Available online at www.isas.org.in/jisas
JOURNAL OF THE INDIAN SOCIETY OF
AGRICULTURAL STATISTICS 66(1) 2012 157-169

Assessment of Zeroes in Survey-Estimated Tables via Small-Area Confidence Bounds

Eric V. Slud
U.S. Census Bureau, Washington DC, and University of Maryland
Received 16 May 2011; Revised 31 August 2011; Accepted 01 September 2011

SUMMARY

Motivated by the problem of 'quality filtering' of estimated counts in U.S. American Community Survey tables, this paper studies methods for placing confidence bounds on zero estimates within demographically cross-classified tables which are estimated from complex surveys. While Coefficients of Variation are generally used in screening the quality of estimated counts, they do not make sense for assessing validity of zero counts. The problem of assessment is formulated here in terms of (upper) confidence bounds for unknown proportions. After summarizing published methods of constructing confidence intervals for proportions based on survey data, we study methods of creating confidence bounds from small-area models including synthetic, logistic, and variance-stabilized (arcsin square root transformed) linear models. The relations between these models and the confidence bounds they generate will be illustrated on demographic (Age-Race-Sex) American Community Survey tables from 2009 data for large (population at least 65000) U.S. Counties.

Keywords : arcsin square-root transformation, Confidence bounds, Effective sample size, Fay-Herriot model, Quality filtering, Synthetic model.

1. INTRODUCTION

As part of its general guidelines on quality of released data, the US Census Bureau regularly checks Coefficients of Variation of survey estimates before their release. The American Community Survey (ACS), the ongoing survey which was designed to replace the decennial-census Long Form last used in 2000, is a particularly prolific source of survey estimates at various levels of aggregation and cross-classification (http://www.census.gov/acs/www). In a process called 'data quality filtering' which is particularly elaborated and documented in ACS, release of some tables of estimates is prevented by rules relating to imprecision of estimates due to extremely small sample sizes (Starsinic 2009). These rules are generally expressed in terms of Coefficients of Variation (CV's) of the survey estimates in individual table entries, although special rules apply to entries with counts of 0, whose CV's are not defined. Generally in ACS, tables are *filtered*, remaining *unpublished* in American Factfinder data releases if a table's median CV among entries not defined as subtotals or totals of other entries is greater than 0.61.

Quality of survey estimates is generally summarized in US government publications in terms of CV's. However, it is recognized (ACS Methodology 2010, Ch. 12, p. 12-4) that CV's for zero estimates or very small counts or proportions are either undefined or not meaningful. The current ACS methodology for defining an artificial 'SE' for cells with 0 estimates (ACS Methodology 2010, p. 12-4) was developed in Navarro (2001): in a cell with population size N within which the proportion $\hat{p}$ is estimated and found to be 0, the convention is to define

$$N \cdot \text{SE}(\hat{p}) = C\sqrt{\text{Avg.Wt}} \qquad (1)$$

E-mail address : evs@math.umd.edu

where Avg.Wt is defined as the larger of the average final person weight and the average ACS housing unit weight, both averages being taken over each State for estimates of cells that do not cut across State boundaries. The constant $C = 20$ was chosen from data analyses in 2001 of ACS 2000 data so that at least 90% of confidence intervals $[0, z_{.05}\, N\, \text{SE}(0)]$ contained the corresponding 2000 census cell-count.

It is the purpose of this paper to compare and develop methods for assessing quality of survey estimates in settings, like those of many ACS tables, where estimated proportions of the surveyed units which have specified attributes are frequently zero or very small. After relating confidence bounds for CV's to confidence intervals, through transformations of the probability scale, in Section 2, we review in Section 3 the principal ideas for confidence interval construction for unknown proportions in a survey setting. We turn then in Section 4 to consider old and new ideas for confidence interval construction in surveys based on models. That leads us to the central topic of the paper, the application of area-level estimation models to the construction of upper confidence bounds (UCB's) for small survey-estimated proportions. The small area estimation methods described in Section 4 are applied to 2009 ACS data in Section 5. Findings and tentative conclusions, with particular reference to proposed methodology for ACS, are summarized in Section 6.

2. CV's AND CONFIDENCE BOUNDS FOR PROPORTIONS

The most common indicators of precision and quality of survey estimators are Standard Errors and, more particularly, relative Standard Errors as a fraction of the expected (or estimated) attribute value. When a positive average attribute value μ for a population is to be estimated from a survey, the relative standard error of the survey estimator $\hat{\mu}$ is

$$\widehat{\text{CV}}(\hat{\mu}) = \text{SE}(\hat{\mu})/\hat{\mu} \tag{2}$$

the estimated Coefficient of Variation (CV) for the estimator. Whether relative or absolute, these standard errors are useful in comparing estimators from different subpopulations or surveys because they are meaningful in constructing confidence intervals. Under the large-sample and large-population assumptions used to justify approximate normality of survey estimators, a level $1-\alpha$ confidence interval for the unknown μ is

$$\text{CI} = \hat{\mu} \pm z_{\alpha/2}\text{SE}(\hat{\mu})$$

$$= \hat{\mu} \cdot (1 \pm z_{\alpha/2}\,\text{CV}(\hat{\mu})) \tag{3}$$

where $z_r = \Phi^{-1}(1-r)$ is the standard-normal $(1-r)$-quantile and is Φ the standard normal distribution function.

Another interpretation of the CV is given by the logarithmic transformation and the Delta Method. That is, under assumptions justifying approximate normality, for positive μ the transformed estimator $\log(\hat{\mu})$ is also approximately normal, with standard error equal to the CV of $\hat{\mu}$, and a confidence interval of approximate level $1-\alpha$ for $\log(\mu)$ is given by

$$\log(\mu) \in \log(\hat{\mu}) \pm z_{\alpha/2}\widehat{\text{CV}}(\hat{\mu})$$

Precision of estimation of μ can be expressed either through bounds on the CI width in the original measurement scale μ through $\text{SE}(\hat{\mu})$, or through bounds on CI width in the logarithmic scale $\log(\mu)$ through $\widehat{\text{CV}}(\hat{\mu})$. It is in this spirit that we discuss precision of estimation of unknown population proportions $\mu = p$ through transformations of p. The primary tool is always the Delta Method, which for a known smooth strictly monotonic transformation $h(p)$, with estimator $\hat{p}$ which is consistent and approximately normal in large samples, says that

$$h(\hat{p}) - h(p) \overset{\mathcal{D}}{\approx} \text{N}(0, (h'(p)\text{SE}(\hat{p}))^2) \tag{4}$$

Our interest here is to study confidence intervals for small p, based on survey estimators $\hat{p}$. Precision of estimation of p could be measured either on the original probability scale or the h-transformed scale. In order to measure precision when p is very close to 0, particular interest centers on the transformation $h(p) = \arcsin(\sqrt{p})$ which is *variance-stabilizing* for binomial sampling, as we now explain.

When the estimator $\hat{p}$ is based on independent Bernoulli(p) sampling, with sample size n, its standard error is $\sqrt{p(1-p)/n}$. Approximate normality holds when n is moderately large, and since the transformation $h(p) = \arcsin(\sqrt{p})$ has derivative $h'(p) = 1/(2\sqrt{p(1-p)})$, equation (4) shows that the transformed estimator $h(\hat{p})$ is approximately normal with variance $1/(4n)$ not depending on the probability p being estimated. Thus, although the normality of $\hat{p}$ is questionable when np is not large, the transformation

h approximately frees the variance of $\hat{p}$ from dependence on p, a very desirable property when p is small.

More generally, the variance of a survey estimator of a proportion p is given as

$$(\text{SE}(\hat{p}))^2 = \text{deff} \cdot \frac{1-n/N}{n}\, p(1-p) \qquad (5)$$

where N is the population size and n the sample size on which the estimator is based, and the so-called 'design effect' deff is defined as the ratio of the variance of the estimator $\hat{p}$ under the actual complex survey design to that under Simple Random Sampling. (Here a factor $N/(N-1)$ is ignored in the formula for sample variance of the attribute indicator over the frame population.) As a result, we have the general approximate distribution for a transformed survey-estimator of proportion,

$$\arcsin(\sqrt{\hat{p}}) - \arcsin(\sqrt{p}) \overset{\mathcal{D}}{\approx} \text{N}\left(0, \text{deff} \cdot \frac{1-n/N}{4n}\right) \qquad (6)$$

The corresponding upper confidence bound (UCB) for p would be

$$p \in [0, \sin^2(\arcsin(\sqrt{\hat{p}}) + z_\alpha\{\text{deff}\,(1-n/N)/(4n)\}^{1/2})] \qquad (7)$$

Exact UCB's in the setting of Simple Random Sampling (without replacement) can be provided from the hypergeometric distribution. See Buonaccorsi (1987) for the test-based UCB, and Wright (1990) for a related bound on its coverage. In many survey applications, including ACS, the *finite population correction* factor $1 - n/N$ in (5)–(7) can be replaced by 1 because n is much smaller than N. Then the square of standard error in equation (6) can be re-expressed as $1/(4\, n_{\text{eff}})$, where the *effective sample size* is defined as $n_{\text{eff}} = n/\text{deff}$. Thus the variance-stabilizing arcsin square-root transformation retains the property of stabilizing the variance of survey estimators of proportions, with the necessary caveat that design effects deff can vary from subpopulation to subpopulation and cell to cell within a cross-classified survey and can therefore not be regarded as quite independent of the proportions p in different cells which possess a specified attribute.

3. BOUNDS ADAPTED FROM BERNOULLI SAMPLING

There is an extensive literature on confidence intervals for unknown binomial proportions in non-survey settings. We focus here on two papers which summarized that literature, Korn and Graubard (1998) and Liu and Kott (2009), in order to make an informed choice, supported by simulations, of methods applicable with survey estimators of proportion, including one-sided intervals or intervals for very small proportions.

As regards confidence interval methods for proportions estimated in Bernoulli, non-survey settings, Korn and Graubard (1998) and Liu and Kott (2009) take different approaches. Korn and Graubard favor the Clopper-Pearson or 'exact binomial' interval, which is reliably conservative because it is based on exact binomial tail probabilities. Liu and Kott instead compare many one-sided intervals, preferring intervals with coverage as close as possible to nominal over a broad range of true proportions (again, under Bernoulli sampling), without preferring conservative to anti-conservative violations of nominal coverage probabilities. They include several proposed modifications of normal-theory-based intervals which follow Brown *et al.* (2001, cited in Liu and Kott) in incorporating small-sample Edgeworth-expansion corrections for skewness of $\hat{p}$. Liu and Kott (2009) found the best among the various methods to be Cai's (2004) modification of the Brown *et al.* proposals, along with a further Kott-Liu (2009) modification. They provide a series of graphs showing the computational performance of these and other intervals, over all p, for moderate n. Their graphs also show that the interval (7) based on $h(p) = \arcsin\sqrt{p}$ through (6) is quite good and slightly conservative *for small p* (say $p \leq 0.2$) but seriously biased and anticonservative for larger p.

These Bernoulli-sampling intervals are made relevant to survey practice in the same way in both Korn and Graubard (1998) and Liu and Kott (2009). These papers, and virtually all other design-based work on confidence intervals for proportions in surveys, propose to take the best available non-survey intervals and apply them in survey contexts by replacing the actual sample size n with the effective sample size n_{eff} as defined above. (Korn and Graubard 1998 also refer to their own earlier work justifying certain t-intervals under strong survey-sampling assumptions. But the main thrust of their paper is to propose the use of effective sample sizes as described above.) The papers of Korn and Graubard (1998) and Liu and Kott (2009) support this approach with simulation studies which

Table 1. Upper 95% confidence bounds ($z = z_{.05}$) from methods compared by Korn and Graubard (1998) and Kott and Liu (2009), in the case $\hat{p} = 0$.

Method	Formula	n			
		20	10	5	3
arcsin sqrt	$\sin^2(z/(2\sqrt{n}))$	0.033	0.066	0.129	0.209
Hall (1982)	$(2z^2 + 1)/(6n)$	0.053	0.107	0.214	0.356
Kott-Liu	$(2z^2 + 1)/(3n)$	0.107	0.214	0.427	0.712
Clopper-Pearson	$(1 + n/F_{2,2n,.05})^{-1}$	0.139	0.259	0.451	0.632

confirm that the idea works well in some stratified-sampling examples.

The specific modifications of the confidence intervals (3) via (5) which Liu and Kott (2009) advocated were selected for overall performance under a range of p's, not particularly for their behavior when $\hat{p} = 0$. There is some interest in comparing these proposed intervals for this specific case of 0 estimates, to the arcsin sqrt interval (7), and to the Clopper-Pearson interval advocated by Korn and Graubard (1998). Table 1 shows this comparison: all of these intervals are considerably more conservative than the arcsin sqrt, which Kott and Liu (2009) already found to be somewhat conservative near $p = 0$. Another more complicated interval due to Cai (2004), highlighted by Kott and Liu, gives corresponding UCB values .097, .185, .343, .530. There is no reason to prefer any of these other intervals to (7) for small or 0 values of $\hat{p}$. Note that in survey contexts, the values n in the Table's UCB formulas would be replaced by n_{eff}.

4. CONFIDENCE INTERVALS FROM MODELS

The purely design-based approach of the previous section to constructing confidence bounds, treats estimation within each cell of a cross-classified survey separately. We consider next a series of model-based methods, which aim to combine knowledge across cells through models.

4.1 Types of Models

The easiest way to derive local-area estimates from models is to assume that the target parameters contain some of the same parameters that can be estimated directly at higher levels of aggregation. This is the idea of *synthetic models*. Rao (2003, Ch. 4) treats this idea in the context of linear models to augment direct estimation of the mean from surveys. Here, in connection with the estimation of demographic proportions at local areas of geography, the most natural type of synthetic model assumption is that certain conditional probabilities are the same for higher levels of aggregation as for lower levels. The specific example we consider below in the ACS concerns the probability $p_{a\,|\,k,c}$ of a randomly selected individual in County c to fall in age-category a given that the individual falls in demographic category k, where $k = (g, r)$ is a composite factor defined from gender g and Race group r. The synthetic assumption would be that this conditional probability is the same as $p(a|k, s)$, the probability that a randomly selected individual in State s falls in age-category a conditioned on falling in demographic category k, where s is the State containing county c. In symbols, this synthetic model assumption can be written in terms of unconditional probabilities within County and State as

$$p_{a,k}^{Cty} = p_k^{Cty} * p_{a,k}^{St} \Big/ p_k^{St} \qquad (8)$$

The next stage of complexity of models that can be used to augment direct small-area estimators is the Structure Preserving Estimation (SPREE) idea of Purcell and Kish (1980) reviewed in detail in Sec. 4.2.5 of Rao (1993). This is a generalization of synthetic models which relies on the existence of at least one cross-classifying variable b which is predictive for a and for which higher-level (e.g., State-level) marginal counts $N_{b,k,s}$ are known with high accuracy from other sources. The SPREE method is essentially a calibration idea which improves upon design-based methods which rely exclusively on direct information about individual cells developed only from the current survey. We do not go into this idea in detail here because it does not seem to offer promise in the ACS context discussed below.

A few authors (Noble *et al.* 2002, Zhang and Chambers 2004) explicitly brought loglinear modeling into Small Area Estimation near the time when Rao's

(2003) survey of the subject found little work to summarize concerning small area applications of loglinear modeling. Both of these cited papers explicitly consider loglinear model generalizations of SPREE estimation, and Zhang and Chambers (2004) are particularly concerned with small-area estimation of survey proportions. However, as in non-survey work, loglinear models have generally been subsumed into generalized linear models to accommodate continuous covariates and also to allow extensions to mixed-effects models, although loglinear or polytomous regression models (Agresti 2002) more readily accommodate constraints of response categorical probabilities summing to 1.

There are other published models with a Bayesian flavor applicable to the estimation of small proportions from surveys. For example, Liu *et al.* (2007) and Chen and Lahiri (2011) develop hierarchical models for Bayesian small-area estimation of survey proportions. We do not pursue Bayesian methods in this paper, although a Bayesian approach could be taken to the Generalized Linear and transformed linear models that we consider next.

We focus on models which explicitly allow both continuous and categorical area-level covariates, and which can be applied in the ACS context, with or without random area effects. Such models arise when the population counts with a specified attribute in cross-classified cells indexed by i have direct weighted estimates Y_i in a current survey, with observed vectors X_i of cell-level predictor variables, where cell i has sample size (or effective sample-size) ν_i and known population size N_i.

With survey data given in the format above, a first simple model is

$$\text{round}(Y_i\ \nu_i/N_i) \sim \text{Binom}(\nu_i, p_i),\quad p_i = (1 + \exp(-X_i'\beta))^{-1} \tag{9}$$

where round(x) denotes the rounded integer value of x. The model is easy to fit and work with, but the sampling variability is addressed only by the Bernoulli trial assumption. Lack of fit might be partly addressed by introducing a random intercept effect at cell level, but such random-effect models are less easy to work with than the mixed-effect version of the next model.

Motivated by the variance-stabilizing relation (6), giving sampling variability a standard form not depending on the unknown proportion, we model transformed cell probabilities arcsin($\sqrt{p_i}$) linearly, as

$$\arcsin(\sqrt{Y_i / N_i}) = X_i'\beta + u_i + \varepsilon_i,$$

$$\varepsilon_i \sim \mathrm{N}\left(0, \frac{1}{4\nu_i}\right),\ u_i \sim \mathrm{N}(0, \sigma^2) \tag{10}$$

where the error terms u_i and ε_i are independent of each other and also across index i. For $\sigma^2 = 0$, (10) is a variance-stabilized linear model; with general σ^2, it is an arcsin-square-root transformed Fay-Herriot (1979) model; and in what follows we also sometimes fit the model without terms ε_i. When there is only one error-term in (10), the model has only fixed effects, and the cell proportion $p_i = \sin^2(X_i'\beta)$ is the target parameter. When both ε_i and the area random effects u_i enter the model, the latter account for otherwise unmodeled differences in proportion between cells. In that case, $p_i = \sin^2(X_i'\beta + u_i)$ is the target parameter in cell i, and the transformed EBLUP predictor (Rao 2003) for p_i based on the model is

$$\hat{a}_i \equiv X_i'\hat{\beta} + \hat{\gamma}_i\,(\arcsin(\sqrt{Y_i/N_i}) - X_i'\hat{\beta}),$$

$$\hat{\gamma}_i = \frac{\hat{\sigma}^2}{\hat{\sigma}^2 + 1/(4\nu_i)} \tag{11}$$

4.2 Confidence Intervals for *p* from Models (9) and (10)

Each of the models (9) and (10) provides confidence or prediction intervals for cell probabilities p_i in moderate to large sample settings, by way of asymptotic standard errors and the Delta method. Software estimating model (9) provides an asymptotic variance-covariance matrix $\hat{\mathbf{V}}_{\hat{\beta}}$ for the maximum likelihood estimator $\hat{\beta}$, from which it is easy to check by the Delta method that the asymptotic variance of the estimator of p_i is given by

$$\mathrm{Var}\left((1+\exp(-X_i'\hat{\beta}))^{-1} - p_i\right) \approx \left\{\frac{\exp(X_i'\hat{\beta})}{(1+\exp(X_i'\hat{\beta}))^2}\right\}^2 X_i'\hat{\mathbf{V}}_{\hat{\beta}}X_i$$

leading to the Confidence Interval expressed in terms of the logistic distribution function $q(x) = e^x/(1 + e^x)$ and $\hat{p}_i = q(X_i'\hat{\beta})$ as

$$p_i \in q(X_i'\hat{\beta}) \pm z_{\alpha/2}\, q\left(X_i'\hat{\beta}\right)\left(1 - q\left(X_i'\hat{\beta}\right)\right)\left[X_i'\hat{\mathbf{V}}_{\hat{\beta}}X_i\right]^{1/2} \tag{12}$$

Similarly, when parameters (β, σ^2) are estimated in model (10) by maximum likelihood, a variance-covariance estimator $\hat{\mathbf{V}}_{\hat{\beta}}$ is also produced, leading to the following Prediction Interval for p_i based on (10)–(11) :

$$p_i \in \sin^2 \left(\hat{a}_i \pm z_{\alpha/2} \left\{ (1-\hat{\gamma}_i)^2 X_i' \hat{\mathbf{V}}_{\hat{\beta}} X_i + (1-\hat{\gamma}_i)\hat{\sigma}^2 \right\}^{1/2} \right) \cap [0, 1] \quad (13)$$

Other, more precise variance estimators could be used in this interval as discussed by Rao (2003), but the number of small areas in the ACS application is large enough that higher order corrections may not be important.

In each of the intervals (12) and (13), replacing the lower endpoint by 0 and $\alpha / 2$ by α provides a one-sided upper-bounding interval.

4.3 Prediction Intervals for Y_i/N_i from Models (9) and (10)

In order to examine the fit of models (9) and (10) on ACS data, we consider also the prediction intervals for Y_i/N_i derived from these models, respectively based on the predictors $(1 + \exp(-\hat{\beta}'X_i))^{-1}$ and $\max(0, \min(\sin^2(\hat{a}_i), 1))$. Large-sample prediction intervals are easily derived by considering the approximate distribution in each case of the difference between Y_i/N_i and the predictor, on the respective logit and arcsin-square-root scales. For the logistic model, the prediction interval obtained by the Delta method is

$$Y_i / N_i \in q(X_i'\hat{\beta}) \pm z_{\alpha/2} q(X_i'\hat{\beta})(1 - q(X_i'\hat{\beta})) \left[1/v_i + X_i' \hat{\mathbf{V}}_{\hat{\beta}} X_i \right]^{1/2} \quad (14)$$

while in the transformed Fay-Herriot model, the interval is

$$\arcsin(\sqrt{Y_i / N_i}) \in \hat{a}_i \pm (1-\hat{\gamma}_i)\, z_{\alpha/2} \{ X_i' \hat{\mathbf{V}}_{\hat{\beta}} X_i + \hat{\sigma}^2 + 1/(4v_i) \}^{1/2} \quad (15)$$

5. ACS DATA ANALYSES

The application of small-area models of the types described in Section 4.1 to produce confidence or prediction bounds for attribute survey proportions p_i within cells i depends strongly on finding adequate explanatory variables X_i. As has been discussed in previous papers (Noble *et al.* 2002, Zhang and Chambers 2004, and Liu *et al.* 2007), models for proportions in large demographically cross-classified surveys are naturally hierarchical in the sense that related proportions at higher levels of aggregation provide natural predictors. The higher-level proportions will often be well estimated by direct design-based survey-weighted estimators.

The U.S. American Community Survey (ACS) provides a rich test-bed for small area models of proportions. In this paper, we restrict attention to proportions within Race-by-Sex groups within Counties which are particularly large, that is, which have 65000 or larger census population.

However, the ACS has one further feature which has a strong impact on demographically cross-classified analyses at the county level, namely that the population totals are controlled (i.e., raked or calibrated) to updated census totals (U.S. Census Bureau Estimates, Overview document 2011) at the level of cells defined through County demographic groups. These controls imply (Asiala *et al.* 2010) that the population in the cross-classified cells defined by County and Sex and 6 mutually exclusive Race/Ethnic groups and 13 Age-groups are constrained to be equal to the corresponding population totals in the updated Census. The Race/Ethnic groups used in defining these cells are: Hispanic, Non-hispanic White, Non-hispanic Black, Non-hispanic American Indian or Alaska Native, Non-hispanic Asian, and Non-hispanic Native Hawaiian or Other Pacific Islander. The age-intervals defining the age-groups used in controls are: 0-4, 5-14, 15-17, 18-19, 20-24, 25-29, 30-34, 35-44, 45-49, 50-54, 55-64, 65-74, and 75+.

In 2009, there were 805 Counties with 65,000+ population. Our source of cross-classified data is the so-called ‘race-iterated’ series of tables considered for publication by the ACS. Two examples of such table series are :

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

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

The age-groups used in the published ACS tables are slightly different from those used in population controls, but the totals for groups 45-54, 55-64 and 65-74 used in example model analyses below are precisely controlled. By contrast, the Race groups appearing in ACS published tables are not exactly the same as those used in controls, except for the two final groups H=Hispanic and I=White alone not Hispanic. The control groups explicitly exclude hispanics from each non-hispanic group, while the first seven Race groups used in the ACS tables do not. The first 7 Race Groups in the ACS race-iterated tables are: A=White alone, B=Black alone, C=American Indian or Alaska Native alone, D=Asian alone, E=Native Hawaiian or Other Pacific Islander alone, F=Some Other Race alone, and G = Two or More Races alone. The word 'alone' enters here because current U.S. censuses and the ACS allow respondents to choose multiple racial classifications.

5.1 Demographic Data Structure and Effective Sample Size in ACS

The models discussed in Section 4.1 are applied here primarily to a data structure defined from the purely demographic tables B01001 mentioned above. (For analyses of the related but different data structure of B17001, with poverty-indicator as response, see Section 5.4 below.) Cells are defined by cross-classifying County by Race by Sex are indexed by $i = 1, \ldots, 805 * 7 * 2 = 11270$. Counts N_i are the total cell populations, and the response variables Y_i are defined as the direct weighted survey estimators of population counts within specified Age-intervals, such as the count of individuals Age 45-54 within cell i.

Table 2. Numbers of zero-population county-by-race-by-sex cells out of 1610 in 2009 ACS in the 65000+ population counties, by Race group.

Race Group	A	B	C	D	E	F	G
# Zero-pop Cells	0	17	167	42	1005	57	2

Predictor variables considered as entries of X_i in modelling Age-Group Count responses by cell for each Race include:

- Race, Sex, State (52) or Region (11) factor (dummy) variables;
- FracWh, FracB, FracAs = fraction of population respectively in A, B, and D race groups in County
- Agefrac = fraction in Age-gp in State by Race by Sex cell
- AgfrRg = fraction in Age-gp in Region by Race by Sex cell
- PCT-URBAN = percent of County in Urban blocks

plus possible interactions of these variables. Predictor fractions are recoded to logit(max(1/(2N), min(x, 1 – 1/(2N)))).

Some of the 1610 County-by-Sex cells for each Race group in the 805 large-population counties have 0 population. Table 2 shows the numbers of zero-population cells (out of 1610) by Race Group. The American Indian and Alaska Native category C and Native Hawaiian and other Pacific Islander category E have many zero cell-populations, heavily depending on region of the U.S. So in developing examples below, we restrict attention to the other 5 race groups and to the 669 counties for which all county-by-sex cells of those 5 race groups have updated-census cell population greater than 70. Within the remaining 5 * 669 * 2 cells, the ACS 2009 sample-sizes ranged from 1 to 33240, with median 54.

Since the County-Race-Sex cell population totals N_i in ACS tables are (almost, except for the slight disparity in defining Race groups discussed above) controlled to the values from the updated census, there is little variability in them even though they are derived from survey data. This effect of population controls on variability persists down to the level of Age-groups within County-Race-Sex cells, as can be seen in the calculated design effects displayed in Table 3. While many complex multistage and survey designs result in design effects ranging from 1.5 to 7 or more, here the design effects are generally of order 0.1 to 0.3. Recall that the reciprocal of the design effect is a multiplicative

Table 3. Design-effects by age-group for County-Race-Sex cells restricted to the (669) Counties with 65000 + population and cells of population > 70 for all ACS Race groups other than C and E.

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

factor converting sample size to effective sample size: this is a range of design effects far different from those used in previous simulation studies like those of Korn and Graubard (1998) and Liu and Kott (2009) in connection with Section 3 to support the use of non-survey Confidence Intervals with effective sample sizes on survey data. For this reason, we use actual sample-size numbers rather than effective sample-sizes as v_i in the models fitted below.

5.2 Results and Model Checks on ACS Data

The data structure created to fit models (9) and (10) to Age-group proportions within County-Sex cells for each specified Race is $\{(Y_i, N_i, \mathbf{X}_i)\}_{i=1}^{1338}$, with cell-index restricted to the 669 counties with counts > 70 for race groups A, B, D, F, and G, and where $\mathbf{X}_i$ are vectors of predictor variables including dummies for sex and region as well as the transformed age-group proportion Agefrac within state-by-sex aggregates. Predictors other than Agefrac, Region, AgefrRg are defined at county level.

The most striking finding in fitting models (9)–(10) in non-white racial groups is the paucity of strongly significant predictor variables other than the Agefrac synthetic variable. Consider, for example, the logistic regression models (9) for Age-Group 45-54 in the Black only and Asian only ACS Racial groups (groups B and D, as in Table 2 and preceding text). In these models, only the Agefrac variable is extremely significant, with respective z-values 15.2 and 17.6, and coefficients 0.99 and 0.98, making the logistic model (9) essentially the same as the synthetic model (8).

By contrast, for Age-group 55-64 several other variables also seem significant, as can be seen in the analysis of deviance, for the complete model selected for the Black only race-group:

ANALYSIS OF DEVIANCE–Logistic Model, Black 55-64

	Df	Deviance	Resid. Df	Resid. Dev
NULL			1337	1983.2
Agefrac	1	490.74	1336	1492.4
FracB	1	49.12	1335	1443.3
PCT_URBAN	1	13.29	1334	1430.0
FracWh	1	7.46	1333	1422.6
FracB:PCT_URBA	1	12.06	1332	1410.5

We illustrate the quite different model selected for the Asian race-group through the coefficients

STANDARDIZED COEFFICIENTS–Logistic Model, Asian 55-64

	Estimate	Std. Error	z value	Pr (>|z|)
(Intercept)	0.1590	0.0990	1.606	0.1084
Agefrac	1.1625	0.0516	22.537	1.8e-112
PCT_URBAN	0.0153	0.0026	5.844	5.09e-09
FracAs	–0.0563	0.0103	–5.465	4.63e-08

In Age-group 65-74, within both Races, the same variables are significant as in Age-group 55-64, except that FracWh is dropped as insignificant in the Black race-group. When predictors other than Agefrac are significant, as in the two models just displayed, the coefficient of Agefrac differs from 1, and the model goes beyond (8).

When the transformed linear models (10) were fitted with u_i but no ε_i terms, slightly different predictors beyond Agefrac were found significant: in Age-group 45-54, only FracB was highly significant in the Black race group, and in the Asian race group nothing beyond Agefrac was significant; in Age-group 55-64, FracB was highly significant in the Black race-group, as was FracAs in the Asian group. Lastly, models (10) with random effects u_i and sampling errors ε_i were fitted in the same three age-groups for each of the Black and Asian race-groups, and showed exactly the same significant variables as in the model (10) model fits without ε_i.

As described in Section 4.3, nominal 95% prediction intervals for sample proportions Y_i/N_i were computed from each of the fitted models, partly as a way of assessing fit of the models by calculating the proportion of such confidence intervals containing the actual observations Y_i/N_i. In the present ACS setting, because of the county-level demographic population controls, the observations Y_i/N_i are guaranteed to be close to the corresponding proportions from the decennial census updated to 2009. Table 4 displays the two-sided coverage and median widths for the 95% confidence intervals (14) for Y_i/N_i derived from model (9) and prediction intervals (15) from the arcsin square-root transformed Fay-Herriot model (10) including both the random effect and sampling error terms. The predictors included in each model, for respective Race

Table 4. Coverage (fraction of 1338 cell proportions Y_i/N_i falling in intervals) and median widths for 95% logistic and arcsin transformed prediction intervals for age-group proportions, presented by race-age groups.

Race	Age Group	Pred.Int.(14)		Pred.Int.(15)	
		Coverage	Med. width	Coverage	Med. width
Black	45-54	85.6%	0.143	85.9%	0.124
Asian	45-54	76.5%	0.253	78.1%	0.227
Black	55-64	82.7%	0.114	83.7%	0.100
Asian	55-64	69.1%	0.196	71.4%	0.163
Black	65-74	71.9%	0.080	73.2%	0.048
Asian	65-74	55.6%	0.134	58.1%	0.084

and Age groups, are precisely those described in the preceding paragraphs and displayed analysis-of-deviance and coefficient tables as being significant. However, for all of these cases the prediction intervals based on models including only Agefrac as a predictor (which were calculated, but are not shown) were different by at most a few per cent in coverage or width from the intervals summarized in Table 4.

Table 4 says that on the ACS demographic table data, the two types of model, (9) and (10), lead to prediction intervals with very similar coverage rates, while the interval widths are clearly smaller for the prediction interval based on the transformed Fay-Herriot model. However, neither model fits well, in view of the great disparity between the nominal coverage rate of 95% and the actual coverage rates seen in the Table. The sense of coverage displayed here, as a summary statistic, is the average coverage over groups of cells, as in Zhang (2007).

There are many ways that the large-sample coverage for prediction intervals based on models like (9) and (10) can fail. In the numerous cells with small sample size, the normal distribution behavior of observed cell proportions in (9) is not valid. Similarly, in (10), for small-sample cells there is no real justification for the normality of the sampling-error terms ε_i, while there was never a compelling reason to assume the cell random effects u_i were normal.

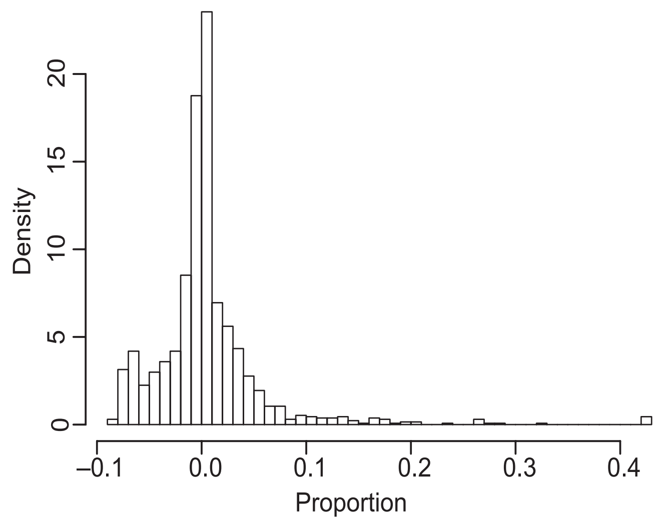

Fig. 1. Histogram of residuals from transformed Fay-Herriot model fit to ACS 2009 Age-group 55-64 within Black Large-county cells.

Checking for non-normality of residuals in either model confirms that the failure of distributional assumptions used in the prediction intervals is a likely reason for the failures of coverage seen in Table 4. This use of histograms of residuals as model diagnostics is not common in works on small-area estimation, but has been used before, for example in the paper of Slud and Maiti (2011). We emphasize the models (10), because these are the models we propose for future use. A histogram of residuals on arcsin-square-root scale from the predictions (11) in the model for 55-64 within Black cells is plotted in Fig. 1. It is typical of many other such histograms, not shown: the empirical distribution of residuals is essentially bimodal (with a small mode near the lower extreme), skewed, and heavy-tailed.

Since the residuals are so non-normal, we propose a modification of the confidence intervals (13) to reflect the skewness and long tails. In place of the symmetric percentage points $\pm z_{\alpha/2}$ appearing in (13), there should instead appear the $\alpha/2$ and $1 - \alpha/2$ quantiles of the approximately standardized residuals

$$\frac{\left(\arcsin\left(\sqrt{Y_i/N_i}\right) - X_i'\hat{\beta}\right)}{\left\{X_i'\hat{\mathbf{V}}_{\hat{\beta}}X_i + \hat{\sigma}^2 + 1/(4v_i)\right\}^{1/2}} \tag{16}$$

This modification of (13) is implemented in the next subsection, where model-based UCB's are applied to cells with estimated proportions of 0.

5.3 Confidence Bounds for Estimated Proportions of 0

In the examples considered above, age-groups 45-54, 55-64, and 65-74 were viewed as subsets of Black and Asian county-by-sex cells within large ACS counties with positive sample-size. In the 2009 ACS data, many cells contained no sampled individuals in these age-groups, as summarized in Table 5. In those cells, it is interesting to compare the UCB's for proportions of the cell-population falling in the respective age-groups derived from single-cell CI's to those based on the model-based CI's. The brief comparison undertaken here is restricted to the single-cell arcsin-square-root UCB (7), using deff= 1, versus the UCB based on the arcsin square-root transformed Fay-Herriot model, through (13) modified by the empirical quantiles of the standardized residuals (16). Table 6 displays the median, upper-quartile, and maximum of the UCB's generated by the two arcsin-square-root transformed methods (7) and (13) for the cells within the separate model-based analyses by race-and age-group. This comparison shows that the single-cell method (7) often leads to much larger UCB's than does the model-based method (13), perhaps unnecessarily large. Note that even the relatively smaller UCB proportions arising from the model-based method will involve some large counts, based on larger cell populations. It is not clear yet whether the model underlying (13) is sufficiently reliable to become a general method of quality assessment for zeroes in purely demographic ACS tables. We next consider analogous models for ACS tables with poverty-status as response-variable.

Table 5. Numbers of county-sex cells out of 1338 for Black and Asian which have 0 sampled individuals in specified age-groups, in ACS 2009.

Race-group	Age-groups		
	45-54	55-64	65-74
Black	99	143	269
Asian	182	282	464

Table 6. Upper confidence bounds for proportions p_i in age-groups within cells for Black and Asian race, where sampled counts in age-group are 0.

		Cell-based UCB			Model-based UCB		
Race	Age-gp	Med	Q3	Max	Med	Q3	Max
Black	45-54	0.106	0.182	0.536	0.136	0.145	0.184
Asian	45-54	0.127	0.207	0.535	0.129	0.144	0.188
Black	55-64	0.072	0.156	0.536	0.068	0.076	0.086
Asian	55-64	0.093	0.158	0.535	0.076	0.089	0.152
Black	65-74	0.054	0.107	0.536	0.027	0.032	0.062
Asian	65-74	0.073	0.128	0.535	0.029	0.036	0.061

5.4 Analyses with Poverty-indicator Response

We consider now an analogous model-fitting exercise and UCB calculation for the ACS table series B17001 in which a Poverty-status indicator is tallied within a County by Race by Sex by Age-Group cross-classification for counties with 65000+ population. The frame population now consists only of those individuals for whom poverty status could be determined, so is slightly smaller than in tables B01001. The response variable of interest is the survey-weighted direct estimate Y_i of the number in poverty within each County × Race × Sex × Age-group cell. There are again many such cells with very small population and sample, so we restrict attention to 10322 cells defined in the following way:

- Counties are restricted to the 650 (a proper subset of the 669 considered before) in which all County × Race × Sex population counts are greater than 70, when Race groups are restricted to A,B,D,F,G;
- Age-groups are restricted to 45-54, 55-64, 65-74 and Race-groups to A, B, D, and County × Race × Sex × Age cell counts are all at least 10.

All of the 10322 cells i defined in this way had population size $N_i \geq 10$ and sample size $v_i \geq 1$. These cells form a subset of the array of 11700 cells with 650 Counties, 3 Race-groups, 2 Sexes and 3 Age-groups. The predictor variables used in fitting models of type (10) for the expected proportions Y_i/N_i were selected from among the following along with their interactions:

- PovSyn, a synthetic variable defined for each cell i as the arcsin-square-root transformed proportion of the State population that is in the same Race, Sex and Age category as cell i;
- dummy variables for Sex, Age-group, Race;

- RacFr = logit-transformed proportion of County population in the same Race category as cell i; and
- URB = arcsin-square-root transformed fraction of County in urban blocks.

The model of type (10) fitted to these data has estimated coefficients and standard errors as follows:

Term	Coef	SE
Const	–0.0154	0.0072
PovSyn	1.0050	0.0180
URB	0.0068	0.0049
RaceB	0.0408	0.0057
RaceD	0.0395	0.0074
RacFr	0.0535	0.0041
URB:RacFr	–0.0219	0.0034

with random effect σ^2 estimated as .00358 with standard error .00011. Other coefficients, for dummy variables for Sex and Age, were only barely significant and not predictively useful. In this instance, although there were several useful predictors beyond the synthetic variable PovSyn, the coefficient for the synthetic variable was indistinguishable from 1. A histogram of the standardized residuals (16) from this model were found to be slightly asymmetrically distributed, skewed to the left, but to have approximately normal tails, with upper 95% quantile of 1.957 in place of the nominal 1.645.

The fitted transformed Fay-Herriot model just described was next applied to construct UCB's for cell population proportions in poverty for those cells where the direct estimates of poverty rates are 0, just as was previously done for cell proportions in Table 6. Table 7 displays, by Race and Age categories, the numbers of cells with direct poverty-rate estimates Y_i/N_i of 0, along with the purely cell-based 95% UCB (7) for poverty rate and the model-based analogue (13). It can be seen in this Table that for those Race × Age categories with large numbers of (mostly very small) cells with $Y_i = 0$, the model-based UCB's are generally much tighter than the cell-based UCB's. In the White or A race group, the cell counts N_i were generally larger, and there were relatively few cells with $Y_i = 0$. The fitted model in the latter cells tended to predict poverty rates sufficiently greater than 0 that the UCB's were actually larger than the cell-based UCB's. These results seem reasonable and suggest a further data setting where model-based UCB's could be used to assess the merit and publishability of direct estimates of 0.

Table 7. Upper 95% confidence bounds for proportion in Poverty with cells defined by (County, Race, Sex, Age), by purely Cell-based calculation (7) or model-based (13).

			Cell-based UCB			Model-based UCB		
Race	Age-gp	ncell	Med	Q3	Max	Med	Q3	Max
White	45-54	5	0.011	0.015	0.016	0.034	0.043	0.063
Black	45-54	296	0.208	0.524	0.536	0.168	0.213	0.446
Asian	45-54	703	0.207	0.524	0.536	0.066	0.097	0.269
White	55-64	15	0.008	0.011	0.023	0.034	0.044	0.066
Black	55-64	353	0.294	0.527	0.536	0.076	0.102	0.336
Asian	55-64	667	0.093	0.158	0.535	0.076	0.089	0.152
White	65-74	65	0.014	0.019	0.037	0.043	0.054	0.092
Black	65-74	367	0.293	0.528	0.536	0.164	0.209	0.620
Asian	65-74	535	0.298	0.530	0.535	0.112	0.146	0.553

6. SUMMARY AND CONCLUSIONS

This paper has expounded the following points:

- Usable methods do exist for Upper Confidence Bounds (UCB's) for estimated proportions 0 based on single cells of tables.
- Extending these methods to surveys requires 'effective sample sizes', which is problematic for surveys like the American Community Survey where population-controls can lead to design effects much less than 1.
- UCB's for ACS cell proportions have been developed based on small-area style models which 'borrow strength' across cells.
- In exploratory analysis of ACS demographically cross-classified tables with age-group as the finest subdividing category, the synthetic-model predictor Agefrac has in almost all cases been the only important predictor in the fitted small area models.
- In ACS tables with Poverty indicator nested within demographically cross-classified age-group as the finest subdividing category, preliminary analysis finds several strongly significant predictors in addition to the synthetic predictor PovSyn.

The outcome of this study is a proposed method based on an arcsin square-root transformed Fay-Herriot model. Such extended random-effect synthetic models allow reasonable estimation of cell-level random effect variances, along with EBLUP prediction and confidence bounds.

Proposed Method: Fit a transformed Fay-Herriot model of the form

$$\arcsin(\sqrt{Y_i/N_i}\,) = b_1 \text{ Synth}_i + u_i + \varepsilon_i \qquad (17)$$

in terms of a synthetic predictor Synth_i, adding in other highly significant predictor terms $\beta' X_i$ when they can be found. The basic model (17) resembles the synthetic model (8) when there are no other strong predictors, but it also 'borrows strength' for estimating variances across cells in different demographic categories and counties. From the EBLUP predictors (11), UCB's can be constructed from data for p_i as the one-sided CI's (13). In case the empirical distribution of the residuals (16) is very non-normal, due to skewness and long tails, the normal percentage point z_α appearing in the UCB from (13) should be replaced by the empirical $1 - \alpha$ quantile of the standardized residuals (16).

This method is simple enough to use in the intended application of upper confidence-bound construction, applicable even when many single-cell Y_i's are 0. 'Effective sample sizes' remain a problem in the context of ACS demographic tables, but may still prove useful in connection with the proposed method in other survey applications. However, additional research is needed, both in ACS and other contexts, to confirm that this method performs satisfactorily in capturing accurate estimates (validated from supplementary surveys or censuses) of the desired small proportions.

ACKNOWLEDGEMENT

The author is grateful to Freddie Navarro of the US Census Bureau, for a clear statement of the methodological problem concerning zeroes in ACS tables, and to Michael Starsinic, also of the Bureau, for providing orientation and access to the data and relevant internal Census Bureau reports. All data analyses described in this paper have been performed in the R (2009) statistical computing platform.

REFERENCES

Agresti, A. (2002). *Categorical Data Analysis*. 2nd ed. Wiley-Interscience, Hoboken.

American Community Survey Methodology (2010). Variance Estimation. http://www.census.gov/acs/www/methodology/methodology-main

Asiala, M., Beaghen, M. and Albright, K. (2010). Evaluating Alternative Population Controls for ACS Estimates. Census Bureau Report.

Buonaccorsi, J. (1987). A note on confidence intervals for proportions in finite populations. *Amer. Statist.,* **41**, 215-218.

Cai, T. (2004). One-sided confidence intervals in discrete distributions. *J. Statist. Plann. Inf.,* **131**, 63-88.

Chen, S. and Lahiri, P. (2011). Inferences on small-area proportions. *J. Ind. Soc. Agril. Statist.*

Fay, R. and Herriot, J. (1979). Estimation of income from small places: An application of James-Stein procedures to census data. *J. Amer. Statist. Assoc.,* **74**, 269-277.

Hall, D. (2000). Zero-inflated poisson and binomial regression with random effects: A case study. *Biometrics,* **56**, 1030-1039.

Hall, P. (1982). Improving the normal approximation when constructing one-sided confidence intervals for binomial or poisson parameters. *Biometrika*, **69**, 647-652.

Korn, E. and Graubard, B. (1998). Confidence intervals for proportions with small expected number of positive counts estimated from survey data. *Survey Methodology*, **24**, 193-201.

Liu, Y. and Kott, P. (2009). Evaluating alternative one-sided coverage intervals for a proportion. *J. Official Statist.*, **25**, 569-588.

Liu, B., Lahiri, P. and Kalton, G. (2007). Hierarchical Bayes modeling of survey-weighted small area proportions. *Proc. Amer. Statist. Assoc.*, Survey Res. Meth. Section, pp. 3181-3186.

Navarro, A. (2001). Estimating standard errors of zero estimates. Internal U.S. Census Bureau Memorandum.

Noble, A., Haslett and Arnold, G. (2002). Small area estimation via generalized linear models. *J. Official Statist.*, **18**, 45-60.

Purcell, N. and Kish, L. (1980). Postcensal estimators for local areas (or domains). *Internat. Statist. Rev.*, **48**, 3-18.

R Development Core Team (2009). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna. http://www.R-project.org

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

Slud, E. and Maiti, T. (2011). Small-area estimation based on survey data from a left-censored Fay-Herriot model. *J. Statist. Plann. Inf.,* **141**, 3520-3535.

Starsinic, M. (2009). Assessment of data release rules on the reliability of multiyear estimates in American Community Survey data products. *Proc. Amer. Statist. Assoc.*, Surv. Res. Meth. Section, pp. 3509-3522.

U.S. Census Bureau Estimates: Overview document (2011). http://www.census.gov/popest/overview.html

Wright, T. (1990). When zero defectives appear in a sample: upper bounds on confidence coefficients of upper bounds. *Amer. Statist.*, **44**, 40-41.

Zhang, L-C. and Chambers, R. (2004). Small area estimates for cross-classifications. *J. Roy. Statist. Soc. Ser.,* **B66**, 479-496.

Zhang, L.-C. (2007) Finite population small area estimation. *J. Official Statist.,* **2**, 223-237.