Drawing Statistical Inferences from International Census Data

Lara L. Cleveland and **Michael Davern**

Minnesota Population Center, University of Minnesota

OBJECTIVE

Using full count census data from 4 countries, we evaluate the impact of sample design on standard error estimates of microdata samples from the IPUMS International.

BACKGROUND

Although census microdata used by social scientists derive from complex samples, researchers commonly apply methods designed for simple random samples. Standard error estimates from clustered and stratified data can differ dramatically from those derived from simple random samples of the same size.

To the extent that the characteristics of individuals are homogeneous within households, household clustering yields standard errors that are greater than would be obtained from a simple random sample of the same size. (Graubard and Korn 1996; Mansen, Hurwitz, and Madow 1953; Kish 1992; Korn and Graubard 1995, 1999). Variables such as race and poverty status tend to be comparatively homogeneous within household, and therefore pose a risk for underestimated standard errors if clustering is ignored.

Stratification in census microdata samples has the opposite effect from clustering: in general, failure to control for the effects of stratification leads to overestimated standard errors. To the extent that the characteristics of individuals or households are homogeneous within strata, the variance within the stratum is decreased and estimates that account for the additional information about the sample have lower standard errors. Household characteristics that reflect local or regional economic status as well as characteristics of individuals like literacy or ethnicity can be homogeneous within geographic strata.

DATA AND METHODOLOGY

We use data from the Integrated Public Use Microdata Series-International (IPUMS International) which consists of the world's largest collection of census microdata. Since some data samples in IPUMS International were drawn from full count census data, we were able to compare nearly perfect estimates of means and standard errors from the full count data to sample estimates from a test set of 4 countries for which we had access to full count data: Bolivia 2001, Ghana 2000, Mongolia 2000, and Rwanda 2002.

Most IPUMS International samples are systematic random samples, typically drawn by selecting every tenth household in the source file after designating a random starting point. Due to the way that census data are collected, we assumed the existence of low-level geographic sorting. We replicated an approach used by Davern et al. (2009) for 4 IPUMS International census samples and created pseudostrata of 10 households, ensuring that each stratum fell entirely within an administrative unit of the country. Using a replicate method of variance estimation, we drew 100 10% replicates from the full count data using a sampling procedure that mimics the procedure used to draw the 10% public use sample and estimated the standard error of the mean around several household and person-level variables. We compared these estimates to estimates using three methods of variance estimation for the 10% public use sample:

- subsample replicate approach,
- Taylor series, using pseudostrata and household cluster complex sample specifications, and
- standard estimation relying on simple random sample assumptions.

If data are clustered by household or geographically stratified, we would expect the standard errors from the subsample replicate and Taylor series estimates (adjusting for geographic stratification and household clustering) to better approximate the standard errors from the "gold standard" estimates than those derived assuming a simple random sample.

Ratios of estimates from both household-level and person-level characteristics are presented in Tables 1 through 4.

Geographic Stratification and Household Clustering in IPUMS International Samples

Systematic sample of households (every n^{th} unit) with proportional geographic representation

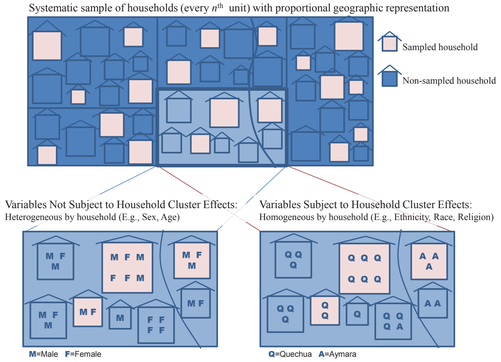

Table 5. Comparison of Clustering and Stratification Effects on Standard Error Estimates in the IPUMS International 10% Bolivian 2001 Census Sample

			Taylor Series Linearization			SRS
Person	Mean	SE (Full Count Replicate)	Accounting for Clustering and Implicit Stratification	**Effect of Clustering** (Adjusting for Strata Only)	**Effect of Stratification** (Adjusting for Cluster Only)	Combined Effect of Clustering and Stratification
Age (mean)	24.7	0.0004	1.1	1.0	1.1	1.0
Sex (%)	49.8	0.0024	0.9	1.1	0.9	1.1
Ethnicity						
Quechua (%)	**30.7**	**0.0053**	**1.0**	**0.6**	**1.4**	**0.8**
Aymara (%)	**25.2**	**0.0047**	**0.9**	**0.5**	**1.4**	**0.8**
Married (%)	26.1	0.0023	1.0	1.0	1.0	1.0
Literate (%)	75.0	0.0025	0.9	0.9	1.0	0.9
Worked (%)	34.4	0.0022	1.1	1.0	1.2	1.0

(Reporting Ratios of Standard Error Estimates from the 10% Sample to Estimates from the Full Count Census using the Subsample Replicate Technique Adjusting for Complex Sample Design Characteristics Independently and Combined)

The decomposition of stratification and clustering effects are illustrated in the figure and reported in Table 5. For variables that are neither geographically stratified nor clustered by household (e.g., age and sex), the method of standard error estimation makes little difference. Ethnicity in Bolivia 2001, however, is both geographically stratified and clustered by household. Specifying both the household cluster as well as the geographic sorting using a pseudostrata variable yields a standard error estimate that best approximates the full count subsample replicate estimate.

RESULTS

Standard Error Computations Comparing Replicate Estimates from Complete count Censuses with Estimates Derived from Sample Data Using Alternative Methods*

Table 1. Rwanda 2002: Standard Error Computations Comparing Replicate Estimates From the Complete Count Census With Estimates Derived From Sample Data Using Alternative Methods

			Ratio of (SE) Estimates Using the Rwanda 2002 10% Sample to Replicate Estimates From the Entire Rwanda 2002 Census		
Selected Characteristics	Parameter Estimate From the Entire Rwanda 2002 Census	Replicate Standard Error Estimates Drawn From the Entire Rwanda 2002 Census	Subsample Replicate Method	Taylor Series Linearization With Pseudo-Strata	Simple Random Sample
Household					
HH Size (mean)	4.71	0.005	0.8	0.9	0.9
Electric Light (%)	4.18	0.034	0.9	0.9	1.3
Toilet (%)	0.38	0.013	0.9	0.9	1.0
Radio (%)	43.11	0.103	0.9	1.0	1.0
Earth Floor (%)	85.28	0.073	0.8	0.9	1.0
Home Ownership (%)	86.41	0.056	1.1	1.1	1.3
Non-relatives (mean)	0.30	0.002	1.1	1.0	1.1
Person			Subsample Replicate	Pseudostrata and HH Cluster	Simple Random Sample
Age (mean)	20.77	0.015	0.9	1.0	1.1
Sex (%)	46.81	0.045	0.9	1.0	1.1
Religion					
Catholic (%)	46.69	0.100	1.0	1.0	0.5
Protestant (%)	26.16	0.077	1.1	1.1	0.6
Married (%)	17.64	0.039	0.9	1.0	1.0
Literate (%)	39.75	0.060	0.9	0.9	0.8
Employed (%)	40.94	0.048	0.9	0.9	1.0

Table 2. Mongolia 2000: Standard Error Computations Comparing Replicate Estimates From the Complete Count Census With Estimates Derived From Sample Data Using Alternative Methods

			Ratio of (SE) Estimates Using the Mongolia 2000 10% Sample to Replicate Estimates From the Entire Mongolia 2000 Census		
Selected Characteristics	Parameter Estimate From the Entire Mongolia 2000 Census	Replicate Standard Error Estimates Drawn From the Entire Mongolia 2000 Census	Subsample Replicate Method	Taylor Series Linearization With Pseudo-Strata	Simple Random Sample
Household					
HH Size (mean)	4.45	0.008	0.9	0.9	1.0
Electric Light (%)	67.53	0.098	1.1	1.0	1.8
Toilet (%)	62.46	0.135	1.1	1.2	1.4
Kitchen as separate room (%)	39.08	0.145	1.0	1.0	1.3
Bathroom (%)	21.74	0.096	1.0	1.1	1.5
Phone (%)	17.01	0.136	1.0	1.0	1.1
Non-relatives (mean)	0.11	0.002	0.9	1.0	1.0
Person			Subsample Replicate	Pseudo-Strata and HH Cluster	Simple Random Sample
Age (mean)	24.57	0.034	1.0	1.0	1.0
Sex (%)	49.47	0.078	0.9	1.0	1.2
Ethnicity					
Khalkh (%)	81.59	0.111	0.9	1.0	0.6
Kazak (%)	4.28	0.047	1.0	1.1	0.8
Married (%)	32.33	0.081	0.9	1.0	1.1
Literate (%)	81.56	0.071	1.1	1.0	1.0
Employed (%)	32.47	0.095	0.9	0.9	0.9

Table 3. Bolivia 2001: Standard Error Computations Comparing Replicate Estimates From the Complete Count Census With Estimates Derived From Sample Data Using Alternative Methods

			Ratio of (SE) Estimates Using the Bolivia 2001 10% Sample to Replicate Estimates From the Entire Bolivia 2001 Census		
Selected Characteristics	Parameter Estimate From the Entire Bolivia 2001 Census	Replicate Standard Error Estimates Drawn From the Entire Bolivia 2001 Census	Subsample Replicate Method	Taylor Series Linearization With Pseudo-Strata	Simple Random Sample
Household					
HH Size (mean)	3.93	0.0046	1.0	1.0	1.1
Electric Light (%)	60.51	0.0536	1.1	1.2	1.9
Toilet (%)	59.48	0.0649	1.0	1.1	1.6
Kitchen as separate room (%)	70.62	0.0882	0.9	1.0	1.1
Phone (%)	21.33	0.0605	1.3	1.1	1.4
Radio (%)	71.17	0.0819	0.9	1.0	1.1
Earth Floor (%)	35.66	0.0519	1.2	1.3	1.9
Home Ownership (%)	62.81	0.0877	1.0	1.0	1.1
Non-relatives (mean)	0.19	0.0012	1.0	1.0	1.1
Person			Subsample Replicate	Pseudo-Strata and HH Cluster	Simple Random Sample
Age (mean)	24.70	0.0004	1.0	1.1	1.0
Sex (%)	49.84	0.0024	0.9	0.9	1.1
Ethnicity					
Quechua (%)	30.69	0.0053	1.0	1.0	0.8
Aymara (%)	25.19	0.0047	0.8	0.9	0.8
Married (%)	26.09	0.0023	0.9	1.0	1.0
Literate (%)	74.99	0.0025	0.9	0.9	0.9
Worked (%)	34.37	0.0022	1.1	1.1	1.0

Table 4. Ghana 2000: Standard Error Computations Comparing Replicate Estimates from the Complete Count Census with Estimates Derived from Sample Data Using Alternative Methods

			Ratio of (SE) Estimates Using the Ghana 2000 10% Sample to Replicate Estimates From the Entire Ghana 2000 Census		
Selected Characteristics	Parameter Estimate From the Entire Ghana 2000 Census	Replicate Standard Error Estimates Drawn From the Entire Ghana 2000 Census	Subsample Replicate Method	Taylor Series Linearization With Pseudo-Strata	Simple Random Sample
Household					
HH Size (mean)	4.99	0.005	1.1	1.0	1.0
Electric Light (%)	43.54	0.042	1.5	1.5	1.8
Toilet (%)	8.49	0.026	1.2	1.5	1.7
Kitchen as separate room (%)	46.17	0.062	1.2	1.2	1.2
Bathroom (%)	23.47	0.046	1.5	1.4	1.4
Non-relatives (mean)	0.14	0.001	0.9	1.0	1.0
Person			Subsample Replicate	Pseudo-Strata and HH Cluster	Simple Random Sample
Age (mean)	23.90	0.013	1.0	1.1	1.0
Sex (%)	49.48	0.035	1.0	1.0	1.0
Ethnicity					
Akan (%)	45.28	0.066	0.9	1.0	0.5
Mole-dagbani (%)	15.25	0.051	1.0	1.0	0.5
Married (%)	29.28	0.029	1.2	1.2	1.1
Literate (%)	34.00	0.038	1.0	1.1	0.9
Worked (%)	42.44	0.038	1.3	1.1	0.9

* Due to the relatively large sample sizes (10%), all sample estimates have been corrected by the finite population correction factor (fpc).

CONCLUSIONS

- For many variables in each country, the ratios for each of the methods of standard error estimation to full count estimate are close to 1.0 and similar to each other, suggesting that the method of standard error estimation does not matter much for these variables.
- Use of pseudostrata improves precision in estimating standard errors for variables that are geographically stratified, especially those dependent upon public utilities or ethnic group membership. For large samples, this correction may not be necessary as failure to do so will yield conservative estimates of standard error and statistical significance. Specifying pseudostrata in Taylor series linearization procedure may be beneficial for analysis of smaller geographic areas. The strata specification is of limited use in Ghana, requiring further work to assess whether the difference in geographic scale of the full count and sample estimates is contributing to the limited utility of the pseudostrata adjustment.
- As expected, household clustering is evident in some person level characteristics. Taylor series adjustments provide a reasonable correction in standard error estimation for household clustering. Research often concentrates on sub-populations which require only one person or one unit (a mother, a school-aged child, a cohabiting couple) from each household. Analyses on such populations are not subject to the effects of household clustering and do not require this type of correction.

ACKNOWLEDGEMENTS

Support for this research was provided by the National Science Foundation and the Minnesota Population Center. We gratefully acknowledge Reiping Huang, Sheela Kennedy, Steven Ruggles, Matt Sobek and Vladimir Vladykin for valuable assistance.