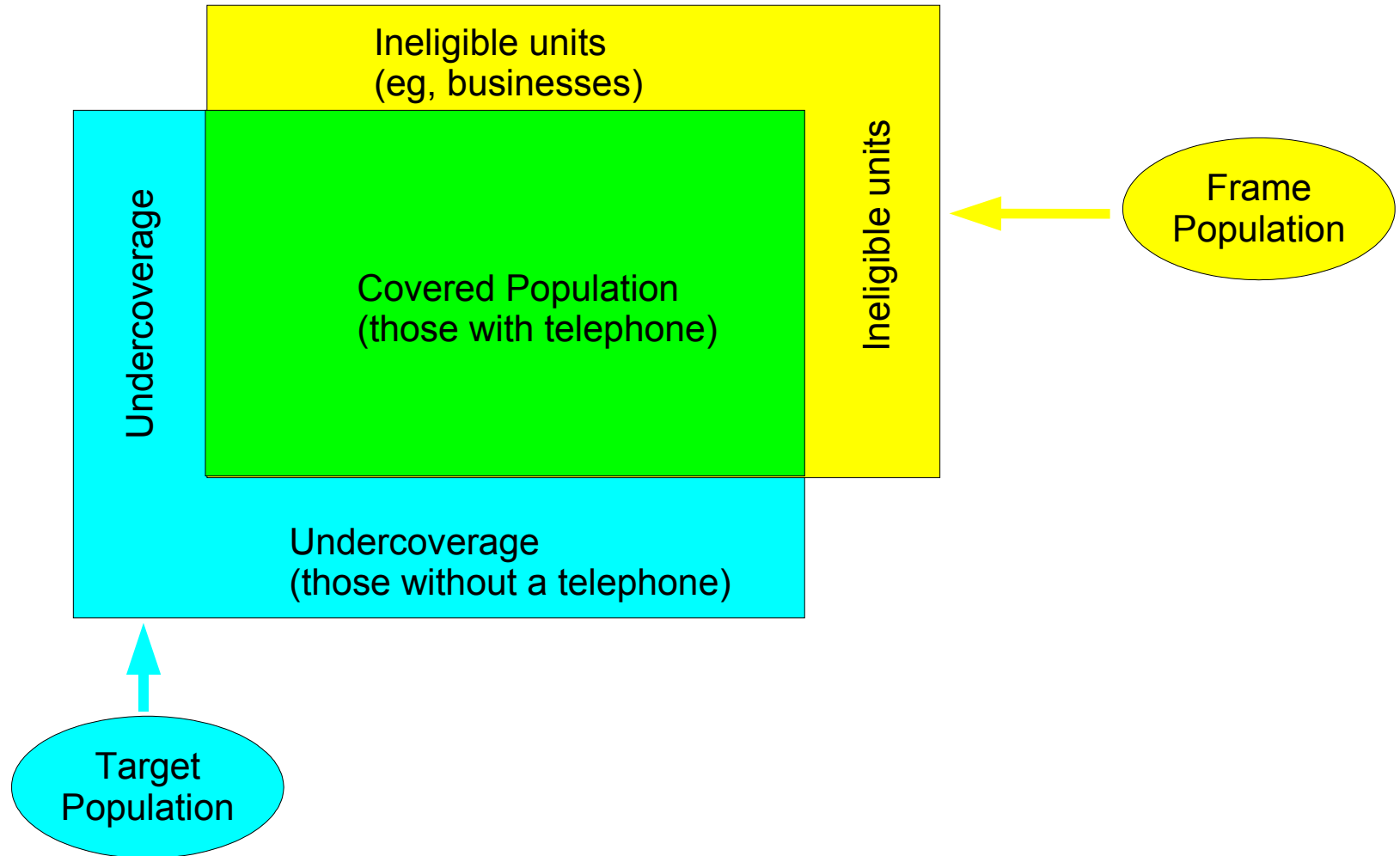


Sampling Frames for Residents

- ♦ In the United States, there is no updated list of residents.
- ♦ However, Sweden does have a population register, an updated list of names and addresses of almost all residents.
- ♦ Sample surveys of the target population of all US residents often use sampling frames of telephone numbers.



Coverage of a Target Population by a Frame



Coverage Bias

- ♦ The error: what proportion of US residents can be reached by telephone AND how different are they from others *on the statistics in question?*
- ♦ Example:
 - ♦ Persons with lower incomes and in rural areas are less likely to have telephones in their homes.
 - ♦ If the survey statistic was the percentage of persons receiving unemployment compensation, it is likely that a telephone survey would underestimate the percentage.
 - ♦ There would be a coverage bias in this statistic.



Coverage Error Defined

- ♦ Coverage error: the difference between statistics based on the population defined by the frame and statistics based on the target population.
- ♦ Coverage error arises when there is a lack of correspondence between the frame and the target population.



Sources of Variation

- ♦ Two sources of variation between the population of inference and the sample frame:
 - ♦ how the eligible population is defined
 - ♦ flaws in the actual listings of the eligible population that comprise the frame.



Purpose of NHANES

Source: <http://www.cdc.gov/nchs/nhanes.htm>

- ◆ Purpose:

- (1) To estimate the number and percent of persons in the U.S. population and designated subgroups with selected disease and risk factors;
- (2) To monitor trends in the prevalence, awareness, treatment and control of selected diseases;
- (3) To monitor the trends in risk behaviors and environmental exposures;
- (4) To analyze risk factors for selected diseases;
- (5) To study the relationship between diet, nutrition and health;
- (6) To explore emerging public health issues and new technologies.

- ◆ Target Population: civilian, noninstitutionalized US population



“Solution” One

Redefine your target population!
(not sure this is a good idea)



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.



More on these next time!



Summary

- ♦ Defining Research Objectives
- ♦ Target Populations and Elements
- ♦ Sampling Frames
- ♦ Coverage Bias

