INSTITUTE FOR SOCIAL RESEARCH • SURVEY RESEARCH CENTER
SURVEY RESEARCH OPERATIONS
UNIVERSITY OF MICHIGAN

Lunch & Learn
Statistical Concepts & Terminology I
Probability Sampling
(Part I)

Raphael Nishimura & Paul Burton

Design, Methodology & Statistical Support

Outline – Part I

- Populations & Sampling frames
- Probability sampling
  - Sampling distribution
  - EPSEM vs non-EPSEM
  - Domain vs Sub-class
- Sampling Techniques
  - Simple Random Sampling
  - Systematic Sampling
  - Stratification
  - Clustering
  - Strata vs Cluster
  - Multi-stage Sampling
- Design Effect & Effective sample size

1) Define Target Population

2) Determine Sampling Frame

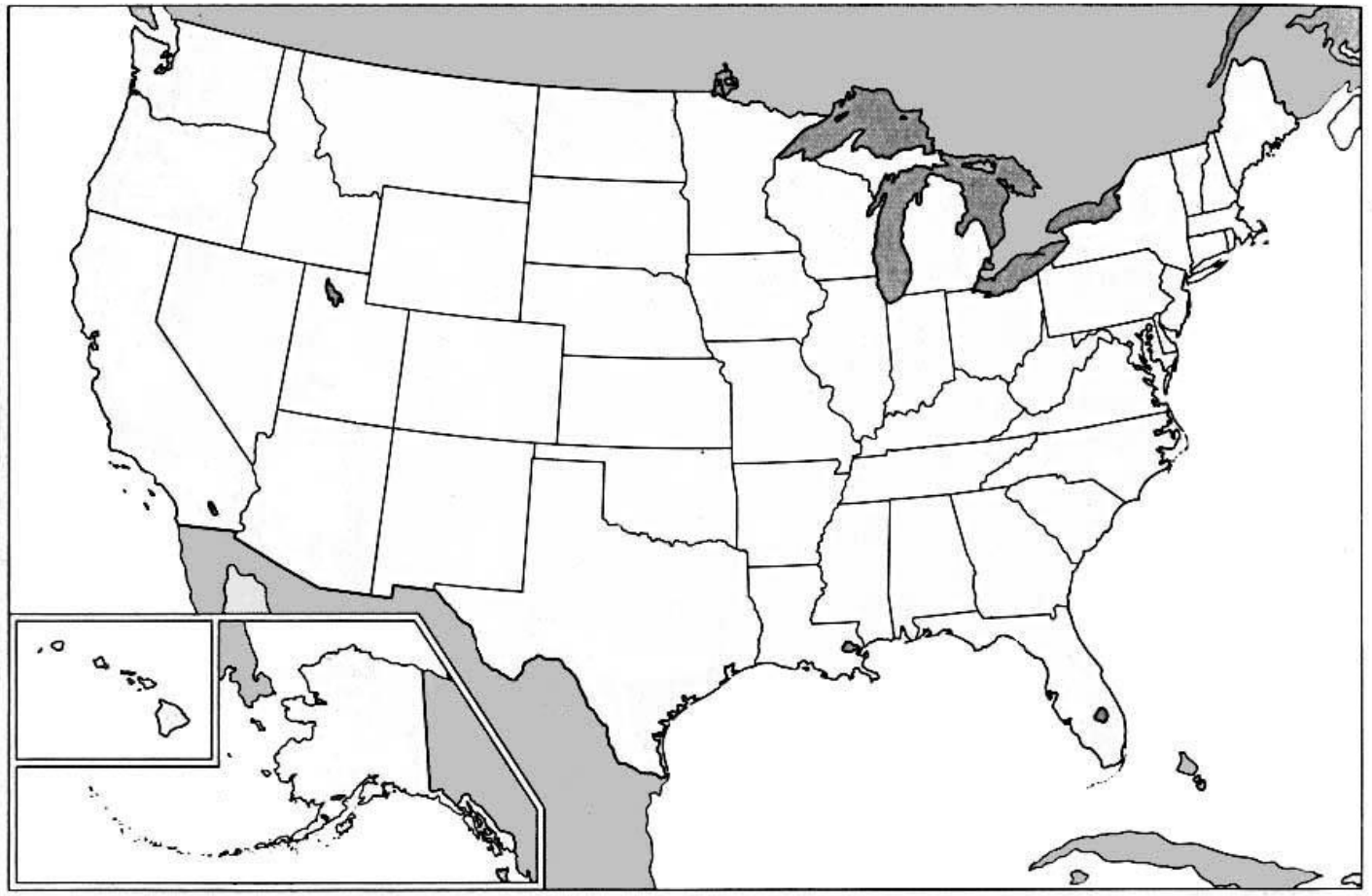
3) Select Sampling Technique

4) Execute Sampling Plan

5) Post-Survey Adjustments

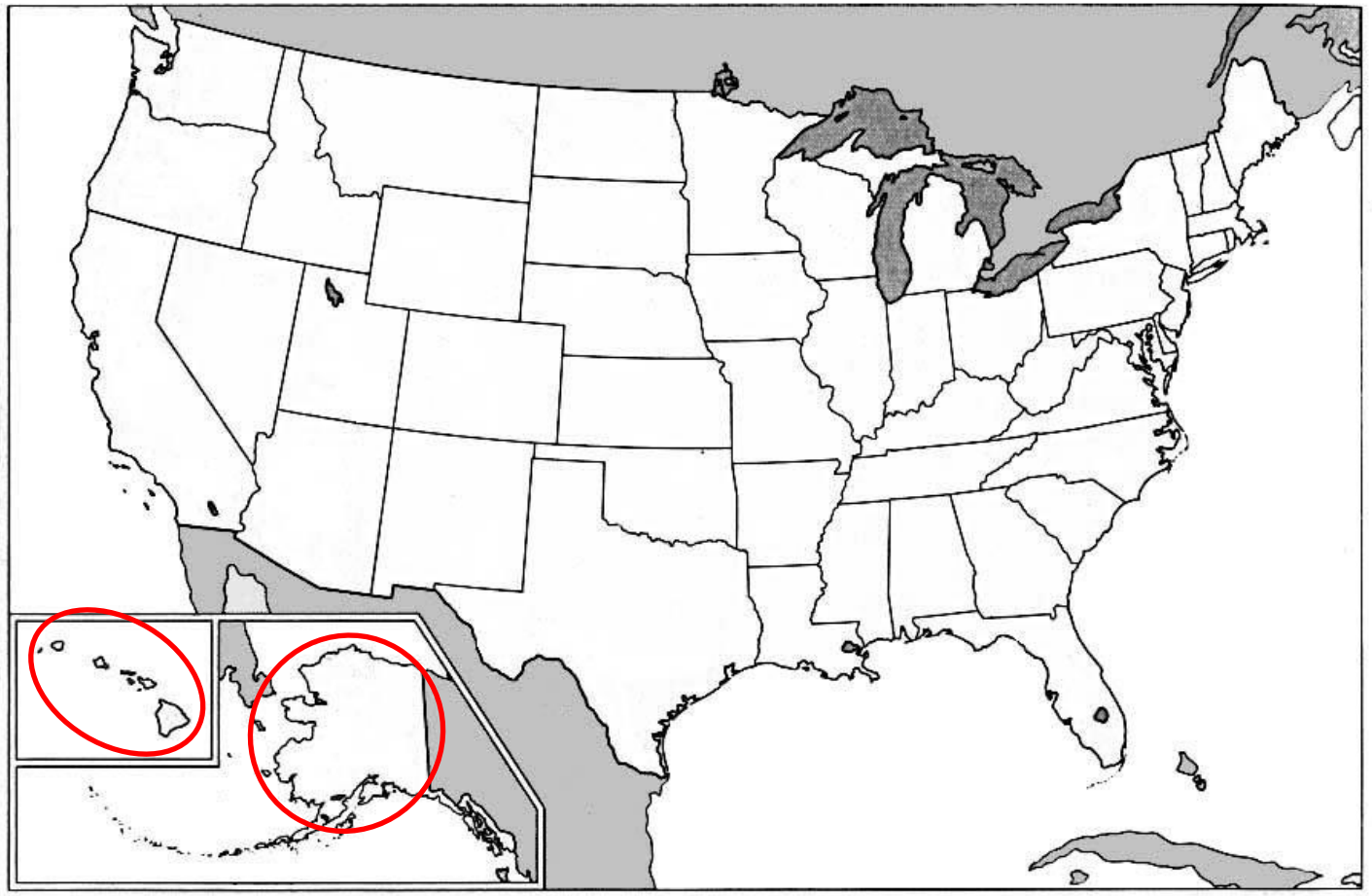
Target Population:

- Finite set of elements for which the survey sponsor wants to make inferences about
- Define in terms of: **Space** (Geography), **Time** (Date or Period) and **Units** (Households, Persons, etc.)



Survey Population:

- The actual population from which the survey data are collected, **given the restrictions from data collection operations**

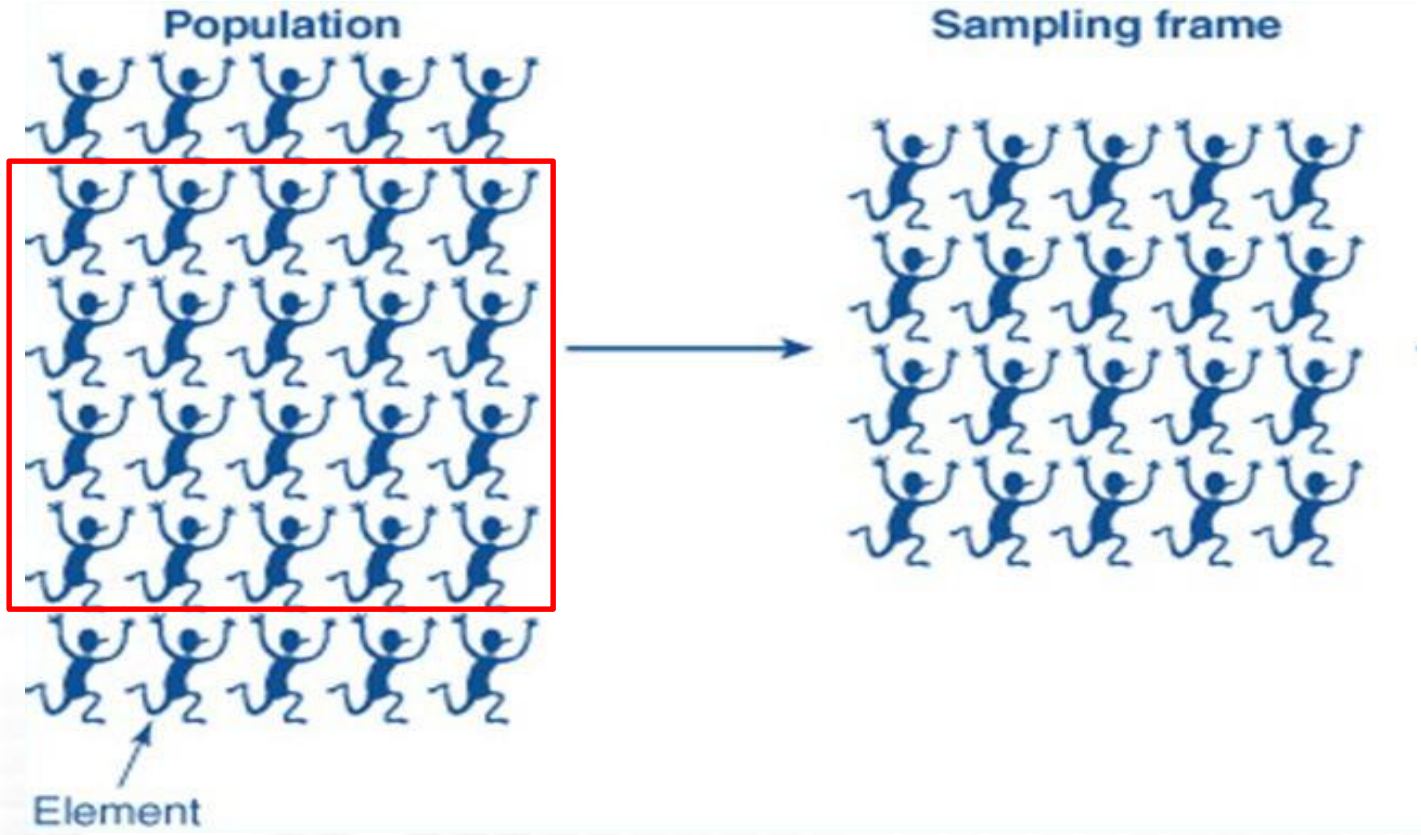


Sampling Frame

- Set of materials used to identify all elements of a survey population to select a sample
- This set of materials can include **lists**, **maps** of areas or **procedures**

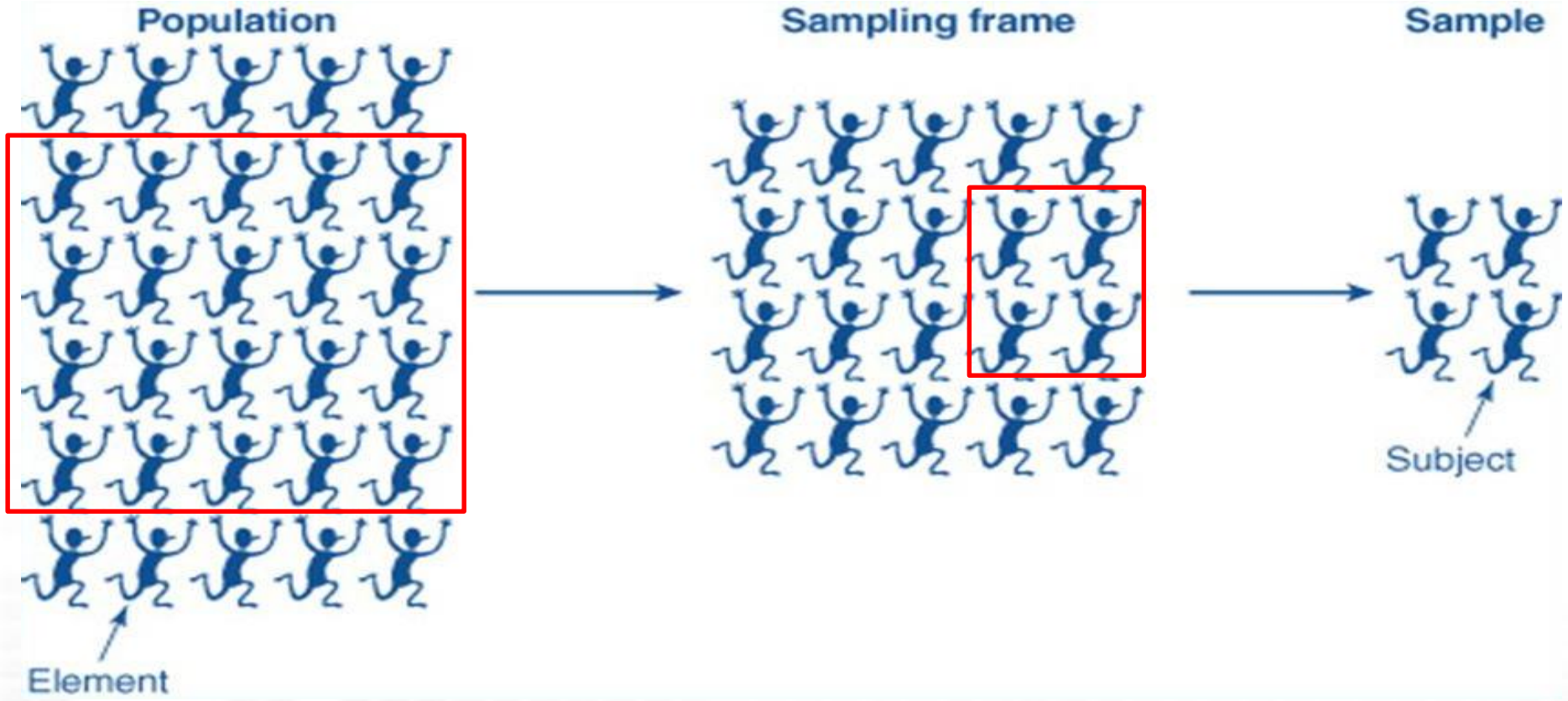
Sampling Frame

- Set of materials used to identify all elements of a survey population to select a sample
- This set of materials can include **lists**, **maps** of areas or **procedures**



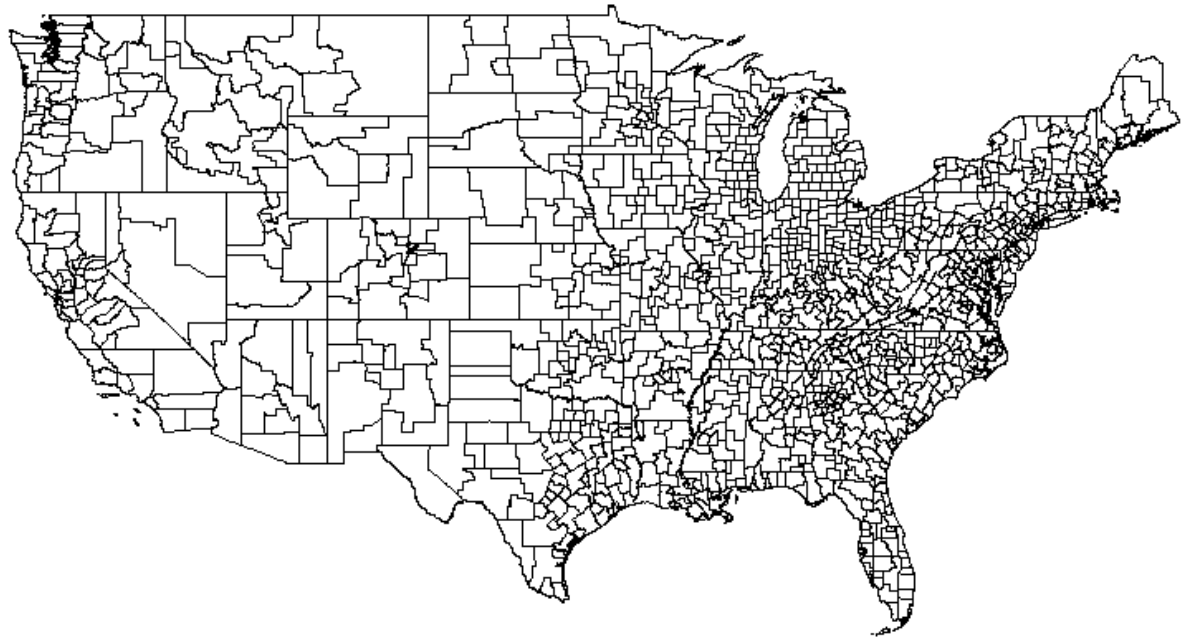
Sampling Frame

- Set of materials used to identify all elements of a survey population to select a sample
- This set of materials can include **lists**, **maps** of areas or **procedures**

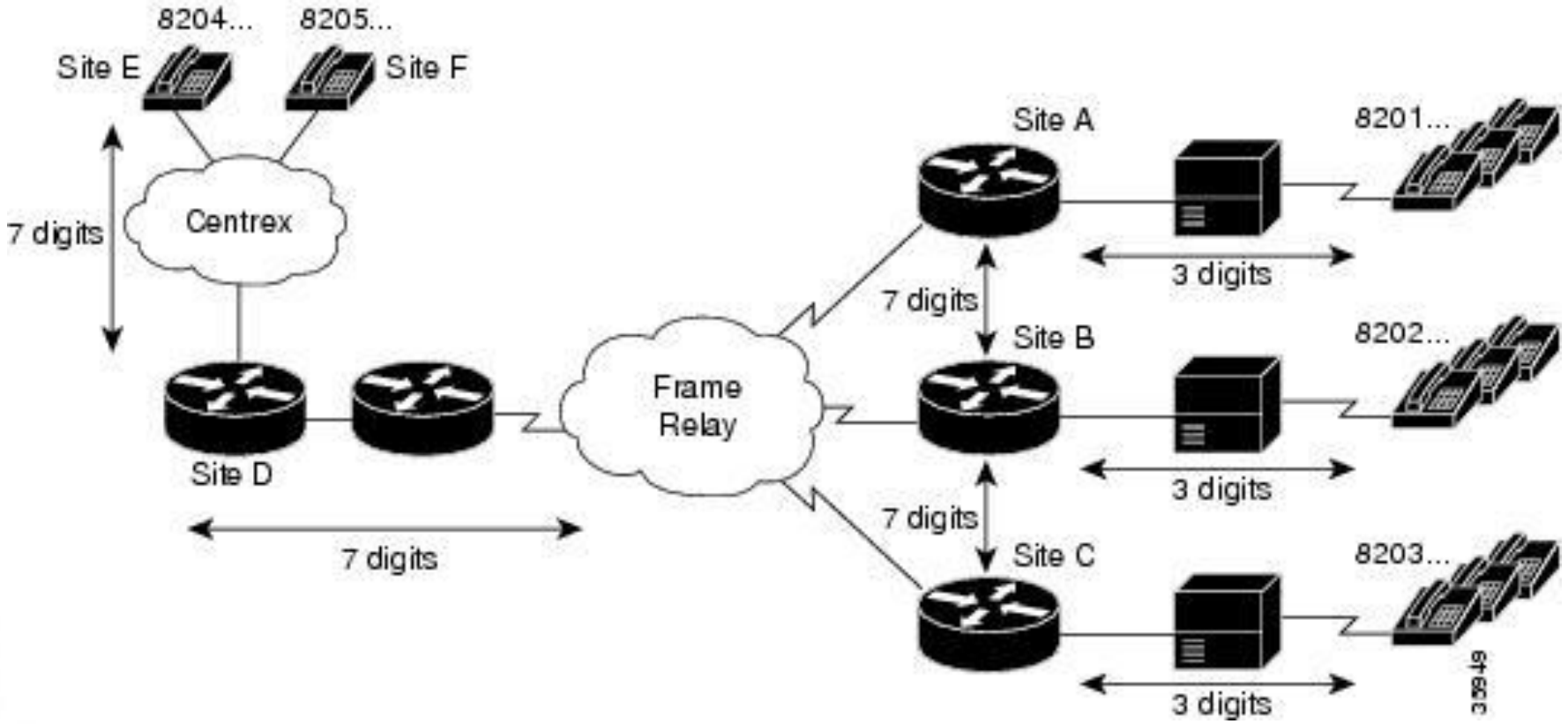


Sampling Frame: ANES 2024 Face-to-Face sample

	PSUID	GEOID	NAME	Total_Housing_Units	Total_Occupi		
599	C55089	55089	Ozaukee County, Wisconsin	39086			
600	C55097	55097	Portage County, Wisconsin	31148			
601	C55101	55101	Racine County, Wisconsin	84490			
602	C55105	55105	Rock County, Wisconsin	70068			
603	C55117	55117	Sheboygan County, Wisconsin	52303			
604	C55127	55127	Walworth County, Wisconsin	53146			
605	C55131	55131	Washington County, Wisconsin	58311			
606	C55133	55133	Waukesha County, Wisconsin	172177			
607	C55139	55139	Winnebago County, Wisconsin	76046			
608	C55141	55141	Wood County, Wisconsin	34549			
609	C56025	56025	Natrona County, Wyoming	36808			
610	G01001	01001	Autauga County, Alabama	24350			
611	G01001	01047	Dallas County, Alabama	18880			
612	G01001	01085	Lowndes County, Alabama	4779			
613	G01002	01005	Barbour County, Alabama	11618			
614	G01002	01011	Bullock County, Alabama	4516			
615	G01002	01067	Henry County, Alabama	9058			
616	G01002	01109	Pike County, Alabama	15977			
617	G01003	01007	Bibb County, Alabama	9002	7927	22293	
618	G01003	01021	Chilton County, Alabama	19438	17302	45014	
619	G01003	01063	Greene County, Alabama	4205	3388	7730	
620	G01003	01065	Hale County, Alabama	7399	6138	14785	
621	G01003	01105	Perry County, Alabama	3933	3360	8511	



Sampling Frame: Random-Digit Dialing

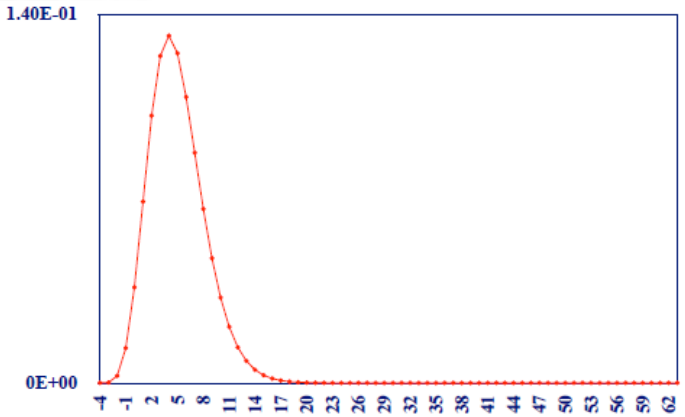


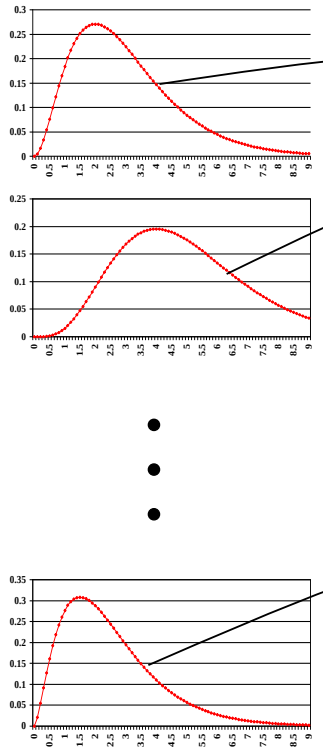
Probability Sampling

- Every element in the population has a ***known*** and ***nonzero*** probability of being selected in the sample
- Involves a randomization device to select samples
- Theoretical framework: repeated realizations of the sample selection
  - *All possible samples* that can be selected from the population with a given sample design
- Statistical inference based on the sampling distribution due to the randomization used on the sample selection
- Sampling Bias = 0 or have data to adjust for “sampling bias”
- Sampling variance (standard error) can be estimated using sampled data

Sampling Distribution

Population distribution



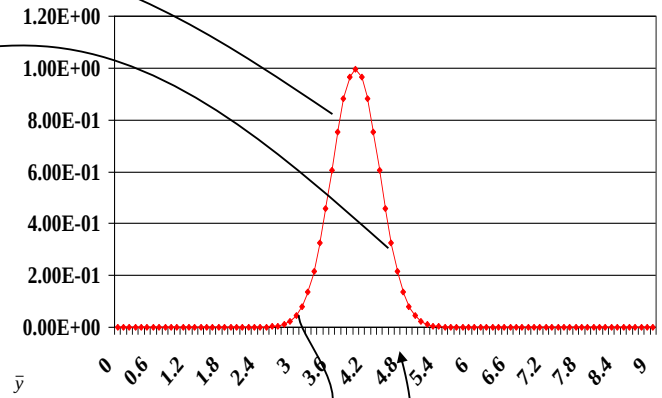


Sample realization 1

Sample realization 2

Sample realization K

Sampling Distribution of Sample Means of Individual Samples



(Sampling Variance called V($\bar{y}$))

(Variance of each sample realization called s^2)

Distributions of *y* Variable from Sample Realizations Samples and the Sampling Distribution of the Mean

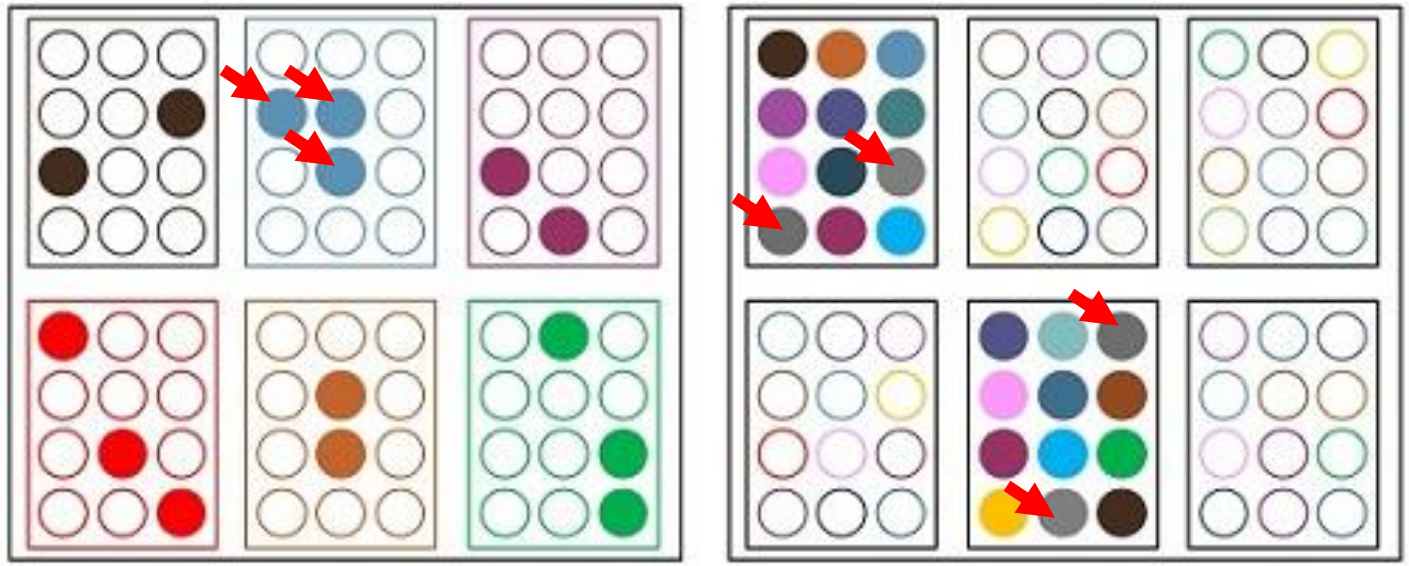
EPSEM vs non-EPSEM

- Equal Probability of Selection Method
  - Most SRO-projects do not use EPSEM designs
- Reasons for not using an EPSEM sample:
  - Within-household selection
  - Probability Proportional to Size (PPS) selection
  - Over-sampling certain subpopulations to obtain more precise estimates for those subgroups
    - HRS: Over-sample Hispanics and African Americans
    - AFHS: Over-sample younger females

Subgroups: Domains vs Subclasses

- Surveys are seldom designed to yield information only for the total population
- **Domains** are subpopulations within the target population for which separate estimates are prepared
  - A subclass is thus a portion of the sample for which inferences are to be made to the totality of subclass elements in the population
- **Subclasses** are subpopulations that the sample is not initially designed to obtained separate estimates
  - However, depending on their (observed) sample size, separate estimates can still be produced for these subgroups

Subgroups: Domains vs Subclasses

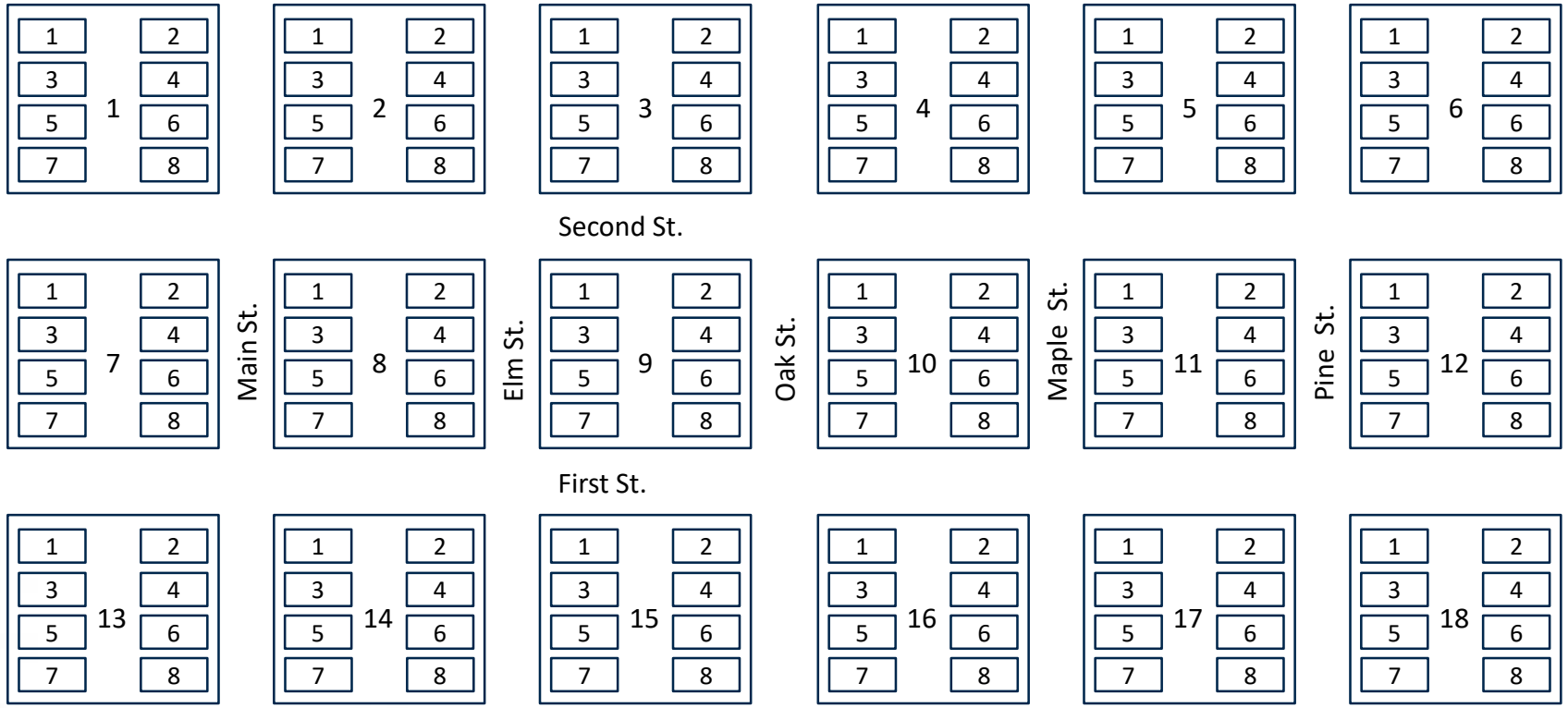


Domains Subclasses

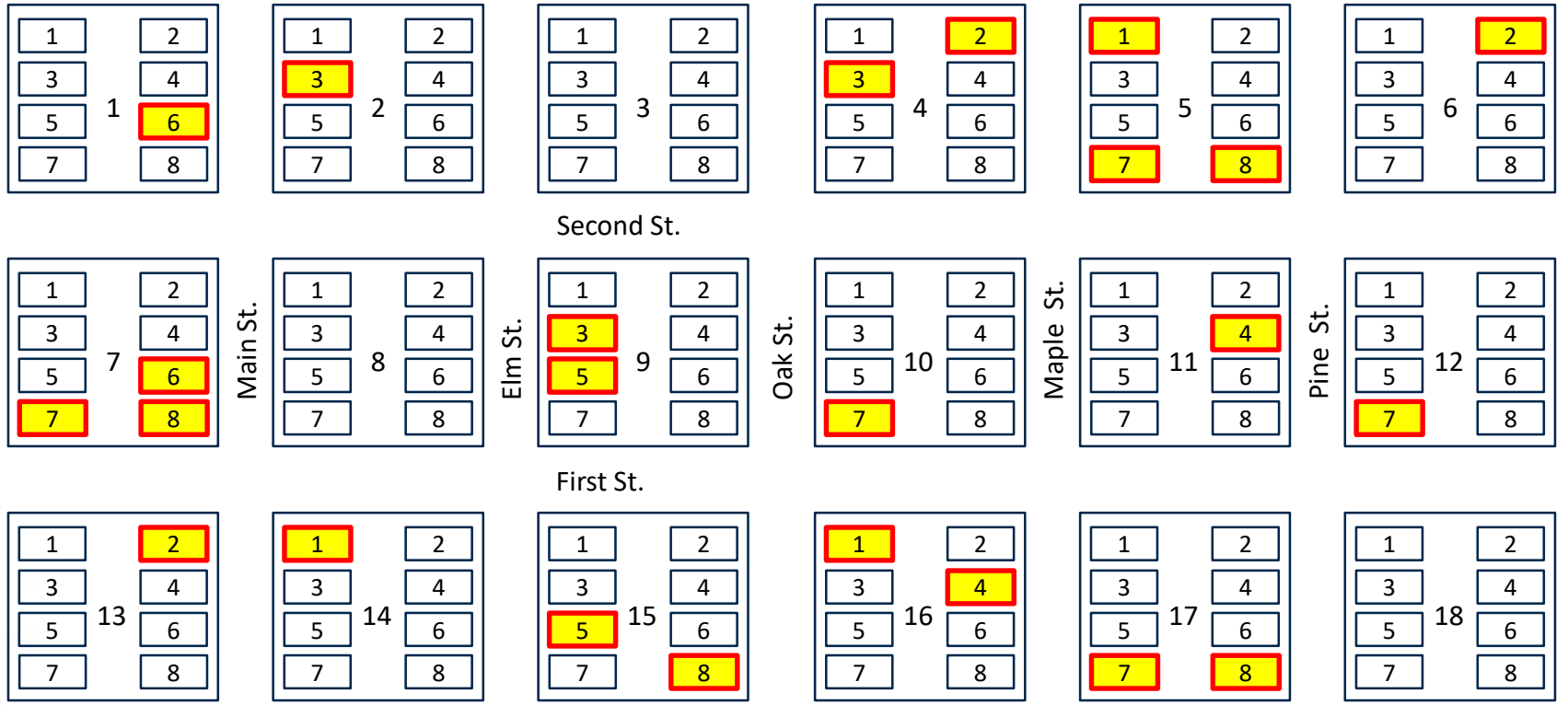
Simple Random Sampling (SRS)

- Theoretical basis for other sample designs
- Rarely used in practice for large scale surveys, even with simple list frames
- Sample of size *n* from population of size *N*
- EPSEM design
- Every combination of size *n* has the same probability of selection
  - Even "bad" samples have an equal chance of being selected
- Other sample designs can achieve smaller sampling error for the same cost or same sampling error for lower cost

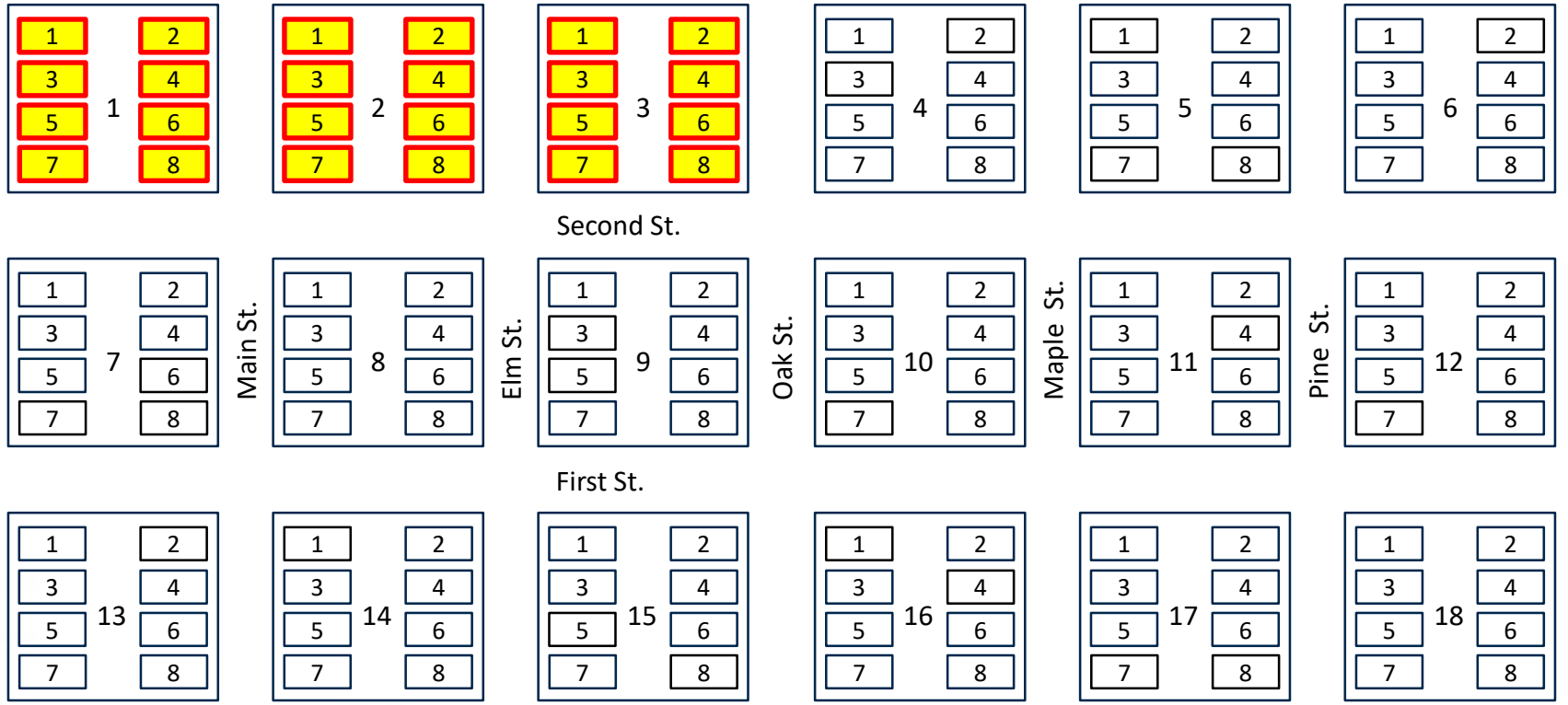
Household population in a small town (*N* = 144)



Simple Random Sample (*n* = 24)



Simple Random Sample (*n* = 24): "Bad sample"



Systematic Sampling

- SRS can be difficult to implement and check
- Systematic sampling is clerically easy to do
- From a starting point, select every k^{th} case in the population
  - $k = N/n$ is called the **sampling interval**
  - The starting point should be a random number between 1 and k

Systematic Sample (n = 24; N = 144; k = 6)

Stratification

- Objective:
  - Can we sample to avoid, or even eliminate, "bad" samples?
  - Can we achieve higher quality data (more precise estimates) for the same cost or same data quality for lower cost?
- Stratification:
  - Group elements into mutually exclusive and exhaustive groups (strata)
    - Strata should be internally homogeneous
    - Strata should differ as much as possible from each other
  - "Auxiliary information" used to create strata
  - Select samples from each and every stratum

Stratification

- Sample allocation:
  - Proportionate allocation (EPSEM)
  - Equal allocation
  - Other disproportionate allocations

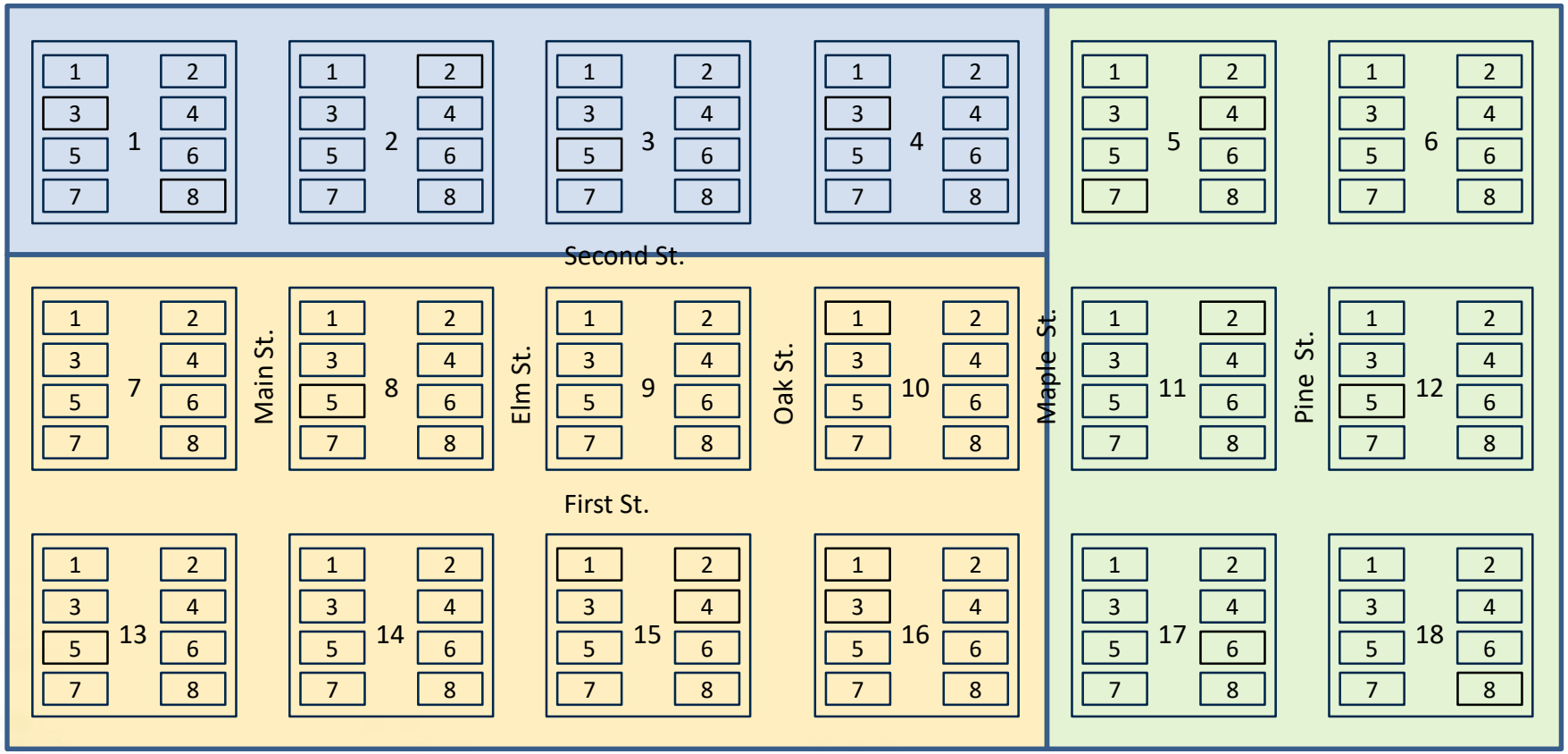
- Advantages of stratified sampling:
  - Gains in precision
  - Administrative convenience
  - Guaranteed presence of important domains
  - Acceptability/credibility
  - Flexibility

Stratification

North region
$N_1 = 32$

East region
$N_2 = 48$

South region
$N_3 = 64$

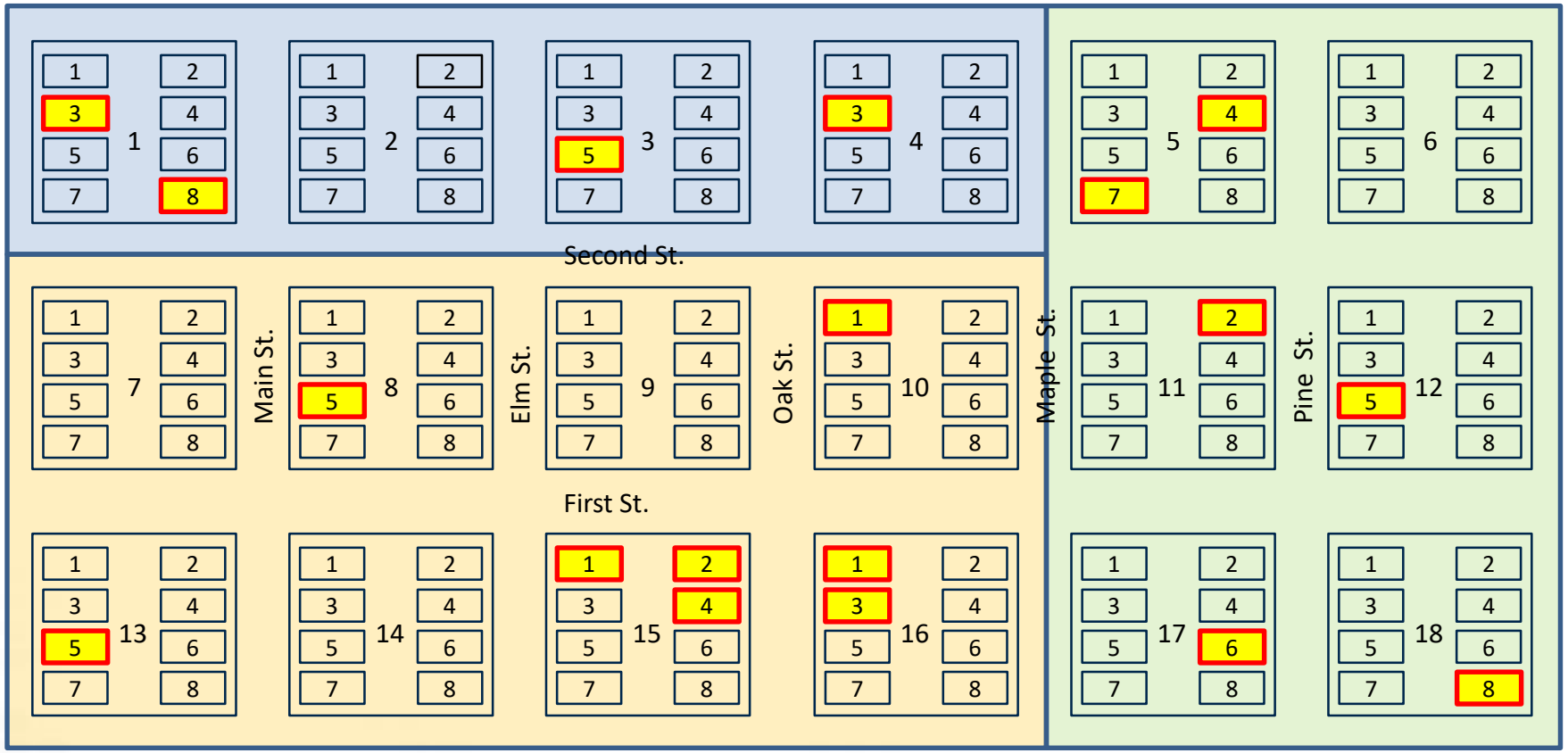


Stratified sample (n = 18)

North region
N_1 = 32
n_1 = 4

East region
N_2 = 48
n_2 = 6

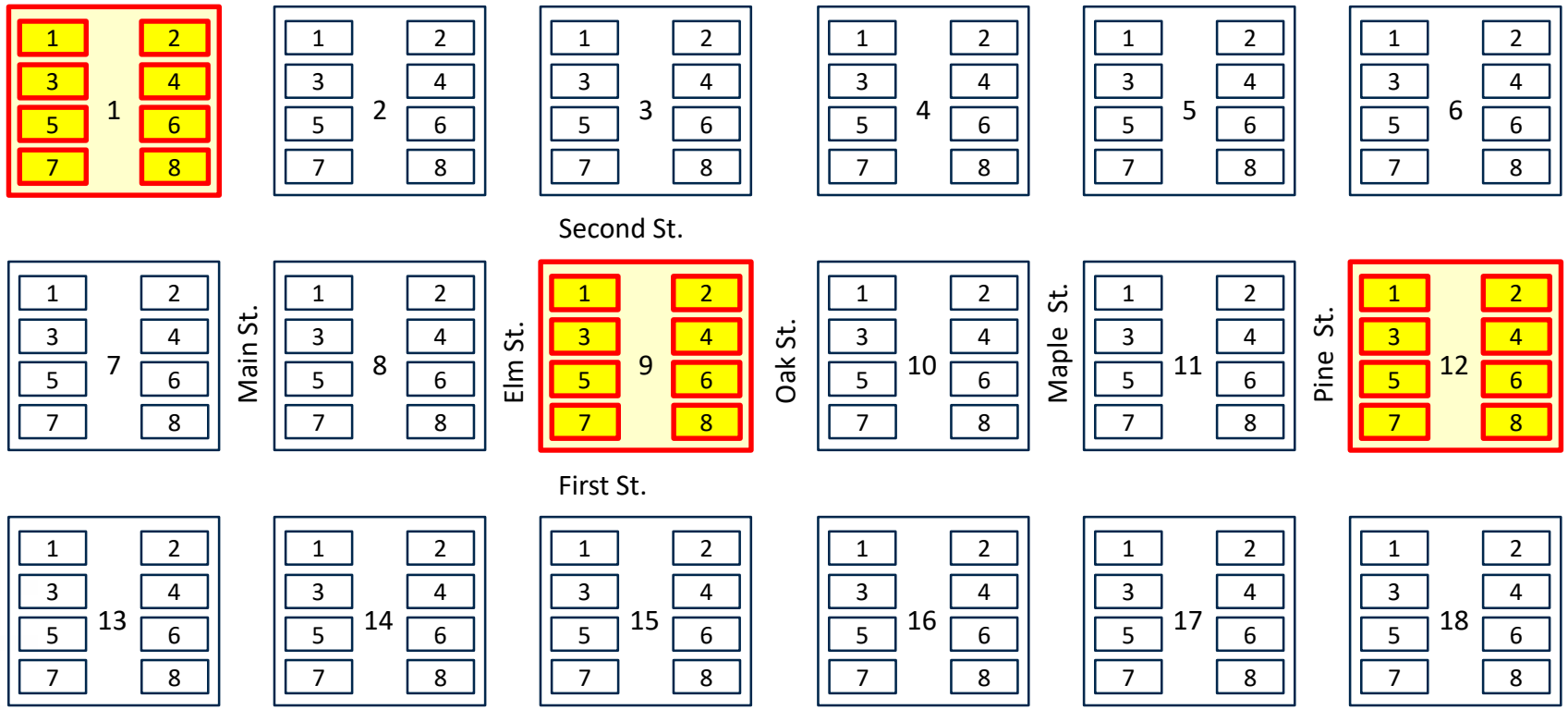
South region
N_3 = 64
n_3 = 8



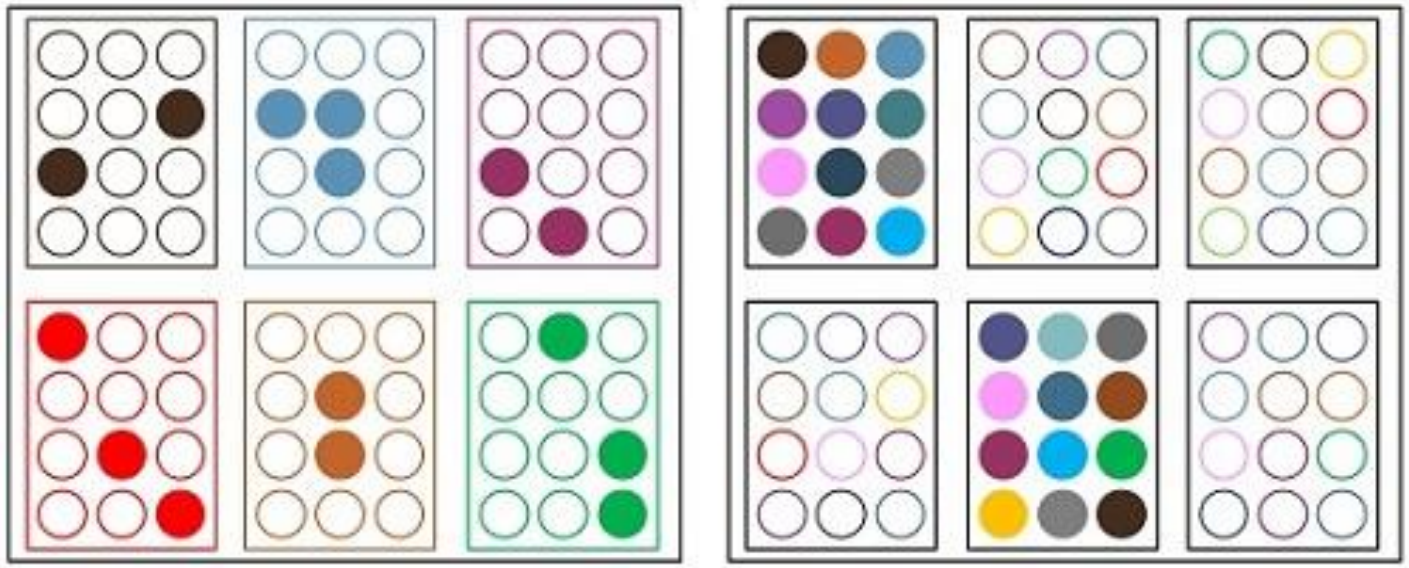
Clustering

- Populations widely distributed geographically
  - Cannot afford to visit *n* units drawn randomly from the entire area
  - Cannot afford to create an element frame
- Clusters:
  - Mutually exclusive and exhaustive groups of elements
  - Usually naturally occurring units:
    - Seldom equal size
    - Typically, internally homogeneous
- Cluster sampling reduces the cost of data collection
  - Select a sample of clusters
  - List and select elements only for selected clusters
- Clustering creates statistical inefficiencies (decreases precision)

Cluster sample (*n* = 24): 3 clusters selected



Strata vs. Clusters



Multi-stage Sampling

- A multi-stage sampling selects the sample sequentially across two or more hierarchical level

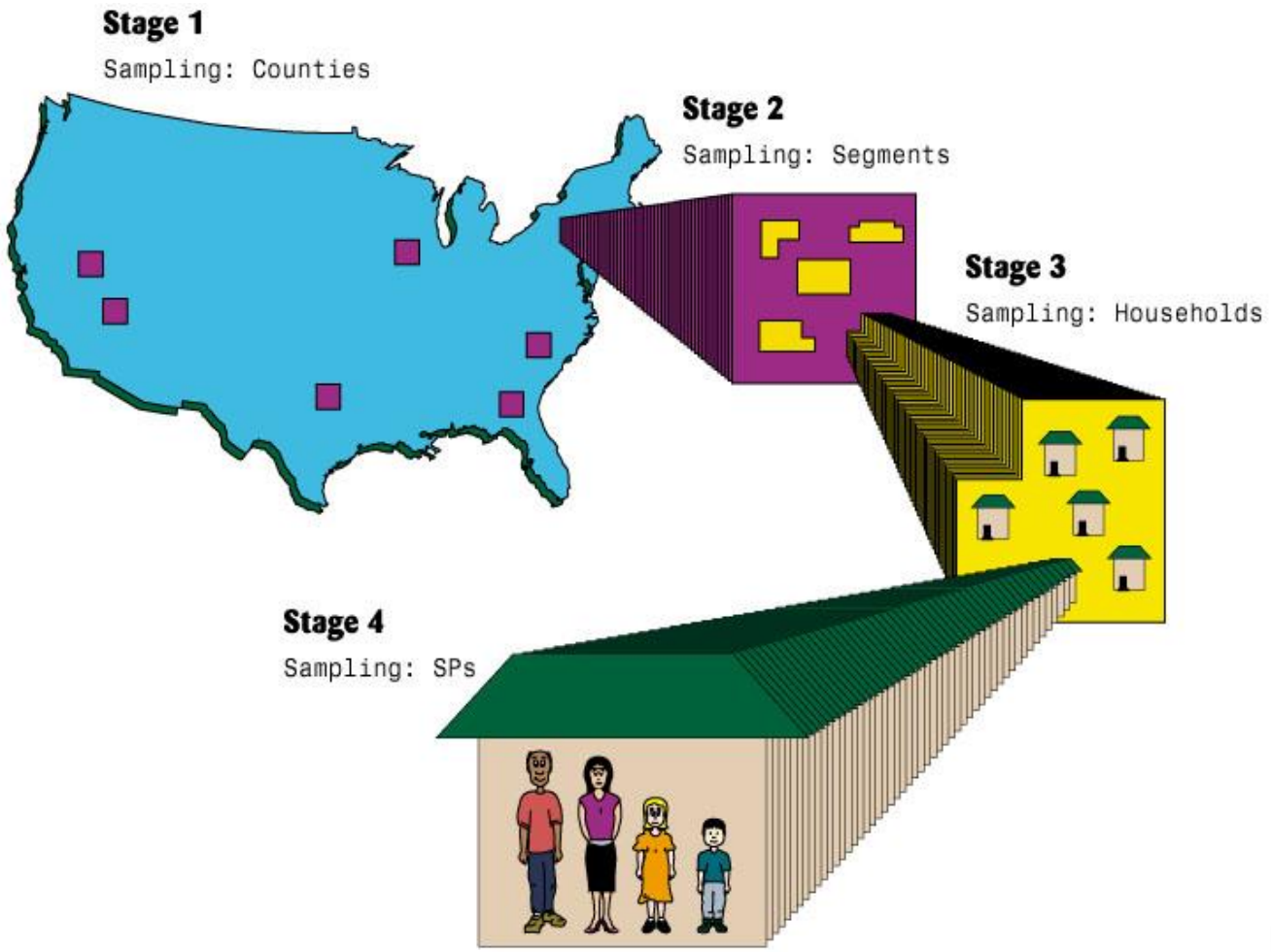
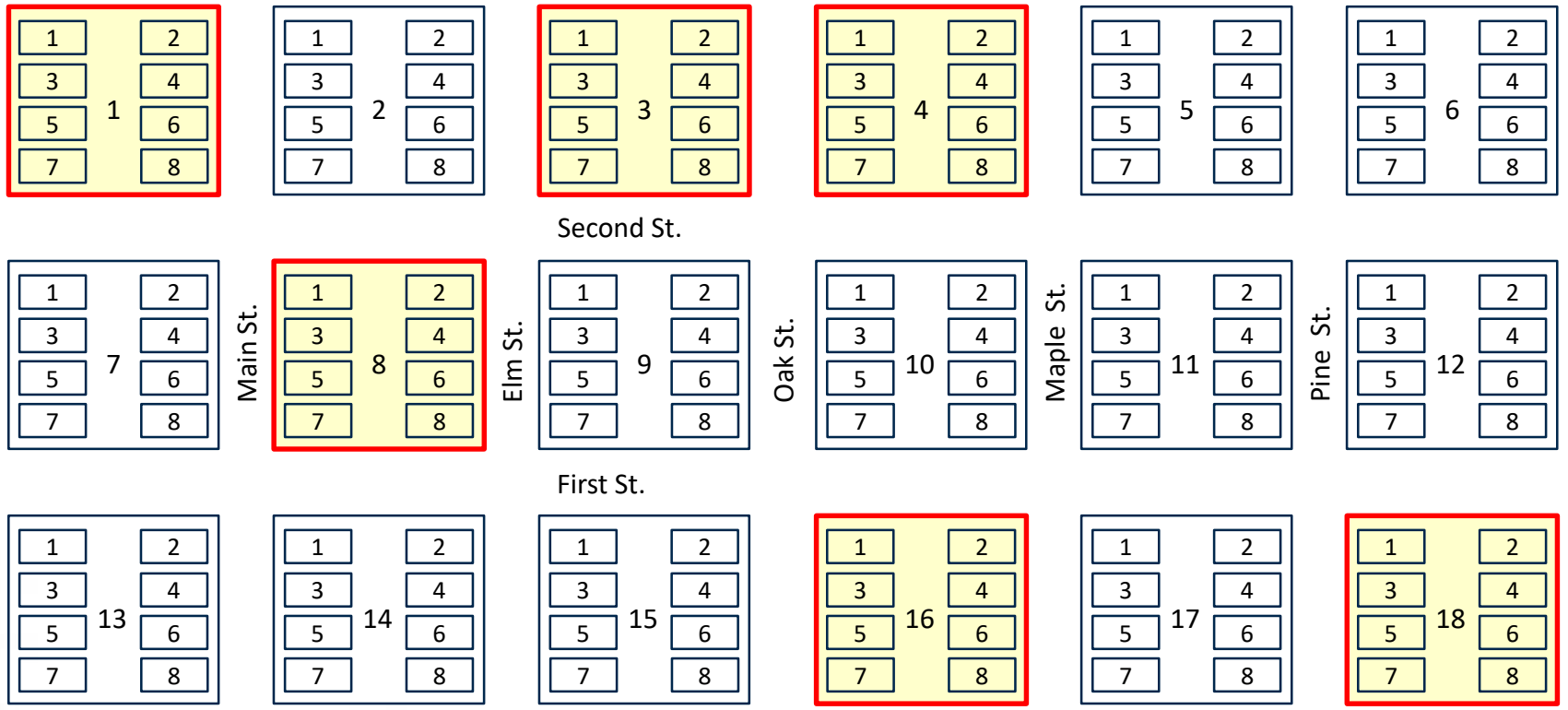


Figure by Leyla Mohadjer (Westat)

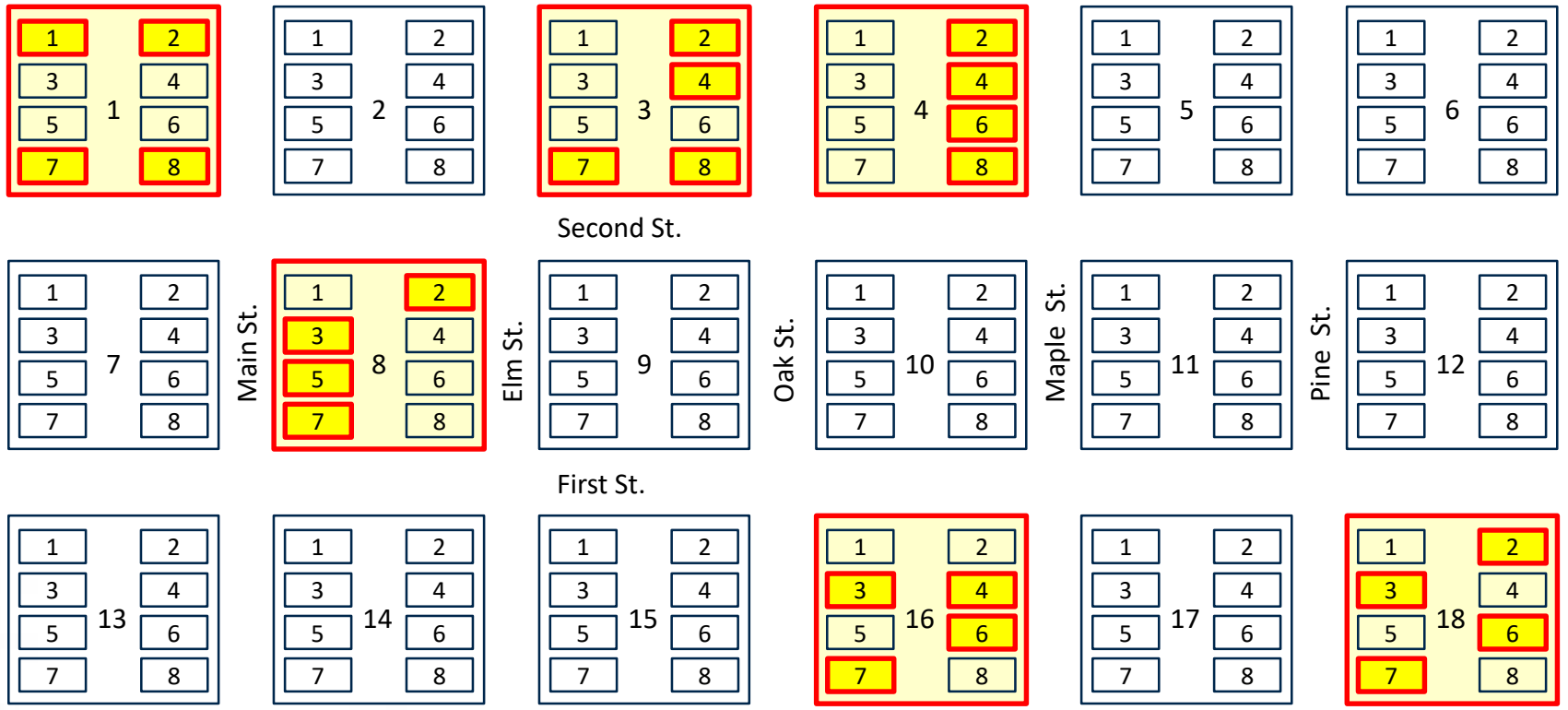
Multi-stage Sampling: Sampling Unit

- **Sampling Unit**: elements or group of elements considered for selection in each sampling stage
  - *Primary Sampling Units* (PSUs): clusters of units selected in the first (primary) stage of sampling
    - Examples: Counties, School districts, Schools
  - *Secondary Sampling Units* (SSUs): units selected in the second stage of sampling
    - Examples: Area segments, Schools, Classrooms
  - *Sample Line*: typically refers to the units (elements) selected in the last stage of sampling
    - Examples: Households, Adults, Students,

Two-stage cluster sample (*n*=24)
First stage: 6 clusters (PSUs) selected



Two-stage cluster sample (*n*=24)
Second stage: 4 elements selected in each PSU



Design Effect (deff) & Effective Sample Size

- The effect of the complex sample design on the quality (precision) of the survey estimates measured as

$$\text{deff} = \frac{\text{Var}(\bar{y})}{\text{Var}_{\text{SRS}}(\bar{y})}$$

  - If **deff < 1** → complex sample is statistically **more** efficient than a SRS
  - If **deff > 1** → complex sample is statistically **less** efficient than a SRS

- Effective sample size: $n_{\text{eff}} = \frac{n}{\text{deff}}$
  - Corresponding SRS sample size to get the same level of precision of the complex sample

INSTITUTE FOR SOCIAL RESEARCH • SURVEY RESEARCH CENTER
SURVEY RESEARCH OPERATIONS
UNIVERSITY OF MICHIGAN

Thank you!

What questions do you have?

Next Lunch & Learn: March 18 (Tuesday)

Statistical Concepts & Terminology I: Probability Sampling (Part II)