# Drawing Statistical Inferences from International Census Data

Cleveland, Lara
*University of Minnesota, Minnesota Population Center*
*50 Willey Hall - 225 19th Avenue South*
*Minneapolis, MN 55455*
*Email: clev0026@umn.edu*

Davern, Michael
*National Opinion Research Center (NORC)*
*One North State Street*
*Chicago, IL 60602*

Ruggles, Steven
*University of Minnesota, Minnesota Population Center*
*50 Willey Hall - 225 19th Avenue South*
*Minneapolis, MN 55455*

Draft submitted in advance of the 2011 ISI World Congress (Dublin, Ireland, August 21-26, 2011)



Census microdata are collected by countries around the globe and contain a wealth of information useful to social science researchers. Although large machine-readable census microdata samples exist for many countries, access to these data has been limited and the documentation has often been inadequate, making cross-country and across-time comparisons difficult. The Integrated Public Use Microdata Series-International (IPUMS International) converts census microdata from multiple countries into a consistent format, supplying comprehensive documentation, and making the data available through a web-based data dissemination system. Although census microdata used by social scientists, like the data in the IPUMS, derive from complex samples, researchers commonly apply methods designed for simple random samples. Using full count data from 4 countries, we evaluate the impact of sample design on standard error estimates of microdata samples from the IPUMS International. We compare standard error estimates from the full count data to estimates from the 10% public use samples using three methods: subsample replicate, Taylor series linearization, and estimates using simple random sample assumptions. We conclude by discussing strategies for obtaining unbiased and efficient estimates of statistical significance.

Like most census microdata, IPUMS samples contain individual level data, clustered by household, and often stratified and differentially weighted. Standard error estimates from clustered, stratified, and differentially weighted 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). Stratification in census microdata samples has the opposite effect from clustering and differential weighting: 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. Most IPUMS-International samples are systematic random samples, drawn by selecting every tenth household in the source file after designating a random starting point. The data are typically sorted according to small geographic areas so that records in resulting samples retain geographic proximity. Therefore, the systematic sample design is equivalent to low-level geographic stratification, even though no explicit stratification may have been carried out.

In recent years, it has become fairly straightforward for users if IPUMS-International data to account for the effects of clustering and differential weighting on estimates of standard errors using Taylor series linearization features available in major statistical analysis software packages. Users need only identify a case weight and a variable identifying the clusters, and the software will automatically adjust the estimates to account for the sample design. Controlling for the effects of implicit stratification, however, is more difficult, since there is no variable in the data identifying the strata. This article proposes a solution that will allow IPUMS-International users to create more reliable estimates.

**Methods**

*Taylor series linearization.* Taylor series linearization has been underutilized by census researchers because the method requires explicit information about strata. The data contain no geographic unit that corresponds to the geographic stratification embedded in the geographic ordering of census records. Davern et al. (2009) used information from a complete machine-readable enumeration of the 1880 U.S. census to develop and test geographic pseudostrata, which are constructed by grouping contiguous records together to simulate small geographic areas. The present paper replicates the Davern et al. (2009) approach for 4 IPUMS-International census samples. To create a proxy for implicit geographic stratification within a subset of IPUMS International samples, we used the ordering of full count data sets along with as much low level geographic information as was available accompanying them. We created pseudostrata of 10 households, ensuring that each stratum fell entirely within an administrative unit of the country (pooling strata at the end of geographic breaks containing fewer than 10 households with the preceding stratum). *Subsample replicate approach.* An alternative to Taylor series variance estimation is the subsample replicate approach (Rust 1985; Wolter 2007; Verma 1993), which involves dividing the sample into subsamples (or replicates) that reflect the complex design of the entire sample.

**Validation**

As in Davern et al. (2009), to validate both the Taylor series linearization with pseudo-strata and the subsample replicate approach, we needed a "true" estimate of variance in the census samples. Since some data samples in IPUMS International were drawn from full count census data, we were able to consult full count census data for nearly perfect estimates for a test set of countries. We used recent census data from four countries for which we had access to well-formatted full count data: Rwanda 2002, Mongolia 2000, Bolivia 2001, and Ghana 2000. 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.[1] We considered these variance estimates the gold standard against which to measure three methods of variance estimation for the 10% public use sample: subsample replicate, Taylor series, and 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 to better approximate the standard errors from the "gold standard" estimates than those derived assuming a simple random sample design.

**Results**

Tables 1 through 4 compare methods for estimating standard errors for each country. The first two columns in each table are based on the 100 10% sample replicates of the full count population. We estimated the variance using the full count "true" mean from the population. Standard errors from the resulting 100 replicate samples are reasonably unbiased

[1] Since we could not draw more than 10 independent samples of size 10%, we approximated our sampling strategy by creating strata of 10 households and randomly drawing one household from each stratum to form 10% samples.

estimates of the standard error that would be expected in a 10% sample. The last three columns in each table contain ratios of standard errors from the 10% sample[2] to standard errors from the full count replicate estimates for each country using the three methods of calculating standard errors described above: subsample replicate, Taylor series linearization, and simple random sample assumptions. Ratios of estimates from both household-level and person-level characteristics are presented in the table.[3] We expect that both our subsample and our Taylor series estimates will more closely approximate the full count replicate estimates for variables that represent characteristics that contain systematic geographic sorting or household clustering than the simple random estimates. Measured against full count standard, the ideal ratio of sample to full count estimate would be 1.0. Ratios under 1.0 indicate underestimated standard errors, and ratios over 1.0 indicate overestimated standard errors.

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

| Selected Characteristics | Parameter Estimate From the Entire Rwanda 2002 Census | Replicate Standard Error Estimates Drawn From the Entire Rwanda 2002 Census | Ratio of (SE) Estimates Using the Rwanda 2002 10% Sample to Replicate Estimates 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 1 shows that, for to the 2002 Census of Rwanda, the average number of persons in a household is 4.71, with a full count replicate standard error estimate of 0.005. The ratio of the 10% Rwanda 2002 replicate standard error estimate to the full count estimate is 0.8, to the Taylor series estimate is 0.9, and to the simple random sample is 0.9, suggesting that the method of standard error estimation does not matter much for this variable. The same pattern applies for the number of non-relatives in the household. The characteristics of these two variables are not highly correlated within geographic strata.

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

[3] We measure household characteristics at the household-level rather than at the individual level because of the effect of household clustering. As demonstrated by Davern et al. (forthcoming), when household characteristics are written across person level records and analyzed at the individual level, standard errors based on a simple random sample assumption are severely underestimated.

For household characteristics frequently used to represent wealth or economic development, subsample replicate and Taylor series estimates using geographic pseudo-strata are close to 1.0 and close to each other. Further, there is little difference between such estimates and simple random estimates for some variables such as radio ownership, floor material and the presence of a flush toilet. For other variables, including electricity and home ownership, the simple random assumption overestimates standard errors. Simple random estimates are 1.3 times larger than the full count estimates for electricity and home ownership. Failure to account for the geographic sorting of the data can lead to inflated standard errors for some household characteristics. The opposite effect is present for select person level characteristics largely due to household clustering. Again, for many characteristics, all three method estimates closely approximate those of the full count replicate method. For characteristics that we expect to cluster by household, like race or religion, we see evidence of clustering in reduced standard error estimates from the simple random sample.

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

| Selected Characteristics | Parameter Estimate From the Entire Mongolia 2000 Census | Replicate Standard Error Estimates Drawn From the Entire Mongolia 2000 Census | Ratio of (SE) Estimates Using the Mongolia 2000 10% Sample to Replicate Estimates 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 |

Tables 2, 3 and 4 show similar patterns for characteristics from additional countries. Ratios indicate that simple random sample assumptions contribute to overestimated standard errors for a subset of household characteristics, particularly those associated with household utilities. Clustering has the expected effect on select individual level characteristics.[4] In Ghana 2000 (Table 4), however, standard error estimates for the utilities and dwelling characteristics are overestimated by all three techniques relative to estimates from the full count replicate data. It is possible that the grouping of household characteristics in Ghana is not well represented by the sampling method, but further investigation is required to determine whether this occurs as a result of scale differences from full count to subsample or other factors.

[4] In Bolivia 2001 (see Table 3), ethnicity does not seem to have the expected clustering. Taylor series linearization enabled us to analyze the effects of clustering and stratification independent of one another. In a separate analysis, we found that the two design elements cancel one another out.

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

| Selected Characteristics | Parameter Estimate From the Entire Bolivia 2001 Census | Replicate Standard Error Estimates Drawn From the Entire Bolivia 2001 Census | Ratio of (SE) Estimates Using the Bolivia 2001 10% Sample to Replicate Estimates 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**

| Selected Characteristics | Parameter Estimate From the Entire Ghana 2000 Census | Replicate Standard Error Estimates Drawn From the Entire Ghana 2000 Census | Ratio of (SE) Estimates Using the Ghana 2000 10% Sample to Replicate Estimates 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 |

**Discussion**

The sample methodology of IPUMS-International samples has the potential to significantly affect the precision of sample estimates. Individuals are sampled as parts of households because many important topics of analysis, such as fertility, household composition, and nuptiality, require information about multiple individuals within the same household. In addition, all of the IPUMS International sample data are implicitly or explicitly stratified. In some cases, as that of ethnicity in Bolivia 2001, the positive effects of stratification outweigh the adverse effects of clustering, but researchers should not rely exclusively on these opposing effects.

IPUMS samples are large, and for the great majority of studies there is little risk of drawing invalid inferences because of underestimated variance. Geographic clustering can lead to overestimated standard errors for a set of variables describing household characteristics, but analysis based on these estimates will be conservative at worst. For studies of weak relationships or small population subgroups, however, there can be risk of misleading estimates of statistical significance. The effects of clustering are of greater concern because underestimated standard errors have the potential to lead to erroneous findings of statistical significance. However, most census research has minimal household clustering because it focuses on particular subpopulations that rarely cluster in households. For example, studies of fertility focus on women of childbearing age, and households typically only have one such woman. The clustering concern can arise with studies of children, since households often include multiple children. When doing analyses of children and other groups likely to appear multiple times in the same household, researchers can adopt strategies to eliminate the redundant cases.

An alternative, thanks to improvements in the analytical power of modern statistical software, is to incorporate information about sample design into estimation procedures. Using Taylor series linearization procedures available in major software programs, IPUMS users can specify the household identifier as the cluster variable (or primary sampling unit) and the weight variable (WTPER) to account for the effects of household clustering and heterogeneous sample weights. The IPUMS staff is developing a new cluster variable that will offer the potential for more refined variance estimates. The new variable will identify geographic clustering as well as household clustering. The staff is also developing a new variable, which will include information of explicit strata whenever such information is available, and will also include geographic pseudostrata for the systematic samples following the procedure described above and in Davern et al. (2009).

For most analyses using IPUMS data, there is little risk of drawing invalid conclusions due to underestimated variance. When examining relationships on the margin of statistical significance, however, it may be wise to adjust for household clustering and weighting as outlined above. These procedures will yield conservative estimates of statistical significance for all IPUMS samples except the few that incorporate geographic clustering. Until the new clustering and stratification variables are available, marginally significant results from those samples should be viewed with caution.

References

Berenson, M. L., D. Levine, and T. Krehbiel. 2005. *Basic Business Statistics: Concepts and Applications.* (10 ed.). Prentice Hall.
Davern, M., S. Ruggles, T. Swenson, J. T. Alexander, J. M. Oakes. 2009. "Drawing Statistical Inferences from Historical Census Data, 1850-1950." *Demography* 46: 429-49.
Graubard, B., and Korn, E. 2002. "Inference for superpopulation parameters using sample surveys." *Statistical Science* 17: 73-96.
Kish, Leslie. 1992. "Weighting for Unequal Pi." *Journal of Official Statistics* 18:129-54.
Korn, E., and Graubard, B. 1999. *Analysis of Health Surveys.* New York: John Wiley & Sons.
Korn, E., and Graubard, B. 1998. "Variance estimation for superpopulation parameters." *Statistica Sinica* 8: 1131-51.
Hansen, Hurwitz, and Madow 1953. *Sample Survey Methods and Theory*. New York: Wiley.
Rust, K. 1985. "Variance Estimation for Complex Estimators in Sample Surveys." *Journal of Official Statistics* 1:381-97.
Verma, V. 1993. *Sampling Errors in Household Surveys.* United Nations National Houshold Survey Capability Programme. U.N. Statistics Division, United Nations.
Wolter, K.M. 2007. *Introduction to Variance Estimation* (2ed.). Chicago: Springer.