VARIANCE ESTIMATES OF THE CORRELATION BIAS ESTIMATES OF THE TWO - GROUP MODEL

Roger Shores and Robert Sands, U.S. Bureau of the Census
Roger Shores, U.S. Bureau of the Census, Washington, D.C. 20233

Key words: correlation bias, dual-system estimation

Introduction

Dual-System Estimation is an example of the familiar capture-recapture method for estimating the number of items that exist in a population. In this case we have counts obtained for the Census and for the Accuracy and Coverage Evaluation (A.C.E.), and the problem is to estimate the number of people missed by both the census and the survey. The Dual System Estimates (DSEs) calculated for poststrata in the A.C.E. in Census 2000 were subject to a particular kind of error called correlation bias. This occurs whenever the probability that an item will be captured in the census is not independent of the probability that it will be captured in the associated coverage survey, and vice versa. There are two possible sources of correlation bias, causal dependence and heterogeneity. Causal dependence refers to the event that the inclusion of an item into either the survey or the census changes the probability that the item will be included in the other. Heterogeneity occurs if subgroups within the population have different probabilities of being included in the survey and the census. If this is the case, then even though there can be complete statistical independence between the survey and the census for a particular subgroup, when the results are aggregated over all of the subgroups, the resulting totals display correlation between the census and the survey. The usual concern with correlation bias is with heterogeneity leading to an underestimate of the population.

Discussion

William Bell (1993) has developed alternative dual system estimators that make use of demographic analysis (DA) and that allow for the calculation of estimates of the correlation bias. The information that is known and that will be estimated can be presented in 2 x 2 tables as:

		A.C.E.		
	Model	In	Out	Total
Census	In	p_{k11}	p_{k12}	p_{k1+}
	Out	p_{k21}	p_{k22}	p_{k2+}
	Total	p_{k+1}	p_{k+2}	1

		A.C.E.		
	Data	In	Out	Total
Census	In	x_{k11}	x_{k12}	x_{k1+}
	Out	x_{k21}		
	Total	x_{k+1}		

In the first table, p_{k**} represents the probability of inclusion for that cell for poststratum k. In the second table, x_{k**} represents the count of people for that cell for poststratum k. For the later table, we have the count for the Out-Out cell as the missing piece of information. This is the number of people missed by both the Census and the A.C.E. The total desired is given by $N_k = x_{k11} + x_{k12} + x_{k21} + x_{k22}$, the true count of the population in poststratum k. The DSE that was used in Census 2000, and that assumes independence between the Census and the A.C.E., is given by

$$\hat{N}_k^I = x_{k(1)} + \hat{x}^I_{k22} ,$$

where $x_{k(1)} = x_{k11} + x_{k12} + x_{k21}$, and

$$\hat{x}^I_{k22} = \frac{x_{k12}\, x_{k21}}{x_{k11}} .$$

A basic assumption made under Bell's methodology is that independence exists for females so that correlation bias is only a potential problem for males. Bell's estimators call for the use of three age categories – 18 - 29, 30 - 49, and 50 and over – and two race categories, Black and Nonblack. There were inconsistencies between the way races were classified between the census and DA that made it impractical to have more than these two race categories. For a given age-race group, the usual DSEs are calculated for females for all poststrata, and then summed over the poststrata to obtain

a national total for females for the age-race group. Then, sex ratios of males to females obtained from demographic analysis are used to calculate "control totals" for males by multiplying the sex ratios by the national totals for females. These control totals are then calculated for each age-race group for males. This control total is assumed to be equal to the sum over the k sub-populations for males within a particular age-race group. The usual DSEs that assume independence are calculated for each poststratum and then summed over all poststrata within a given age-race group.

The control total for males for a given age-race group is given by

$\hat{N}^{DA} = r^{DA} \hat{N}_f^I$, where

r^{DA} is the sex ratio of males to females for a given age-race group, and

$\hat{N}_f^I$ is the DSE for the independence case for females for that group.

Each model was created by taking a different function of the 2 x 2 table of probabilities and assuming it to be constant across poststrata within an age-race group. The function estimate is calculated such that, when individual DSEs within a given age-race group are summed, the resulting aggregation is equal to the comparable DSE determined from DA. The difference between the control total and the usual independence-assumption DSE for each age-race group is calculated. This provides an estimate of the correlation bias at the national level for the age-race groups. The correlation bias for each of the alternative estimators is given by this difference, allocated proportionateley according to assumptions for the estimator, across poststrata within the age-race groups.

When this methodology was applied to the 2000 A.C.E., some of the findings were:

- There was significant correlation bias for adult Black males, ranging from -4.7 percent for males 50 and over to -8.1 percent for males 30-49.
- Correlation bias for Nonblack males 30 and over was small.
- There were serious inconsistencies between the DA and A.C.E. estimates for Nonblacks 18-29. These made it difficult to use the DA results to calculate correlation bias estimates, so the assumption was made that there was no correlation bias for that group.

The model for which we will provide variance estimates of the correlation bias for the DSEs calculated from the model is known as the two-group model. In the two-group model it is assumed that in each poststratum people belong to one of two groups, such as "hard-to-count" and "easy-to-count" persons. It is further assumed that there is a parameter, η, that is constant across poststrata within each age-race group. It is important to understand that in this context it is not actually possible to divide the population into two discrete groups. Nevertheless, through the use of this methodology it is possible to develop a useful estimation scheme.

The capture probabilities, as represented by the 2 x 2 tables for the two groups, are assumed to differ. Let the proportion of the population in group I be given as $\pi = N^I / N$. The inclusion probability for a given cell in the 2 x 2 table representing groups I and II combined is given by $p_{ij} = \pi p^I_{ij} + (1 - \pi) p^{II}_{ij}$. Now, let $\alpha_1 = p^{II}_{1+} / p^I_{1+}$ and $\alpha_2 = p^{II}_{+1} / p^I_{+1}$. If we take the expected value of the usual independence assumption DSE for a particular poststratum, we have

$E\left(\hat{N}_k^I\right) = N_k \eta$, where the parameter for this model is given by

$$\eta = \frac{[\pi + (1-\pi)\alpha_1][\pi + (1-\pi)\alpha_2]}{[\pi + (1-\pi)\alpha_1\alpha_2]}.$$

If we knew η, we could estimate the population with

$$\hat{N}_k = \frac{\hat{N}_k^I}{\eta}.$$

We cannot calculate η, but it turns out that we can obtain an estimate of η by taking the expected value of the sum of the independence assumption DSEs.

Now, we let

$\hat{N}^I = \sum_k \hat{N}_k^I$

represent the independence estimate of the male population for the particular age-race group.

Taking the expected value of the summation over the k poststrata of the independence assumption DSEs, we get

$$E\left(\sum_k \hat{N}_k^I\right) = \sum_k \eta N_k = \eta \hat{N}^{DA}.$$

This provides an estimator of η of

$$\hat{\eta} = \left(\frac{\hat{N}^{I}}{\hat{N}^{DA}} \right),$$

remembering that the control total is assumed to be equal to the sum over the k poststrata within the age-race group. Now, let $\Delta = \hat{N}^{DA} - \hat{N}^{I}$. That is, Δ represents the discrepancy between the national control total and the national DSE for males within a particular age-race group.

The resulting alternative DSE estimator is given by

$$\hat{N}_{k}^{\hat{\eta}} = \hat{N}_{k}^{I} + \Delta \left(\frac{\hat{N}_{k}^{I}}{\hat{N}^{I}} \right).$$

It follows that the correlation bias is

$$C_{k} = \hat{N}_{k}^{I} - \hat{N}_{k}^{\hat{\eta}} = -\Delta \left(\frac{\hat{N}_{k}^{I}}{\hat{N}^{I}} \right).$$

The correlation bias is simply the proportionate allocation of Δ over the poststrata within a particular age-race group. The relative correlation bias is:

$$C_{k}^{R} = -\Delta \frac{\left(\frac{\hat{N}_{k}^{I}}{\hat{N}^{I}} \right)}{\hat{N}_{k}^{\hat{\eta}}},$$

which can be shown to reduce to

$$= \frac{-\Delta}{r^{DA} \hat{N}_{f}^{I}}.$$

Results

The Dual System Estimates of the two-group model for the six age-race categories are:

Age	Black	Nonblack
18 - 29	-7.2	2.7
30 - 49	-8.4	-0.25
50+	-4.9	-1.23

The calculation of the variance estimates of the correlation bias estimates followed the methodology used to calculate the variance estimates for the person Dual System Estimates in the A.C.E. That methodology recognized that the use of a standard stratified jackknife variance estimator was not appropriate because of the multi-phase nature of the sampling scheme employed for the A.C.E. Instead, a modified jackknife variance estimator for a reweighted expansion estimator, developed by Kim(2000), was used.

The variance estimates of the correlation bias estimates for the two-group model were:

Age	Black	Nonblack
18 - 29	1.185×10^{-4}	1.370×10^{-5}
30 - 49	3.247×10^{-5}	2.769×10^{-6}
50+	3.368×10^{-5}	2.656×10^{-6}

The standard deviation estimates of the correlation bias estimates were:

Age	Black	Nonblack
18 - 29	0.0109	0.0037
30 - 49	0.0057	0.0017
50+	0.0058	0.0016

Comparing these standard deviations with the correlation bias estimates, we have, for the Black 18 - 29 category, for example, a correlation bias estimate of -7.2 percent with a standard deviation of about 1.1 percent.

Future Research

Plans for future research include:

- Calculating the variance estimates for other alternative dual-system estimators.
- Calculating variance estimates of differences between the alternative dual-system estimates.
- Conducting multiple comparison Z test for differences between estimates, using the Bonferroni technique or a similar method to control type I error.

- Investigating the use of different methods for assessing differences between estimates overall, rather than for a single age-race category.

References

Bell, W.R. (1993), "Using Information From Demographic Analysis in Post-Enumeration Survey Estimation", *Journal of the American Statistical Association*, 88, 1106 - 1118.

Kim, J., Navarro, F., and Fuller, W. (2000), "Replication Variance Estimation for Multi-Phase Stratified Sampling", internal Census Bureau document.

Wolter, K.M. (1986), "Some Coverage Error Models for Census Data", *Journal of the American Statistical Association*, 81, 338 - 346.