



# Sampling Design and Sampling Error

Sona R. Saha, MPH

Division of Epidemiology

UC Berkeley School of Public Health

Email: [ssaha@berkeley.edu](mailto:ssaha@berkeley.edu)

*Center for Infectious Disease Preparedness  
UC Berkeley School of Public Health*



# Sampling Design Overview



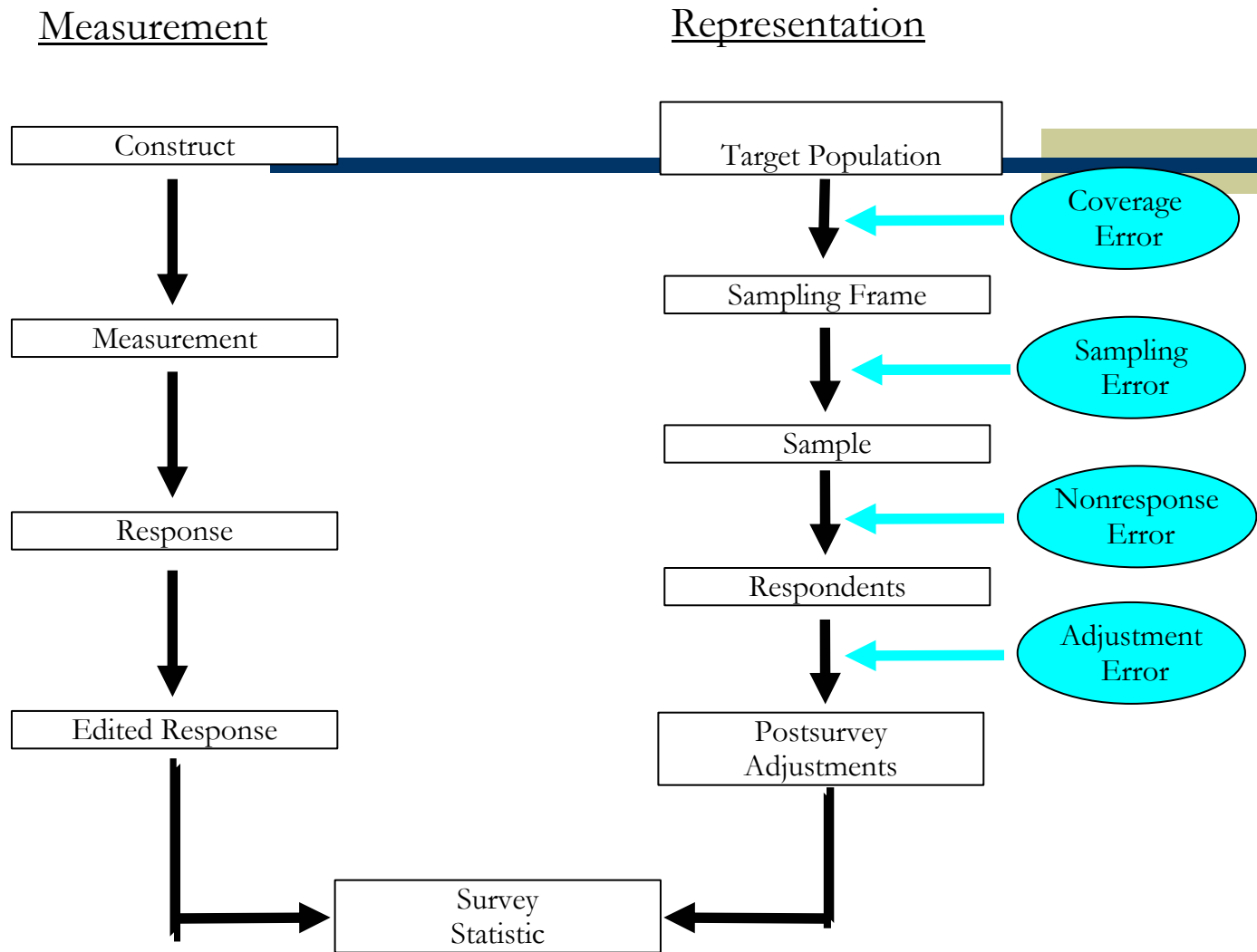
- ◆ Sample Selection Implications
- ◆ Sampling Designs
  - Simple Random Sampling
  - Cluster Sampling
  - Stratified Sampling
  - Systematic Selection
- ◆ Field Applications



# Sample Selection

- ◆ How a sample is constructed can affect:
  - Quality of data
  - Ability to make valid inference to population of interest
- ◆ Be aware of potential selection biases that may influence composition of sample
  - Distribution of age, gender, geography, etc.
  - Self selection, healthy responders, etc.

# Survey Quality Perspective



# Sonoma Water Evaluation Trial

UC Berkeley School of Public Health  
(PI: Jack Colford, MD PhD)

- ◆ Randomized Controlled Trial of Tap Water Treatment
  - Estimate rate of gastrointestinal illness (GI) among elderly individuals over 55 in Sonoma County
  - Estimate the attributable risk of GI illness due to drinking water in study sample
  - Daily symptom collection for 1 year
- ◆ RDD Cross-sectional Survey (based on FoodNet)
  - Estimate rate of GI illness among elderly individuals over 55 in Sonoma County
  - Telephone survey asking symptoms in past 1 month
  - *Background rate of illness to compare for the trial*
  - Estimate % tap water drinkers in Sonoma County
  - *Informs PAR of AR from RCT*

# SWET Participant Selection

## ◆ RCT


- Target Population- elderly individuals in US who consume municipal tap water
- Sampling Frame
  - Existing NIH cohort for Aging study in Sonoma, CA
  - Retirement community lists
  - Sonoma County population telephone lists
- Sample
  - Purged for households on wells
  - Geographically staged
  - Convenience sample

## ◆ RDD Survey

- Target Population- elderly individuals in US
- Sampling Frame
  - Households in Sonoma County with at least one individual over 55 years of age identified through population telephone list
- Sample
  - List assisted random digit dial, random selection
  - Weighted by zip code to match geographic coverage of trial

# SWET Demographics

<i>Variable</i>	<i>RCT Total N (%)</i>	<i>RDD Survey N (%)</i>	<i>Community Demographic s (%)</i>
Gender			
Male	308 (43.3)	566 (34.9)	41.2
Female	403 (56.7)	1054 (65.0)	58.8
Ethnicity			
White (non-Hispanic)	682(95.9)	1543 (95.25)	74.5
Native American	4(0.6)	5 (0.31)	1.2
Hispanic	12(1.7)	26 (1.06)	17.3
African American	1(0.1)	9 (0.56)	1.4
Asian	7(1.0)	13 (1.12)	3.1
Other	5(0.7)	24 (1.48)	2.5



---

# Basic Features of Sample Selection

---

- ◆ A list, or combinations of lists, of elements in the population (Sampling frame)
- ◆ Chance or random selection from list(s) to avoid to minimize known and unknown biases
- ◆ Some mechanism that assures key subgroups of the population are represented



# Probability Sample

- ◆ Probability samples rely on some random procedure applied to the sampling frame
  - (or list of elements- eg. phone numbers, addresses..)
  - Assign non-zero chances of selection to each element
- ◆ Selection probabilities need not be equal for all elements depending on your research goal
  - May want to over-sample hard to reach subgroups (eg. over 75, youth, homeless,etc.)



# Sources of Sampling Error

- ◆ As the entire sampling frame is not measured in any sample, errors will be present
  - Bias- systematic failure to observe some elements due to sample design
  - Variance- each sample taken from the element pool or frame can be different
    - Precision- low variance of estimate (survey statistic), high precision



# Sampling Error Function of Sample Design

---

---

- ◆ How large a sample is selected
- ◆ How the selection probability varies for each element in the sampling frame
- ◆ Whether elements are drawn independently or in groups (clusters)
- ◆ Whether the sample is designed to control the representation of key subgroups (stratified to ensure adequate representation)



---

# Simple Random Sampling

---

- ◆ SRS assigns equal probability of selection to element in sampling frame
- ◆ Random numbers can be applied directly to lists
- ◆ Sample without replacement, same element cannot be selected twice

# Cluster Sampling

- ◆ Cost of simple random sample can be prohibitive
- ◆ Cluster samples select groups of elements jointly
  - Eg. by city blocks, zip codes, or other parameter of interest
  - Homogeneity within cluster
    - Evaluate within cluster homogeneity (correlated data)
    - Evaluate between cluster homogeneity (rate of homogeneity,  $\rho_h$ )
  - Increased sampling variance relative to element samples (SRS)



# Classroom Example



- ◆ Obtain list of 10,000 4<sup>th</sup> grade classrooms in the US
- ◆ Each class has approx. 25 students
- ◆ Take SRS of 100 classrooms
- ◆ Interview all students in class (cluster),  $N=2500$
- ◆ Students within each cluster (class) may be more alike than students in different classes across country



# Stratified Sampling



- ◆ Stratification techniques can be used to assure sufficient representation of population subgroups
- ◆ Each element in the frame can be sorted by some variable into groups or “strata”
- ◆ Each strata is sampled separately either by SRS or some other procedure (SRS in some strata, clusters within strata, etc.)



# Systematic Selection

- ◆ Devise some procedure to select every  $k$ th element from the list.
- ◆ Simplified way of stratified sampling aiming to ensure equity across strata
- ◆ Determine population size relative to sample size, compute  $k$  as ratio of population to sample size
- ◆ Sort population into strata, choose random number from 1 to  $k$ , and every  $k$ th element thereafter.



# Telephone Surveys : FoodNet



- ◆ The Centers for Disease Control and Prevention's (CDC) Foodborne Diseases Active Surveillance Network (FoodNet) Population Study is designed to precisely estimate acute diarrheal illness in the United States, and the frequency of important exposures.
- ◆ FoodNet population survey data are useful in determining the prevalence and severity of self-reported diarrheal illness, common symptoms associated with diarrhea, and the proportion of persons with diarrhea who seek care.
- ◆ Exposures that might be risk factors for foodborne illness, such as the consumption of potentially “risky” foods or recent travel out of the United States, are included as questions on the survey instrument to be asked in conjunction with illness questions.



# FoodNet Sample Design 1



- ◆ FoodNet provides for a disproportionate random sample of telephone-equipped households with approximately 16,000 interviews during the 12-month interviewing period for nine states (CA, CO, CT, GA, MD, MN, NY, OR, TN).
- ◆ The sample design for this survey specified a list-assisted, random digit dial (RDD) sample of telephone-equipped households in the nine states.
- ◆ The list-assisted RDD procedure ensures that households with telephone numbers that have been assigned since the publication of the current directories, as well as households with deliberately unlisted numbers, are sampled in their correct proportions.

# FoodNet Sample Design 2

- ◆ List-assisted RDD samples are generated by first preparing and maintaining an up-to-date list of all current operating telephone exchanges (three-digit prefixes) in the area codes of each of the eight states.
- ◆ These telephone exchanges, when combined with all four-digit numbers from 0000 to 9999, constitute the set of all possible working telephone numbers, both residential and non-residential.
- ◆ This set of all possible telephone numbers is then arranged in ascending order by exchange and suffix, and divided into blocks of 100 numbers each.
- ◆ Cross-reference directories were utilized to determine which of these blocks contained at least one listed residential number (active blocks). The active blocks were then combined to create the sampling frame, and numbers were systematically sampled from this frame.
- ◆ Weighting is used to compensate for differential selection probabilities at household and respondent levels, and to ensure for distribution of the weighted sample with the population distribution of key demographic variables.



# Summary



- ◆ Samples vary by the following key features of sample design:
  - The number of selected units on which stats are computed (sample size, higher  $N$ , lower variance)
  - Use of stratification which sorts the frame into separate groups and samples them independently (usually decreases sample variance)
  - Use of clustering, which samples groups of elements into the sample at the same time (usually increases variance, creates within cluster homogeneity)
  - Assignment of selection probabilities

# Survey Process Perspective

Groves et al. *Survey Methodology* 2004

