Discussion:
Techniques to Identify and Find Small Populations

Krista J. Gile

University of Massachusetts, Amherst

January 18, 2018

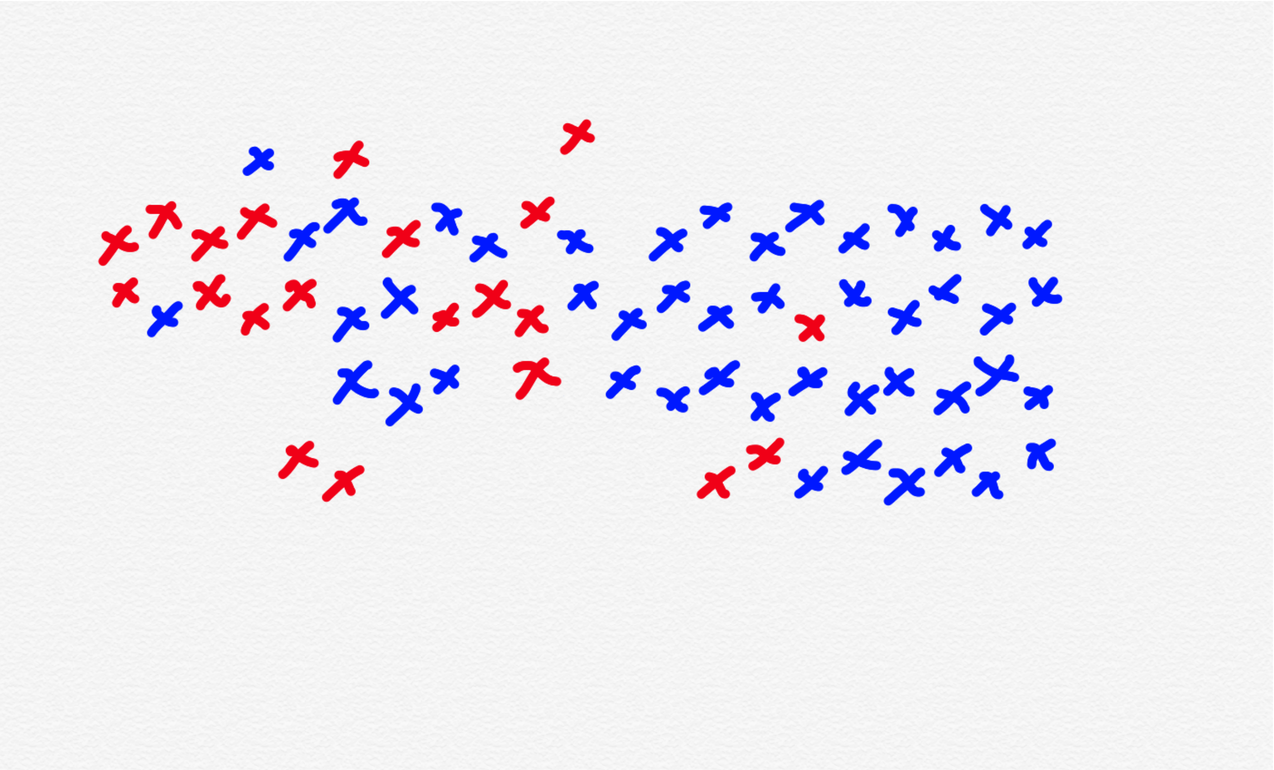
Goals:

- Identify and Find Small Populations for Health Research
- Make statements about the whole small population
  - Population Size
  - Population Proportions
  - Associations/multivariate
- Quantify uncertainty about the small population
  - Confidence Intervals
  - Testing: are difference over time/location/population real?

Methods:

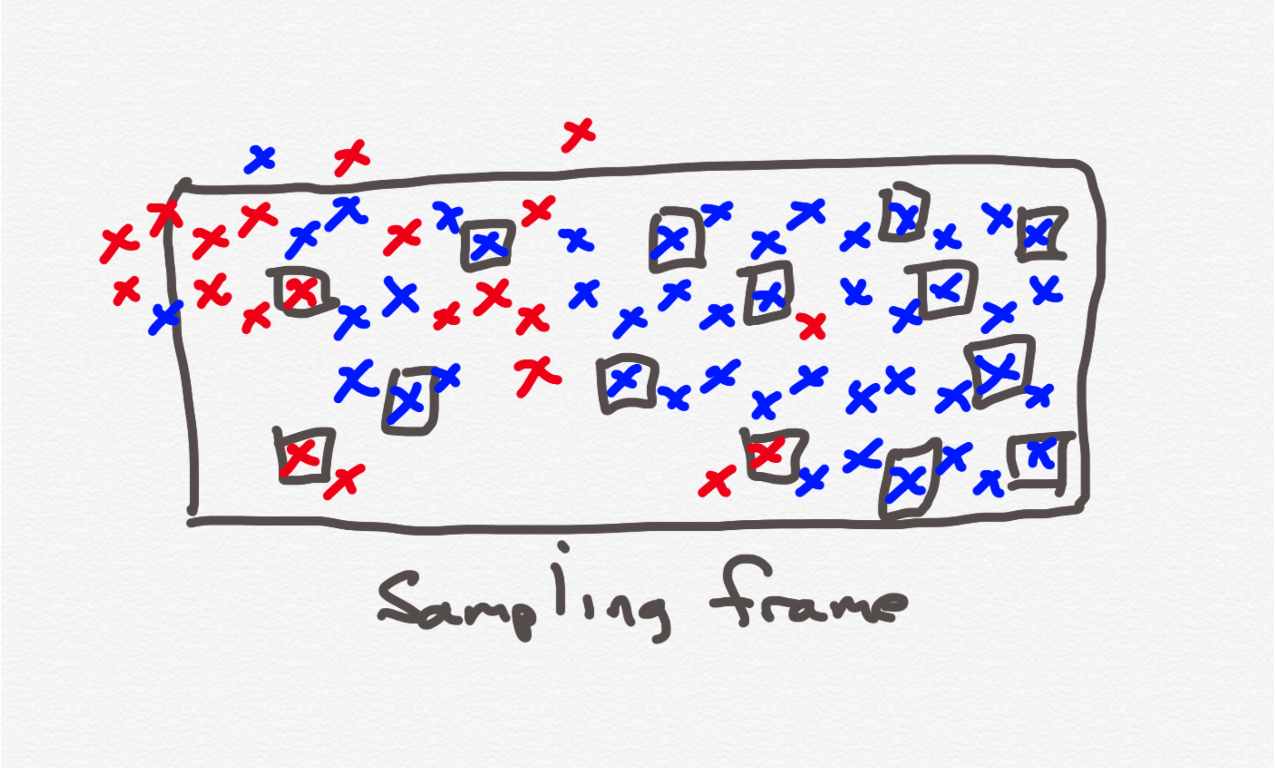
- Probability Sampling
- Respondent-driven Sampling
- Venue-based sampling
- Online sampling

Illustrations: Population



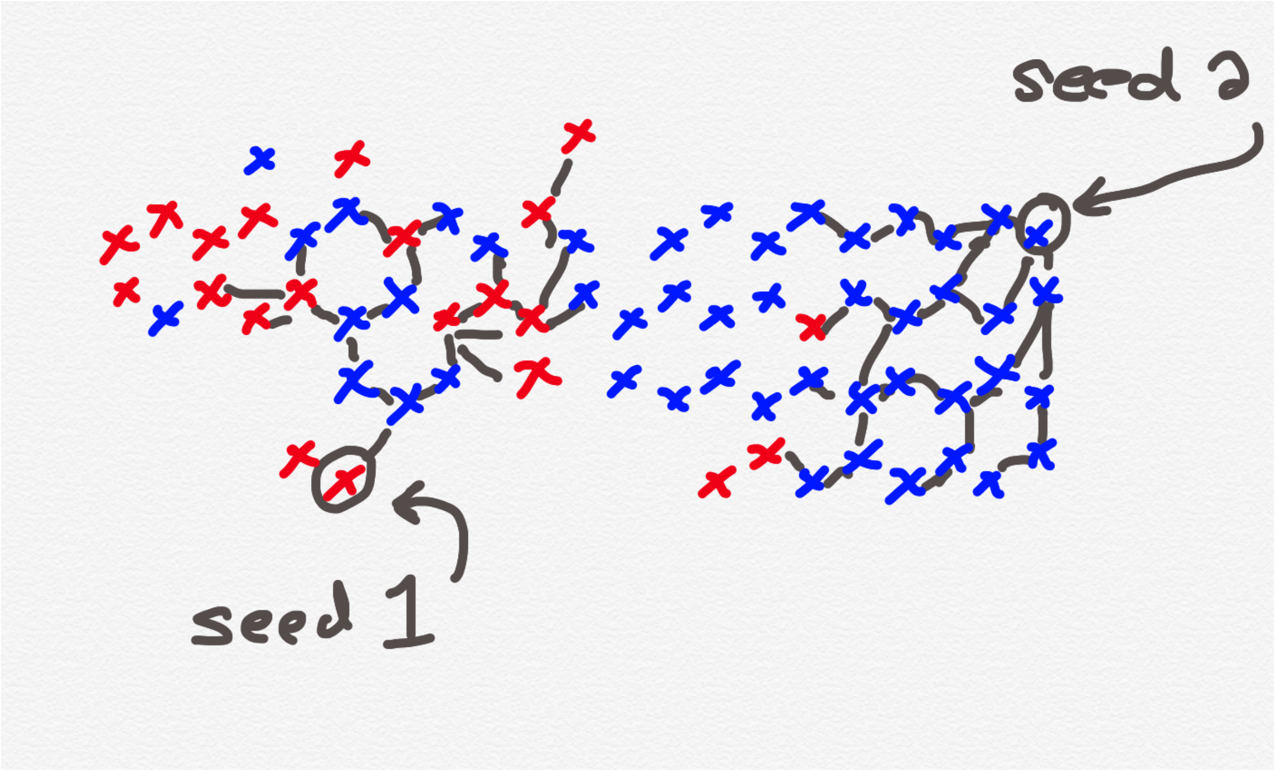
Red = high-risk (LGBTQI), Native Hawaiian (AANHPI), women (homeless)

Illustrations: Probability Sampling



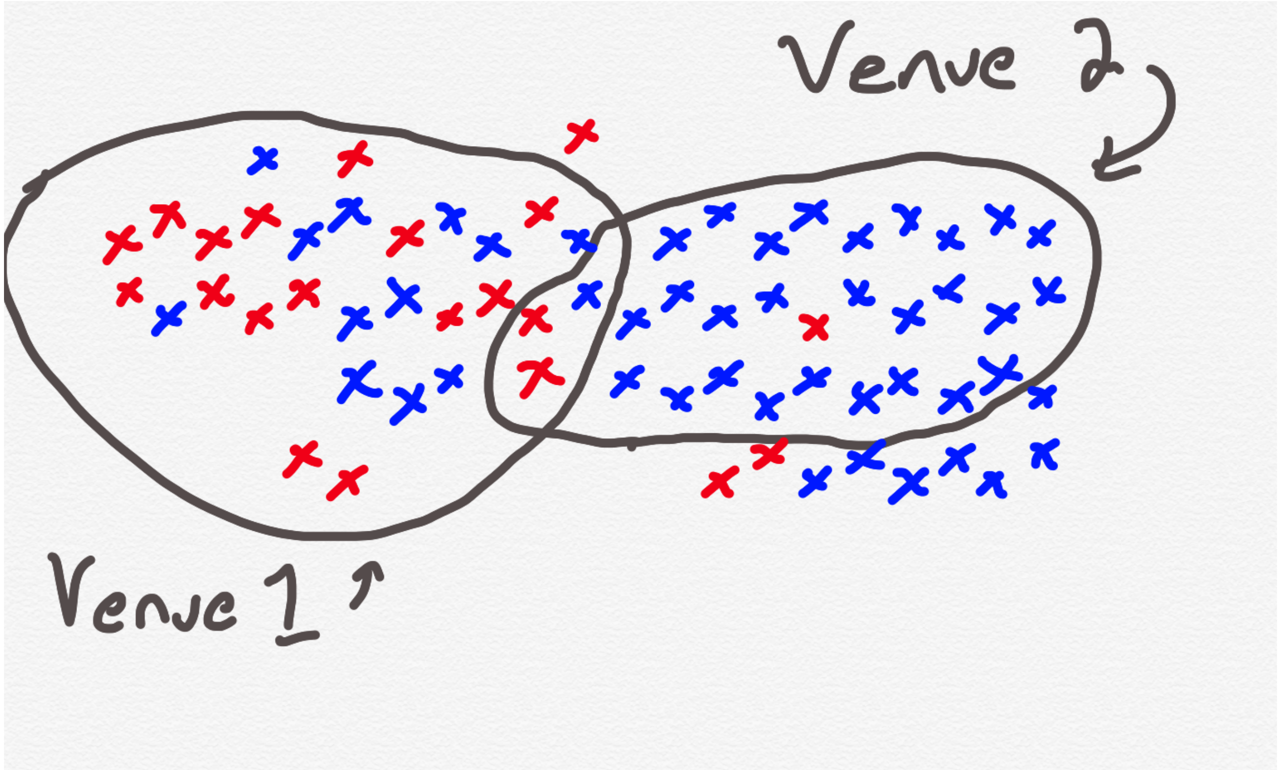
Who is excluded?

Illustrations: Respondent-driven Sampling



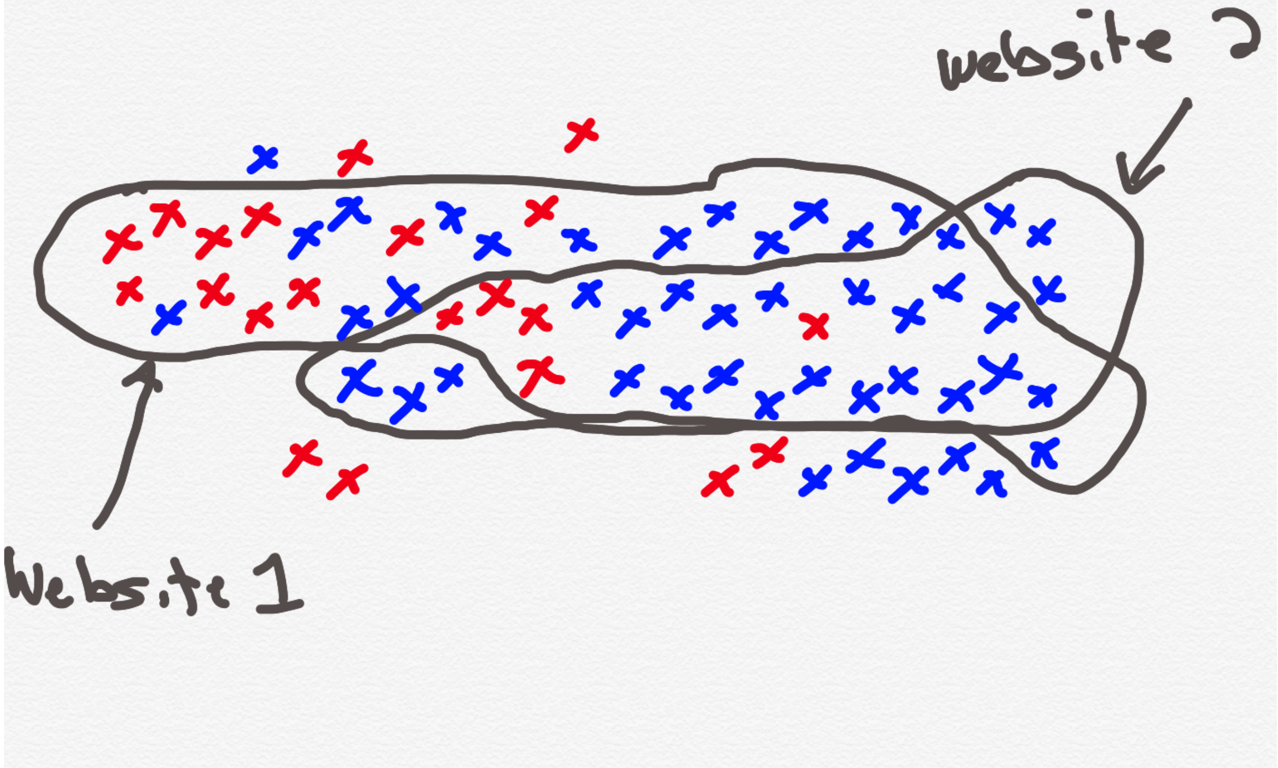
How are seeds found? Who is recruited?

Illustrations: Venue-based Sampling



Sampling unit is venue-time. Who is excluded from venues? Who is over-represented?

Illustrations: Online Sampling



Who will see the ad? Who will click?

Points of Comparison:

- Formative research
- Role of rapport
- Sampling frame (and who is excluded)
- Differential sampling rates within frame
- Knowledge of differential sampling rates
- Sensitivity to decisions of participants
- Methods for statistical inference (point estimates, confidence intervals)
- Dependence between sampled individuals
- Populations not suitable

Formative Research

Probability	Extensive: Sampling frame of full population of interest
Respondent-driven	Moderate: Choose diverse seeds, set up study site
Venue-based	Extensive: times and locations of congregation, arrange for surveys
Online	Moderate: Identify online locations of community

The more you know, the more you can learn.

Role of rapport

Probability	To identify sampling frame, get participation
Respondent-driven	Trust: Find seeds, get participation, get recruitment
Venue-based	Trust: Find times/locations, get access, get participation
Online	Find websites, draw participation

Getting truth requires trust. Want to ask the right question, and get an answer.

Sampling Frame (and who is excluded)

Probability	Whoever falls within the frame
Respondent-driven	Connected to (large component of) social network
Venue-based	Frequent targeted venues
Online	Visit targeted sites

We can only learn about who we can find.

Differential sampling rates within frames

Probability	Controlled by design
Respondent-driven	Based on network connections
Venue-based	Based on venue use
Online	Based on website use and clicking

Who is over-represented? Under-represented?

Knowledge of differential sampling rates

Probability	Known by design
Respondent-driven	Ask number of ties (some limitations)
Venue-based	Ask about use (controversy, many methods)
Online	Ask about online use (how to assess tendency to click?)

Can we adjust for over/under representation?

Sensitivity to decisions of participants

Probability	Non-response to direct contact
Respondent-driven	Who gets coupons, non-response to recruiter
Venue-based	Non-response to physical interaction
Online	Non-clicking

We can ask, but we can’t control or coerce behavior.

Methods for statistical inference
(point estimates, confidence intervals)

Probability	Excellent: Gold Standard
Respondent-driven	Several available, dependent on assumptions
Venue-based	Venue-time, person weights, No consensus method
Online	Post-stratification? No consensus method

What can we say beyond the people we actually see?

Dependence between sampled individuals

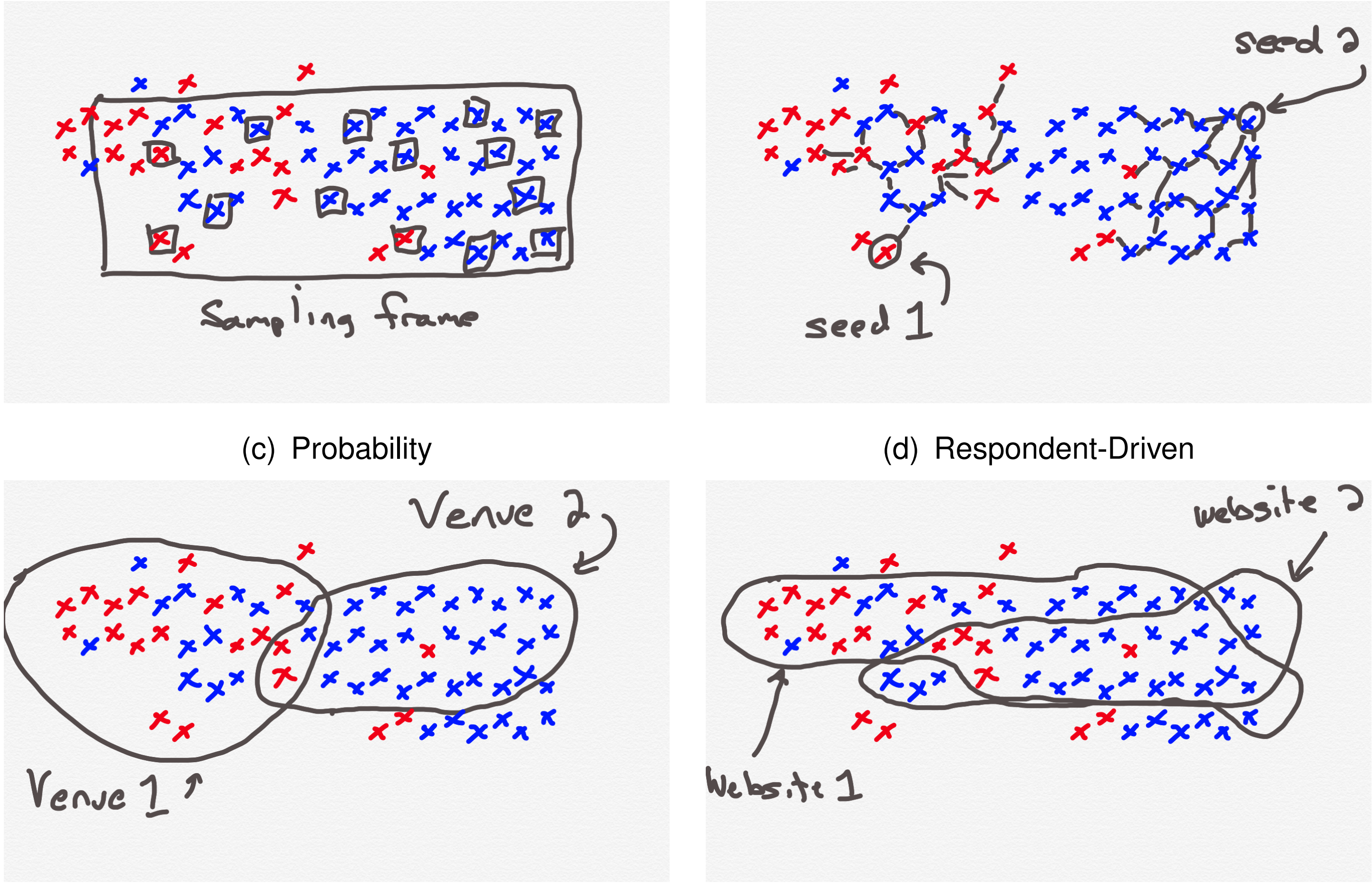
Probability	Low, by design
Respondent-driven	High, by pairs
Venue-based	Moderate, but many per sampled venue-time
Online	Low

How much new information does each person add?

Populations not suitable

Probability	Cannot form suitable sampling frame (transgender)
Respondent-driven	Not well connected by network (AANHPI, cross- group)
Venue-based	Do not congregate in known/accessible venues (children with autism)
Online	No online community/low internet usage/unlikely to click (homeless)

Four Methods



(c) Probability

(d) Respondent-Driven

(e) Venue

(f) Online

Major Advantages

Probability	Straightforward valid inference
Respondent-driven	Reaches unknown parts of population, approximate valid inference
Venue-based	Valid (non-person-based) sampling frame
Online	Ease of implementation, cost

Major Concerns

Probability	Depends on good sampling frame
Respondent-driven	Depends on well-connected population and respondent behavior
Venue-based	Unequal representation of individuals, may exclude some
Online	Depends on clicking

Conclusions

- Probability Sampling is ideal if possible, if the sampling frame is adequate.
- Respondent-driven sampling provides methods for treating the sample as probability sample, relies on strong assumptions
- Venue-based and online don’t allow for inference (uncertainty, intervals)
- Venue-based probability sample on venue-times, also unequal individual rates
- Venue-based less sensitive to non-response than online, but the sampling frame may not be as complete.

Discussion

- What is the best we can do for sampling weights for venue-based sampling? Venues? Frequency?
- How can we know about who we are missing in an online sample?
- Can we leverage multiple of these methods in the same population?
  - Combine venue-based and online sampling, treating websites as additional venues?
  - Use methods as multi-list (capture-recapture) methods for population size, characteristics.
- Sensitivity of self-identification: LGBTQI, homeless, some non-white US populations
- Amazon and Political Campaigns can be ‘greedy’: results are more important than fairness. Researchers and health services need to be more careful.