Jean Opsomer
Westat

August 20, 2018

Outline

1. Design-based survey estimation and inference
2. Constructing estimators
3. Weighting by observed response probabilities
4. Case study: constrained estimation of response probabilities
5. Conclusions

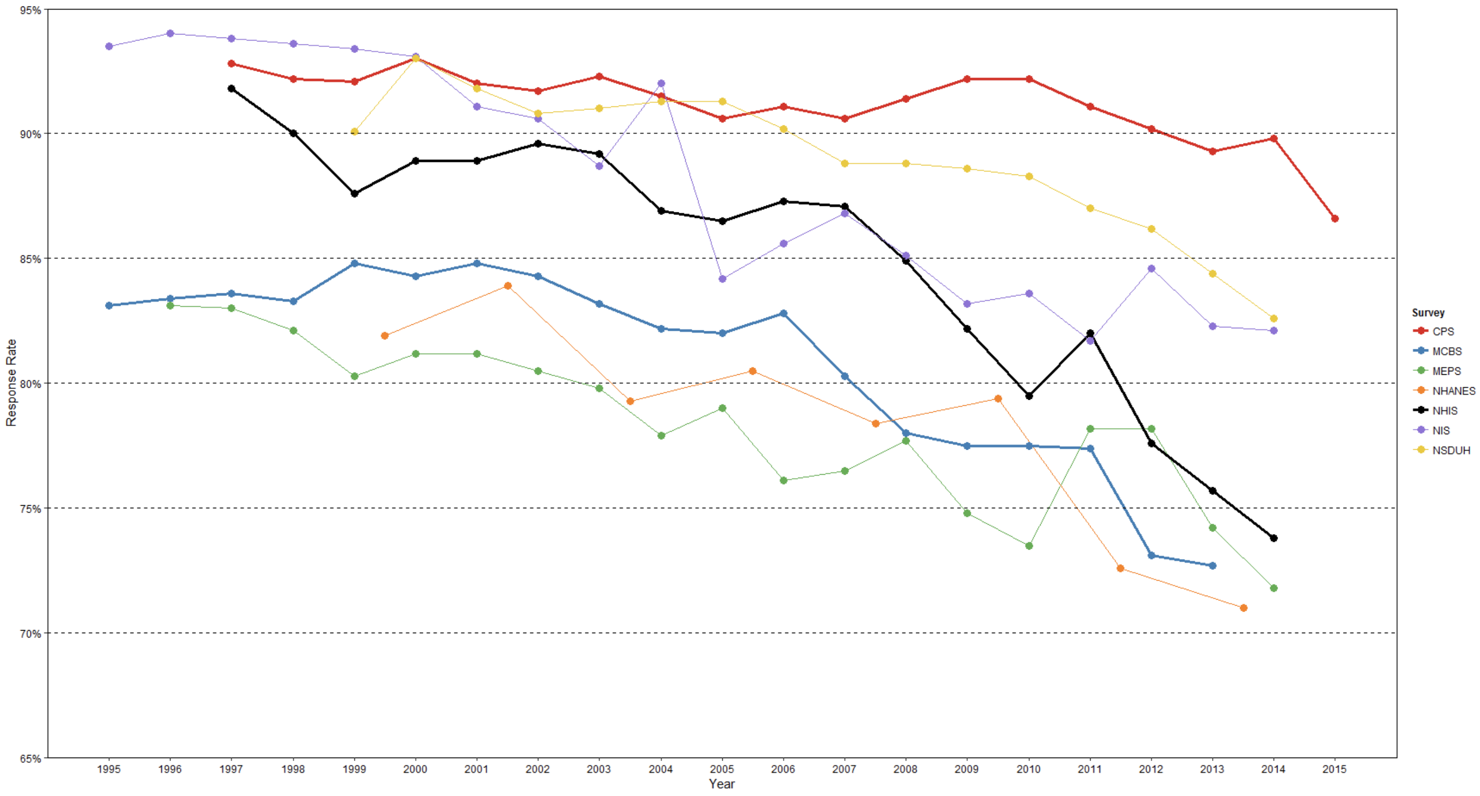
1. Survey modes of estimation and inference

- "Typical" social science survey:
  - large-scale data collection effort conducted on behalf of government agency, using complex multi-stage design
  - output: summary tables and/or weighted datasets
- Key concept: target of inference is specific finite population, e.g. all infants born in US hospitals in 2018, *not* characteristics of a model
- This traditionally leads to *design-based inference*: population treated as fixed but unknown, only randomness comes from sampling design

Design-based inference

- Design-based inference is conceptually attractive
  1. assumption-free inference, because design is known
  2. model-free tools available to quantify sampling variability
  3. enables access to high quality datasets for analysis
     - variables available in their original form
     - analyses do not have to be pre-specified
- But:
  1. estimators based on design often inefficient
  2. high nonresponse "breaks" known design assumption

Growing nonresponse issue in official surveys



(Czajka and Beyler, 2016)

Nonresponse mitigation

- Nonresponse is seen as important practical and research issue in human population surveys
- Tourangeau and Plewes (2013), *Nonresponse in Social Science Surveys: A Research Agenda*, National Academies Press
- Important on-going research on reducing nonresponse, especially in adaptive multi-mode approaches
- Nevertheless, nonresponse rates are generally expected to continue to increase

Model-based solution?

- Model-based inference: build model for target variables; once estimated, allows full set of model-based techniques including prediction of population quantities of interest

  1. maximize efficiency, subject only to inherent variability of data (and skill of modeler)
  2. bypass "nuisance" random processes: sampling design, response mechanism

- But:

  1. labor-intensive
  2. sample selection effect can invalidate results

Nonresponse adjustments in design-based estimation

- Thriving area of research within survey statistics, including by many SMURF participants
- Options:
  - ignore
  - apply model-assisted ideas and rely on relationships between variables to correct for nonresponse
  - explicit modeling of response mechanism
  - double-robust approaches
  - etc
- Theory is well understood (still room for improvement!)

Nonresponse adjustments in estimation (2)

- Conceptually, nonresponse is treated as "add-on" to design randomness
- We continue to appeal to classical design properties to justify design-based (weighted) approach to survey inference to users of survey data
- Is this counter-productive?

Reframing design-based inference

- "Generalized design-based" (?) relies on combination of design *and models* to account for selection process of obtaining data
  - sampling design, response mechanism, other selection steps (e.g. response-driven sampling)
  - → "selection probability" is longer assumption-free, but does not claim to be
- Key aspects:
  1. finite population still treated as fixed target of inference
  2. avoids modeling of survey variables to extent possible

Generalized design-based inference

- Modeling selection process: design known, but other components need to be modeled
  - access to paradata, frame data and confidential unit-level data can lead to better models and creation of weights that can be released
  - survey specialists focus on developing and fitting selection models, no need to be subject-matter specialist
  - weights are presented as result of careful modeling, instead of modified design inclusion probabilities (similar to output in other areas of statistics)

Generalized design-based inference (2)

- Modeling data selection process instead of data can still result in inefficient inference, since model is “generic” w.r.t. survey variables
- Improving efficiency
  - model-assisted approaches continue to apply
  - weighting by empirical response probabilities
- In both cases, efficiency gains depend on relationship between model variables and survey variables

What about item nonresponse?

- Could in principle be handled as a selection problem and modeled as such
- But: "Swiss cheese" nonresponse makes this often not practical
- Approaches:
  - explicit modeling (e.g. multiple imputation, regression imputation)
  - implicit modeling (e.g. hierarchical and/or fractional hot-deck imputation)

What about item nonresponse? (2)

- Goal of imputation: create pseudo-data that look like original data
- Pro: allows survey users to continue using data as if obtained under selection-only approach
- Cons:
  - requires modeling of survey variables
  - can increase variability of estimators
- Might be preferable to leave this to subject-matter analysts?

2. Constructing generalized design-based estimators

- Finite population: $U = \{1, 2, \dots, k, \dots, N\}$
- Survey variables

  $y_k =$ target variables (unknown outside sample, fixed)
  $x_k =$ auxiliary variables (known, fixed)

- Target population parameters: totals, means, proportions, e.g.

$$T_y = \sum_{k \in U} y_k$$

- Sample: $s \subset U$, obtained by selection mechanism $p(s)$

Generalized design-based estimation

- Sample membership indicator (random)

$$I_k = \begin{cases} 1 \text{ if } k \in s \\ 0 \text{ otherwise} \end{cases}$$

- Selection probabilities

$$\begin{aligned} p_k &= \Pr(k \in s) = \mathsf{E}(I_k) \\ p_{kl} &= \Pr(k, l \in s) = \mathsf{E}(I_k I_l) \end{aligned}$$

  – traditional: p_k, p_{kl} known

  – if generalized design-based: well-defined quantities, to be estimated/predicted

Generalized design-based estimation (2)

- Specifications of selection probability (unit nonresponse case)

$$p_k = \pi_k \, r(x_k)$$
$$p_{kl} = \pi_{kl} \, r(x_k) \, r(x_l)$$

  - π_k, π_{kl} are "pure" design inclusion probabilities
  - $r(x) = r(x;\theta)$ is unknown function of auxiliary variable(s)
    - usually: x is multivariate and categorical
    - usually: $r()$ is parametric

Generalized design-based estimation (3)

- Inverse-probability weighting estimator

$$\widehat{T}_y = \sum_{k \in s} \widehat{w}_k y_k = \sum_{k \in U} \frac{I_k}{\widehat{p}_k} \, y_k = \sum_{k \in U} \frac{I_k}{\pi_k \, r(x_k; \widehat{\theta})} \, y_k$$

- $\widehat{T}_y$ does not behave like "oracle" 2-phase estimator

$$\tilde{T}_y = \sum_{k \in s} w_k y_k = \sum_{k \in U} \frac{I_k}{\pi_k \, r(x_k; \theta)} \, y_k$$

  - model dependent
  - no longer exactly unbiased, even if model is correct
  - often includes additional variance terms

3. Weighting by observed response probabilities

- Response homogeneity group (RHG) model
  - common nonresponse adjustment in practice
  - flexible and efficient "all-purpose" approach, as approximation to more complicated underlying model
  - closely related to post-stratification
- Revisit efficiency of RHG (Särndal *et al*, 1992, Ch. 15.6)

RHG model

- Selection process
  - sample s drawn according to sampling design $p(s)$
  - conditional on s, units respond independently with unknown probabilities that are equal within groups s_g ($s = \cup s_g$)
- Selection probabilities

$$
\begin{aligned}
p_k &= \pi_k\, \theta_g && \text{for all } k \in s_g \\
p_{kl} &= \pi_{kl}\, \theta_g\, \theta_{g'} && \text{for all } k \in s_g, l \in s_{g'}
\end{aligned}
$$

- Groups s_g can be sample-dependent and θ_g are unknown parameters

RHG estimation

- Notation
  - $R_k = 1$ if unit k responds, 0 otherwise
  - $n_g = \sum_{s_g} 1$: sample size in s_g
  - $m_g = \sum_{s_g} R_k$: respondent sample size in s_g
  - r_g = subset of respondents in s_g

RHG estimation (2)

- If θ_g known, classical 2-phase estimator

$$\tilde{T}_y = \sum_{g=1}^{G} \sum_{r_g} \frac{1}{\pi_k \, \theta_g} \, y_k$$

- Properties

$$\mathsf{E}(\tilde{T}_y) = T_y$$

$$\mathsf{Var}(\tilde{T}_y) = \sum \sum_{U} \Delta_{kl} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} + \mathsf{E}_p \left(\sum_{g=1}^{G} \frac{1-\theta_g}{\theta_g} \sum_{s_g} \frac{y_k^2}{\pi_k^2} \right)$$

$(\Delta_{kl} = \pi_{kl} - \pi_k \pi_l)$

RHG estimation (3)

- RHG estimator

$$\widehat{T}_y = \sum_{g=1}^{G} \sum_{r_g} \frac{1}{\pi_k \widehat{\theta}_g} y_k = \sum_{g=1}^{G} \sum_{r_g} \frac{1}{\pi_k \frac{m_g}{n_g}} y_k$$

- Properties

$$\begin{aligned} \mathsf{E}(\widehat{T}_y) &= T_y \\ \mathsf{Var}(\widehat{T}_y) &\approx \sum\sum_U \Delta_{kl} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} \\ &\quad + \mathsf{E}_p \left(\sum_{g=1}^{G} \frac{1-\theta_g}{\theta_g} \sum_{s_g} \left(\frac{y_k}{\pi_k} - \frac{\sum_{s_g} y_k/\pi_k}{n_g} \right)^2 \right) \end{aligned}$$

RHG efficiency

- Compare

$$\mathrm{Var}(\widehat{T}_y) \approx \sum\sum_U \Delta_{kl} \frac{y_k}{\pi_k}\frac{y_l}{\pi_l} + \mathrm{E}_p\left(\sum_{g=1}^{G} \frac{1-\theta_g}{\theta_g} \sum_{s_g} \left(\frac{y_k}{\pi_k} - \frac{\sum_{s_g} y_k/\pi_k}{n_g}\right)^2\right)$$

$$\mathrm{Var}(\tilde{T}_y) = \sum\sum_U \Delta_{kl} \frac{y_k}{\pi_k}\frac{y_l}{\pi_l} + \mathrm{E}_p\left(\sum_{g=1}^{G} \frac{1-\theta_g}{\theta_g} \sum_{s_g} \frac{y_k^2}{\pi_k^2}\right)$$

RHG efficiency (2)

- Using observed response probabilities $\widehat{\theta}_g = m_g/n_g$ is equivalent to ratio-type estimator
  - efficiency gains relative to Horvitz-Thompson estimator
- Gains depend on:
  - correctness of response model
  - homogeneity of y_k/π_k within groups
- Gains can offset efficiency losses due to (modest) model departures

4. Special case: RHG under natural ordering

- We consider estimator of RHG type

$$\widehat{T}_y = \sum_{g=1}^{G} \sum_{r_g} \frac{1}{\pi_k \, \widehat{\theta}_g} \, y_k = \sum_{g=1}^{G} \sum_{r_g} \frac{1}{\pi_k \, \frac{m_g}{n_g}} \, y_k$$

with r_g, s_g defined by values of ordinal variable x

- Assume response probability monotone in x:

$$x_k \leq x_l \Rightarrow r(x_k) \leq r(x_l)$$

and for simplicity, rewrite as

$$\theta_1 \leq \ldots \leq \theta_G$$

Constrained estimation of response probabilities

- Can be set up as design-weighted or unweighted problem; consider unweighted here
- Estimators $\widehat{\theta}_1^c, \ldots, \widehat{\theta}_G^c$ are solution to

$$\min \sum_{g=1}^{G} \sum_{s_g} n_g (R_k - \theta_g)^2 \quad \text{subject to } \theta_1 \leq \ldots \leq \theta_G$$

  with $R_k = 1$ if unit k responds, 0 otherwise
- If minimizer satisfies constraint,

$$\widehat{\theta}_g^c = \frac{m_g}{n_g} = \widehat{\theta}_g$$

Constrained estimation of response probabilities (2)

- If constraint is binding,

$$\widehat{\theta}^c_g = \frac{m_{g_1:g_2}}{n_{g_1:g_2}} = \widehat{\theta}_{g_1:g_2}$$

with $g_1 \leq g \leq g_2$

- In general,

$$\widehat{\theta}^c_g = \max_{g_1 \leq g} \min_{g \leq g_2} \frac{m_{g_1:g_2}}{n_{g_1:g_2}}$$

and $\widehat{\theta}^c_{g_1} = \ldots = \widehat{\theta}^c_{g_2}$ (Brunk, 1955)

$\rightarrow$ automatic determination of response homogeneity groups, by pooling neighboring groups

Constrained RHG estimator

- Estimator

$$\widehat{T}_y^c = \sum_{g=1}^{G} \sum_{r_g} \frac{1}{\pi_k \, \widehat{\theta}_g^c} \, y_k = \sum_{g'=1}^{G_s^*} \sum_{r'_g} \frac{1}{\pi_k \, \widehat{\theta}_{g'}^c} \, y_k$$

  with G_s^* sample-dependent, determined by pooling

- We study its theoretical properties
  - classical design-based asymptotic framework ($N \to \infty$, sequence of designs p_N, asymptotic normality, etc)
  - assuming constrained RHG model holds in population

Theoretical results

1. Response probability estimator $\widehat{\theta}_g^c$ is consistent for θ_g w.r.t. design and response model
2. RHG estimator $\widehat{T}_y^c$ is consistent for T_y w.r.t. design and response model
3. Let $\widehat{V}_s^* =$ linearized variance estimate treating the pooled groups $\{r_{g'}, g' = 1, \ldots, G_s^*\}$ as fixed. Then,

$$\frac{\widehat{T}_y^c - T_y}{\sqrt{\widehat{V}_s^*}} \rightarrow \mathcal{N}(0, 1)$$

Simulations: setup

- Population: $N = 10,000$, 5 equal-sized groups U_g with

$$y_k \sim \mathcal{N}(1+g, 1) \quad \text{for } k \in U_g$$

- Sampling design: SRSWOR with $n = 400$
- Response mechanism: RHG with

$$R_k \sim \text{Ber}(\theta_g) \quad \text{for } k \in U_g$$

  and we consider different specifications of $\theta_1, \ldots, \theta_5$
- 10,000 replications

Simulations: setup (2)

- Estimators of population mean

$$
\begin{aligned}
\bar{Y} &= \text{unconstrained RHG estimator} \\
\bar{Y}^c &= \text{constrained RHG estimator} \\
\bar{Y}_{HT} &= \text{Horvitz-Thompson estimator, true } \theta_g \\
\bar{Y}_{HA} &= \text{Hájek (ratio) estimator, true } \theta_g
\end{aligned}
$$

with

$$
\begin{aligned}
\bar{Y}_{HT} &= \frac{\sum_g \sum_{s_g} y_k/(\pi_k \theta_g)}{N} \\
\bar{Y}_{HA} &= \frac{\sum_g \sum_{s_g} y_k/(\pi_k \theta_g)}{\sum_g \sum_{s_g} 1/(\pi_k \theta_g)}
\end{aligned}
$$

Simulations: scenarios

- Scenario 1: high response rate, monotone

$$(\theta_1, \dots, \theta_5) = (0.5, 0.6, 0.7, 0.8, 0.9)$$

- Scenario 2: medium response rate, monotone

$$(\theta_1, \dots, \theta_5) = (0.3, 0.4, 0.5, 0.6, 0.7)$$

- Scenario 3: low response rate, monotone

$$(\theta_1, \dots, \theta_5) = (0.2, 0.25, 0.3, 0.35, 0.4)$$

- Scenario 4: equal-probability (monotone)

$$(\theta_1, \dots, \theta_5) = (0.5, 0.5, 0.5, 0.5, 0.5)$$

Simulations: scenario 1

$$(\theta_1, \ldots, \theta_5) = (0.5, 0.6, 0.7, 0.8, 0.9)$$

	Rel. Bias (%)	Scaled MSE
$\bar{Y}$	-0.014	–
$\bar{Y}^c$	-0.21	1.04
$\bar{Y}^{HT}$	-0.034	8.52
$\bar{Y}^{HA}$	-0.009	3.08

Simulations: scenario 2

$$(\theta_1, \ldots, \theta_5) = (0.3, 0.4, 0.5, 0.6, 0.7)$$

	Rel. Bias (%)	Scaled MSE
$\bar{Y}$	0.011	–
$\bar{Y}^c$	-0.247	1.06
$\bar{Y}^{HT}$	-0.034	5.82
$\bar{Y}^{HA}$	0.106	2.90

Simulations: scenario 3

$$(\theta_1, \ldots, \theta_5) = (0.2, 0.25, 0.3, 0.35, 0.4)$$

	Rel. Bias (%)	Scaled MSE
$\bar{Y}$	-0.004	–
$\bar{Y}^c$	-0.586	1.12
$\bar{Y}^{HT}$	-0.011	11.00
$\bar{Y}^{HA}$	0.155	2.97

Simulations: scenario 4

$$(\theta_1, \ldots, \theta_5) = (0.5, 0.5, 0.5, 0.5, 0.5)$$

	Rel. Bias (%)	Scaled MSE
$\bar{Y}$	-0.001	–
$\bar{Y}^c$	-1.772	2.51
$\bar{Y}^{HT}$	0.005	11.07
$\bar{Y}^{HA}$	-0.012	2.97

Simulations: findings

- Applying RHG estimation at smallest scale possible appears to be most efficient
- Using external knowledge about response probabilities not sufficient to offset this
- Not shown: effects disappear if y_k *iid* across RHG groups

5. Conclusions

- Generalized design-based inference
  - corresponds to current "best practice" in survey organizations, but is hidden behind nominal design-based approach
  - claim: should be explicitly recognized and advocated
- RHG (and PS) provide good all-purpose approach for constructing efficient estimators in social surveys, since most variables are categorical

CONTACT: JeanOpsomer@westat.com