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

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

Paul Burton & Raphael Nishimura

Design, Methodology & Statistical Support

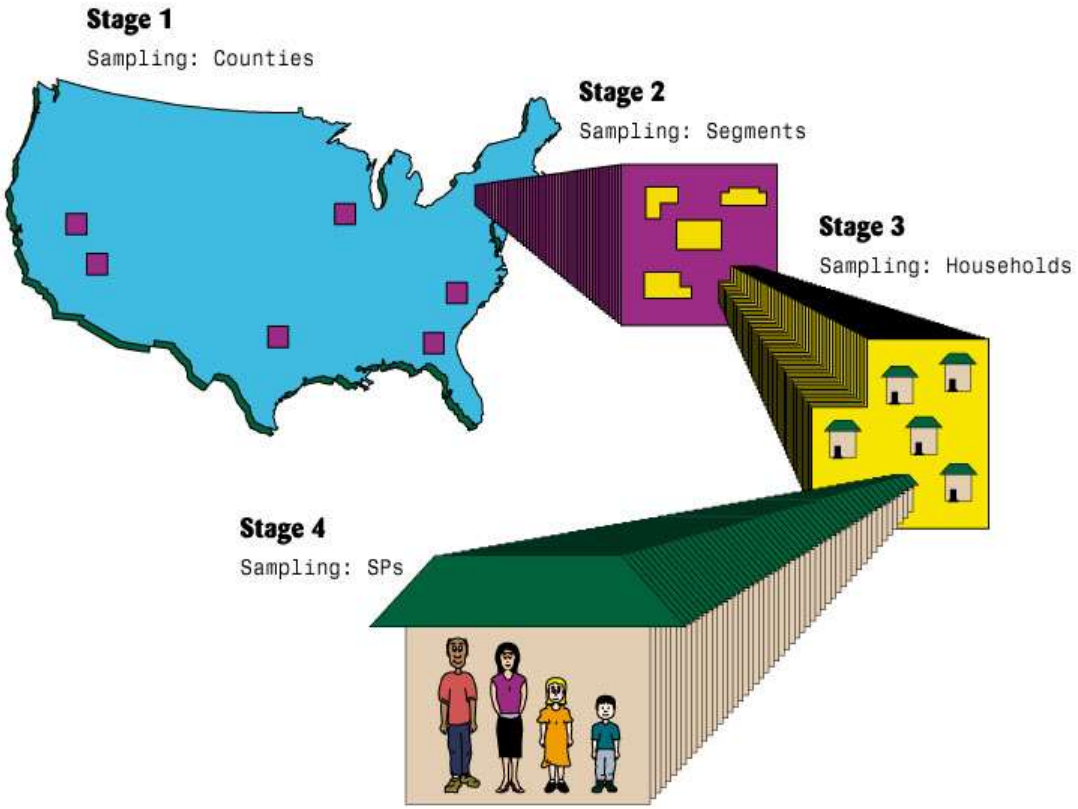
Outline – Part II

- Area probability sampling
- Address-based sampling
- Within-household selection methods
- Half-Open Interval Procedure
- Random-Digit Dialing
- Sample Release vs Sample Replicate
- Probability sampling vs Non-probability sampling

Area Probability Sampling

- Direct application of multistage cluster sampling
- Divide target population area into geographic units
  - Select sample of areas
  - Repeat across smaller areas until target units are selected
- Most often used for Face-to-Face studies – but other applications are appropriate
  - Agricultural or school based surveys
  - Any application in which data needs to be collected where the units are physically located

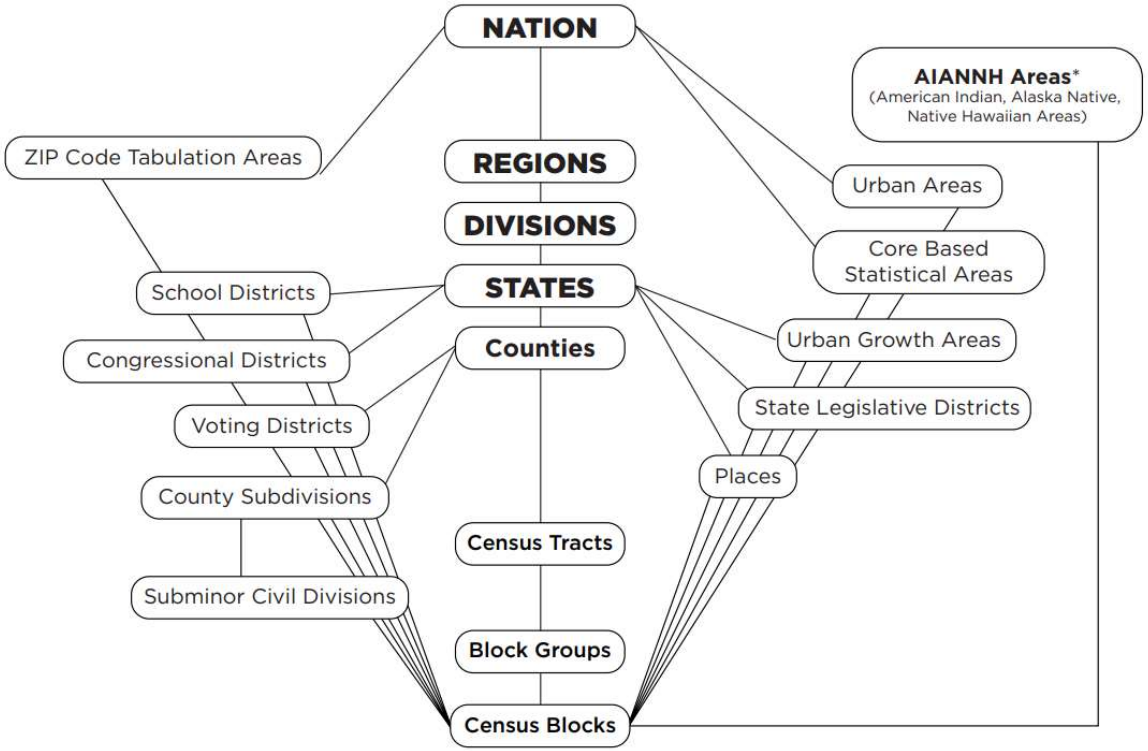
Area Probability Sampling



Area Probability Sampling

- Geographic units in this procedure are usually preexisting
  - Administrative units or divisions
    - State, Counties, Cities, etc.
  - Units constructed for statistical usage
    - Census geographies such as tracts/blocks/block groups, enumeration areas, etc.
  - Units created for other purposes
    - School districts, voting/electoral districts, postal codes (zip codes), etc.
- Areas used for stratification and/or sample units

Area Probability Sampling



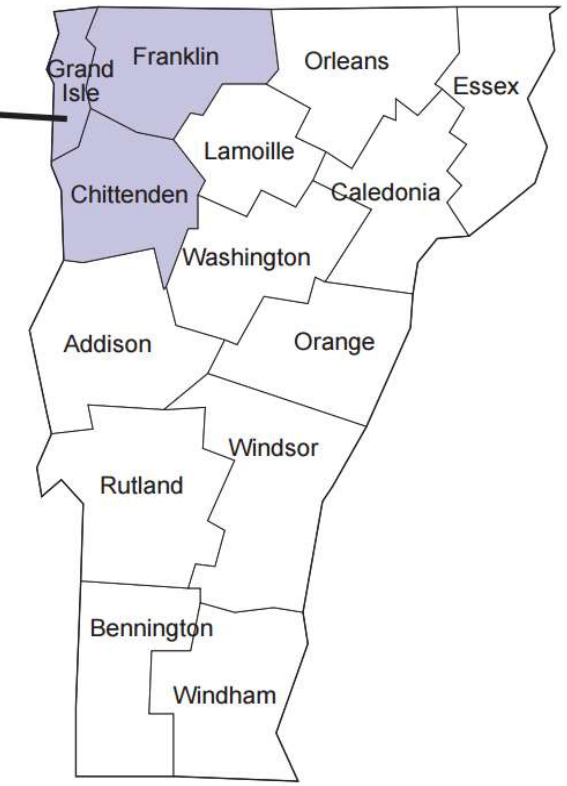
* Refer to the "Hierarchy of American Indian, Alaska Native, and Native Hawaiian Areas."

Area Probability Sampling

- SRO Examples
  - In multi-stage designs, such as the design used by the Health & Retirement Study, the preliminary stage(s) of selection are often done at the area level
  - In HRS, the PSU's are selected as counties and/or metropolitan statistical areas
    - 58 PSU's in the 2016 and 2022 HRS designs

Vermont: Burlington MSA and Counties

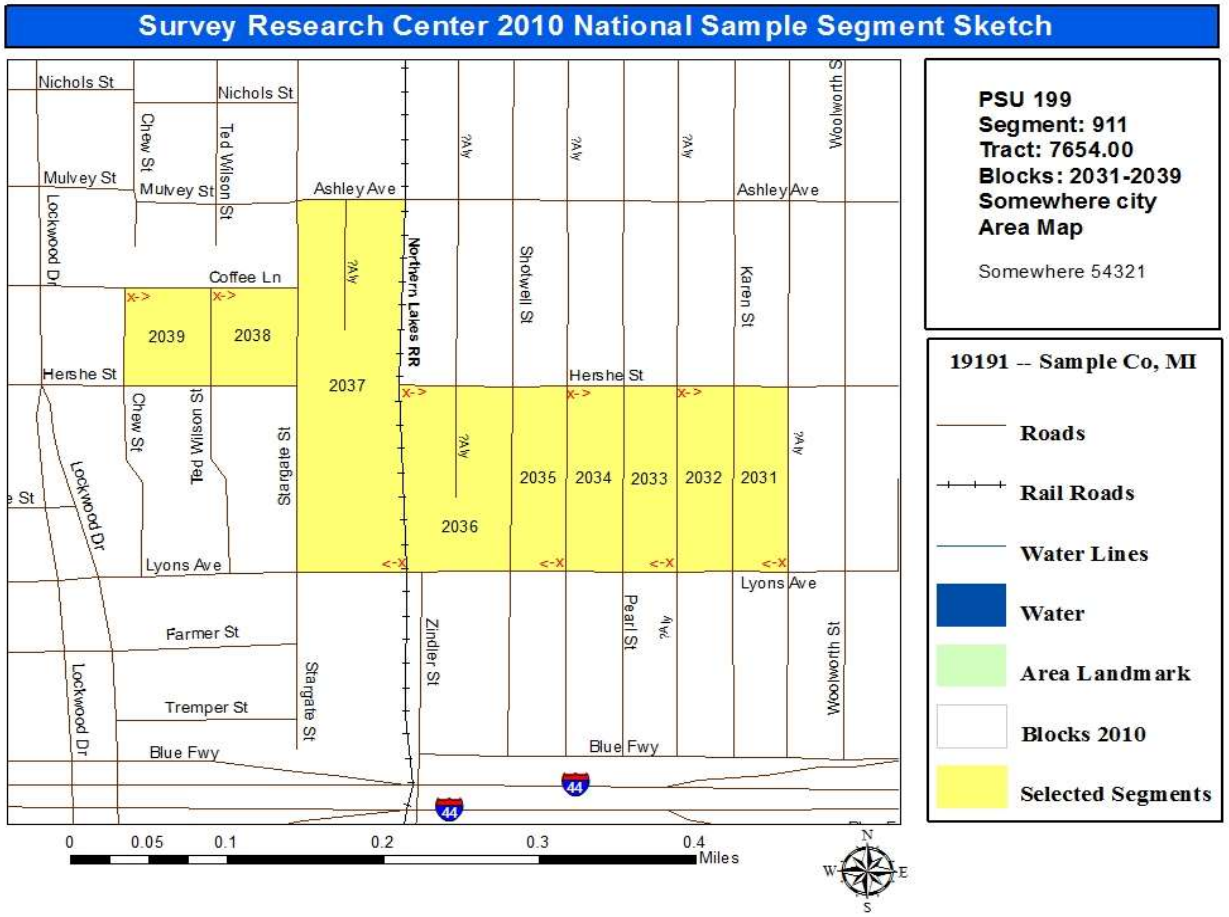
Burlington-South Burlington, VT



Burlington-South Burlington MSA

Area Probability Sampling

- In HRS, the secondary sample units (SSU's) are selected as groups of Census block groups
  - Probability of selection proportional to predicted # of eligible households
  - Increased operational efficiency



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

Survey Research Center HRS 2022

Last Updated : 30 June, 2021

PSU : 800

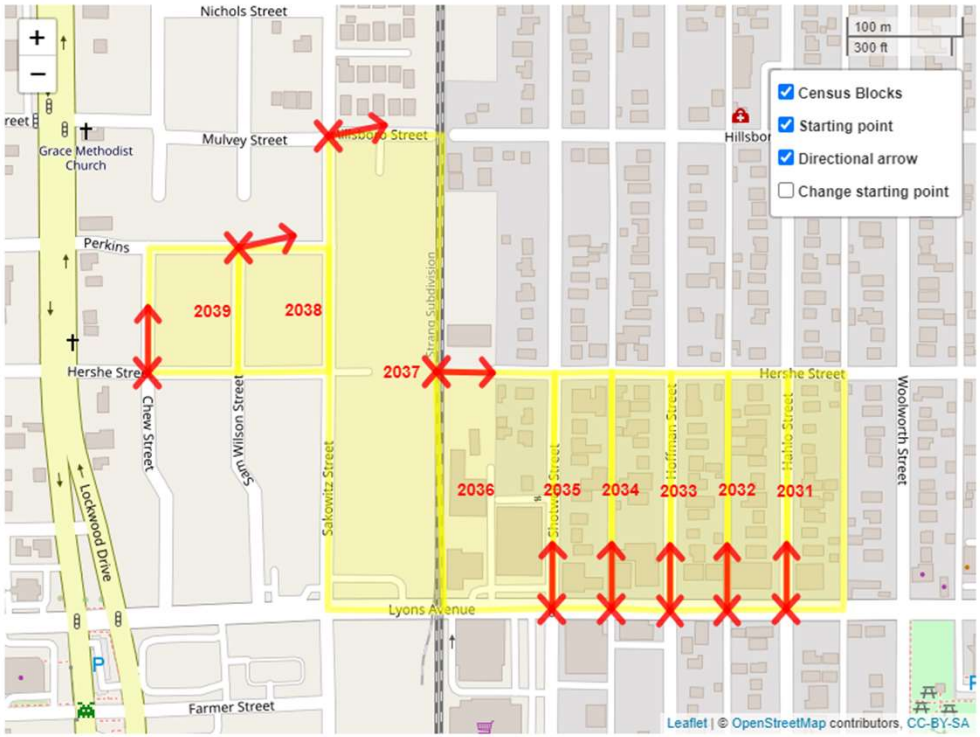
Segment : 101

Tract : 765400

Place : Sample Country

Maps for all and each block

All | 2031 | 2032 | 2033 | 2034 | 2035 | 2036 | 2037 | 2038 | 2039



Address based sampling

- A method of selecting sample addresses from a list
  - Residential addresses from a city or county
  - USPS DSF
  - Probability sampling methods used
  - Commercially available address-level data helps identify eligible households
    - Age, race, gender of household members
    - 50%-60% matching rate
    - Not 100% accurate
- Selection can be done using any method that has been previously discussed

Address based sampling

- In practice, ABS is often last, or 2nd-to-last, step in the sample selection process
- In HRS, DSF is purchased for all selected SSU's and used as frame
  - MSG data used to create eligibility predictions
- The American Family Health Study used a two-stage ABS sample design
  - Stratified sample design using commercial data
  - Over sampled based on race and likely eligibility
- The ANES Fresh Web component also utilized a ABS sample design with no clustering
  - USPS DSF with Census data used for stratification in a stratified random sample routine
  - Commercial data and voter file data used for stratification based on turnout and predicted vote choice

Within-household selection methods

- Study targets are often individuals, not households
  - Possible that there are multiple study eligible individuals within a selected household
- All study eligible individuals within a household should have a chance at being selected
  - Need the full roster of eligible individuals in the HH

Within-household selection methods

- Selection methods can be random or pseudo-random
- Random method example:
  - Kish Roster Method
    - List all eligible members of HH from oldest to youngest, by male/female
    - Manually use selection tables to select respondent
    - Software now computes this procedure

Within-household selection methods

- Examples of pseudo-random selection methods
  - Last/next birthday
  - Troldahl-Carter Selection Method
    - Use selection matrices to select the oldest/youngest male/female

2. **This questionnaire is for the (youngest/oldest) (female/male) age 18 or older living at the address above—or, if there are no (females/males) here, the (youngest/oldest) person age 18 or older. Please answer this questionnaire only if you are this person. Are you this person?**

☐ I am the (youngest/oldest) adult (female/male)
☐ I am the (youngest/oldest) person
☐ I am someone else — *Stop, give to that person.*

Within-household selection methods

- Each eligible person within the household could be given the same probability of selection
- This is not necessary, and sometimes, is not the ideal way
- Likely eligible individuals can be given a higher chance of selection
  - In the AFHS project, teenaged individuals were selected with a higher probability
  - This is a type of oversampling of less common groups done in order to increase the group's sample count

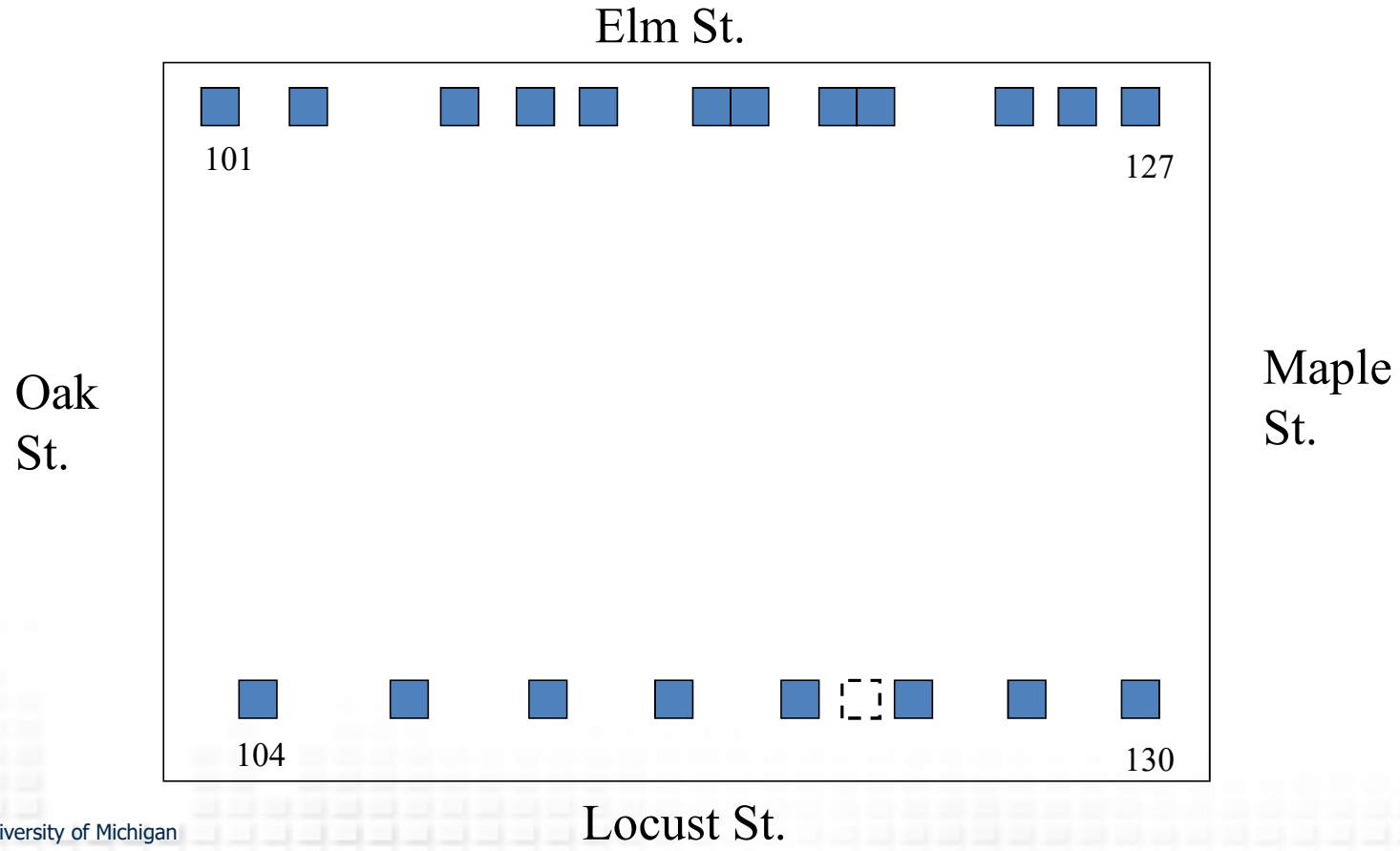
Half-Open Interval Procedure

- The address frame used in address based sampling is sometime imperfect
  - Time dependent
  - Units may be missed!
- Field interviewers, when first visiting a selected address, check the address for the following:
  - Additional units on property
  - Split address
- These situations would be considered “Type II” address updates

Half-Open Interval Procedure

- Field interviewers also check the “next address” on the frame
  - “Next address” may not be part of the selection
  - Checks for new units between the selected address and the “next address” on the frame
  - New unit(s) between the selected address and the next frame address are considered “Type I” address updates
    - New addresses are added to the frame and to the selection
    - Rules for multiple missed addresses
- Done to address coverage concerns

Half-Open Interval Procedure



Random Digit Dialing (RDD)

- Method used in phone surveys to create a random sample from a population of individuals with phones
- Different levels of “randomness”
  - Area codes are not random
  - Completely random within an area code
  - Completely random within an area code and exchange
  - Very inefficient

Random Digit Dialing

- Mitofsky-Waksberg method
  - Improved efficiency
  - Takes advantage of non-random assignment of telephone numbers
  - 2-Stages
    - Stage 1 – primary list of ‘purely random’ numbers called
      - If number is residential, then a ‘100-bank’ cluster of numbers is created
        - (309) 965-2479 is called an is residential, the ‘100-bank’ would be:
          - (309) 965-2400 to (309) 965-2499
    - Stage 2 – subsample a from the 100-bank and begin calling
      - If residential, attempt interview
      - If non-residential, replace with another number from the 100-bank
      - Continue until a certain number of residential phone numbers has been reached

Random Digit Dialing

- List-assisted designs
  - Use commercial data or phone directories to identify active/residential number
    - Improved efficiency
  - Replace inactive/non-residential numbers by adding +1 to the last digit

Random Digit Dialing

- Historically, phone numbers were truly geographically based
  - Not as true with the ubiquity of cell phones
- Cells phone use has complicated things
  - Usually attached to a person, not an address
  - Important to consider depending on the target population
  - Commercial data can provide indicators if a phone number is a cell number or landline
- An example of RDD use in production at SRO is the recently retired phone protocol for the SCA project

Sample Release vs. Sample Replicate

- Sample release is a production term
- Sample replicate refers to a specific aspect of the sample design
- Sample replicates are divisions of the sample that are interchangeable from a design perspective
  - Within each replicate, the stratification of the overall design is represented
    - Designed so that if a replicate were removed, the remaining sample would have the identical proportion of design elements as the initial total
- Sample releases are groups of sample that are distributed to the field/interviewers for production
  - Ideally, these would be groups of replicates
    - This allows for better control of the sample from a design perspective
    - Flexibility for changing protocols using adaptive design
  - Releases can be across replicates, but this introduces some risks

Probability sampling vs Non-probability sampling

- Probability sampling – method of sample creation where every element has a non-zero, positive probability of selection
  - Discussed in detail in Part 1
  - Minimal post data collection adjustments needed
  - Less restrictive assumptions
  - Can be expensive

Probability sampling vs Non-probability sampling

- Non-probability sampling – method of sample creation where no random selection process is used to select elements into the sample
  - Often less expensive
  - Sometimes necessary for hard to reach populations
  - Examples:
    - Convenience sample – Volunteer web panels
    - Sample matching - Quota sampling
    - Network sampling - Snowball sampling
  - Strong assumptions required to make population inferences

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

Thank you!

What questions do you have?

Next Lunch & Learn: April 4 (Friday)
Statistical Concepts & Terminology II: Non-Probability Sampling