Quality Assessment of Zeroes in ACS Tables

Eric V. Slud, Univ. of Maryland & Census Bureau

OUTLINE

I. Problem Setting: CV's and other measures of data quality

II. Confidence Intervals for survey proportions

III. Model-based approach: Small Area models for Proportions
— synthetic vs GLM vs Fay-Herriot style models

IV. Data Illustration with ACS 2009 Data

V. Summary and Conclusions

Confidence Intervals & Data Quality Filtering

Common Approach: require estimates $\widehat{\mu}$ to have

$$\widehat{CV}(\widehat{\mu}) = \widehat{SE}(\widehat{\mu}) \,/\, \widehat{\mu} \;\le\; 0.2$$

Rationale is based on Confidence Intervals:

$\widehat{\mu} \pm z_{\alpha/2}\, SE(\widehat{\mu})$ on original scale

$\log(\widehat{\mu}) \pm z_{\alpha/2}\, SE(\widehat{\mu})/\widehat{\mu}$ on $\log(\mu)$ scale

if $\mu = p$: $asin(\sqrt{\widehat{\mu}}) \pm \dfrac{z_{\alpha/2}\, SE(\widehat{\mu})}{2\sqrt{\widehat{\mu}(1-\widehat{\mu})}}$ on $asin\sqrt{\mu}$ scale

CV-bound Standard requires log-scale CI half-width $\le z_{\alpha/2}\,(0.2)$

Approach Based on Transformed Proportions

Study CI's for $\widehat{p}$ applicable to small p, in survey (ACS) data.

Standards could be set for CI widths for p or transformed p.

In large samples, **delta method** for $h(p)$ gives

$$h(\widehat{p}) - h(p) \approx \mathcal{N}\Big(0,\, (h'(p)\, SE(\widehat{p}))^2\Big)$$

and for survey data (ignoring *fpc*)

$$h(\widehat{p}) - h(p) \approx \mathcal{N}\Big(0,\, \text{deff}\ (h'(p))^2\ \frac{p(1-p)}{n}\Big)$$

Re-express using **effective sample-sizes** $n_{eff} = n/\text{deff}$.

Variance-stabilizing $h(p) = asin(\sqrt{p})$ gives $h'(p) = 1/\sqrt{p(1-p)}$.

Confidence Intervals for Survey Proportions

Studied by Korn and Graubard (1998), Liu and Kott (2009).

Main idea for surveys: to take good *iid* CI and replace n by n_{eff}.

Korn & Graubard favor Clopper-Pearson, conservative interval based on exact binomial tail probabilities.

Liu & Kott compare many **one-sided** intervals, including modifications in spirit of Brown et al. (2001) with small-sample Edgeworth correction for skewness of $\widehat{p}$. Best are found to be a Cai (2004) and Kott-Liu (2009) interval, with interval based on $h(p) = asin\sqrt{p}$ good (*for small p only*) but slightly conservative.

Upper Confidence Bounds for $\hat{p} = 0$

Consider the upper CI bounds which arise for $\hat{p} = 0$, $z = z_{.05}$

Name	Formula	$n = 20$	$n = 10$	$n = 5$	$n = 3$
asin sqrt	$\sin^2(z/(2\sqrt{n}))$	.033	.066	.129	.209
Cai (2004)	$\frac{z}{6n}\sqrt{2z^2 + 7}$	.048	.097	.193	.322
Kott-Liu	$(2z^2 + 1)/(6n)$	.053	.107	.214	.356

NB. Values n here would be n_{eff} in practice.

ACS Approach to Confidence Bounds for $\hat{p} = 0$

ACS Design and Methodology, p. 12-4
A. Navarro Memo, 2001

Criterion: $N \cdot SE(\hat{p})$ for $\hat{p} = 0$ is defined as $C\sqrt{Avg.Wt}$

Avg.Wt = max of Average ACS HU weight and
Average final person weight
(averages over State for within-state estimate)

N = population size from which $\hat{p}$ was estimated.

Constant C = 20 was chosen in 2001 so that $\geq$ 90% of CI's $[0,\ z_{.05}\, N\, SE(0)]$ contained the 2000 census cell-count.

Propose to use *synthetic or small-area models* in order to find upper confidence bounds for small p's from ACS data.

The small cells in ACS Tables all subdivide larger demographic cells which are well estimated.

Data Structure in ACS Tables

Examples, for 2009 data on 805 Counties with 65,000+ pop'n:

(1) (**B01001**) Population by Race (7 mutually exclusive groups), Sex, and Age (14 groups), by County (805);

(2) (**B17001**) Poverty status (income above/below Pov level in last 12 months) by Race (7 groups), Sex, Age (13 groups) within County (805).

Synthetic & Small-Area Models for Proportions

Response variable: count Y_i of Group (e.g., Age 45-54) within County by Sex cell, $i = 1, \ldots, 805 * 7 * 2 = 11270$ (separate analysis for each Race)

Predictors:

- Race, Sex, St (52) or Region (11) factors, cell i
- FracWh, FracB, FracHsp by County
- Agefrac = fraction in Age-gp in St by Race by Sex cell
- AgfrRg = fraction in Age-gp in Region by Race by Sex cell
- PCT-URBA = percent of County in Urban blocks
- plus possible interactions

Predictor fractions recoded to $\text{logit}\Big(\max(\frac{1}{2N}, \min(x, 1 - \frac{1}{2N}))\Big)$

Comparisons of Different Models

Synthetic Model: $i = (a, s, r), \qquad p_{asr}^{Cty} = p_{a|sr}^{St} * p_{sr}^{Cty}$

Logistic Model: $Y_i \sim \text{Binom}(\nu_i, p_i)\ , \quad p_i = plogis(\mathbf{X}_i'\beta)$

$\nu_i =$ actual or effective sample size

Transformed Linear Model: $asin(\sqrt{Y_i/\nu_i}) = \mathbf{X}_i'\beta + u_i + \epsilon_i$

$\epsilon_i \sim \mathcal{N}(0, \frac{1}{4\nu_i})\ , \quad u_i \sim \mathcal{N}(0, \sigma^2)$

With $\sigma^2 = 0$: a variance-stabilized linear model, **but** with general σ^2 : an Arcsin-Sqrt Fay-Herriot (1979) type model

Effective Sample Sizes and Cell Pops in ACS

Restrict attention to (669 out of 805) of 65000+ pop Counties with 7 Age-Gp by Race min CellPop > 70 (except for Amer-Indian/Alaskan and Hawaiian/Pacific race-gps).

	Min.	1stQ	Med	Mean	3rdQ	Max.
SampSiz	1	16	54	489	406	33240

DESIGN EFFECTS BY AGE-GP

	45-54		55-64		65-74
Min.	0.0152	Min.	0.0155	Min.	0.0098
1stQ	0.1602	1stQ	0.1195	1stQ	0.1379
Median	0.2308	Median	0.1844	Median	0.2179
Mean	0.2584	Mean	0.2120	Mean	0.2441
3rdQ	0.3291	3rdQ	0.2822	3rdQ	0.3339
Max.	2.4653	Max.	0.8481	Max.	0.8710

Model Fits on ACS Data — Examples

Logistic Model, AgeGp 4, Race `Black`:

only `Age4frac` signif., coef. $= 0.99$.

similarly for Race `Asian`

Transformed Linear Model, AgeGp 5, Race `Black`:

`Age5frac, FracB` highly signif

similarly `Age5frac, FracAs` for Race `Asian`

Transformed Fay-Herriot Model, AgeGp 5, Race `Black`:

`Age5frac, FracB` both highly signif

similarly `Age5frac, FracAs` for Race `Asian`

CI's from ACS Age-group models

Fixed-effect logistic models: using Δ-method SE for $\hat{p}_i$

in models for AgeGp 4 &5 , Races Black & Asian:
CI's resp. cover 86, 83, 77, 68 `pct` of estimated Y_i/ν_i

Fixed-effect transf'd linear: Δ-method SE for $asin(\sqrt{\hat{p}_i})$

in models for AgeGp 4 &5 , Races Black & Asian:
CI's resp. cover 90, 89, 96, 96 `pct` of estimated Y_i/ν_i

`(no` $1/n_i$`'s were used in these fits)`

Fay-Herriot arcsin sqrt: Δ-method SE for $asin(\sqrt{\hat{p}_i})$

in models for AgeGp 4 &5 , Races Black & Asian:
CI's resp. cover 86, 84, 78, 71 `pct` of estimated Y_i/ν_i

`(may reflect need to correct the` n_i`'s)`

CI's from Transformed Models, Continued

Numbers of 0-count cells out of 1338 in AgeGp 4 &5 ,
Races Black & Asian: respectively 99, 143, 182, 282

`Upper Conf Bds for 0 cells in 4 combination Age-Gps` $\times$ `Races:`

	Min	1stQ	Med	Mean	3rdQ	Max.
AgeGp4, Black:	.004	.356	.450	.462	.588	.708
AgeGp5, Black:	.000	.218	.286	.321	.374	.708
AgeGp4, Asian:	.135	.276	.350	.370	.463	.708
AgeGp5, Asian:	.000	.194	.269	.295	.377	.708

Must still tally numbers of census cell-proportions which are covered, to check comparability with current ACS method.

Extended Synthetic Models for ACS

Proposal: continue to use Transformed FH Model of the form

$$asin(\sqrt{Y_i/\nu_i}) \;=\; b_1\, \texttt{Agefrac}_i \;+\; u_i \;+\; \epsilon_i$$

with additional predictor terms when they can be found. This is like the synthetic model except that it also 'borrows strength' for estimating variances across cells in different counties !

This seems simple enough to use in the **intended application of upper-confidence-bound construction**, applicable even when some (many ?) single-cell Y_i's are 0.

Summary & Conclusion

• Some usable methods exist for Upper Confidence Bounds for Zero-Estimated Proportions.

• Extending these methods to surveys requires 'effective sample sizes', which is problematic for ACS because of pop-controls.

• Explored CI's for ACS cell proportions based on models 'borrowing strength' across cells: small area style models.

• Proposed a method based on arcsin sqrt transformed Fay-Herriot model. Preliminary analysis suggests the predictor will usually be restricted to a synthetic-model transformed proportion; these models allow reasonable estimation of cell-level random effects. 'Effective sample sizes' remain a problem.

References

ACS Design & Methodology, sec. 12-4: Variance Estimation, and A. Navarro memos 2001

Hall, D. (2000), Zero-Inflated models. *Biometrics* **56**

Purcell, N. and Kish, L. (1980) SPREE estimators. *ISI Rev.*

Korn, E. and Graubard, B. (1998) CI's. *Surv. Meth.* **24**.

Liu, Y. and Kott, P. (2009), CI's. *J. Official Stat.* **25**

Rao, J. N. K. (2003) **Small Area Estimation**, Wiley.