

# IS A PROBABILITY SAMPLE REQUIRED?

---

Lynne Stokes  
Statistical Science Department  
Southern Methodist University

# Outline

- Definition of probability sample
- Background on the National Children's Study and the probability sampling question
- Various research perspectives and how they affect data collection needs
- Examples: Widening your research perspective
- How to think about whether your study needs to consider probability sampling

# What is a probability sample?

- A sample selected in such a way that every member of the sample has a known chance of selection (and every member of the population has a nonzero chance of selection)
  - Does not mean equal probability of selection is needed; weighting for unequal selection is fine, as long as randomness is used for selection

# What is the advantage of a probability sample over non-probability (convenience) sample?

- Reduces risk of selection bias (allows for unbiased estimators of population parameters)
- Produces results that are generalizable to the U.S. population (allow for known margins of error from population parameters)
  - External validation

IMPROVING THE HEALTH  
OF AMERICA'S CHILDREN



- Congressionally mandated (Children's Health Act of 2000) to examine effect of environment on development
- Designed to follow 100K children from conception (or before) to age 21 and have findings generalizable to US population
- Joint effort of NIH, EPA, CDC
- Planning began in ~2001

# Current NCS Sampling Plan

- ~ 41K children selected via a probability sample at birth (in hospitals)
- ~ 41K children selected via a probability sample prenatally
- ~ 8K subsequently born siblings, selected pre-conceptually, of the probability sample children
- Convenience sample of 10K children, split between pre-conceptual nulliparous and “hurricane sample”

# How much?

- Cost about \$1 billion so far
- No data for the main study has been collected yet
- Sampling plan still not completely determined
- NAS panel convened to review (again) the generalizability of the study

# Why so hard to pin down data collection method?

One reason is the varying scientific perspectives of the partners:

Includes physicians (NIH), chemists and biologists (EPA), epidemiologists and public health scientists (CDC/NIH)

# Data collection in the sciences

- Medicine
  - Clinical trials: no random selection of patients; randomize patients to treatments within sites (RCT's; provide internal validity); may have multi-site centers
- Bench scientists (Chemistry/Biology)
  - Lab: Control everything except factor of interest; no random selection or maybe even randomization of rats

# Data collection in the sciences (con't)

- Epidemiology
  - Observational data from surveillance systems, also no randomization, but use modeling
- Public Health
  - Also use RCT's, but worry about external as well as internal validity; Probability samples (e.g., BRFSS, National Youth Fitness Survey, National Ambulatory Surgery Survey, ....)

# What kind of study was the NCS?

- Mandated to study “health disparities”
- Ability to study “emerging environmental factors” (e.g., fracking)
- Much interest in causal relationships (e.g., house dust and asthma)
- Public health scientists (mostly) won: nationally representative random (probability) sampling

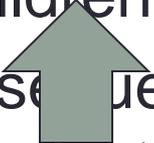
# What does a probability sample look like for the NCS?

- Every birth in the U.S. during a particular 4-year period must have a non-zero selection probability, and every sample birth must have a calculable probability of selection

# Problem

- Nationally representative probability sample of births is very expensive and logistically difficult, especially for collecting preconception data

# Current NCS Sampling Plan

- ~ 41K children selected via a probability sample at birth (in hospitals) 
- ~ 41K children selected via a probability sample prenatally
- ~ 8K subsequently born siblings of the probability sample children selected pre-conceptually 
- Convenience sample of 10K children, split between pre-conceptual nulliparous and "hurricane sample" 

## An example (from Ellenberg 2010)

- What is the risk of low birth weight children to develop CP?
- NCCP, Precursor study to NCS, ~ 50K children from 15 medical centers
- Not a probability sample
- Blacks overrepresented (about 50/50 W/B); Income above avg for W and below avg for B

# Incidence of CP by birth weight

Exposure birthweight	CP	
	Yes	No
Low	1.53%	98.47%
Normal	0.31%	99.69%

Relative risk = RR =  $1.53\% / 0.31\% = 4.9$

## Incidence of CP differs by race

Exposure birthweight	White CP		Black CP	
	Yes	No	Yes	No
	Low	1.02%	1.03%	2.57%
Normal	0.26%	0.26%	0.36%	99.64%
	RR=	3.9	RR=	7.2

## RR of low birth weight must be adjusted

Exposure birthweight	CP	
	Yes	No
Low	1.53%	98.47%
Normal	0.31%	99.69%

Relative risk = RR = 1.53%/0.31% = 4.9

$$\begin{aligned}
 RR &= (\%W) * RRW + (\%B) * RRB \\
 &= (0.88) * 3.9 + (0.12) * 7.2 = 4.2
 \end{aligned}$$

# So what's the problem?

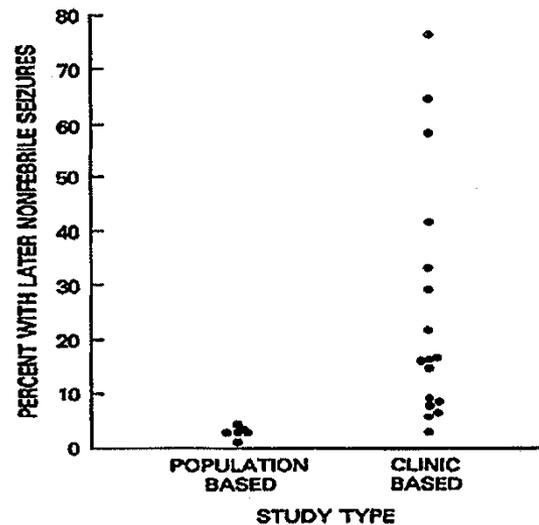
- Suppose RR is moderated by an unknown or unobservable variable
- Then there will be no information that allows an adjustment
- Estimator of RR (an association measure) will be biased
- Only protection is make sure sample is “representative” (i.e., we know probability of selection so proper weights can be assigned)
- Probability sampling is insurance against unknown correlates that you can't adjust for in your model

## Another example: treatment of febrile seizures

- In the 1970's and 80's, risk of recurrent febrile or non-febrile sz following a febrile seizure was believed to be high (about 1 in 3), based on clinic based studies, though there was large variability between studies
- Aggressive tx of preventative fever-lowering agents at onset; prophylactic tx with anticonvulsants was standard of care in the US

# Population Studies

- More recently, data from population studies (probability samples) have become available and the story is VERY different



# Population Studies

- More recently, data from population studies (probability samples) have become available and the story is VERY different
- Why? Something different about children presenting at research clinics that made them non-representative; not recognized
- Current standard of care: no preventative use of anti-fever or anticonvulsants, except in selected patients

# Do you need a probability sample?

- How variable are the explanatory variables in the target population and how well do you understand their affect?
  - If not variable and/or perfect, you don't need probability sample
- Are you primarily interested in theory testing (internal validity) or scale-up (external validity)?
  - If theory testing, you probably don't need probability sample, but may need diversity of settings
- Are you interested in accurately estimating prevalence/cost/rates of  $\Delta$  in population?
  - If yes, you probably need a probability sample

## Do you need to use the weights?

- Do you know all the factors that caused the data to be differentially weighted (e.g., prob's of selection and nonresponse adjustments)?
  - If yes, you can include them as predictors and then the model will be unbiasedly estimated without weights
  - If no, you must use weights to ensure unbiasedness, in case response differs by unequally represented subgroups
- Do you want a simple model (few predictors)?
  - If yes, you can use the weights and not bother with complicating your model with extra predictors
  - If no, and the answer to the 1<sup>st</sup> question was yes, you can get away without using the weights

# Summary

- Two kinds of randomness for studies: treatment and subject selection
- Each have a role in science
- Know when you need each
- Are not mutually exclusive
- Warning: Probability sampling is expensive (usually more so than randomizing treatments)

# References

- Ellenburg, J. (2010) NCS: Establishment and protection of the inferential base, *Statistics in Medicine*, 29, 1360-1367.
  - This issue has several papers on the NCS
- Michael, R.T. and O'Muircheartaigh, C. (2008) Design priorities and disciplinary perspectives: the case of the National Children's Study, *J.R. Statistical Society A*, 171, 465-480.

## CALL FOR PROPOSALS TO ADD QUESTIONS TO THE 2016 GSS

The General Social Survey (GSS) invites proposals to add questions to its 2016 survey. Proposals ...need not be accompanied by funding .... They will be judged with their scientific merit as a primary consideration. The deadline for submissions is March 15, 2014.

The GSS is a **nationally representative survey of non-institutionalized adults in the United States**, conducted primarily via face-to-face interviews. A National Science Foundation (NSF) award provides foundational support for the GSS.