Provider-Based Sampling for the National Children's Study: Some Thoughts

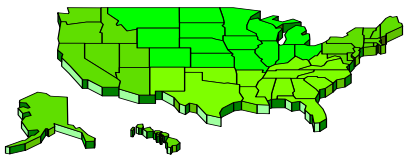
Michael R. Elliott

Department of Biostatistics, Program for Survey Methodology, University of Michigan

Michigan Alliance for the National Children's Study (MANCS)

Current Study Design:

Multi-Stage Area Probability Sample



Nation

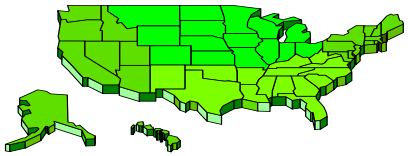
Sample of PSUs

Sample of Segments

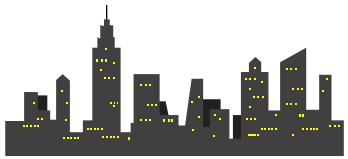
Sample of Housing Units

- **3,141 PSU's (primary sampling units: counties)**
- **105 PSUs (by PPS)**
  - 79 metropolitan and 26 non-metropolitan
- **10-15 segments per PSU (approx. equal size)**
- **250 newborns per year per PSU (16-25 per segment)**

Proposed Study Design: First Stage Area Probability Sample, Second Stage Physician Practice Sample



Nation



Sample of PSUs

Sample of Practices

Sample of Births

- **3,141 PSU's (primary sampling units: counties)**
- **105 PSUs (by PPS)**
  - 79 metropolitan and 26 non-metropolitan
- **25 practices per PSU (by PPS)**
- **10 newborns per year per practice (by simple/systematic random sample)**

Sampling Providers: PPS sample

- Birth certificates: they aren’t just for Donald Trump and President Obama!
- Birth certificates contain the name of the person present and responsible for the delivery of the baby.

Sampling Providers: PPS sample

- Can use birth certificate data to obtain a <u>sampling frame</u> of providers that has an <u>associated size measure</u>.
  - Use to obtain a probability proportional to size (PPS) sample of providers in the same way that a PPS sample of PSUs (counties) was obtained.
  - Sampling a constant number of births from each provider yields a equal probability of selection (epsem) sample of births from each PSU:

$$\pi_{ij} = \pi_{j|i}\pi_i = \frac{n}{N_i} \times \frac{N_i}{N} = \frac{n}{N}$$

    - π_{ij} is the probability that the *j*th birth at the *i*th provider is sampled
    - $\pi_{j|i}$ is the probability that the *j*th birth at the *i*th provider is sampled given the *i*th provider is sampled,
    - π_{ij} is the probability that the *i*th provider is sampled
    - n is the number of births sampled at the provider
    - N_i is the number of PSU births at the ith provider
    - N is the number of births in the PSU

Sampling Providers: PPS sample

- At a high level the story is done, but there are a several twists and turns that need to be considered
  - Some providers will have less than *n* PSU births.
    - Sample size will be too small.
    - Epsem design destroyed: use weights
  - Obtaining a subsample when births are greater than *n*.
  - Refusals at the provider and mother level.
  - Would prefer to sample <u>practices</u>, not <u>providers.</u>

Providers with less than *n* PSU births

- A provider can contribute no more births than he or she actually has – if the number "required" is $n > N_i$, then only N_i will be contributed.
- In order to obtain the required sample size in the PSU, increase the number required where the number of births allows to $\tilde{n}$, where $\tilde{n}$ is estimated by determining $s(\tilde{n}) \geq n$, where $s(\tilde{n})$ is the expected mean number of births sampled:

$$s(\tilde{n}) = \sum_i \left(\frac{N_i}{N} \right) \left[\tilde{n} I(N_i \geq \tilde{n}) + N_i I(N_i < \tilde{n}) \right]$$

Providers with less than *n* PSU births

- With this twist, we lose the epsem design, but the representativeness of the sample can be reestablished using sampling weights.
- Probability of selection is now $\pi_{ij} = \pi_{j|i}\pi_i = \dfrac{n_i}{N}$ where $n_i = \begin{cases} \tilde{n} \text{ if } \tilde{n} \le N_i \\ N_i \text{ if } \tilde{n} > N_i \end{cases}$.
- Sampling weight is given by $w_{ij} = \dfrac{1}{\pi_{ij}} = \dfrac{N}{n_i}$
  - Note weight is constant within a given provider.

Providers with more than *n* PSU births

- Usually we will have the opposite problem: large providers with far more than *n* births in the PSU each year.
  - Some type of random sampling scheme must be employed.
- Sample women at <u>first visit</u> to provider.
- Sample based on time: e.g., if $\frac{\tilde{n}}{N_i} = \frac{1}{4}$, sample one week per month, with week (ideally) chosen randomly.
  - Adjust as needed depending on actual production rate
  - May need to account for seasonality in births if severe
- Other random sampling mechanisms can be employed.

Dealing with non-response

- The above assumed perfect response.
- Although our experience with provider-based recruiting is very positive, response will not be perfect.
- Two options:
  - Adjust number of providers and mothers sampled to be

$m^* = m / r_m$ and $n^* = n / r_n$, where r_m is the expected response rate for the providers and r_n is the expected response rate for the mothers. (Experience thus far in Wayne County MI suggests $r_m = r_n = 0.8$.)

  - Replace refusing providers and mothers with another sampled provider/mother (PPS in case of provider), but keep track of refusals so that response rate can be determined.

PPS sample of practices

- Vast majority of providers will be situated in a practice with other providers → far easier to work with a sample of practices, not providers.
  - Unfortunately, this data not contained in birth certificate.
- In small PSUs, network of practices might be worked out directly, and provider measures of size summed to obtain practice measure of size.
  - Draw PPS sample of practices directly as proceed as with PPS sample of providers.

PPS sample of practices

- In large PSUs – actually, in PSUs of any size – can obtain the probability of selection of a given practice after sampling the providers <u>only with the names of the providers in the sampled practices</u> – no need to work out the full network of providers:

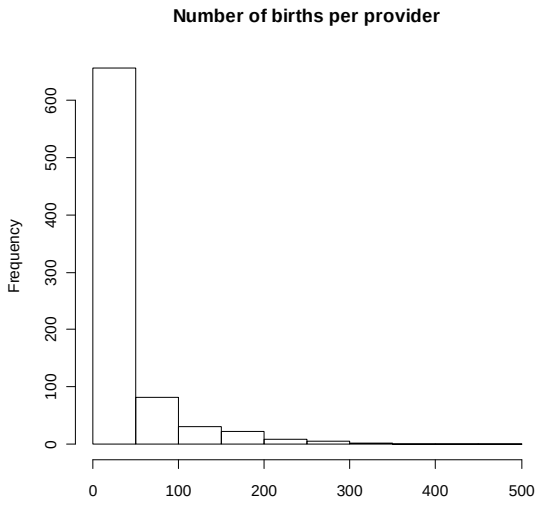
  Pr(selection of *k*th practice)=1-Pr(*k*th practice not selected)=$\theta_k = 1 - \prod_{i=1}^{n_k}(1-\pi_i)$

  where *i* indexes the providers in the *k*th practice and $\pi_i = \frac{N_i}{N}$ is the PPS selection probability for the ith provider as before.

- Maintain an epsem sample by obtaining $n_k = N\theta_k$ births from each provider.
  - Same issues about adjusting for providers unable to supply n_k PSU births.

Simulation Study: Wayne County, MI

- Use 2008 birth certificate data to draw sample of 2009 births using provider PPS sample.
  - Ignore issue of provider vs. practice since simulation study
- 808 unique provider names delivered 99.4% of births
- 50% of births from 75 providers, 90% of births from 273 providers:



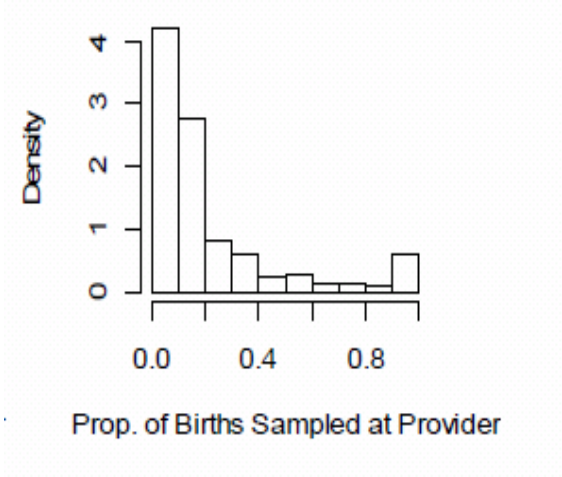
Simulation Study: Wayne County, MI

- Draw sample PPS samples of births as described: let $\tilde{n} = 11$ to accommodate sampling of providers with few PSU births; use weights to adjust undersampling of small providers.
- Race of child and age of mother was follows (200 simulations):

	True	Mean	10th Percentile	90th Percentile
Race				
White	45.9	45.1	31.8	56.5
Black	46.6	47.4	35.8	60.1
Asian	3.7	3.5	1.1	7.0
A/P	0.3	0.4	0.4	1.3
Other	3.3	3.5	1.2	7.0
Age				
<21	19.3	19.7	13.8	26.7
21-25	25.9	26.2	19.3	32.7
26-30	26.2	26.0	20.1	31.7
31-35	18.8	18.4	12.1	24.2
36-40	8.3	8.4	5.0	12.7
>40	1.5	1.4	0.2	3.1

Simulation Study: Wayne County, MI

- Fraction of births to be sampled with in each clinic was distributed as follows:



- Median proportion was .113, but highly variable (.02 to 1.0).

Simulation Study: Wayne County, MI

- Obtained sample representative of 2009 race/mother age using 2008 provider list.
  - Substantial turnover between 2008 and 2009 (274 providers delivering infants in 2009 that did not appear on 2008 list).
  - Turnover mostly in small size providers: 2008 providers delivered 94.2% of births in 2009.

Other Issues

- Movers: women who switch practices after being sampled exactly analogous to movers in the area probability sample.
  - If uncommon, live with oversample of "movers"
  - If common:
    - Drop women once they leave sampled practice unless they move to another sampled practice
    - Try to follow, but adjust for increased probability of selection at both practices.
- Stratification: as in area probability sample, can stratify based on factors known for the entire population:
  - Age of mother
  - Race of mother/child
  - Geographic region