Sampling Design and Variance Estimation for the National Survey of WIC Participants, Wave 3

Stas Kolenikov

FCSM Research and Policy Conference
Washington, DC : March 7, 2018

BOLD THINKERS DRIVING REAL-WORLD IMPACT

National Study of WIC Participants, Wave 3

Study goals/surveys within the study:

- census of state agencies
- sample of local agencies
- sample of WIC participants: program experience
- sample of recently certified WIC participants: certification error (rates, dollars)

Focus on the latter as the most complicated component (at least as far as sampling goes)

BOLD THINKERS DRIVING REAL-WORLD IMPACT

Sampling design considerations

- Precision requirements on certification error, per RFP
- Interest in FNS regions (7 total, do not align with Census Divisions)
- Natural hierarchy of states/ITOs, local agencies, clinics, participants to sample

Sampling design decisions

- Stratification by region
  - larger regions broken to smaller strata, total of 13 strata
  - state $\subset$ strata
- 1825 local agencies combined into 215 PSUs to minimize FI travel time
  - Clerical review of PSU boundaries
  - From 1 to 62 agencies per PSU
  - 30 PSU to be selected
  - 2 to 3 PSU/stratum to be selected
  - no certainty PSUs

Sampling design decisions, ctd

- LWAs are SSUs
  - some really tiny LWAs are combined to ascertain the minimum size of an SSU
  - some large LWAs (think Los Angeles) are certainty SSUs
- ITOs are separated out into its own stratum
  - 2 ITOs are sampled, resulting in $\approx$ 3x higher sampling rates

Sampling and design variance estimation

- 1st stage sampling: high entropy 2 or 3 PSUs (done)
- 2nd stage sampling: high entropy 1 to 3 LWAs (done)
- Participants: SRS within 2nd stage (coming soon)
- Variance estimation: high entropy rescaled balanced bootstrap

Let's go over each component one at a time!

Unequal probability sampling

Brewer and Hanif (1982): ~50 methods.

High entropy (conditional Poisson) sampling was Hajek's Holy Grail, but it was not computationally available in his time. Things have changed now!!

Ideas due to Hajek (starting from late 1960s), computing due to Chen, Dempster and Liu (1994), self-contained description due to Tille (2006).

Exponential sampling

Exponential sampling design:

$$p_{\mathrm{EXP}}(\mathbf{s}, \lambda, \mathcal{Q}) = g(\mathbf{s}) \exp\left[\lambda' \mathbf{s} - \alpha(\lambda, \mathcal{Q})\right]$$

where $\mathbf{s} = (s_1, \ldots, s_k, \ldots, s_N)$ is the vector of sample frequencies (random variables), λ is the parameter determining selection probabilities, $\mathcal{Q}$ is the support, $g(\cdot)$ is

$$g(\mathbf{s}) = \prod_{k \in \mathcal{U}} \frac{1}{s_k!}$$

and $\alpha(\lambda, \mathcal{Q})$ is the normalizing constant (Definition 45 of Tille 2006).

Exponential sampling, ctd

Why bother?

- Unifying framework
- Exponential families are nice:
  - mean = selection probabilities is the first derivative of $\alpha(\lambda, \mathcal{Q})$
  - variance is the second derivative of $\alpha(\lambda, \mathcal{Q})$

Entropy

$$I(p) = -\sum_{\mathbf{s}\in\mathcal{S}_n} p(\mathbf{s}) \ln p(\mathbf{s}),\ \mathcal{S}_n = \{\mathbf{s} : s_k \in \{0,1\}, \sum_{k\in\mathcal{U}} s_k = n\}$$

Higher entropy = sampled units tend to come in unpredictable ways

- SRSWOR is the high entropy design on the space of fixed sample size EPSEM designs (Tille 2006, Sec. 5.8)

Lower entropy = sampled units come in predictable ways

- Systematic sampling: few possible samples, groups of units go hand in hand, many $\mathbf{s} \in \mathcal{S}_n$ have zero probabilities of selection driving entropy down

Unequal probability sampling

If you know λ, you can obtain selection probabilities as:

$$\pi_k(\lambda) = \frac{\sum_{\mathbf{s}\in\mathcal{S}_n} s_k \exp \lambda'\mathbf{s}}{\sum_{\mathbf{s}\in\mathcal{S}_n} \exp \lambda'\mathbf{s}}$$

$$\pi_{kl}(\lambda) = \frac{\sum_{\mathbf{s}\in\mathcal{S}_n} s_k s_l \exp \lambda'\mathbf{s}}{\sum_{\mathbf{s}\in\mathcal{S}_n} \exp \lambda'\mathbf{s}}$$

However, the reverse step of determining λ given the target vector π is difficult, and was the stumbling block for Hajek.

Chen et al (1994) developed an iterative procedure to compute the target parameters λ.

Conditional Poisson sampling

Compute $\mu_k = \frac{n \exp \lambda_k}{\sum_{l \in \mathcal{U}} \exp \lambda_l}$

Rejective procedure (Algorithm 5.5 of Tille 2006):

1. Select unit k with probabiliy μ_k
2. If sample size $\sum_{k \in \mathcal{U}} \neq n$, go back to step 1; otherwise use the selected sample.

(Rejections happen $O_p(\sqrt{2\pi n})$ times.)

Sequential and draw-by-draw procedures are also available. Speed advantages depend on the structure of probabilities (and not as much on n and N)

The Good, The Bad and The Ugly

Method	Fixed # of sample PSUs?	Precision of point estimates	Complexity of sampling	Complexity of var. estimation
PPSWR	No	Good	Low	Low
Durbin-Brewer	Yes	Good	Medium	Medium
Systematic PPS	Yes	Excellent	Low	Medium
Chromy's Method	Yes	Very good	Medium	Medium
CPS	Yes	Good	High	Low

The Good, The Bad and The Ugly, ctd

Method	Unbiased var. estimates?	Stable var. estimates?	Bias in WR var. estimator	Benefit of FPC
PPSWR	Yes	Yes	None	No
Durbin-Brewer	Yes	Often no	Moderate	Hard
Systematic PPS	No	Depends on details	Generally severe	No
Chromy's Method	Yes	No	Generally severe	Hard
CPS	Yes	Yes	Negligible	Yes

Variance estimation for high entropy designs

Sen-Yates-Grundy: compute $\pi_k(\lambda), \pi_{kl}(\lambda)$, proceed to

$$v_{\mathrm{SYG}} = \frac{1}{2} \sum_{k \neq l} \frac{\pi_k \pi_l - \pi_{kl}}{\pi_{kl}} \left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l} \right)^2$$

However, since high entropy designs are "close" to SRSWOR, it is tempting to find a computationally effective representation like

$$v_{\mathrm{approx}} = \sum_{k=1}^{n} \text{something like} \frac{1}{n} \left(\frac{y_k}{\pi_k} - \frac{\text{some sort of a total}}{n} \right)^2$$

High entropy approximations (Preston and Henderson 2007)

Hajek (1964) approximation for CPS:

$$v_{Haj} = B \sum_{k \in s} (1 - \pi_k) \Big(\frac{y_k}{\pi_k} - A_s \Big)^2, A_s = \frac{\sum_{k \in s} (1 - \pi_k) \pi_k^{-1} y_k}{\sum_{k \in s} (1 - \pi_k)}$$

and Hajek's original scaling factor was $B = n/(n-1)$, but can be refined to incorporate π_k.

Hartley and Rao (1962) followed up by Brewer (2001, 2002), Brewer and Donadio (2003):

$$v_{Br} = \sum_{k \in s} c_k^{-1} (1 - c_k \pi_k) \Big(\frac{y_k}{\pi_k} - \frac{\hat{Y}_{HT}}{n} \Big)^2, c_k = \frac{n-1}{n - \pi_k}$$

Rescaled bootstrap (Preston and Henderson 2007)

Brewer variation: draw $1 \leq m_h \leq n_h$ units SRSWOR from stratum h (sampling indicator $= \delta_{hk}$):

$$w_{hk} = \pi_k^{-1}\Big[1 - \gamma_h\sqrt{1-\pi_k} + \frac{\gamma_h}{n_h}\sum_{l=1}^{n_h}\sqrt{1-\pi_l}$$

$$-\frac{\gamma_h}{m_h}\sum_{l=1}^{n_h}\delta_{hl}\sqrt{1-\pi_l} + \frac{\gamma_h n_h}{m_h}\delta_{hk}\sqrt{1-\pi_k}\Big]$$

$$\gamma_h = \sqrt{\frac{m_h}{n_h - m_h}}$$

Balanced bootstrap

Actually, let's talk about BRR first. In BRR, *balance* means:

1. Each unit is selected the same number of times;
2. Each pair of units (across strata) is selected the same number of times.

This is true for the set of all possible pairings, but it is also ensured for a much smaller set of BRR replicates based on Hadamard matrices (McCarthy 1969).

In *balanced bootstrap* (Graham et al 1990; Nigam and Rao 1996), the balance property is likewise extended to the bootstrap samples. You need to use other sorts of orthogonal devices such as *balanced incomplete block designs*.

Mixed orthogonal arrays

MOAs are generalizations of Hadamard matrices for the number of categories greater than 2 (and hence applicable to designs with PSU/stratum greater than 2).

$OA(N; s_1^{k_1}, \ldots, s_v^{k_v}; t)$ is an array $N \times k, k = \sum k_v$

- the first k_1 columns have symbols $\in \{0, 1, \ldots, s_1 - 1\}$
- the next k_2 columns have symbols $\in \{0, 1, \ldots, s_2 - 1\}$, etc.
- in any $N \times t$ subarray, every possible t-tupple occurs an equal number of times.

Introduced to the theory of design-based inference by Wu (1991) and Sitter (1993).

Finding the OA

The NSWP-III design has the structure of $2^9 3^4$, so we need to find the MOA of at least order.

A textbook length treatment: Hedayat, Sloan and Stufken (1999).

Library of orthogonal arrays: `http://neilsloane.com/oadir/`

One possibility:
`http://neilsloane.com/oadir/MA.36.6.2.3.4.2.9.txt`

- $OA(36; 6^2 3^4 2^9; 2)$ with 36 runs,
  - a set of 36 replicate weights (c.f. 30-13=17 design degrees of freedom).
- only the columns with 2 and 3 are utilized.

Bring it all together!

1. Combine LWAs into PSUs.
2. Compute measures of size. (average of stratum-wide shares of the five participant categories), π_k.
3. Compute λ_k.
4. Draw the sample using conditional Poisson sampling algorithm.
5. Pick up OA, use the indicated entries to *drop* units (so that 1 out 2 PSUs is subsampled in strata with 2 PSU/stratum, and 2 out of 3 PSUs are subsampled in strata with 3 PSU/stratum).
6. Compute PSU-level replicate weights.

Cross-check

```
nswp3bs <- svrepdesign(data=nswp3.sample1,type="bootstrap",
  weights=~pwt,repweights="rwt[0-9]+")
cat.total <- svytotal(~`Pregnant women`+`Breastfeeding women`+
  `Postpartum women`+`Infants`+`Children`, design=nswp3bs)
truth <- colSums(agencies[,c("Pregnant women","Breastfeeding w
  "Postpartum women","Infants","Children")])
estimate <- coef(cat.total)
cbind(truth,estimate,confint(cat.total,df=13))
```

```
##                          truth  estimate     2.5 %    97.5 %
## Pregnant women          885377  874856.6  847787.2  901926.1
## Breastfeeding women     673710  668921.8  595082.0  742761.7
## Postpartum women        610118  601371.3  546343.8  656398.7
## Infants                2114308 2089481.2 2051879.7 2127082.7
## Children               4861107 4750656.2 4582309.4 4919002.9
```

Parting words

- The study is not in the field yet, so no real results to show
- Extend rescaled high-entropy bootstrap to two stages
  - John Preston has papers on both high entropy approximate bootstrap (Preston and Henderson 2007) and multistage bootstrap (Preston 2009)
- An alternative: Kim and Wu (2013) replicate weight *exact* representation of the Δ variance estimation operator
  - Can be done for the 30×30 matrix corresponding to PSUs, or for the 53×53 matrix of SSUs
  - requires specific scaling for each pseudo-value, more difficult to implement than bootstrap/BRR

Thank you

- Contact: stas_kolenikov@abtassoc.com
- References: http://www.citeulike.org/user/ctacmo/tag/fcsm2018_kolenikov
- I benefited from discussing this work with Dave Judkins (Abt Associates)

BOLD THINKERS DRIVING REAL-WORLD IMPACT