Small Area Confidence Bounds on Small Cell Proportions in Survey Populations

Aaron Gilary, U.S. Census Bureau

Jerry Maples, U.S. Census Bureau

Eric V. Slud, U.S. Census Bureau
Univ. Maryland–College Park

DC-AAPOR / WSS

Outline

- General Problem
- Cell-Based vs. Model-Based Methods
- Alternative Definitions of *Effective Sample Size*
- Data Example – Erroneous Enumeration of HUs in Census Coverage Measurement study (CCM)
- Numerical Results
- Conclusion and Future Plans

General Problem

Setting: given direct (ratio) estimates for small proportions $\widehat{\pi}_i$ at level of "cell" or domain i (e.g., county)

Problem: specify upper confidence bound for $\widehat{\pi}_i$

- `either` bound in transformed measurement scale $h(\pi_i)$ from data within domain i
- `or` model-based bound connecting values π_i across domains using predictors $\mathbf{x}_i$

Approach: using *effective* sample-sizes n_i^*, adapt binomial/SRS estimator $\widehat{Var}(\widehat{\pi}_i) = \widehat{\pi}_i\,(1-\widehat{\pi}_i)/n_i^*$

Applications

- demographic tables in **ACS**,

  the American Community Survey;

- area-level Erroneous Enumeration rates in

  CCM, Census Coverage Measurement;

- and rates in other Census Bureau surveys.

A Good Cell-Based Method

`Liu and Kott 2009, Survey Methodology`

π_i = true proportion

n_i^* = effective sample size, $y_i^* \sim$ Binom(n_i^*, π_i)

$$\widehat{\pi}_i = \frac{y_i}{n_i} = \frac{y_i^*}{n_i^*} \qquad \textit{direct estimator}$$

Transformation: $asin(\sqrt{\widehat{\pi}})$ `(Variance-stabilizing)`

centered at $asin(\sqrt{\pi})$, Var $\approx 1/(4n_i^*)$

Obtain UCB on arcsin scale,

transform back to prob. scale by $sin^2(x)$

Ideas of Model-Based Methods

π_i includes $\begin{cases} \text{predicted part} & \eta_i = \mathbf{x}_i' \beta \\ \text{unmodeled random component} \end{cases}$

```
(1) Fay-Herriot:
```
$$\pi_i = sin^2(\eta_i + u_i), u_i \sim N(0, \sigma_u^2)$$

```
(2) Logistic:
```
$$\pi_i = \frac{e^{\eta_i + v_i}}{1 + e^{\eta_i + v_i}}, v_i \sim N(0, \sigma_v^2)$$

```
(3) Beta-Binomial:
```
$$\pi_i = \text{Beta}\left(\frac{\tau\, e^{\eta_i}}{1 + e^{\eta_i}}, \frac{\tau}{1 + e^{\eta_i}} \right)$$

Parameters σ_u^2 , σ_v^2 , $(1 + \tau)^{-1}$ quantify *imprecision* of π_i in terms of η_i

Model-Based Methods, Continued

Recall $\quad y_i^*/n_i^* \; = \; \widehat{\pi}_i$

`Fay-Herriot:`

$$\text{arcsin}(\sqrt{\widehat{\pi}_i}) \;\sim\; \mathbf{N}(\text{arcsin}(\sqrt{\pi}_i),\, \frac{1}{4n_i^*})$$

`Logistic & Beta-Binomial:`

$$y_i^* \;\sim\; Binom(n_i^*, \pi_i)$$

Estimation in Model-Based Methods

Point Predictor for π_i

generally different from direct estimator

BLUP: $E(\pi_i \mid y_i^*)$

EBLUP: substitute MLE for β & (σ_u, σ_v or τ)

Upper Confidence Bound for π_i

based on $\widehat{Var}(EBLUP)$

Defining Effective Sample Size n_i^*

If variance $\frac{V_i}{n_i} = \frac{\pi_i(1-\pi_i)}{n_i^*}$ of direct π_i estimator in area i is reliably estimated and sampling fraction $f \ll 1$:

$$\text{DEFF}_i = \frac{V_i}{\pi_i\,(1-\pi_i)} \quad , \qquad n_i^* = \frac{n_i}{\text{DEFF}_i}$$

What if $\widehat{V}_i$ is erratic or π_i too small ?

`Proposal 1:` Area size via higher-level DEFF

With $\widehat{V}/n = \widehat{V}(\widehat{\pi})$, and DEFF at higher (e.g., State) level

$$\text{DEFF} = \frac{\widehat{V}}{\pi(1-\pi)} \quad , \qquad n_i^* = \frac{n_i}{\text{DEFF}}$$

Effective Sample Size, Continued

`Proposal 2:` Eff. size modified by sampling weights

w_{ik} indiv.-level sampling weights within area i

$$n_i^* = \frac{n_i}{\text{DEFF}} \cdot \left(\frac{\sum_k w_{ik}^2}{(\sum_k w_{ik})^2} \right) \Big/ \left(\frac{\sum_{j,k} w_{jk}^2}{(\sum_{j,k} w_{jk})^2} \right)$$

Effective sizes for survey CIs : Liu & Kott 2009

in Bayesian analysis: Chen et al. 2011, Malec 2005

Data Example

- Census Coverage Measurement (CCM):
  - evaluation of Census performance.
  - finds Erroneous Enumeration (EE) rates for counties in sample.
  - 170k Housing Units (HUs) and 1,728 counties.
  - Census publishes detailed estimates for 128 counties with pop. $\geq$ 500k

Specific CCM Details

- National EE rate $\approx$ 2.7% for metro areas (Olson, 2012).
- For our 128 counties, P($\widehat{\pi}_i = 0$) $>$ 0.

EE Rates (*i* = county, *k* = HU):

$$\widehat{\pi} = \frac{\Sigma_{j,k} W_{jk} Y_{jk}}{\Sigma_{j,k} W_{jk}} \quad , \qquad \widehat{\pi}_i = \frac{\Sigma_k W_{ik} Y_{ik}}{\Sigma_k W_{ik}}$$

Specific CCM Modeling Details

- Area- or cell-level models, all covariates from Census.

BIC model-selection penalizes max logLik by $k\,\ln(n)$

($k =$ # regr. coef's, $n =$ total sample size).

Our selected model had 5 predictor variables:

- state EE rate, a synthetic estimator;
- area rate of single-unit households;
- area rate of large multi-unit households;
- area rate of urban households;
- area enumeration rate.

Variability Across Counties

n^* = Proposal 2 Eff. S.S., $n^\dagger$ = Proposal 1 Eff. S.S.

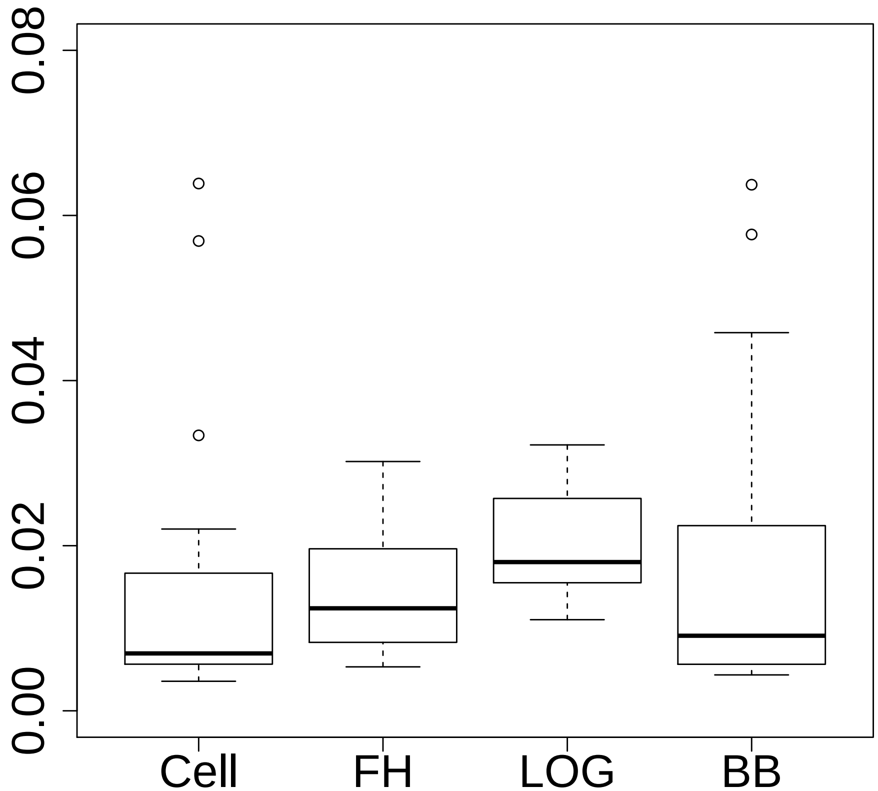
		Mean	**1st Q**	**Med.**	**3rd Q**
`St`	DEFF	6.8	2.7	3.8	6.9
`Cou`	n^*	34.6	8.5	15.8	34.2
`Cou`	$n^*/n^\dagger$	1.2	1.1	1.1	1.2

NB: n^*, $n^\dagger$ based on counties with $\geq$ 20 HUs in sample.

Results

- Inclusion of n^* led to wider, more conservative estimates for the UCBs.
- The Fay-Herriot and Logistic methods had slightly higher means and medians for UCBs for areas with $\widehat{\pi} \approx 0$, but the Cell and Beta-Binomial methods had the highest maximum UCBs.
- Overall, all four methods created upper bounds within a similar range.

UCB of Production Counties where $\widehat{\pi}_i$ = 0



Conclusion & Future Plans

Conclusions:

• For our project, we chose the Cell-Based method. With the results so close, a simpler method without a specified regression model was preferrable.

• We used the Proposal 2 n_i^* for accurately capturing the variance because it did not assume SRS or equal weights within area. Results were more conservative which was a priority in this case.

Future plans:

• extend approach to other applications

• test performance of DEFF under different assumptions?

• generalized R package for Census Bureau use?

References

Chen, C., Lumley, T. and Wakefield, J. (2011), *The Use of Sampling Weights in Bayesian Hierarchical Models for Small Area Estimation.* U. Wash. Stat. Dept. Tech. Rep. #583.

Fay, R. and Herriot, R. (1979, JASA).

Liu, Y. and Kott, P. (2009, Jour. Official Stat.)

Malec, D. (2005, Jour. Official Stat.) *Small Area Estimation ... Housing Units.*

Maples, J., Bell, W. and Huang, E. (2009 ASA Proc.) GVF's for area-level variances

Olson, Doug (2012). *2010 Census Coverage Measurement Report* on Modeling. `DSSD #2010-G-13`.

Prentice, R. (1986, JASA) Beta-binomial regression

Slud, Eric V. (2012). *Small-Area Confidence Bounds . . . in Tables.* Jour. Indian. Soc. Agric. Stat.

Thank You!

Email:

aaron.j.gilary@census.gov

A Cell-Based Method

Variance-Stabilizing transformation to *arcsin sqrt* scale
of estimated area proportion $y_i/n_i \;=\; y_i^*/n_i^*$

For area *i* : π_i = true proportion, n_i^* = eff. samp. size

$$y_i^* \;\sim\; Binom(n_i^*, \pi_i) \;\approx\; \mathcal{N}(n_i^*\,\pi_i,\, n_i^*\,\pi_i\,(1-\pi_i))$$

$$\text{arcsin}\,\sqrt{\frac{y_i^*}{n_i^*}} \;\approx\; \mathcal{N}(\text{arcsin}\,\sqrt{\pi_i},\; \frac{1}{4n_i^*}\,) \qquad \Delta\textit{-method}$$

Transformed scale 90% `CI:` $\text{arcsin}\,\sqrt{y_i^*/n_i^*} \pm 1.645/\sqrt{4\,n_i^*}$

Transform back to the probability scale by $\sin^2(x)$

Models which Borrow Strength across Areas

`Notation:` $\widehat{\pi}_i = \frac{y_i}{n_i} = \frac{y_i^*}{n_i^*}$, $\widehat{a}_i = \arcsin\sqrt{\widehat{\pi}_i}$, $\eta_i = \mathbf{x}_i'\,\beta$

$\mathbf{x}_i$ `area-level observed predictors`

$u_i \sim \mathcal{N}(0, \sigma_u^2)$, $v_i \sim \mathcal{N}(0, \sigma_v^2)$ `random effects`

(1) Fay-Herriot : $\widehat{a}_i = \eta_i + u_i + \epsilon_i$, $\epsilon_i \sim \mathcal{N}(0, \frac{1}{4n_i^*})$

(2) Logistic Random-Intercept : $y_i^* \sim \mathrm{Bin}(n_i^*, \frac{e^{\eta_i+v_i}}{1+e^{\eta_i+v_i}})$

(3) Beta-Binomial : $y_i^* \sim \mathrm{Bin}(n_i^*, \pi_i)$, $\pi_i \sim \mathrm{Beta}(\frac{\tau\, e^{\eta_i}}{1+e^{\eta_i}}, \frac{\tau}{1+e^{\eta_i}})$

Targets for small-area prediction: $\sin^2(\eta_i + u_i)$ in **(1)**; $\frac{e^{\eta_i+v_i}}{1+e^{\eta_i+v_i}}$ in **(2)**; and π_i in **(3)**.

Fay-Herriot Model UCB

Mod1 (FH): EBLUP $\widehat{\pi}_i = \sin^2(\widehat{\theta}_i/n_i^*)$, based on $\theta_i = \eta_i + \mu_i$,

$$\widehat{\theta}_i \;=\; \widehat{\gamma}_i\, y_i + (1 - \widehat{\gamma}_i)\, \mathbf{x}_i'\widehat{\beta} \quad , \quad \widehat{\gamma}_i = \frac{\widehat{\sigma}_u^2}{\widehat{\sigma}_u^2 + (4n_i)^{-1}}$$

$$UCB_i \;=\; \sin^2\left(\frac{1}{n_i^*}(\widehat{\theta}_i + \frac{z_\alpha}{2}((1-\widehat{\gamma}_i)\,\widehat{\sigma}_u^2 + (1-\widehat{\gamma}_i)^2 \mathbf{x}_i'\,\widehat{V}_{\widehat{\beta}}\,\mathbf{x}_i\}^{1/2})\right)$$

Fay and Herriot (1979), Slud (2012)

NB: includes sample variability of $\widehat{\beta}$, not $\widehat{\sigma}_u^2$

Rao (2003) has more inclusive formulas for related $\widehat{\text{mse}}$

Logistic Random-Intercept Model-Based UCB

Mod2 (LgstRI): EBLUP $\widehat{\pi}_i \;=\; \dfrac{g(y_i^* + 1,\, n_i^* + 1,\, \widehat{\eta}_i,\, \widehat{\omega}^2)}{g(y_i^*,\, n_i^*,\, \widehat{\eta},\, \widehat{\omega}^2)}$

where $\widehat{\eta}_i = \mathbf{x}_i' \,\widehat{\beta}$, $\widehat{\omega}^2 = \widehat{\sigma}_v^2 + \mathbf{x}_i' \,\widehat{V}_{\widehat{\beta}}\, \mathbf{x}_i$

$$g(k, n, \eta, \omega^2) \;=\; \int \frac{e^{(\eta+\omega z)k}}{(1 + e^{\eta+\omega z})^n}\, \phi(z) dz \;, \quad \phi(\cdot) \sim \mathcal{N}(0, 1)$$

$$UCB_i \;=\; \widehat{\pi}_i \;+\; 1.645 \left[\frac{g(y_i^* + 2,\, n_i^* + 2,\, \widehat{\eta},\, \widehat{\omega}^2)}{g(y_i^*,\, n_i^*,\, \widehat{\eta},\, \widehat{\omega}^2)} \;-\; \widehat{\pi}_i^2 \right]^{1/2}$$

NB: includes sample variability of $\widehat{\beta}$, not $\widehat{\sigma}_v^2$.

Jiang and Lahiri 2006, Slud 2012

Beta-Binomial Model-Based UCB

Mod3 (Beta-Bin): Estimation via Posterior , $\mu_i = \frac{e^{\eta_i}}{1 + e^{\eta_i}}$,

$$\pi_i \,|\, y_i^* \;\sim\; \text{Beta}(\tau\,\mu_i + y_i^*,\; \tau\,(1 - \mu_i) + n_i^* - y_i^*)$$

Empirical Bayes $\widehat{\pi}_i = \dfrac{y_i^* + \widehat{\tau}\,\widehat{\mu}_i}{n_i^* + \widehat{\tau}}$

- Bootstrap approach to UCB