Probability Sampling Methods for Small Populations

Marc N. Elliott, RAND

Outline

- Purpose of sampling/probability samples
- Screening, including the use of inexpensive modes such as mail.
- Disproportionate stratification at both the cluster and element level –
- Network sampling

Purpose of Sampling

- To make valid and reliable inferences about characteristics (parameters) of a large population of interest from a smaller sample
- A single survey may be used to make inferences about multiple parameters for multiple subpopulations
- Parameters can be means, proportions, regression coefficients, among other things
- Statistics are calculated on the sample
- Sampling links the sample to the population

Two Main Types of Samples

- **Probability sample**
  - **You control or otherwise know the (nonzero) probability of inclusion for all members of the population (don’t have to be equal)**
  - **Statistical inference valid**
  - **Mail survey from list, RDD, etc.**
- Judgment/convenience sample
  - Volunteerism or unsystematic approach makes probabilities of inclusion unknown
  - Statistical inference not valid
  - Mall intercept, inbound calls, etc.
  - Sometimes the only way to get sample

Many Types of Probability Samples

- Simple random sample (SRS)
- Systematic sample
- Stratified sample
- Cluster sample
- Etc.

Some Probability Sampling Approaches for Rare Populations

- Screening with initially inexpensive approaches
- Oversampling at the cluster level
- Oversampling from an incomplete frame with substantial coverage of the population
- Network sampling

Screening

- Start with inexpensive outreach (mail or add-on to existing phone survey) in large probability sample
- Use high-sensitivity/low specificity screener
- Follow up with more expensive and more specific measurement in stages 2+
- Application-Estimate Prevalence of Rare condition: Interstitial Cystitis / Painful Bladder Syndrome (IC/BPS)

Three-stage sceening for IC/BPS

- Population survey by telephone to identify women with IC/PBS and interview them

Stage 1: Screen households for a women with bladder symptoms as part of regular commercial omnibus survey – ORC Caravan

Stage 2: RAND SRG calls households that screened positive in Stage 1 to speak with woman and screen specifically for IC/PBS

Stage 3: If positive, complete disease impact interview Collect urine sample and test

Application of oversampling at the HH level: Cambodians in Long Beach, CA

- A mental health study of Cambodians age 35-75 who lived in Cambodia under Pol Pot (1975-1979) and immigrated to the US prior to 1993 was fielded Oct 2003-Feb 2005 (see Marshall et. al., JAMA, 2005).
- Area sample from the largest and highest-density Cambodian refugee community in the US
  - Five contiguous Census tracts with 15,000 HHs
  - 12% of households in the defined area contained an eligible resident.

Expert Classification of HHs

- In the first stage, randomly sampled 37% of all blocks (80 of 217 Census blocks)
- A local community expert rapidly classified all 5555 individual residences as *likely* (18%) or *unlikely* to contain an eligible resident
- Used externally (publicly) observable cultural indicators
  - Footwear outside front door
  - Buddhist altars on front porch
  - Lemon grass, bamboo, or banana trees in the front yard
- This could be done at a brisk walk, without breaking stride, so was quick and inexpensive

Using the Expert Classification

- Select a sampling rates for the two strata to reduce costs while meeting ESS targets
  - Fewer sampled from *unlikely* HH stratum results in:
    - Greater eligible yield per sampled unit and lower costs
  - Greater sampling weights for eligibles in the strata and smaller ESS for a given sample of eligibles
  - Sample all the households in *likely* HH strata

Cambodian Example

- *a*=12.1% eligibles
- Classified 18% of HHs as likely
  - Include all in screening sample
  - 58% contained eligibles
- Classified 82% of HHs as unlikely
  - Included 25% in screening sample; weights of 4 for eligibles (*s*=4)
  - 2% contained eligibles
- Sensitivity=86.4%; Specificity =91.4%
- DEFF=1.26
- Screening costs reduced 57%; Screening costs per ESS reduced 46%
  - Response rate was 88%, so went from 9.4 to 4.0 approaches per complete

Incomplete Non-Geographic Frames as Oversampled Strata

- Primary sample is a standard probability sample
- Supplement-to be treated as an oversampled stratum-is inexpensive but incomplete
- EXAMPLE-Surveying Chinese Americans (1% of US)
- Run name and address lists through a surname/address technique for generating racial/ethnic probabilities
  - or commercial ethnicity lists based on surnames and phone books
- Lower cost per case
- Can be combined with a representative sample through weighting
- Cases that appear on both lists (on the incomplete frame) are oversampled relative to cases accessible only from 1 source

Incomplete Non-Geographic Frames: Requirements

- Lists must have good coverage (sensitivity) to limit design effects
- Lists must have good specificity/positive predictive value to save screening costs.
- Feasible for Chinese
  - Surname lists have 70% sensitivity at a positive predictive value of 75%
  - Commercial lists have about 33% coverage
- Poorer coverage causes design effects to rise rapidly, so the supplement should be no larger than the core sample of the target group
  - E.g., 200 Chinese completes from surname lists to supplement 200 core completes
- To be cost-effective, total cost per complete in the target population must be at least 30-40% less in the listed sample than the traditional sample

Network Sampling-1

- In most household sample surveys, respondents report only about themselves or their households
  - One-to-one reporting rule
- In network sampling, households are permitted report about themselves and other households to which they are linked
  - One-to-many reporting rule
  - Higher screening yield
  - Less expensive to identify eligible households
- Rarely used but a probability approach
- Key is knowing how many links exist so you can calculate the probability of someone being reached by any means

Network Sampling-2

- Start with a general probability sample
- Screen each selected household/member for target population
- Ask each household about defined links where they would know eligibility-e.g., adult sibling households : presence of target population member- I
  - If Yes
    - Request contact information for sibling household
    - Or take direct report
    - Determine how many other households *could have* reported the identified sibling household