Thoughts on Innovative Approaches for Sampling Special Populations

Robert Santos

Urban Institute

2013-2014 AAPOR President

Acknowledgement

- Thank the authors & organizer
- Papers were interesting and provocative
- Imagine a session like this 10-15 years ago at a FCSM?

Outline

- Comments on each presentation
- Discuss some ‘big picture’ themes
- Parting thoughts on relevance to federal statistical system

A New Source of Local Health Data

- Objective – *complement* BRFSS via county level estimates
- Use ‘Big Data’ (FB Likes) to predict health outcomes via OLS
- Results for FL seemed promising

A New Source of Local Health Data

- Liked the word complementary (instead of replacement)
  - Can't see this replacing current practice
- But BRFSS goal is surveillance:
  - Public health emergencies (e.g. flu-like illness)
  - Emerging health issues (smoking, obesity, diabetes, mental health)
  - Let's think about that a moment…

A New Source of Local Health Data

- Surveillance: look for meaningful changes over time
- More salient assessment of model performance is:

ability to detect (policy relevant) changes when they occur

A New Source of Local Health Data

So, rather than examining:

$$\mu_t \;\; at\ time\ t$$

Consider:

$$\Delta_t = \;\; \mu_t - \mu_{t-1}$$

Or an indicator of whether or not:

$$|\Delta_t| \; > \; k$$

- Can detect change even when μ_t incur same bias

A New Source of Local Health Data

- Why focus on detecting change?
  - Many BRFSS health outcomes are fairly static
  - Best predictor is that of (t-1)
- Suggestion:
  - Assess by looking to BRFSS counties with meaningful outcome changes
  - Otherwise risk failing at most critical time

A New Source of Local Health Data

- Use of imprecise predictor data
  - Likes = clustered transactions by subset of persons
    - If you “Like” once, you probably “Like” multiply
    - How to address underlying structure?
    - Possible Latent model?
  - ACS & survey based predictors:
    - Subject to sampling error
  - Underlying assumptions of OLS don’t hold
- Consider errors-in-variables regression

A New Source of Local Health Data

- Can't yet conclude 'statistically valid'
  - More appropriate model needed
  - Replicate: Establish an empirical knowledge base
- Reality check -- reframe the task:
  - Instead of: How can FB data be leveraged?
  - What about: How can existing data be used…?
  - Be open to other approaches (SAE)

Grassroots (GR) Data Collection

- Motivation: falling survey response in the face of rising costs in the BRFSS
- Research Q (to me): Is GR better than OL Panel?
- Results mixed but favored GR relative to OL

Grassroots Data Collection

- First issue: What is representativeness?
  - Capturing observations in a way that allows generalizing to a population
- "Online panels are not representative"
- But more sources yield 'closer' representation?
  - Logic is not clear given current knowledge (e.g., noncoverage, uneven availability of users)
- GR has issue of face validity issue to tackle

Grassroots Data Collection

- Representativeness, cont'd: SurveyMonkey traffic
  - traffic density map to supports representativeness
  - But same can be said of disabled, rental housing, commuters… logic may not hold
- Not surprising – SM sample was unbalanced
- Post-stratification itself is not sufficient to invoke model-based statistical inference from nonprob
- Findings
  - SM: 5 out of 8 closer to BRFSS (almost a coin flip)

Grassroots Data Collection

- Paper favors government sites for recruitment
  - Yes there is salience, but this defines self-selection
  - Expect high nonparticipation
- Bottom line: unable to conclude these are valid statistical estimates

Use of Nonprobability Samples

- Goal – explore nonprob surveys for "evidence based medicine" (EDM)
  - EDM helps decide a patient's treatment given
    - (1) patient circumstance and
    - (2) external scientific findings
- Fundamental Q: can nonprobability surveys contribute to EBM?
- Quick Answer: It depends

Use of Nonprobability Samples

- What insights CAN nonprob Web survey to contribute?
  - Could be useful for surveillance
    - Gaps in treatment (regional avoidance of certain therapies)
    - Unusual successes or failures
    - Alternative medicinal approaches
    - Clustering of incidence geographically or by subpopulation
    - Life course of disease
  - Idea is to find patterns & relationships
- What type of insight should Web survey NOT contribute?
  - Population estimates (e.g., X% of PIDD population receive treatment Y; Z thousand cases in the U.S.)
  - Prevalence estimates (e.g., 1 in X thousand people have PIDD)

Use of Nonprobability Samples

- Nonprobability surveys may have potential for EBM (for PIDD & other rare diseases)
- But be careful not to overstating results
- It's important to do your 'due diligence' to explore internal and external validity
  - But successful validity analyses are necessary, not sufficient (unknown biases lurking)

Big Picture Comments

- Good illustration of ‘fit for purpose’
  - Complementary county level estimates (OLS)
  - Seek cost efficiency (river sampling)
  - Inform individual care of PIDD cases via EBM (web panel)
- Success is not as important as the discovery process
  - Think out of the box
  - Leverage big data
  - Establish empirical base of knowledge

Big Picture Comments

- Our industry faces challenges
  - Probability surveys aren't what they used to be
  - Need to 'do more with less'
- Empirical research has a role in adopting innovations
  - If it works, then it works (Replications needed!)
  - Theory can catch up (e.g., post-stratification) & may already exist
  - But does that mean Fed Stat System rushes to adopt?

Online and Social Media Caution

- Technology & social interactions evolve rapidly
- FB, Twitter, Google, etc. enjoy high popularity
- BUT a new product or practice could totally change the landscape:
  - What works today may not work tomorrow
  - Data easily or cheaply accessible may suddenly become inaccessible or prohibitively expensive

Implications for Federal Statistical System

- Federal Statistical Agencies should set the standard for statistical quality
  - Have a responsibility to the public to "get it right"
  - Billions of dollars and major policy decisions that affect lives ride on these products
- Virtually every federal survey contract requires "valid statistical estimates"
- Nonprobability *population surveys* currently do not meet this standard
  - (Do all probability populations surveys?)

Implications for Federal Statistical System

- FSAs should be risk adverse:
  - Using social media & commercial portals to data assumes access and affordability in perpetuity
  - Can't be one takeover, IPO or corporate executive decision away from crippling their production capacity
- Yet agencies need to explore data quality and efficiency through innovation

And Remember OMB Guidelines!

- "Agencies must… [use] generally accepted statistical methods (e.g., probabilistic methods that… provide… sampling error)
- "… nonprobability sampling methods… must be justified statistically and… measure estimation error" (Standard 1.2)
- proposed nonprobability designs must "… demonstrate that units not in the sample are impartially excluded on objective grounds…" (Standard 1.2.3)

Final Thought

- Remember ‘fit for purpose’
- Innovative nonprobability surveys have lots of potential
  - Can offer valuable insights for policy decisions
  - Highly economical (good ROI)
  - Not all policy decisions need ‘valid statistical point estimates’
- We would be well served to explore the utility of nonprobability survey’s “fuzzy statistics”

THANK YOU!

Rob Santos
rsantos@urban.org