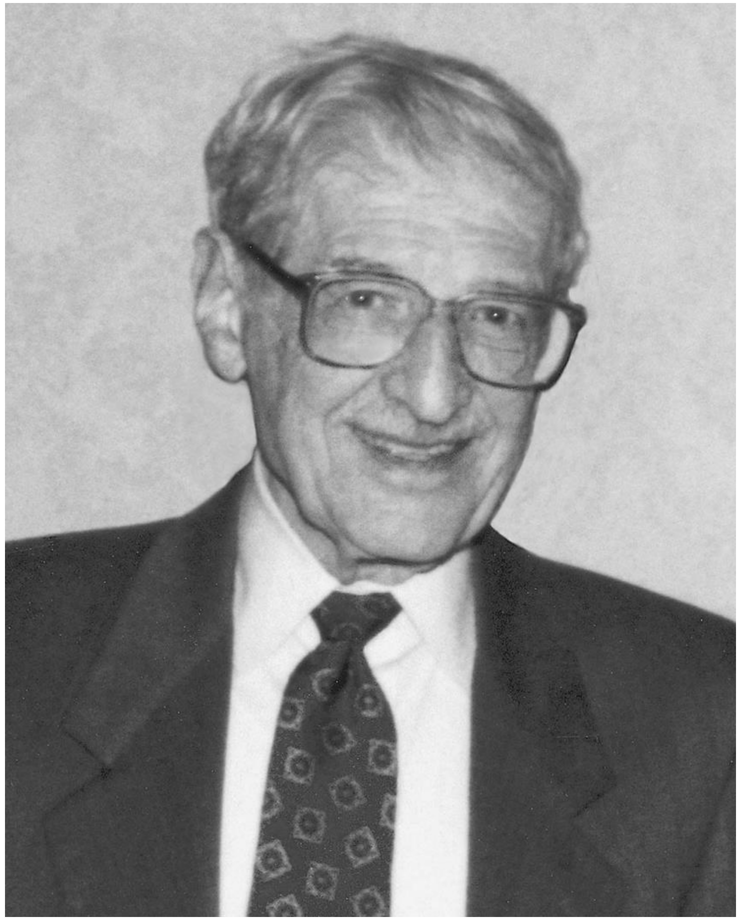
Multiple-Frame Surveys for a Multiple-Data-Source World

2021 Waksberg Lecture

Sharon L. Lohr

www.sharonlohr.com

Joseph Waksberg



- U.S. Census Bureau, 1940-1973
- Westat, 1973-2006
- Themes
  - Coverage
  - Better estimates for lower cost
  - Use all available data resources
- Multiple-Frame Surveys

Outline

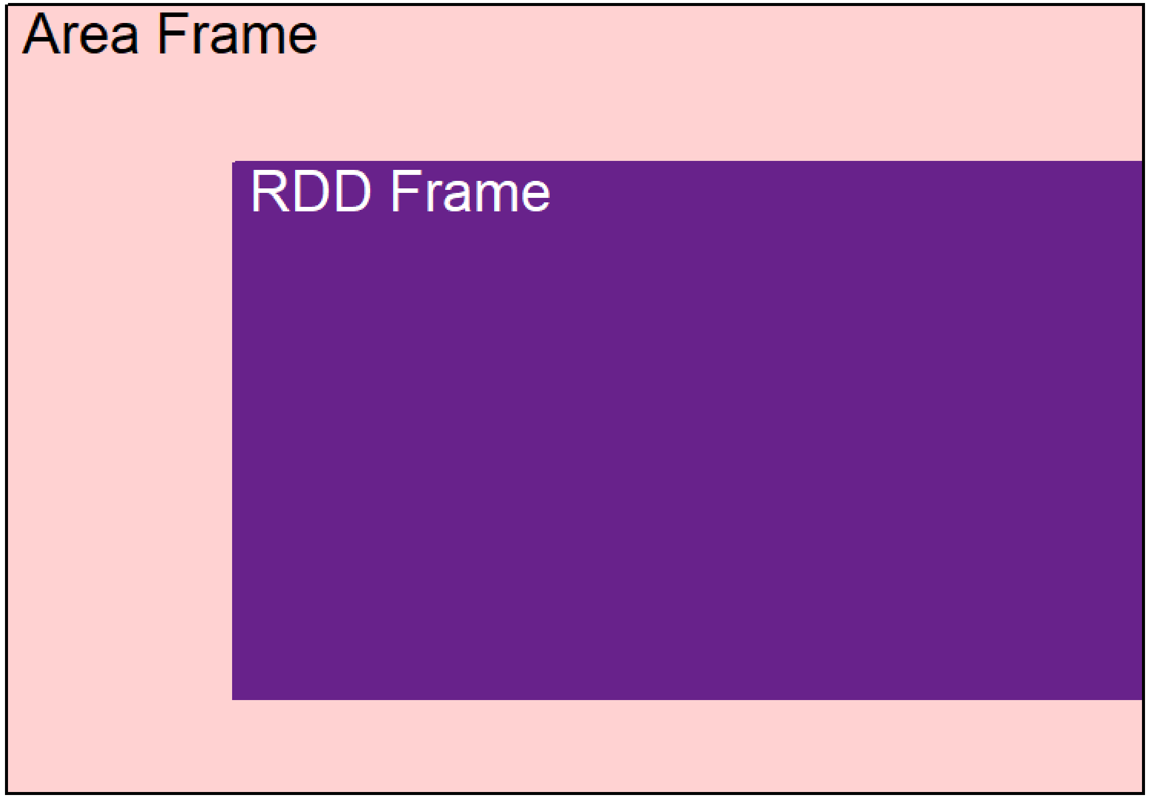
- National Survey of America's Families (Waksberg survey)
- Assumptions for classical multiple-frame (MF) surveys
- Calibration of MF surveys
- MF surveys as organizing principle for combining data
- Implications for design

National Survey of America's Families (1997)

- Waksberg et al. (1997, JSM Proceedings)
- Motivation: Evaluate effects of 1996 welfare program changes
- US civilian noninstitutional population under age 65
- Emphasis: Families with children below 200% of poverty
- National estimates plus separate estimates for each of 13 states
- Goal: Effective sample size of 800 poor children in each state
- Oversample poor families with children (about 1 in 8 families)

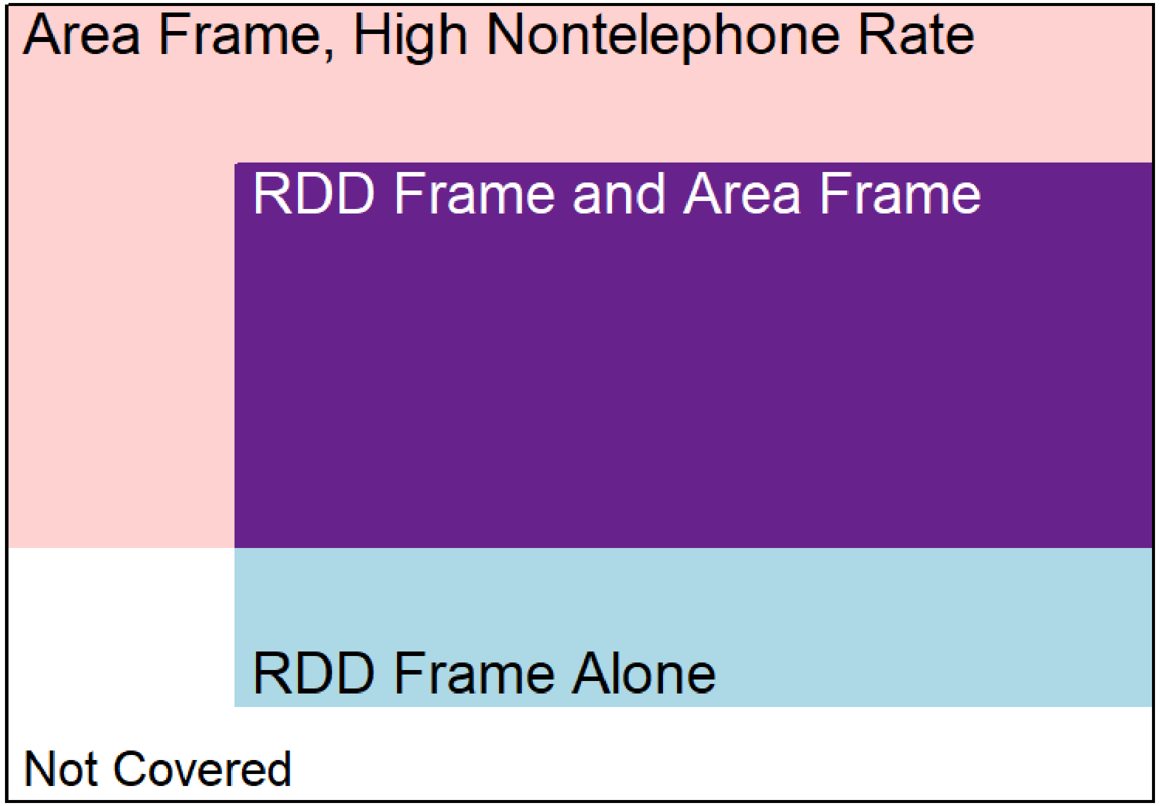
NSAF Design Options

- Area Frame
  - Full coverage
  - Expensive to screen for poverty
- RDD Frame
  - Much less expensive to screen
  - But thought that 20% of poor families have no telephone
  - Do nontelephone families differ from telephone families?
- Dual Frame



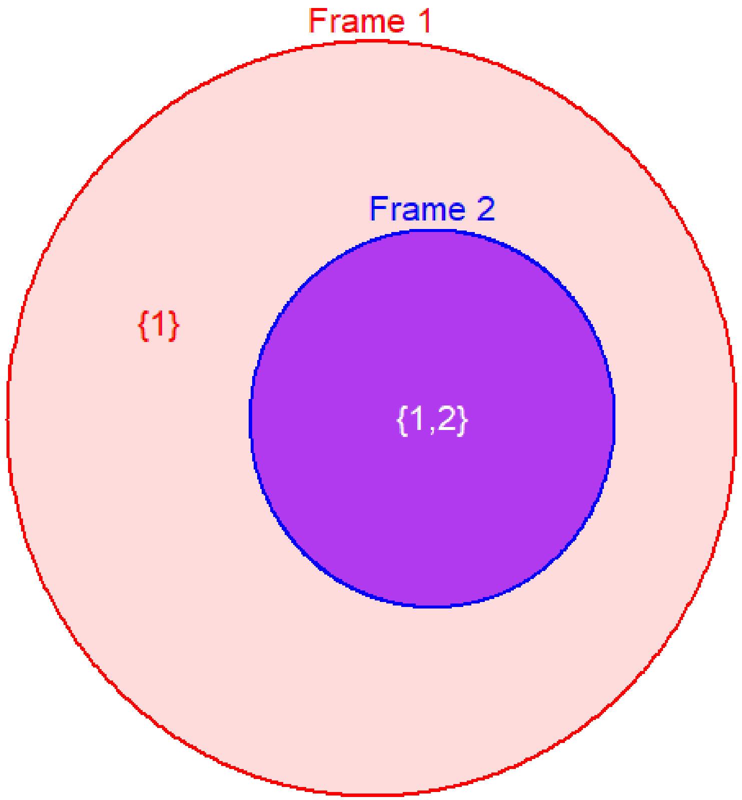
NSAF Design

- Restrict area frame to census blocks with high nontelephone
- Main sample from RDD frame
- Independent sample from area frame, screen out telephone HHs
- Screening dual-frame survey
- If screening accurate, this is stratified sample



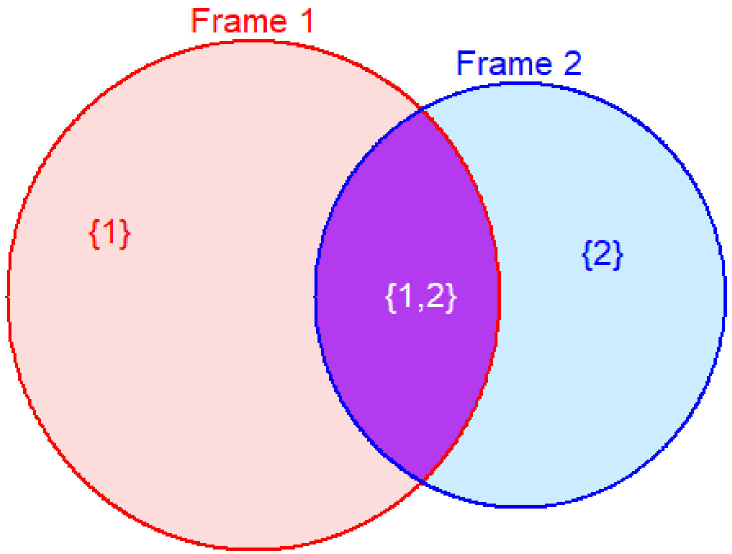
Some Multiple-Frame Designs

Frame 1 Complete, 2 Incomplete



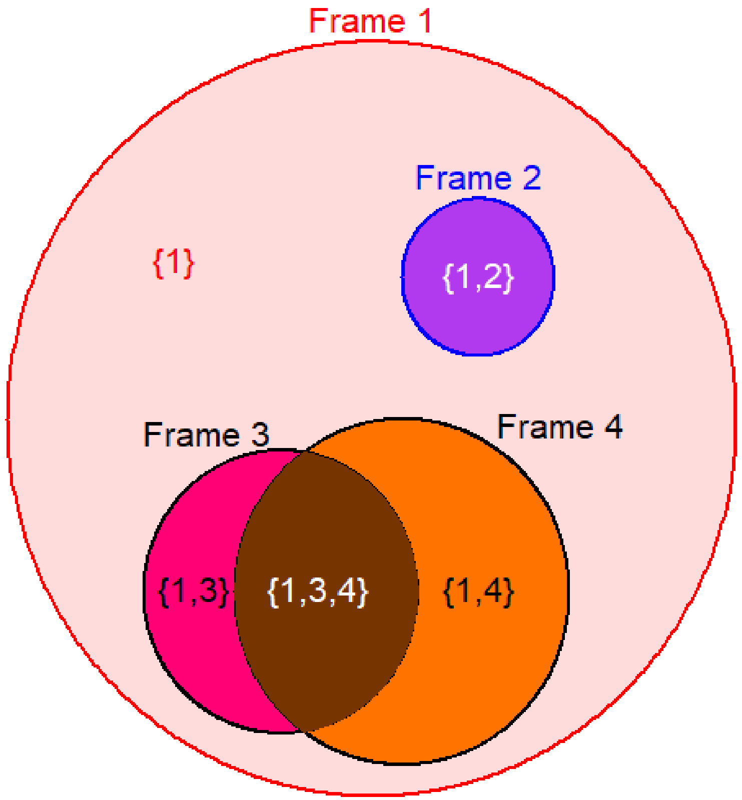
Domains {1}, {1,2}

Frames 1 and 2 Overlap



Domains {1}, {1,2}, {2}

Frame 1 Complete, Others Incomplete



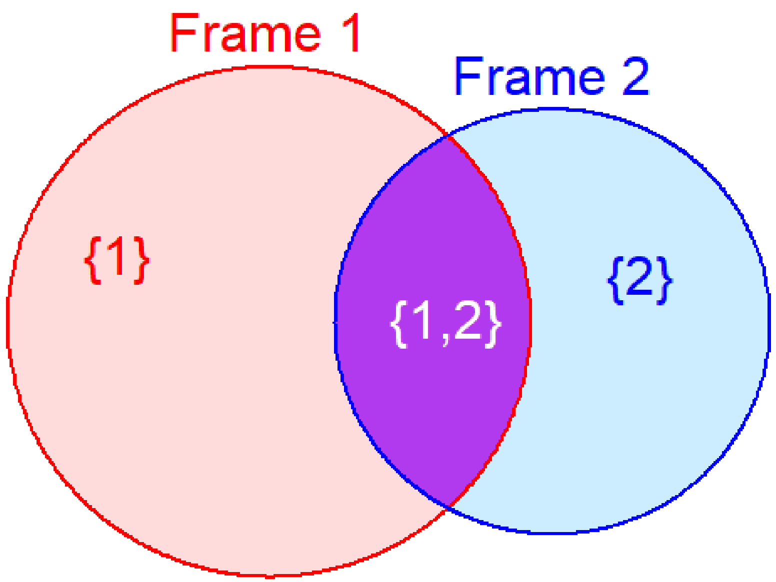
Domains {1}, {1,2}, {1,3}, {1,4}, {1,3,4}

Assumptions for Classical MF Surveys

- Union of frames **covers** population
- Full-response **probability sample** taken from each frame
- Samples from frames are selected **independently**
- **Domain membership known** for each sampled unit
- Estimators of population totals in each domain are **unbiased**
- **No measurement error**. y_i (Sample j) = y_i (Sample k)

Estimation for Classical MF Surveys

- If assumptions met, main problem is to account for overlap
- Domain {1,2} in both samples
- Adjust weights for multiplicity
- Lots of estimators
- See Lohr (2011) for review



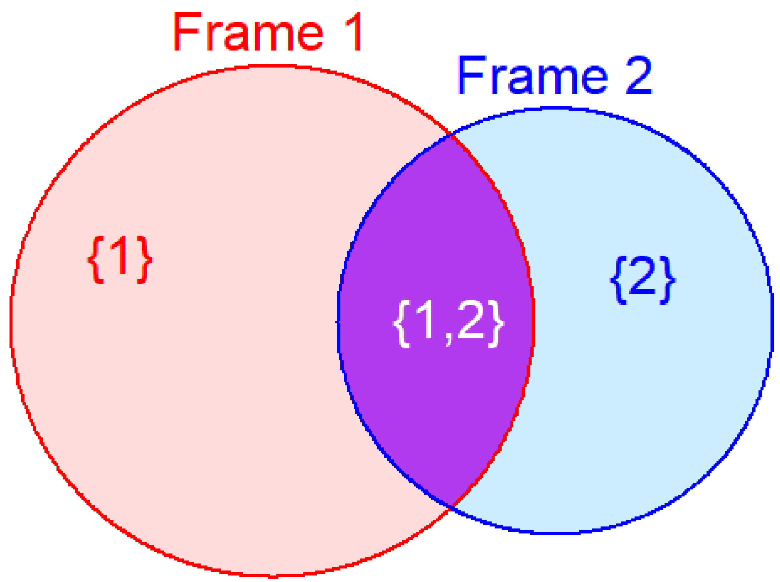
Some Estimators for Population Total Y

- Optimal (Hartley, 1962)

$$\hat{Y} = \hat{Y}_{\{1\}} + \hat{Y}_{\{2\}} + \theta \hat{Y}_{\{1,2\}} + (1-\theta)\, \hat{Y}_{\{1,2\}}$$

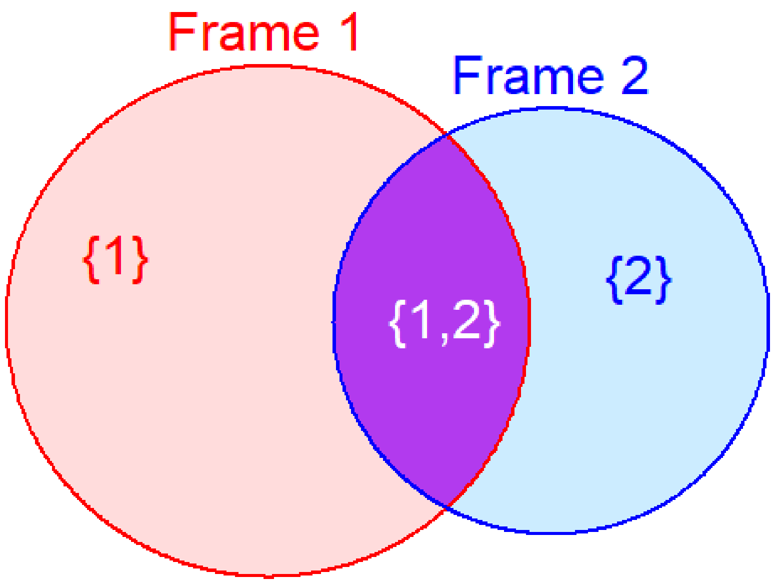
θ chosen to minimize $V(\hat{Y})$

- Screening, $\theta = 0$ or 1
- $\theta = 1/2$
- Effective sample size, $\theta = \tilde{n} \,/\, (\tilde{n} + \tilde{n}\,)$
- Weight by estimated overall selection prob



Adjust Weights for Multiplicity

- Start with sampling weights w_i, w_i
- Simple multiplicity adjustment, $\theta \in [0,1]$
- $\widetilde{w}_i = \begin{cases} w_i, & i \in \{1\} \\ \theta\, w_i, & i \in \{1,2\} \end{cases}$
- $\widetilde{w}_i = \begin{cases} w_i, & i \in \{2\} \\ (1-\theta)\, w_i, & i \in \{1,2\} \end{cases}$
- Weights reduced in overlap domains



Calibration

- Skinner (1991) raking
- Ranalli et al. (2016), general calibration
- Auxiliary vector $\boldsymbol{x}$ with known population totals $\boldsymbol{X}$
- Start with multiplicity-adjusted weights, $\tilde{w}_i$, $\tilde{w}_i$, …
- Calibrated weight, Frame 1:
- $c_i = \tilde{w}_i \left[1 + \left(\boldsymbol{X} - \widehat{\boldsymbol{X}}\right)' \left(\sum_{i \in S_1} \tilde{w}_i \boldsymbol{x}_i \boldsymbol{x}_i' + \sum_{i \in S_2} \tilde{w}_i \boldsymbol{x}_i \boldsymbol{x}_i'\right)^{-1} \boldsymbol{x}_i\right]$
- Repeat for all samples

Calibration Considerations

- InfoU and InfoS (Särndal & Lundström, 2005)
- InfoU: known for population and for every respondent
- InfoS: known for every member of selected sample
- MF: have InfoU and InfoS for each sample, and for merged samples
- National Survey of America's Families
  - Rich auxiliary information for area frame
  - Little auxiliary information for RDD frame
- Or, may have
  - Little auxiliary information for complete frame
  - Rich auxiliary information for incomplete list frame

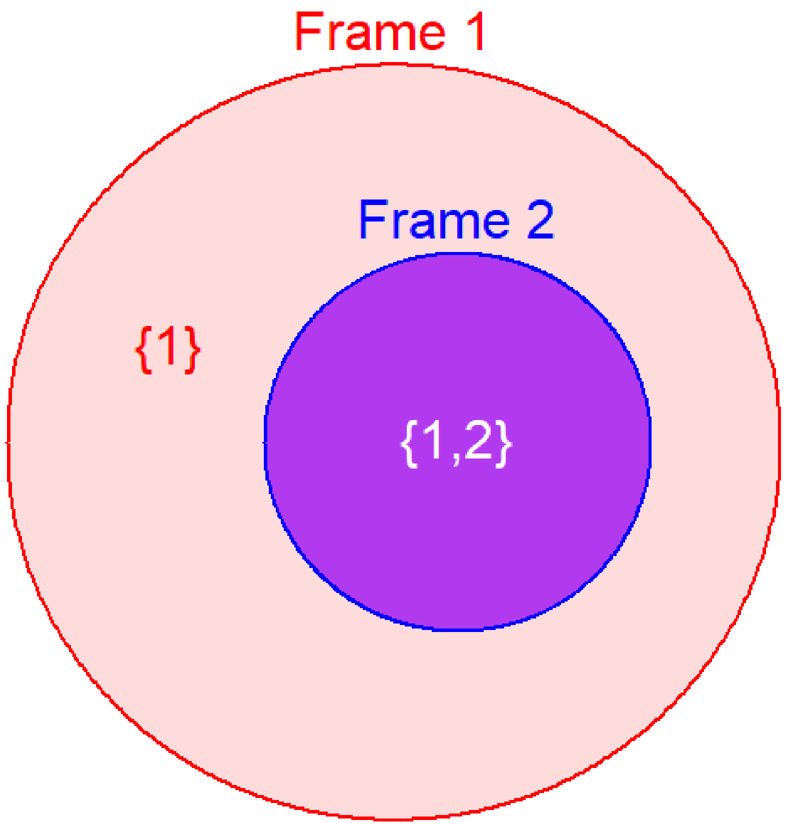
Multi-Step Calibration

More robust to model misspecification (Haziza & Lesage, 2016)

1. Calibrate individual samples to InfoS (nonresponse adjustments)
2. Calibrate individual samples to InfoU (poststratification)
3. Calculate multiplicity weight adjustments (Calibration may change relative effective sample sizes)
4. Calibrate to InfoU for full population

Special Case of MF Survey

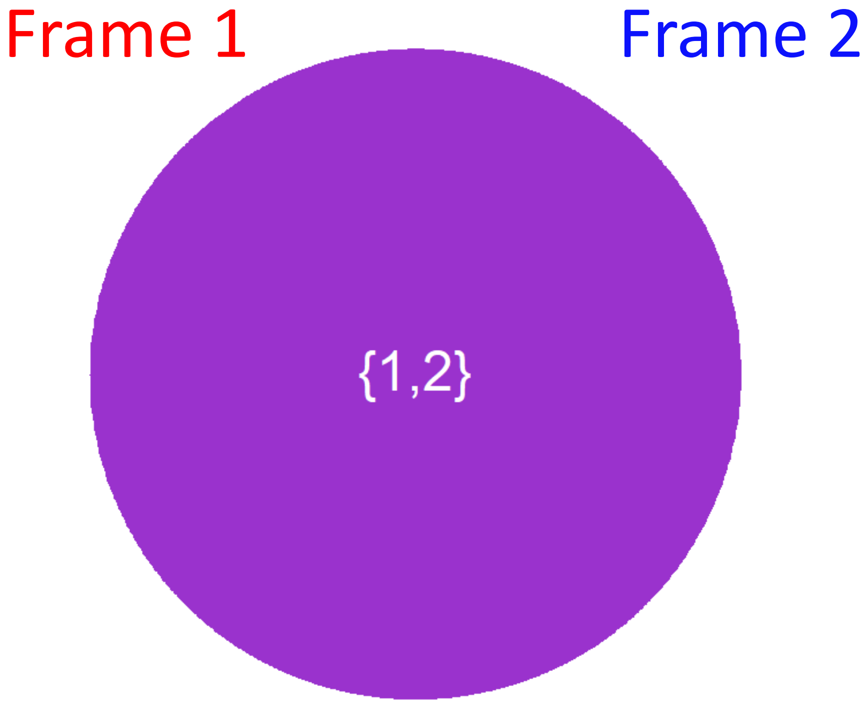
- Sample from Frame 2 is a census
  - Administrative records
  - Convenience sample
- Lohr (2014); Kim and Tam (2020)
- Undercoverage from Frame 2 remedied by Frame 1
- If MF assumptions met, statistical properties come from Sample 1 design; Sample 2 has no sampling error



Beyond Classical MF Surveys

- Framework for data integration methods by relaxing assumptions
- Some data sources are not probability samples
- Citro (2014); Lohr & Raghunathan (2017); Zhang & Chambers (2019); Thompson (2019); Beaumont (2020); Yang & Kim (2020); Rao (2021); many more
- Small area estimation
- Mass imputation
- Capture-recapture estimation

Small Area Estimation



MF Assumptions

✓**Coverage**

✓**Probability sample**

✓**Independent samples**

✓**Domain membership known**

Unbiased estimates (Sample 1)

No measurement error (Sample 1)

$$\theta \hat{Y}_a + (1-\theta) x_a{}' \hat{\beta}$$

θ varies across areas

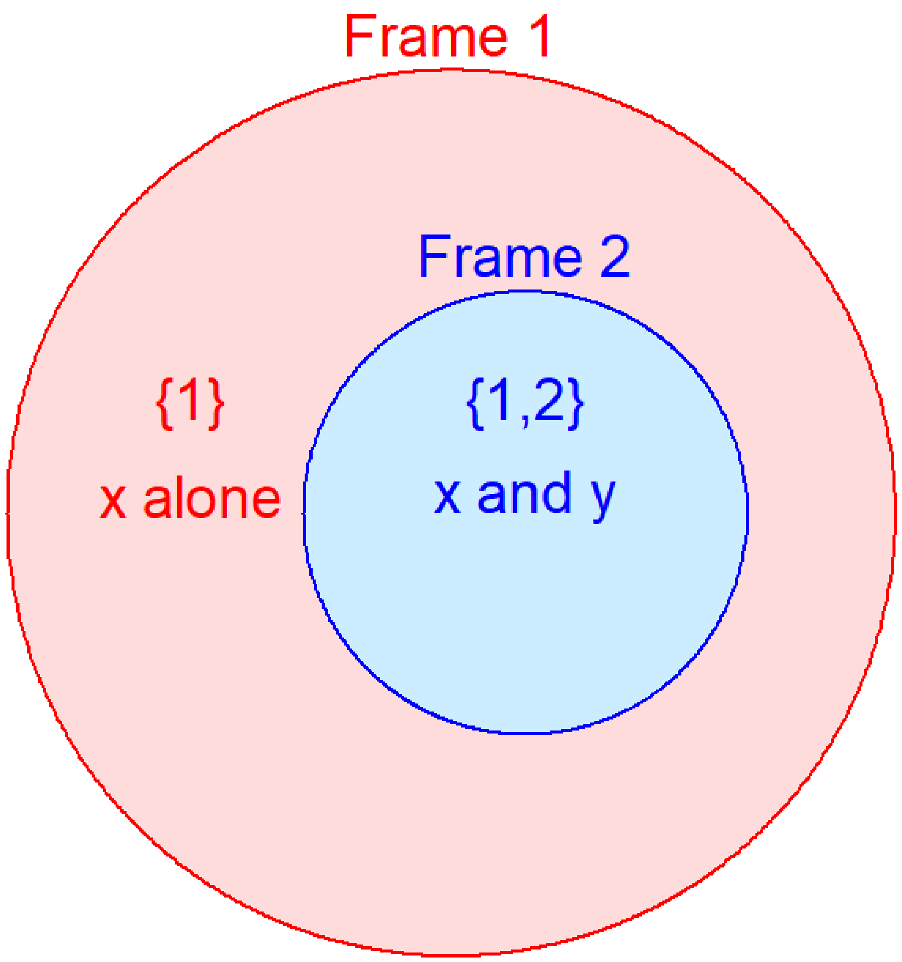
Mass Imputation and Sample Matching

- Want to estimate Y
- Sample 1 measures **x**
- Sample 2 measures y and **x**
- Prediction model from Sample 2

$$\tilde{y} = \hat{g}(\boldsymbol{x})$$

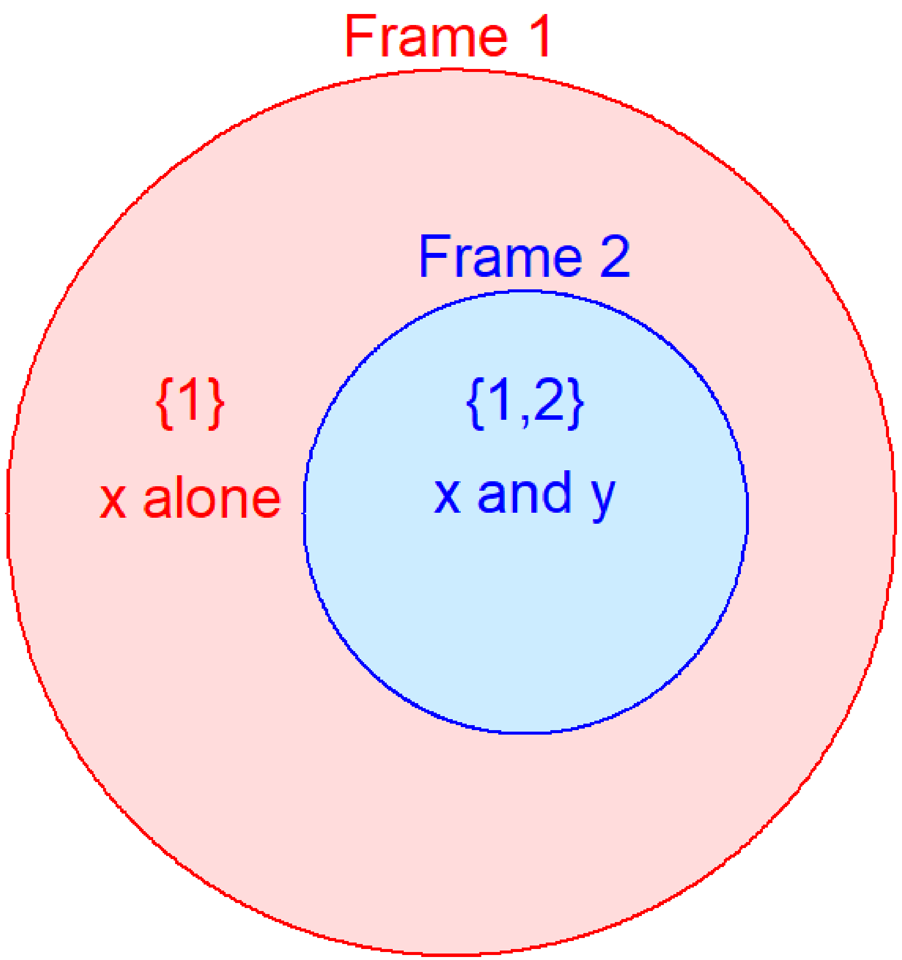
$$\tilde{Y}_{\{d\}} = \sum\nolimits_{\{d\}} w_i \tilde{y}_i$$

$$\hat{Y}_{imp} = \tilde{Y}_{\{1\}} + \theta \tilde{Y}_{\{1,2\}} + (1-\theta)\hat{Y}_{\{1,2\}}$$



Mass Imputation, Sample Matching

- Rivers (2007)
- Kim & Rao (2012)
- Chipperfield et al. (2012)
- Bethlehem (2016)
- Kim & Tam (2020)
- Yang et al. (2021)



Mass Imputation

MF Assumptions

✓**Coverage**

✓**Probability sample**

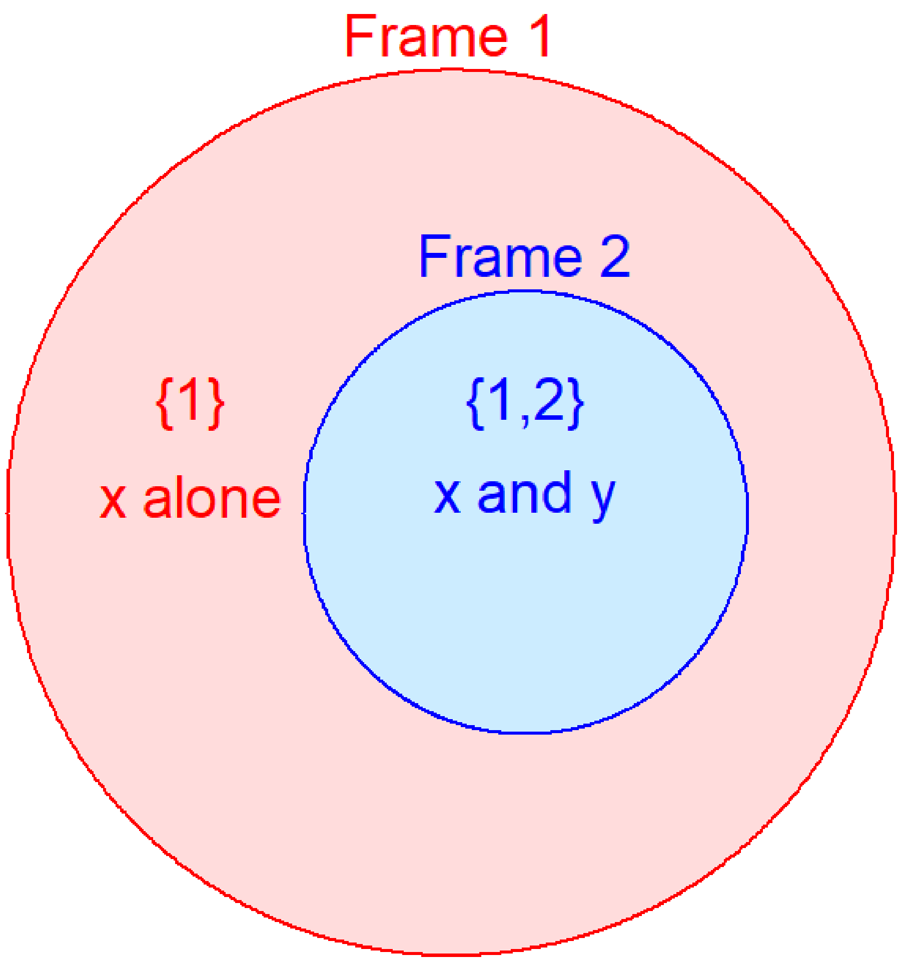
✓**Independent samples**

✓**Domain membership known**

Unbiased estimates (Sample 2)

No measurement error (Sample 2)

? Model applies to domain {1}



Generic Theorem

- If $y = g(\boldsymbol{x})$ is true prediction model, then estimates computed from imputed data have Good Properties
  - Approximately unbiased
  - Variance depends on sampling variances and model

- But what if model is wrong?
- Y. Lu (2014)
  - Regression in MF surveys
  - No reason to believe relationship is same across domains

Imputation and the NSAF

- Estimate percentage of children in poverty
- Pretend poverty not measured in area frame and impute it
- Imputation models fit to RDD sample using demographic variables

Imputation Model 1		
RDD	Area	Full
38.6	30.5	38.1

Imputation Model 2		
RDD	Area	Full
38.6	51.9	39.5

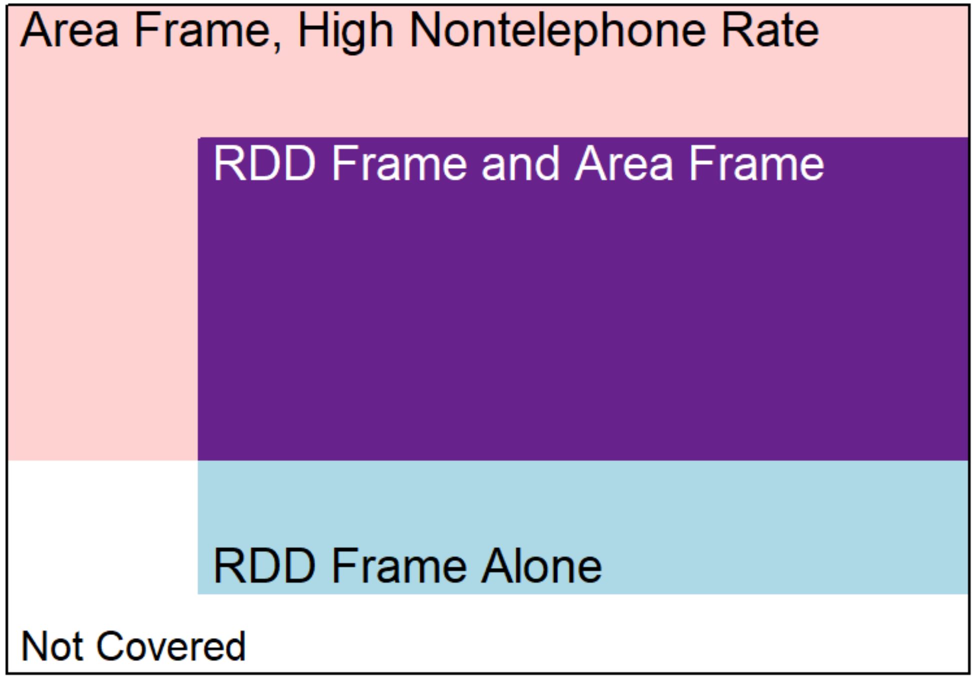
Imputation and the NSAF

Imputation Model 1			Imputation Model 2			Actual Data		
RDD	Area	Full	RDD	Area	Full	RDD	Area	Full
38.6	30.5	38.1	38.6	51.9	39.5	38.6	93.4	42.2

- Lack of telephone highly associated with poverty
- That association cannot be estimated from RDD sample
- Auxiliary information not rich enough to predict *y*
- Without area frame sample, no way to detect the bias

Domain Misclassification

- Know domain for RDD frame
- Area frame: “Is there a working telephone in this household?”
- If no, hand respondent cell phone to talk to CATI interviewer
- 7% excluded at CATI interview because really had telephone
- Area-frame HHs who said they have telephone but did not?



Source: https://commons.wikimedia.org/wiki/File:Motorola_Power_PAK_Bag_Phone.JPG

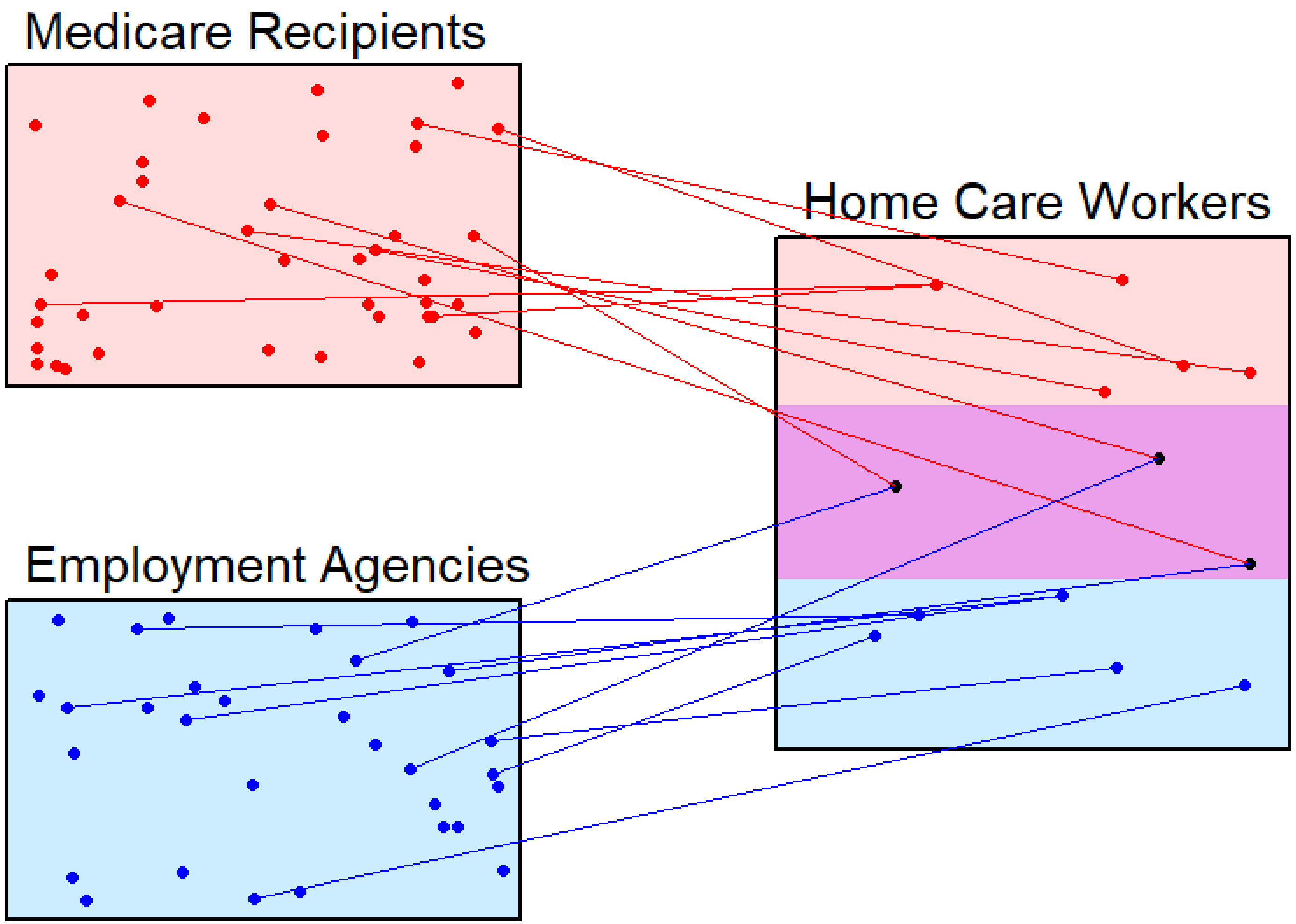
Domain Misclassification

- Even small amount of domain misclassification can lead to bias
- Bias depends on
  - Differences among domain means
  - Misclassification probabilities
- Remedies and diagnostics?
  - Estimate misclassification probabilities from external source (Lohr, 2011)
  - Estimate probability unit i belongs to domain d (Kim & Tam, 2020)
  - Match sample with high-quality probability sample to evaluate frame overlap (Dever, 2018)

Indirect Sampling, Capture-Recapture

- Lavallée & Rivest (2012)
- Individual frames contain links to members of target population
- Alleva et al. (2020) proposed using multiple frames to estimate number of people infected with SARS-CoV-2
  - Frame 1: general population frame
  - Frame 2: persons with verified infections
  - Look at contacts of infected persons in both samples

Indirect Sampling



- Sampling frames contain different types of units
- Units in frames can be linked to multiple units in target population
- Adjust for multiplicity of
  - Links to individual frames
  - Multiple frame links
- Can use to estimate population size

Design of Data Collection Systems

- Hartley (1962) derived n_1, n_2, θ to minimize $V(\hat{Y})$
  - MF design helps when Frame 2 cheap to sample and overlap domain large
- Area frame + RDD frame
  - Biemer (1984), Choudhry (1989), Lepkowski & Groves (1986)
- Nonsampling errors
  - Brick et al. (2011), B. Lu et al. (2013), Lohr & Brick (2014)

Multiple Goals

- Estimate key population quantities with sufficient accuracy
- Assess nonsampling errors from different data sources
- Provide information to improve future data collections
- Be adaptable for future needs
  - Take advantage of new data sources
  - Continuity of time series
  - Will today's data sources be available tomorrow?

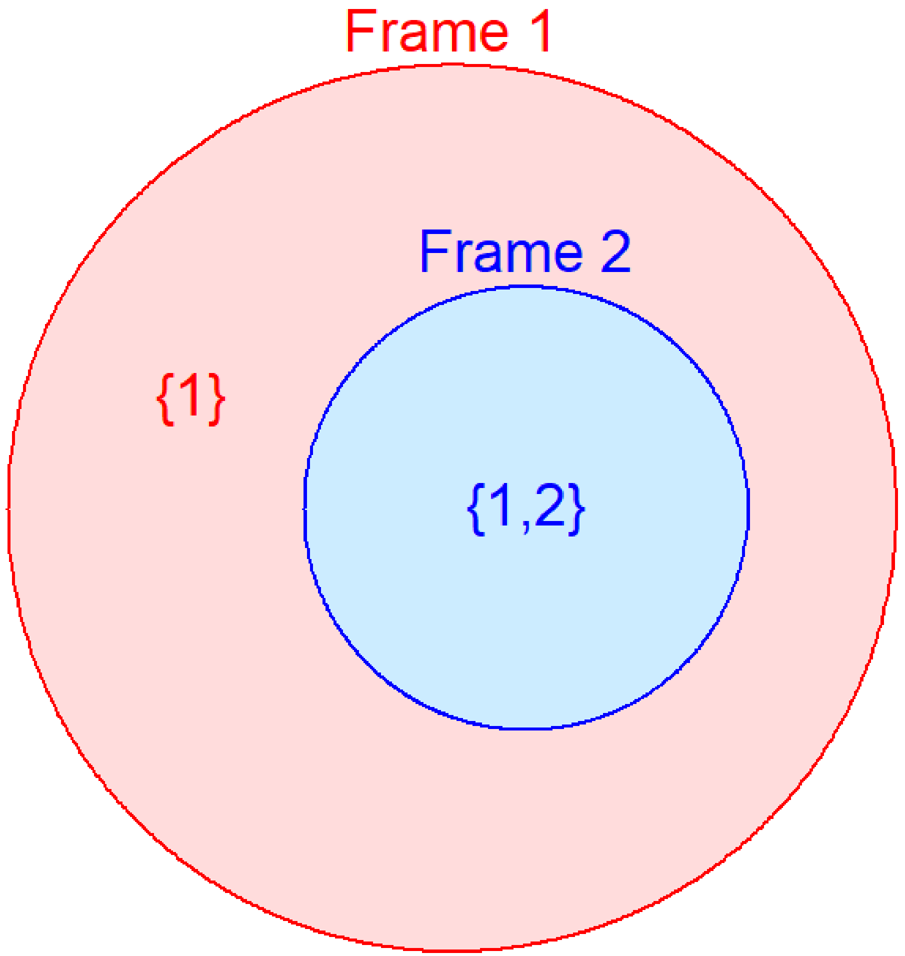
Design Issues

- Quality and stability of data sources
  - Classical MF theory assumes fixed frames
  - What if frame changes over time (web-scraped prices)?
- Measurement of domain membership
  - Collect rich auxiliary data
  - Robust designs?

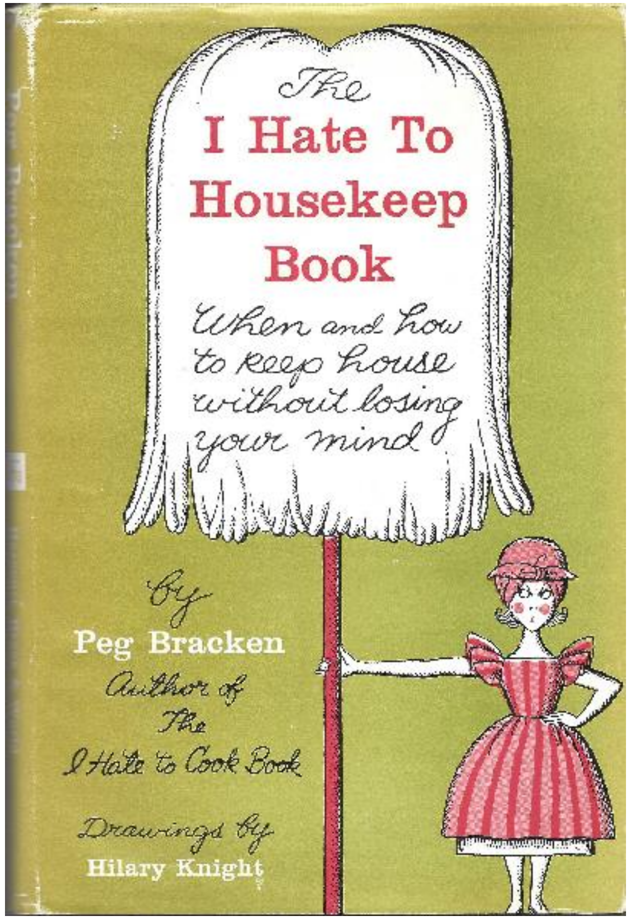
Design Issues

- Does union of frames provide full coverage?
- Relative amounts of information for different domains
  - Greatly unequal weights
  - $w_i = 1, w_i = 6000$
  - Equity

Census of Frame 2



Rule # 3 for Random Housekeepers



Each time you give the house a good going-over, start with a different room.

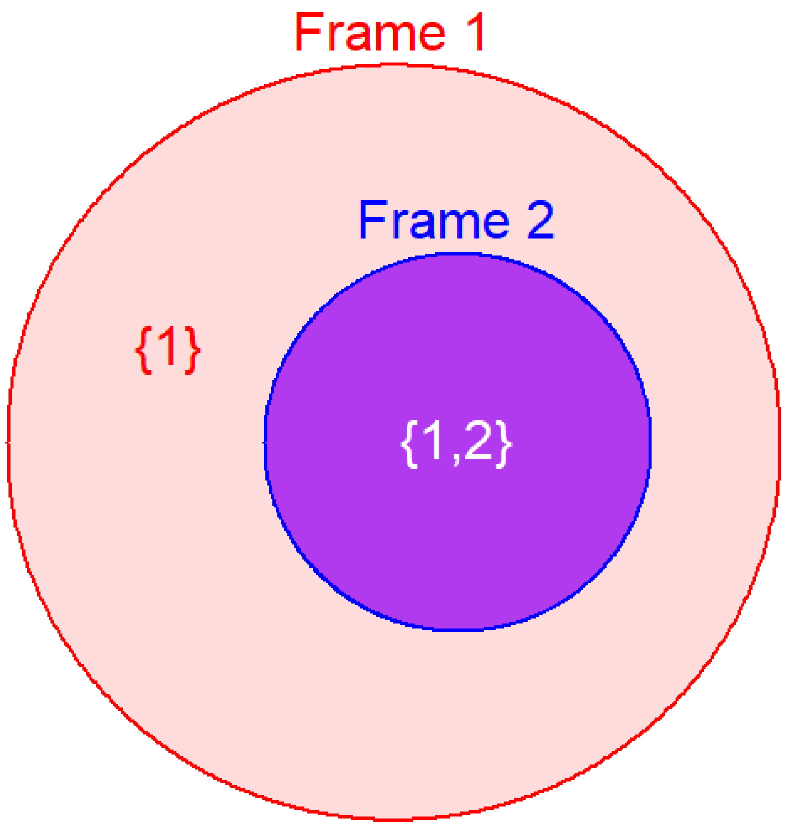
It is quite likely that you'll peter out, you know, after a few hours' slogging, and this rule insures that you will at least peter out in a different place each time. (If you stopped in the *same* place, year after year, for instance just before you got to the back bedroom, you would eventually have to saw it off.)

Peg Bracken (1962)

Design Issues

- Redundancy
  - If we have census of Frame 2, optimal design for Frame 1 screens out Frame-2 units
  - But what if measurement errors?
  - Or bias?

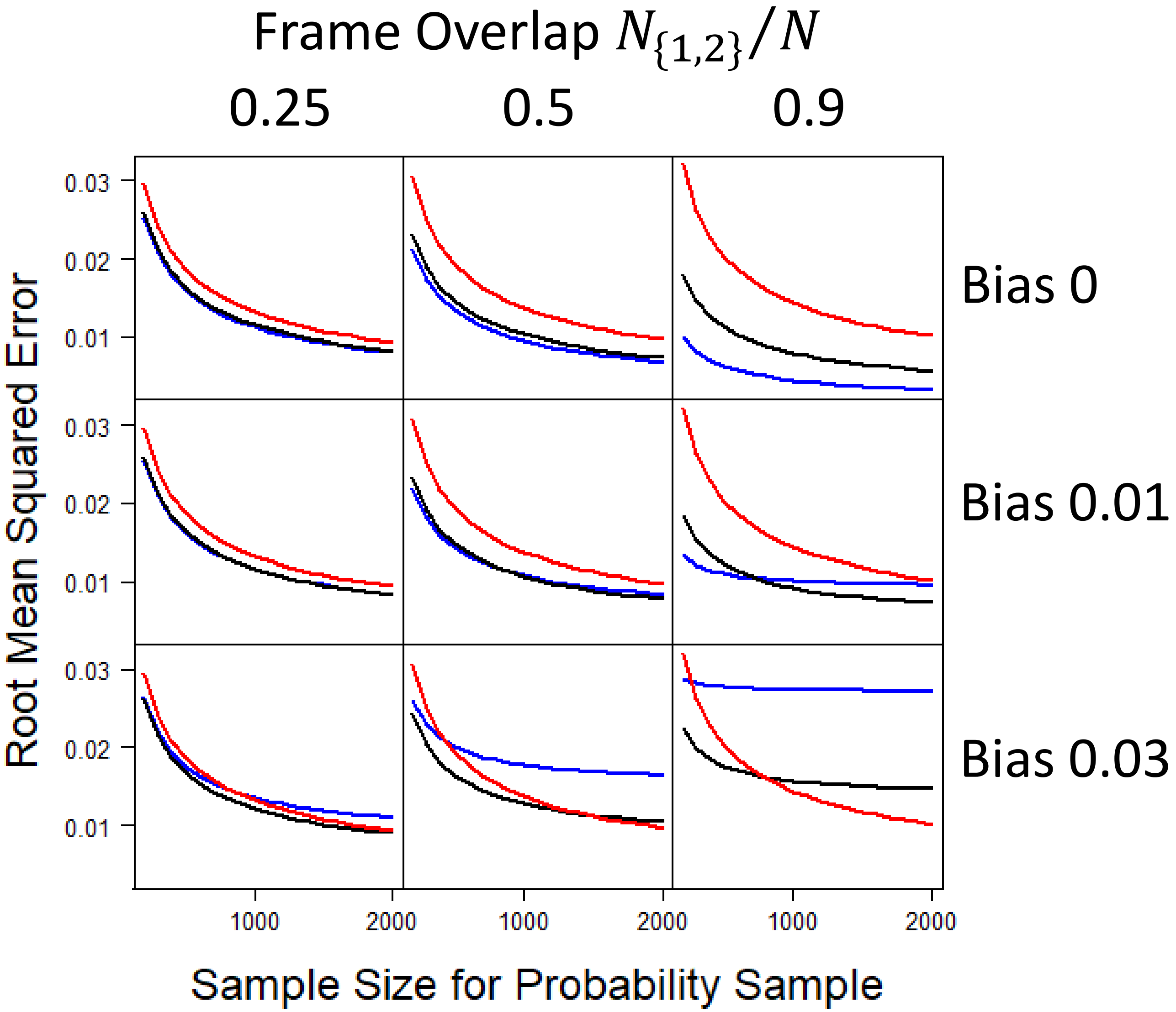
Census of Frame 2



Redundancy

- Frame 1 complete, SRS
- Frame 2 incomplete, census
- $p_{\{1\}} = 0.2$
- $p_{\{1,2\}} = 0.3$ (bias in Frame 2)
- $\hat{Y}_{\{1\}} + \theta \hat{Y}_{\{1,2\}} + (1-\theta)\, \hat{Y}_{\{1,2\}}$

— $\theta = 1$ (only Frame 1)
— $\theta = 1/2$
— $\theta = 0$ (only Frame 2)



Design Issues

- Robustness to design assumptions

  "Do not treat statistical procedures as mechanical operations; be prepared for the unexpected" (Waksberg, 1998)

- Rich auxiliary information
  - Design
  - Domain membership

Multiple-Frame Surveys

- Organizing structure for designing and evaluating data systems
- Waksberg: Sampling statisticians should

“think not only about the specific questions that are asked,
but the broader aspects of these questions:
whether the questions make sense and can be solved,
or whether they should be modified or changed.”

Thank you!

Slides and References
www.sharonlohr.com