Summary of "*IPUMS-Redesign Supplement*," a project funded by the National Institute of Child Health and Human Development (HD 043392-03S1), 9/2004-9/2006

Principal Investigator: Steven Ruggles
Co-Investigators: Michael Davern (Project Director), J. Trent Alexander, J. Michael Oakes
Research Fellow: Tami Swenson

Virtually all quantitative data used by social scientists derive from samples that incorporate clustering, stratification, and weighting adjustments. Such data can yield variances that differ dramatically from a simple random sample of the same size. Researchers using a sample of census microdata, however, usually apply methods designed for simple random samples, and the resulting p-values and confidence intervals are inaccurate.

Accurate variance estimation is critical to scientific research. Tests of statistical significance require appropriate variances in order to make valid statistical inferences. If an estimated variance is too small then a researcher is more likely to reject a true null hypothesis. On other hand, if an estimated variance is too large then the researcher is less likely to reject a false null hypothesis. Because census microdata samples are among the most widely-used sources for social science and policy research, the need for reliable variance estimation is critical and urgent.

To improve variance estimation using the IPUMS we will:

- Construct and document new variables necessary to exploit variance estimation software for all U.S. census microdata. These include a stratification variable for the period from 1960 to 2000 and clustering and subsample-replicate variables for the earlier censuses.

- Evaluate the reliability of variance estimates in the censuses using real-world applications, based on comparison of analyses that use Taylor series linearization, subsample-replicate estimates, and Census Bureau design factors.

- Based on the results of this evaluation, develop user-friendly documentation and recommendations for variance estimation using each statistical package, with examples of typical analyses and sample programs, and disseminate these materials through the IPUMS data access system (http://ipums.org).

Clustering, Stratification and Census Microdata Samples

All the census microdata sample designs include clustering and stratification, so the simple formulas ordinarily used to calculate standard errors assuming a simple random sample do not yield reliable results.

The census microdata samples are clustered by household, and this significantly increases the variance associated with most individual-level variables. Clustering by household is necessary because many important topics of analysis require information about multiple individuals within the same unit. Thus, the number of independent observations in each census file is the number of households, not the number of individuals. This has implications for sample efficiency. Standard errors in cluster samples depend on both the size of the sampled clusters and on the homogeneity of variables within clusters (Kish 1992; Hansen, Hurwitz, and Madow 1953; Graubard and Korn 1996; Korn and Graubard 1995, 1999). In the worst case, with perfect homogeneity within clusters, the standard errors for variables would be inversely proportional

to the square root of the number of clusters rather than the number of individuals. Thus, variables such as race and poverty status, which are comparatively homogeneous within households, have underestimated standard errors if clustering is ignored. Conversely, for variables that are heterogeneous within clusters such as age and sex, clustering may have little effect on sample precision.

The loss of efficiency resulting from clustered design is partially counterbalanced by stratification. In particular, since 1960 the U.S. Census Bureau has employed increasingly elaborate stratified multistage sample designs when creating publicly-available microdata files. Such procedures can yield low standard errors, especially for variables that are explicitly stratified, such as race, household size, and group quarters status. The IPUMS samples for years prior to 1960 were designed to capitalize on geographically-sorted source materials, which enhanced precision through implicit geographic stratification.

Census Bureau Design Factors

When the Census Bureau produced the first public use microdata samples, computing resources were scarce and statistical software was rudimentary. It was therefore impractical for researchers to calculate variance estimates that accounted for the complex design of the samples. Therefore, the Census Bureau calculated "design factors" (originally termed "standard error adjustment factors") for specific variables, and researchers were advised to multiply their conventional standard errors by the adjustment factor to account for the complex sample design (US Census Bureau 2003). The original IPUMS project developed comparable adjustment factors for the earlier census years.

The strategy of using design factors to correct for complex sample designs has several serious weaknesses:

- The design factors are national averages and are not valid for subpopulations.[1] In many instances, the adjustments required for particular age groups, racial groups, or other subpopulations differ dramatically from the national averages.

- The needed adjustments are not uniform across categories of the same variable. To give just one example, the true design factors for the "Head" and "Spouse" categories of the household relationship variable are always much lower than the design factors for "Child" or "Boarder," since the effects of clustering are much more pronounced for the latter categories. The Census Bureau, however, publishes only one design factor for this variable, representing the average of all the categories.

- The design factors are not suitable for multivariate analyses. The Census Bureau documentation for the 1980 microdata sample recommended that when adjusting the standard errors of a crosstabulation, users should simply choose the largest adjustment factor, but there is no theoretical or empirical justification for this approach (U.S. Census Bureau 1982).

- The design factors published by the Census Bureau produce erratic results when compared to estimates calculated adjusting for the actual sample design used to select cases in the Current Population Survey (Davern et al. 2003).

The great majority of IPUMS-based research involves complex regression models that control for many covariates (http://www.ipums.org/usa/research.php). The Census Bureau's design factors are inappropriate

[1] For 2000, the Census Bureau for the first time provided state-level design factors as well as national estimates, but no estimates are available for other population subgroups.

for the exploration of associations among variables and are especially problematic when performing complex analyses. It would be impossible for the decennial census technical documentation to provide guidance for all possible types of analyses and dependent variables. Thus, researchers need to be able to produce standard errors tailored to their particular analyses (Jones et al. 2004).

Alternatives to the Census Bureau's Design Factors

Although statisticians have been working to develop software for estimating reliable variance estimates from complex surveys since the 1970s, these tools were not practical for most researchers until the 1990s. The first software products suitable for multivariate analyses of complex samples were standalone applications such as SUDAAN and WESVAR (Brick and Morganstein 1996; Brogan 1998; Lepkowski and Bowles 1996). In the past few years, however, similar procedures have been incorporated into the three most widely used general-purpose statistical packages, SAS, Stata, and SPSS (SAS 1999; Stata 2001; SPSS 2003). With this development, it is reasonable to hope that variance estimation procedures that are appropriate to complex sample designs will soon become the norm in social science and public health research. Our goal is to make that transition as painless as possible for IPUMS users.

We will evaluate the two principal classes of methods for variance estimation under complex sample designs: replicate methods and Taylor series estimates of standard error (Wolter 1985; Rust 1985; Verma 1993).[2] The replicate approach divides a sample into subsamples (or replicates) that reflect the complex design of the entire sample. Thus, each subsample incorporates the same stratification and clustering used to select the sample as a whole. We can then estimate standard error by calculating any given statistic for each subsample, and measuring the standard deviation of the statistic over the subsamples. In its basic form, replication requires large samples, since the method presumes a substantial number of subsamples that each have sufficient cases to calculate the statistic of interest. To avoid this problem, statisticians have developed a range of variant appoaches that make replication techniques feasible when using smaller samples (e.g. Rao and Shao 1992; Kalton 1977). In the case of the census microdata samples, however, these variants are usually unnecessary since for most analyses the sample sizes are adequate for true replication analysis. For the censuses of 1960 through 2000, the Census Bureau divided each sample into 100 subsample replicates, thus simplifying application of this method to those census years. Although most of the major statistical packages do not include procedures to automatically calculate variances and confidence intervals using replicate methods, the approach can be implemented in every major package. One of our goals is to provide researchers with specific instructions and examples of programming instructions to enable them to apply replicate methods consistently and easily.

In most cases, the Taylor series linearization approach yields results that are comparable to replicate methods (Kish and Frankel 1974; Krewski and Rao 1981; Dippo and Wolter 1984; Weng, Zhang, and Cohen 1995; Hammer, Shin and Porcellini 2003). The Taylor series approach derives a linear approximation of variance that is used to correct standard errors and confidence intervals for statistics of interest. Taylor series linearization is more computationally efficient than replication methods, and the

[2] Statisticians are currently developing model-based variance estimates (e.g., Little 2003) to improve upon the design-based variance estimates we propose to examine, but there is not a specific algorithm available in statistical packages to implement them. The model-based variance estimates are beyond our current scope of work for two reasons (1) the standards for implementing model-based variance estimates are not set for routine statistical analysis, and (2) its still not clear how model-based variance estimates will be used in complex sample surveys like the Census (other than for small-area estimates) (Kalton 2002).

procedure is more widely available in statistical packages. The method requires full information about stratification, clustering, and weighting used in the sample design, and we intend to provide these variables whenever feasible to simplify implementation. A significant disadvantage of the approach, however, is that it is not designed to adjust for stratification that is not explicitly identified in the data. This may pose difficulties when the method is applied to census data for the period prior to 1960, since those samples incorporate implicit geographic stratification. Moreover, there are potential problems for data from 1960 to 1980, since those samples are affected by ratio estimation procedures described below.

Census Bureau Sample Designs 1960-2000

Although the Census Bureau modified the sample designs in every decade from 1960 to 2000, these census years share the same key components. The creation of the samples for these census years consisted of three steps:

1. The Census Bureau systematically selected households to receive the long form. In 1980, for example, every sixth address received a long form in minor civil divisions with over 2,500 persons, and every second address received a long form in smaller places.

2. Households and individuals were assigned weights based on a multi-stage procedure. In each small geographic unit (termed "Smallest Weighting Areas") the Census Bureau calculated the ratio of long-form households to the complete count of households. These ratios—or initial weights—were then adjusted in three or four stages to control for such characteristics as household size and type, householder status, age, sex, race, and Spanish origin. The Census Bureau used these weights to construct small-area estimates of long-form characteristics (e.g. Summary File 4).

3. The Census Bureau selected the public use microdata samples from the weighted long-form data using a stratified systematic selection procedure. Within each state, housing units were divided into strata. The number of strata used grew dramatically over time, from 38 in 1960 to 34,080 in 2000.

For 1990 and 2000, the Census Bureau systematically selected cases from each stratum without regard to the weights, and the original weights as constructed in step 2 were preserved in the data. The public files include adequate information to classify virtually all cases by stratum.[3] Clustering in the public files can be readily identified through the household identification number. These three pieces of information—on weights, stratification, and clustering—are sufficient to take advantage of the Taylor series variance estimation procedures built into the major statistical packages.

In the period from 1960 to 1980, however, the Census Bureau used a somewhat different approach. Within each stratum, the cumulative sum of weights for each housing unit was calculated, and a household was selected for inclusion in the sample each time the cumulative sum passed a multiple of 100. For the 1980 five-percent sample, this procedure was repeated five times and the samples were merged. This procedure yields self-weighting samples at 1-in-100 or 1-in-20 densities, and no weights were included in the public use microdata samples in those years.

Without information on weights in the 1960 to 1980 samples, Taylor series estimates of variance may not yield reliable results. Although the self-weighting samples produced for the 1960 to 1980 censuses are convenient, they mask some of the sample design information. Because the cases were selected in proportion to weights calculated during the multi-stage ratio estimation procedure (step 2), the samples

[3] As discussed below, in a few instances, it will be necessary to infer values for strata that cannot be fully identified in the microdata, but we expect this will have minimal impact on variance estimates.

realize some of the gains in sampling efficiency that would have resulted had the population been stratified into the ratio estimation groups before sampling. The increased precision, however, cannot be captured by a stratification variable, and this could lead to overstated standard errors using the Taylor series approach. One of the goals of this research is to estimate the magnitude of this effect.

Historical Sample Designs, 1850-1950

The sample designs for the pre-1960 censuses differ fundamentally from those of more recent censuses because they were drawn from census enumerator manuscripts instead of from machine-readable files. Explicit stratification was not feasible, but the organization of historical census enumeration forms allowed implicit geographic stratification. Unlike recent mail-in U.S. censuses, the pre-1960 censuses were created through direct enumeration: an enumerator went from house to house to interview residents in person. A byproduct of this enumeration method is that the census forms are sorted according to the sequence of enumeration within each enumeration district. In practice, this means that the files are geographically organized within districts.

The systematic samples of the historical censuses capitalize on this low-level geographic sorting. By ensuring a representative geographic distribution of sampled cases, they are equivalent to extremely fine geographic stratification with proportional weighting. Since many economic and demographic characteristics are highly correlated with geographic location, this implicit stratification yields substantially greater precision than a simple random sample of households.

The one-percent historical census microdata samples created at the University of Minnesota for the censuses of 1850 through 1930 all employ the same basic sample design, with minor variations to accommodate differences in source materials and innovations in data-entry technology. In general, within each enumeration district we generate a random starting point between one and five, and then designate every fifth page thereafter as a sample page. Thus, for example, if the starting point is three, we designate the third, eighth, and thirteenth pages, continuing in that fashion until the end of the district. On each sample page, we randomly select sample points. Households are included in the sample whenever the first person in the household falls on a sample point. [4]

By capitalizing on implicit stratification, this design yields high precision. The difficulty is that the stratification is implicit: there is no geographic unit in the data that corresponds precisely to the geographic stratification embedded in the data. This may pose problems for accurate variance estimation.

Plan of Work

Our goal is to make it easier for researchers to obtain reliable variance estimates when using data from the IPUMS. We plan three specific activities:

1. **Construct and document a new IPUMS variables describing stratification, clustering, and subsample replicates.**

2. **Compare variance estimates derived from replicate methods, Taylor series linearization, and Census Bureau design factors to determine the simplest approach that will yield reliable results for typical multivariate analyses performed with the IPUMS samples.**

[4] The 1940 and 1950 samples employed slightly different criteria for selecting households, but the effect of implicit geographic stratification is similar. The pre-1940 samples also incorporate slightly larger clusters than do the 1940-2000 samples. As described below, we will create a new cluster variable for these early census years to capture the effects of the larger clusters.

3. **Based on these analyses, develop and disseminate user-friendly documentation and recommendations for variance estimation in each statistical package, with sample programs and examples of typical analyses.**

Literature Cited

Brick, J.M., and Morganstein, D. 1996. WesVarPC: Software for Computing Variance Estimates from Complex Designs. *Proceedings of the 1996 Annual Research Conference*, pp. 861-866. Washington, DC: U.S. Bureau of the Census.

Brogan, D. 1998. *Software for sample survey data, misuse of standard packages.* In *Encyclopedia of Biostatistics*, Volume 5 (P. Armitage and T. Colton, eds.). New York: Wiley, pp. 4167-4174.

Davern, M., Jones, A., Lepkowski, J., Davidson, G., and Blewett, L.A. 2003. Standard Error Estimation and the Current Population Survey: Various Approach Produce Unstable Estimates." SHADAC Working Paper. Minneapolis, MN: University of Minnesota.

Dippo, C.S. and K.M. Wolter. 1984. A Comparison of Variance Estimators Using the Taylor Series Approximation. *ASA Proceedings of the Section on Survey Research Methods*, pp. 112-121. Arlington, VA: American Statistical Association.

Graubard, B.I., and E.L. Korn. 1996. Survey inference for subpopulations. *American Journal of Epidemiology*, 144(1): 102-106.

Hammer, H., Hee-Choon Shin and L.E. Porcellini. 2003. A Comparison of Taylor Series and JK1 Resampling Methods for Variance Estimation. *Proceedings of the Hawaii International Conference on Statistics*, pp. 1-9.

Hansen, M.H., Hurwitz, W. and W. Madow. 1953. *Sample Survey Methods and Theory*. New York: Wiley and Sons.

Jones, Arthur Jr., Michael Davern, James Lepkowski, Gestur Davidson, Lynn A. Blewett. 2004. "Estimating Standard Errors for Regression Coefficients Using the Current Population Survey's Public Use File." Presented at the American Statistical Association Annual Meeting, Section on Survey Research Methods and Section on Government Statistics, August 2004, Toronto, Ontario.

Kalton, G. 1977. Practical Methods for Estimating Survey Sampling Errors. *Bulletin of the International Statistical Institute*. 47(3): 495-514.

Kalton, G. 2002. Model in the Practice of Survey Sampling (Revisited). *Journal of Official Statistics*. 18(2):129-154.

Kish, L. 1965. *Survey Sampling*. New York: Wiley and Sons.

Kish, L. 1992. Weighting for Unequal P_i. *Journal of Official Statistics*. 8(2): 183-200.

Kish L. and M.R. Frankel. 1974. Inference from Complex Samples. *Journal of the Royal Statistical Society* B(36), 1-37.

Little, R.J.A. 2003. To Model or Not To Model? Competing Modes of Inference for a Finite Population. The University of Michigan Department of Biostatistics Working Paper Series, University of Michigan School of Public Health. Paper 4. http://www.bepress.com/umichbiostat/paper4.

Korn, E.L. and B.I.Graubard. 1995. Examples of Differing Weighted and Unweighted Estimates from a Sample Survey. *American Statistician*, 49(3), 291-295.

Korn, E.L., and B.I. Graubard. 1999. *Analysis of Health Surveys*. New York: Wiley.

Krewski, D., and J.N.K. Rao. 1981. Inference from stratified samples: Properties of Linearization, Jackknife and Balanced Repeated Replication Methods. *Annals of Statistics*. 9: 1010-1019.

Lepkowski, J.M. and J. Bowles, 1996. Sampling error software for personal computers. *Survey Statistician*, 35, 10-17.

Mills, R. 2002. *Health Insurance Coverage in the United States for 2001*. Washington, DC: US Census Bureau.

Rao, J. N. K. and J. Shao 1992. Jackknife variance estimation with survey data under hot deck imputation. *Biometrika.* 79: 811-822.

Ruggles, S. 1995. Sample Designs and Sampling Errors. *Historical Methods* 28 (1): 40-46.

Ruggles, S., M. Sobek, M.L. King, C. Liebler, and C.A. Fitch. 2003. IPUMS Redesign. *Historical Methods* 36:9-21.

Ruggles, Steven and M. Sobek. 2003. *Integrated Public Use Microdata Series: Version 3.0.* Minneapolis: Minnesota Population Center.

Rust, K. 1985. Variance Estimation for Complex Estimators in Sample Surveys. *Journal of Official Statistics*. 1(4):381-397.

SAS. 1999. *Documentation for SAS Version 8*. Cary, NC: SAS Institute, Inc.

SPSS. 2003. *Correctly and Easily Compute Statistics for Complex Sampling*. Chicago, Illinois: SPSS Inc. http://www.spss.com/complex_samples/

Stata. 2001. *Reference Manual.* College Station Texas: STATA Press.

U.S. Census Bureau. 1964. Census of population and housing, 1960 public use sample: one-in-one-thousand sample. Washington, D.C.: U.S. Government Printing Office.

U.S. Census Bureau. 1982. Public Use Microdata Samples of Basic Records from the 1980 Census: Description and Technical Documentation. Washington, D.C.: U.S. Government Printing Office.

U.S. Census Bureau. 2003. *Technical Documentation for the Public Use Microdata Sample: 2000 Census of Populations and Housing*. Washington DC: US Census Bureau.

Verma, V. 1993. *Sampling Errors in Household Surveys*. United Nations National Household Survey Capability Programme, UN Statistics Division, United Nations, New York.

Weng, S.S., Zhang, F, and Cohen, M.P. 1995. Variance Estimates Comparison by Statistical Software. *ASA Proceedings of the Section on Survey Research Methods,* pp. 333-338. Arlington, VA: American Statistical Association.

Wolter, K.M. 1985. *Introduction to Variance Estimation*. Springer-Verlag, New York.

Woodruff, R.S. 1971. A Simple Method for Approximating the Variance of a Complicated Estimate. *Journal of the American Statistical Association.* 66: 411-14.