Respondent Driven Sampling: Introduction and Applications

Sunghee Lee
University of Michigan

Federal Committee on Statistical Methodology Research and Policy Conference

March 7, 2018

Outline

Introduction

Application

Health and Life Study of Koreans (HLSK)

Summary

Introduction

Respondent Driven Sampling (RDS)

Network Sampling vs. RDS

RDS Inferences

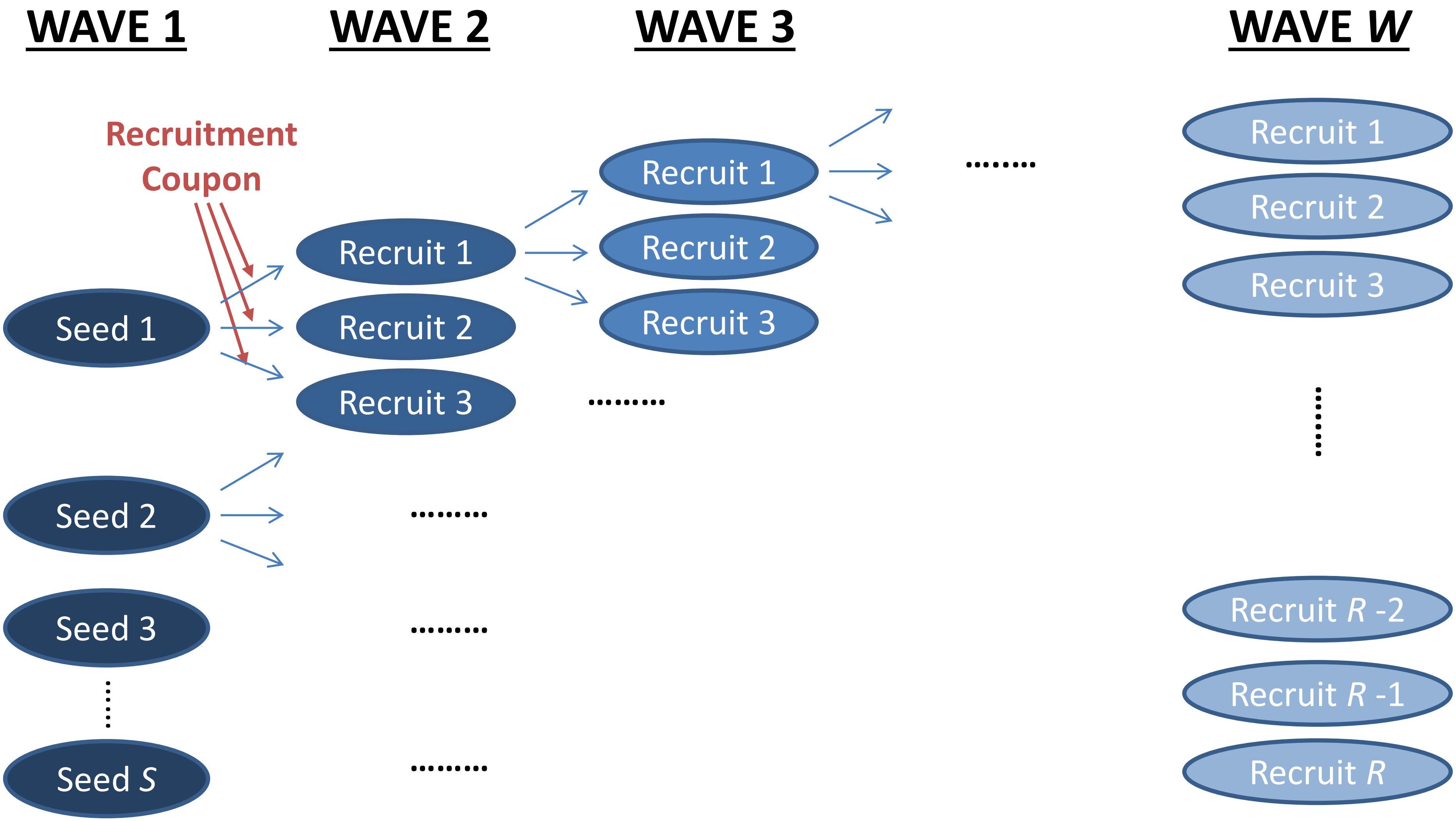
Respondent Driven Sampling – 1

- Growing interest in studying hard-to-reach, rare, elusive, hidden populations
  - HIV at-risk population: Sex workers, IDUs, MSMs
  - LGBT populations
  - Recent immigrants
- No clear and practical solution with probability sampling
  - High screening costs
  - Hesitant to be identified

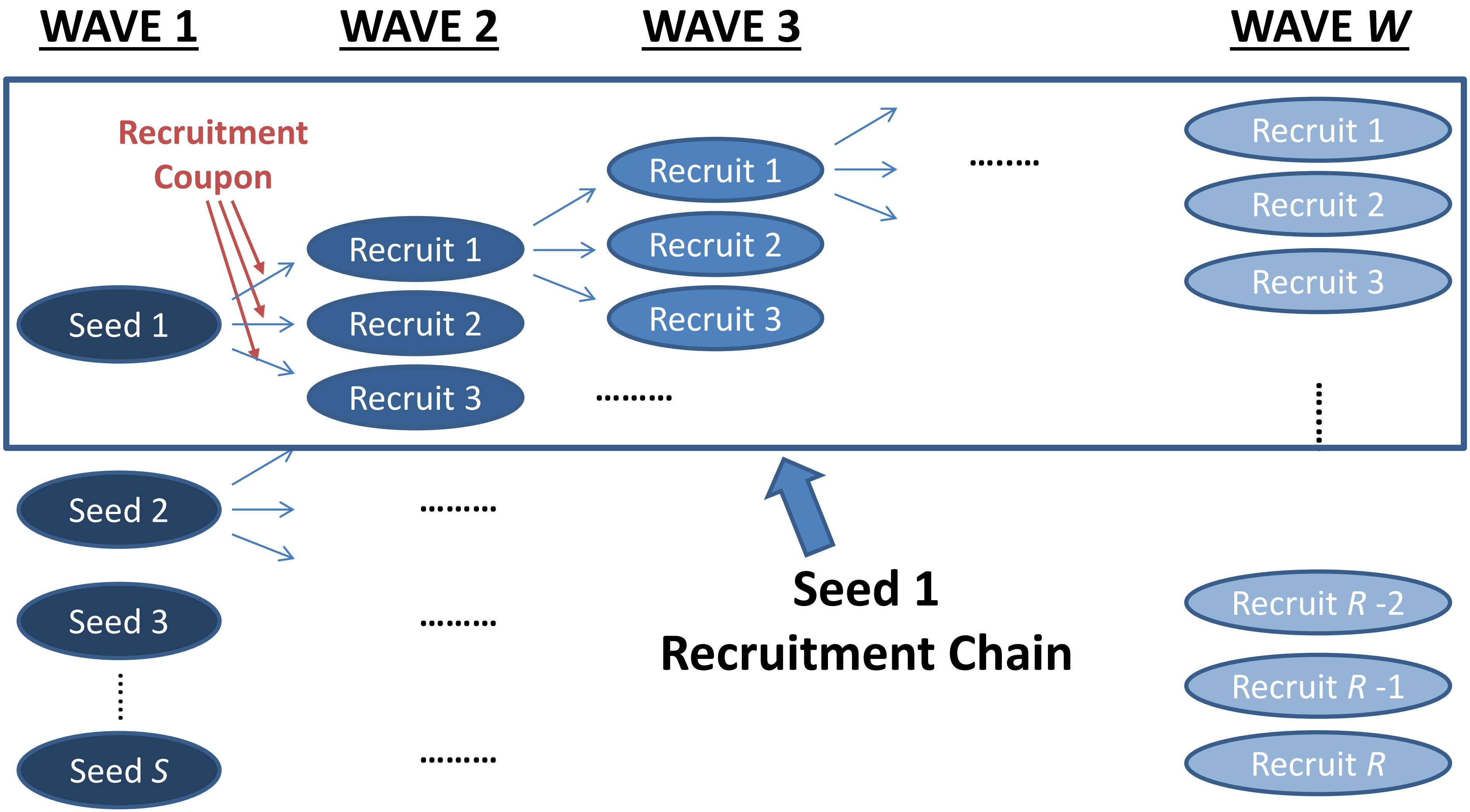
Respondent Driven Sampling – 2

- Proposed by Heckathorn (1997, 2002)
- Popular usage in public health (~$100 million research funds by NIH as of 2011)
- Exploits social networks among rare population members for sampling purposes
  - Sampled members also play a role of a recruiter
  - Incentivized recruitment from own network through coupons and this continues in waves/chains
  - Recruitment assumed to be random within each individual’s network and to follow memory-less Markov chain and reach equilibrium

Respondent Driven Sampling – 3



Respondent Driven Sampling – 4



WAVE 1 **WAVE 2** **WAVE 3** **WAVE *W***

Recruitment Coupon

Seed 1

Recruit 1

Recruit 2

Recruit 3

Recruit 1

Recruit 2

Recruit 3

Recruit 1

Recruit 2

Recruit 3

Seed 2

Seed 3

Seed *S*

Recruit *R* -2

Recruit *R* -1

Recruit *R*

Seed 1

Recruitment Chain

Network/Multiplicity Sampling

- Sirken (1972, 1975)
- Sample from a sample’s network
    - Conduct an interview with a sample
    - Roster eligible kinship members with contact information
    - Sample from the roster

Network Sampling vs. RDS

Similar:

- Rely on social networks

Different:

- Network specification
  - NS: biological siblings, immediate family members
  - RDS: jazz musicians
- Who selects the sample
  - NS: researchers
  - RDS: study participants with coupon
- Selection probability
  - NS: Known
  - RDS: (Mostly) Unknown

RDS Inferences

Issues

1. Nonprobability
   - Within network selection probability may be computed (e.g., # recruits/network size), but
   - Unclear coverage of "network"
   - Measurement error in "network size"
   - With or without replacement?
   - Seed selection probability unknown
2. Dependence
   - Recruiters and recruits are similar
3. None beyond univariate statistics

RDS Inferences: Point estimator

- For binary variables

RDS-I: $\hat{p}_B^{RDS-I} = S_{AB}\bar{\tilde{d}}_A / \left(S_{AB}\bar{\tilde{d}}_A + S_{BA}\bar{\tilde{d}}_B\right)$

RDS-II: $\hat{p}^{RDS-II} = \sum_{i\in S}\left(\tilde{d}_i^{-1} y_{i_i}\right) / \sum_{i\in S}\tilde{d}_i^{-1}$

SS (Gile): $\hat{p}^{G} = \sum_{i\in S}\left(\hat{\pi}\left(\tilde{d}_i\right)^{-1} y_i\right) / \sum_{i\in S}\hat{\pi}\left(\tilde{d}_i\right)^{-1}$

- S_{AB}: proportion of ties (i.e., connections) that cut across A and B (e.g., the proportion of female peers among all peers recruited by all male participants)
- $\bar{\tilde{d}}_A = \sum_{i\in A}\tilde{d}_i / n_A$
- $\tilde{d}_i$ is degree reported by respondent i

  Large degree → high selection probability → small “weight”
- n_A is the sample size of A
- y_i: Outcome variable
- $\hat{\pi}\left(\tilde{d}_i\right)$: estimated population distribution of degrees through successive sampling

RDS Inferences: Sampling Variance – 1

- Naïve estimator
- Direct estimator by Volz-Heckathorn ($\hat{v}^{VH}$)
  - Not usable (requires full network information for all individuals in the population)
  - Only for proportions
  - Assumes first-order Markov process
    - Dependency only between immediate recruiter-recruits
    - Dependency static across chains and waves

RDS Inferences: Sampling Variance – 2

- Bootstrap by Salganik ($\hat{v}^S$)

  1. Group non-seeds by characteristics of recruiter (e.g., recruited by male vs. female)
  2. Randomly sample a seed
  3. Sample a non-seed from the group based on the seed in 2
  4. Sample a non-seed from the group based on the non-seed in 3
  5. Continue this until the bootstrap sample size equals to *n*

  - Only for proportions
  - Assumes first-order Markov process only on the inference variable

RDS Inferences: Sampling Variance – 3

- Bootstrap based on recruitment chains

  1. Randomly sample a seed and preserve its entire recruitment chain

  2. Continue until the bootstrap sample size equals to *n*

  - Can be used for all statistics across all variables
  - Do not assumes first-order Markov process

한인 건강생활 연구
Health and Life
Study of Koreans

Application:
Health and Life Study of Koreans (HLSK)

Funded by the National Science Foundation (GRANT NUMBER SES-1461470)

HLSK

- Targets foreign-born Korean American adults in
  - Los Angeles County
  - State of Michigan
- Web-RDS survey

  http://sites.lsa.umich.edu/korean-healthlife-study/
  - Unique number required for participation
  - Incentive payment through checks
- Target n=800 (currently ~600)
- Benchmarks from American Community Survey

HLSK Formative Research

- 3 rounds of focus group discussions
  - ~30 participants; 2 rounds in Korean and 1 in English
  - Discussion focused on
    - Web surveys

      → URL, Web site contents, etc.
    - Concept of RDS
    - Coupons

      → Up to 2 coupons

      → "Expire" in 2 weeks
    - Level of incentives

      → $20 for main, $5 for follow-up, $0 for recruitment

HLSK Data Collection

- Started with 12 seeds in LA in June 2016
- MI added in November 2016

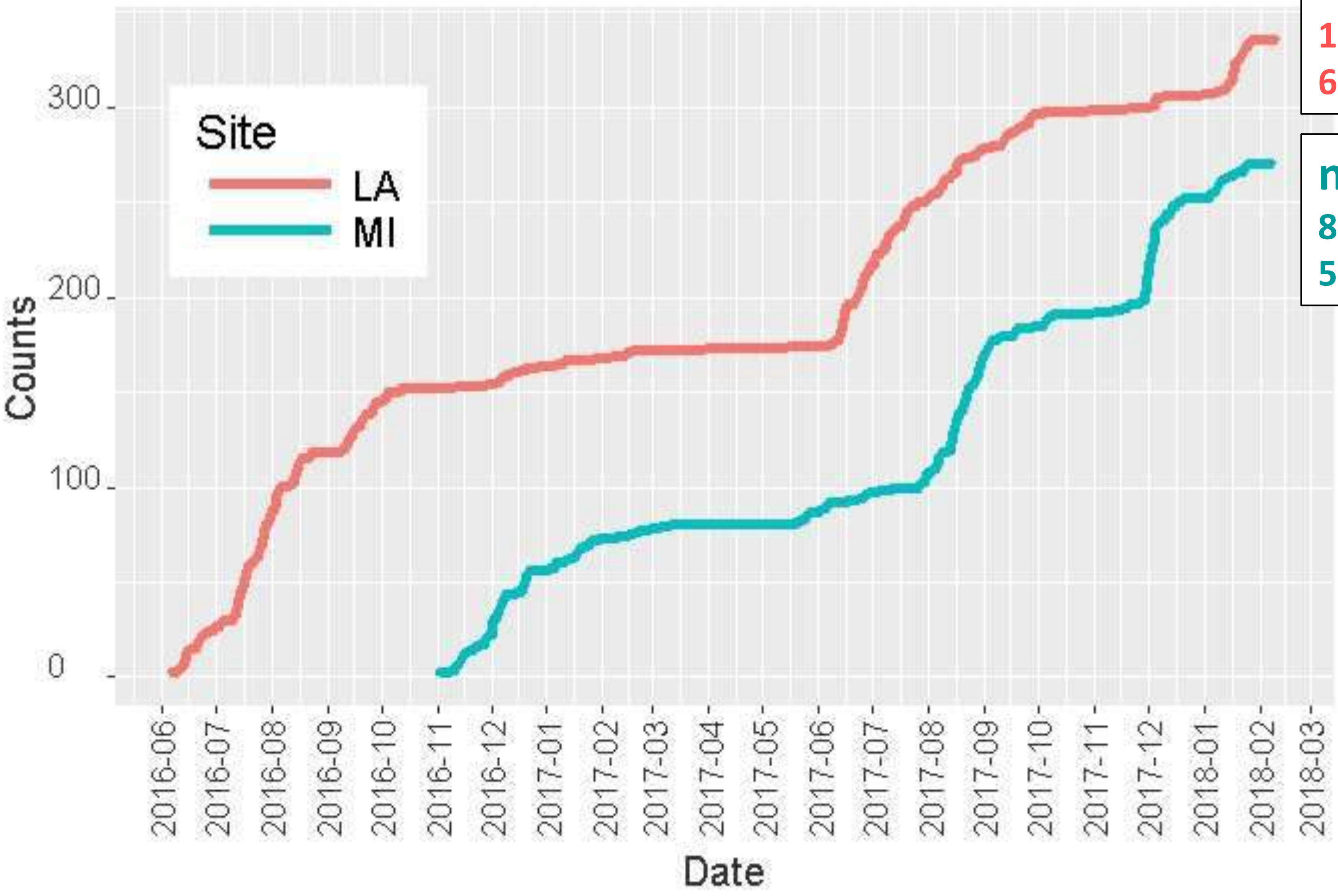
- LA seeds (initially)
  - Recruited through referral
  - Balanced on gender, age, dominant language
  - In-person introduction about the study

→ It became clear the protocols would not work

  - Provide recruitment incentives
  - Add more seeds

HLSK Data Collection Progress

한인 건강생활 연구
Health and Life Study of Koreans



n=336
123 seeds
638 coupons

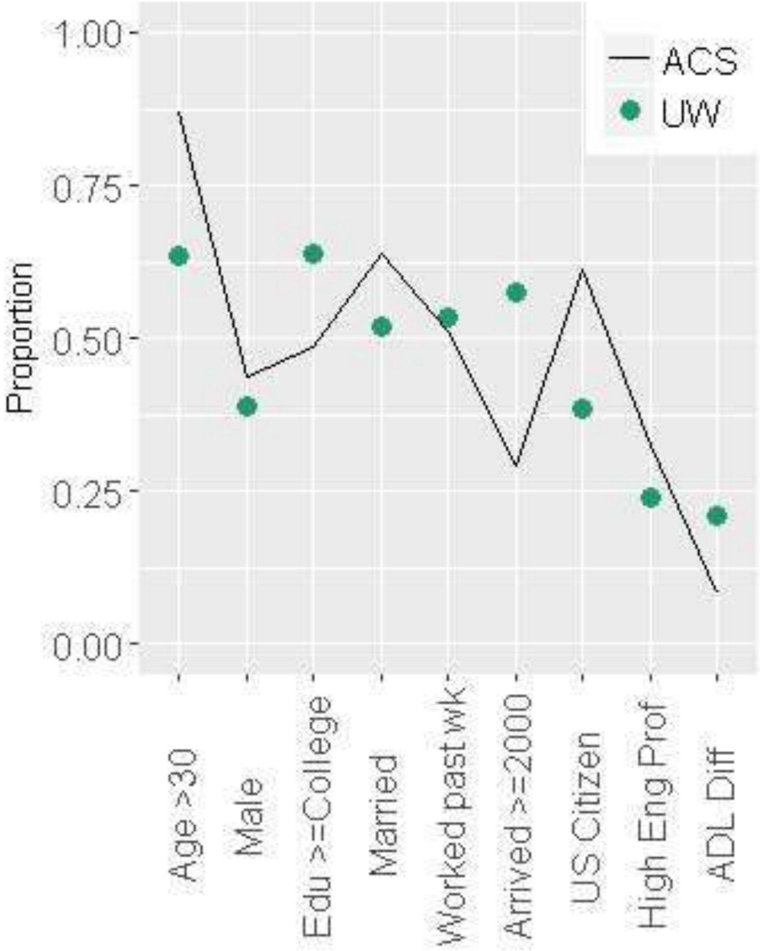
n=270
88 seeds
519 coupons

HLSK vs. ACS – 1

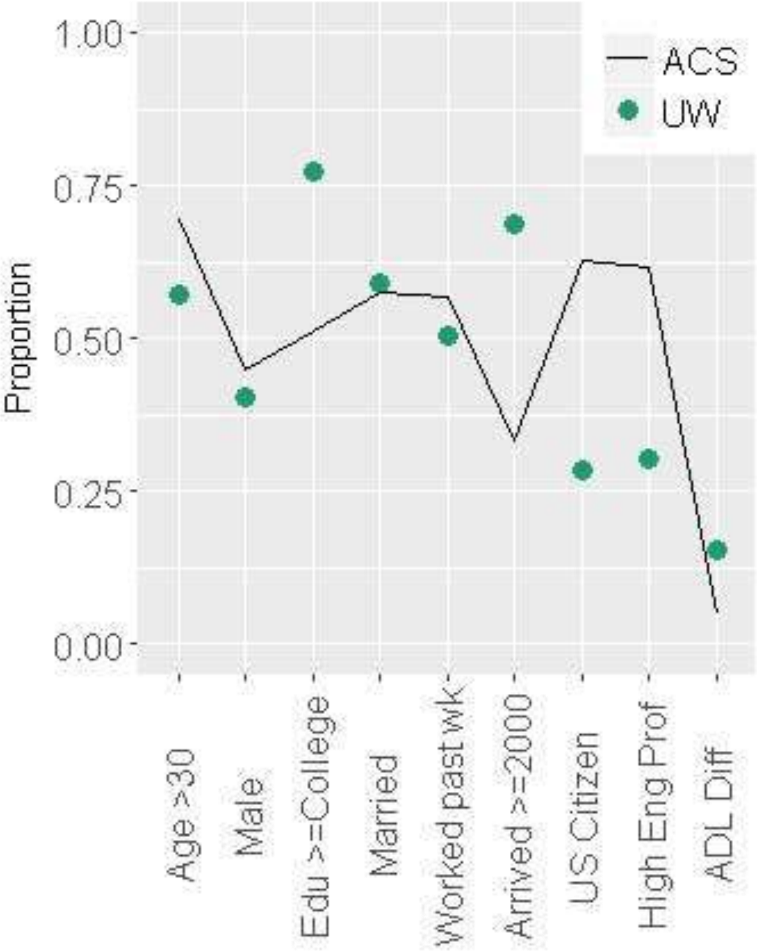
- American Community Survey 2011-2015 data
- HLSK sample estimates
  - Unweighted (UW)
  - RDS-I
  - Weighted: RDS-II
  - Weighted: Post-stratification (PS) by age, sex, educ
  - Weighted: RDS-II + PS

HLSK vs. ACS – 2

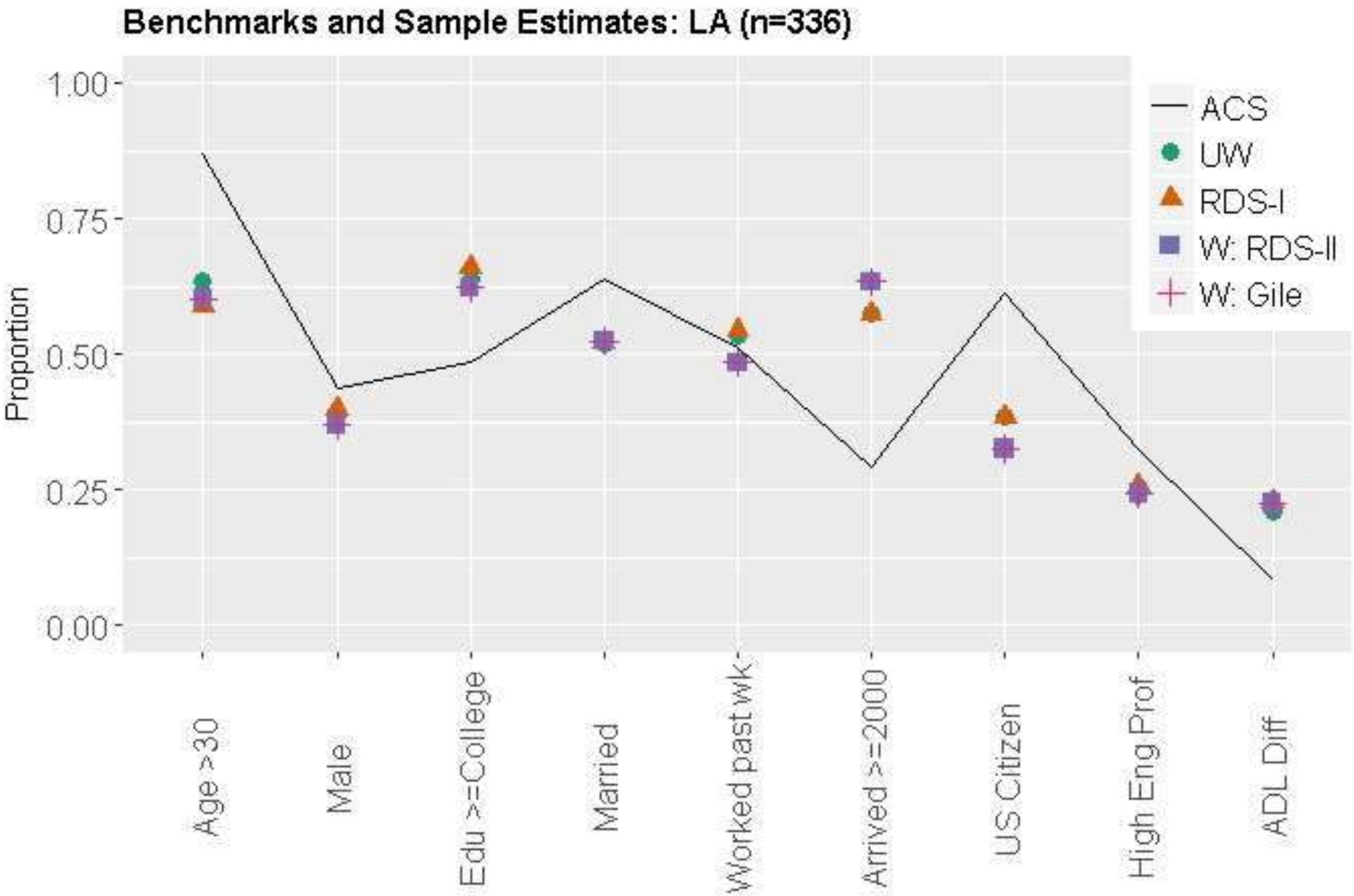
LA Estimates (n=336)



MI Estimates (n=270)



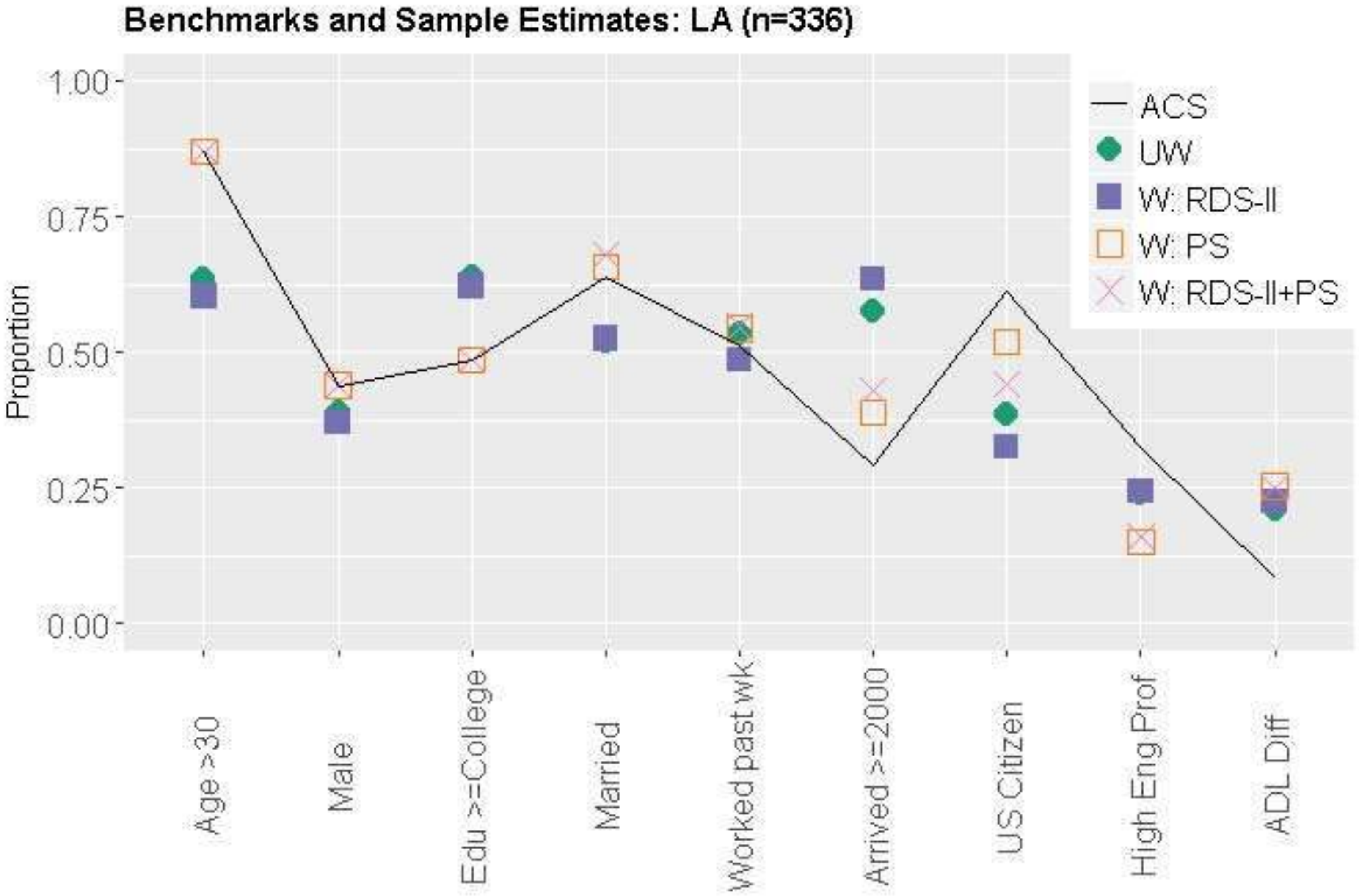
HLSK vs. ACS – 3



Benchmarks and Sample Estimates: LA (n=336)
Proportion
1.00
0.75
0.50
0.25
0.00
ACS
UW
RDS-I
W: RDS-II
W: Gile
Age >30
Male
Edu >=College
Married
Worked past wk
Arrived >=2000
US Citizen
High Eng Prof
ADL Diff

한인 건강생활 연구
Health and Life Study of Koreans

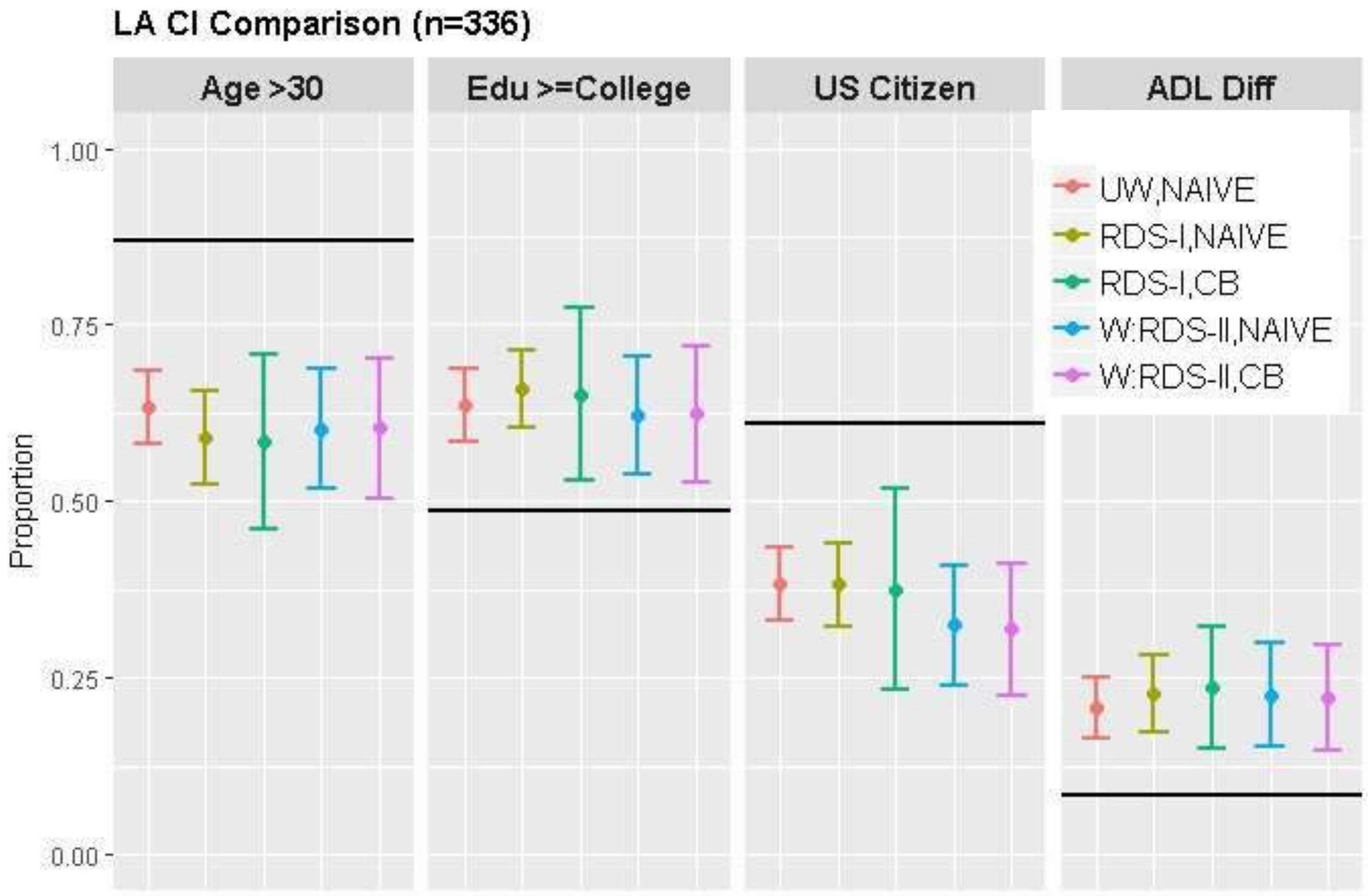
HLSK vs. ACS – 4



HLSK vs. ACS – 5

- HLSK sample estimate CI
  - Unweighted (UW), Naïve
  - RDS-I, Naïve
  - RDS-I, Chain-bootstrap (CB)
  - Weighted: RDS-II, Naïve
  - Weighted: RDS-II, CB

HLSK vs. ACS – 6


LA CI Comparison (n=336)
Age >30
Edu >=College
US Citizen
ADL Diff
Proportion
1.00
0.75
0.50
0.25
0.00
UW,NAIVE
RDS-I,NAIVE
RDS-I,CB
W:RDS-II,NAIVE
W:RDS-II,CB

Summary

What did we learn? – 1

- Non-cooperation is an issue for generating long chains (memorylessness unlikely)
- Had to improvise to make RDS “work”
- Sample size (hence, chain length) is a random variable affected by many (mostly unknown) factors
- Inferences unclear and limited

What did we learn? – 2

- YET, difficult-to sample groups can be recruited
  - highly-educated young recent immigrants
  - low Korean density areas (e.g., MI UP)

Where should we go?

- Non-cooperation is critical for
  - meeting theoretical assumptions (hence, inferences)
  - study design
  - replications of the same study

- Yet to be addressed in the literature and accounted for in inferences

Thank you

sungheel@umich.edu

References

- Heckathorn, D.D. 1997. “Respondent-Driven Sampling: A New Approach to the Study of Hidden Populations.” Society for the Study of Social Problems, 44(2): 174–199.
- Heckathorn, D.D. 2002. “Respondent-Driven Sampling II: Deriving Valid Population Estimates from Chain-Referral Samples of Hidden Populations.” Social Problems, 49(1): 11–34.
- Lee, S. 2009. “Understanding Respondent Driven Sampling from a Total Survey Error Perspective.” Survey Practice, 2(6): 1-6.
- Lee, S., Suzer-Gurtekin, Z.T., Wagner, J. and Valliant, R. (2017). “Total Survey Error and Respondent Driven Sampling: Focus on Nonresponse and Measurement Errors in the Recruitment Process and the Network Size Reports and Implications for Inferences.” *Journal of Official Statistics*, 33(2): 335-366.
- Sirken, M.G. 1972. “Stratified Sample Surveys with Multiplicity.” Journal of American Statistical Association 67: 224–227.
- Sirken, M.G. 1975. “Network Surveys of Rare and Sensitive Conditions.” Advances in Health Survey Research Methods, NCHSR Research Proceedings 31. Hyattsville, MD: National Center Health Statistics.