MINNESOTA POPULATION CENTER

MPC

UNIVERSITY OF MINNESOTA

Working Paper Series

Drawing Statistical Inferences from Historical Census Data

Michael Davern, Steven Ruggles, Tami Swenson and J. Michael Oakes
Minnesota Population Center
University of Minnesota

June 2007

Working Paper No. 2007-02

ABSTRACT

Virtually all quantitative microdata used by social scientists derive from samples that incorporate clustering, stratification, and weighting adjustments (Kish 1992, 1965). Such data can yield standard error estimates that differ dramatically from a simple random sample of the same size. Researchers using historical U.S. census microdata, however, usually apply methods designed for simple random samples. The resulting p-values and confidence intervals could be inaccurate and could lead to erroneous research conclusions. Because U.S. census microdata samples are among the most widely-used sources for social science and policy research, the need for reliable standard error estimation is critical. We evaluate the historical microdata samples of the IPUMS project from 1850-1930 in order to determine (1) the impact of sample design on standard error estimates and (2) how to apply modern standard error estimation software to historical census samples. We exploit a unique new data source from the 1880 census to validate our methods for standard error estimation and then we apply this approach to the 1850-1870 and 1900-1930 decennial censuses. We conclude that Taylor series estimation can be used effectively with the historical decennial census microdata samples, and should be applied in research analyses that have the potential for substantial clustering effects.

INTRODUCTION

Decennial census microdata are a key component of social science infrastructure. Census microdata are among the most frequently used data sources in the leading journals of population, economics, and sociology; indeed, during the past decade census microdata have been used more frequently in the pages of *Demography* than any other data source.[1] Most of these publications use the Integrated Public Use Microdata Series (IPUMS), which makes large, nationally-representative samples of every surviving census from 1850 though 2000 freely available to scholars in harmonized format through a user-friendly data access system with comprehensive documentation (Ruggles et al. 2004). Since 1995, over 25,000 researchers have registered to use the IPUMS data extraction system, and they have produced 2,000 publications and working papers.

Tests of statistical significance require appropriate standard errors in order to make valid inferences. If an estimated standard error is too small then a researcher is more likely to reject a true null hypothesis. On other hand, if an estimated standard error is too large then the researcher is less likely to reject a false null hypothesis. Because statistical tests are at the core of quantitative research in the social sciences, having appropriate standard error estimates is vital. Without reliable standard error estimates, the cumulative process of scientific research rests on a foundation of unstable inferences.

Census microdata samples are individual-level data clustered by household, and often incorporate stratification and differential probabilities of selection (and differential non-response) resulting heterogeneity in sample weights. The clustering of individuals within

[1] In the decade from 1997 through 2006, U.S. census microdata were used in 54 *Demography* articles, substantially more than any other data source.

households can significantly increase standard errors of estimates because the number of independent observations is less than the number of actual observations in each census file. Standard errors in cluster samples depend on both the size of the sampled clusters and on the homogeneity of variables within clusters, measured by an intraclass correlation coefficient (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 people. Thus, variables such as race and poverty status, which tend to be 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 (Kish 1992). The IPUMS samples for years prior to 1960 were designed to capitalize on geographically-sorted source materials, which enhance precision through implicit geographic stratification. Such procedures can lower standard errors, especially for variables that are highly correlated with geography

This paper examines the impact of complex sample designs on standard error estimates using IPUMS historical U.S. census microdata samples for 1850 through 1930. We compare standard errors computed using a simple random sampling assumption – the usual way of computing standard errors using the historical U.S. census data – to estimates that take the complex sample design into account. We develop and test a new variable that allows us to apply modern standard error estimation software to IPUMS census microdata. Based on the results of

this evaluation, we develop recommendations for standard error estimation when using the IPUMS samples from 1850 through 1930.

METHODS

Sophisticated methods for standard error *estimation* are now available in easy to use procedures in statistical packages. In particular, easy-to-use Taylor series estimation procedures are now incorporated into most statistical analysis packages (SAS 1999; Stata 2001; SPSS 2003), and these procedures yield reliable estimates (Kish and Frankel 1974; Krewski and Rao 1981; Dippo and Wolter 1984; Weng, Zhang, and Cohen 1995; Hammer, Shin and Porcellini 2003). Researchers frequently use these products for analysis of survey data, but they are seldom applied to decennial census microdata. In part, this is because the census microdata samples do not include all the variables required to take advantage of the new software algorithms.

Implicit Geographic Stratification

The publicly available IPUMS *samples* created at the University of Minnesota for the censuses of 1850 through 1930 all employ the same basic design, with minor variations to accommodate differences in source materials and innovations in data-entry technology. In general, samples are taken within each enumeration district by generating a random starting point between one and five, and then designating 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.

The sample designs for these early censuses differ fundamentally from those of more recent censuses because they were drawn from microfilm images of the original census

enumerator manuscripts instead of from machine-readable files. Explicit stratification was not feasible, but the organization of historical census enumeration forms incorporated 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 enumeration manuscripts 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 can yield substantially greater precision than a simple random sample of households. By capitalizing on implicit stratification of the pages, this design yields higher precision for many estimates.

Pseudo Strata and Taylor Series Linearization

Taylor series linearization is the easiest and most widely-used method for estimating variance with complex sample designs, but it is not designed for samples with implicit stratification. Because the stratification is implicit, there is no geographic unit in the data that corresponds precisely to the geographic stratification embedded in the page ordering of the data. This poses a major problem for Taylor series linearization, since the method requires *explicit* information about strata.

To create a proxy for the implicit geographic stratification of the historical IPUMS samples, we summarized microfilm page numbers to construct pseudo-strata. These pseudo-

strata reflect the implicit geographic order of the microfilm reels. The strata are calculated by creating sequential values every 100 to 250 microfilm pages. The number of pages included in each stratum varied from census to census, and were calibrated to ensure that 70 percent of the strata contained at least 10 sampled households.

Subsample Replicate Approach

An alternative to Taylor series variance estimation is the subsample-replicate approach (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. Each subsample incorporates the same stratification and clustering used to select the sample as a whole.[3] Iterative computer procedures are then used to estimate standard errors. It has not been established, however, that the subsample replicate method is reliable in samples that incorporate implicit geographic stratification. Subsample replicate estimates could be biased if the degree of

---

[2] Statisticians are currently developing model-based variance estimates (e.g., Little 2003) to improve upon the design-based variance estimates we 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 for complex sample designs like the census (Kalton 2002).

[3] The IPUMS subsample standard error estimates were calculated using the subsample variable on the 1850-1930 IPUMS 1% samples that systematically divides the IPUMS sample into 100 replicate sub-samples. For the historical IPUMS samples, the values are assigned sequentially from 0 to 99 for each household. The IPUMS replicate standard errors are constructed by calculating the mean for each subsample separately and then averaging the 100 subsample means. The standard deviation of these 100 replicates is the IPUMS sub-sample replicate standard error estimate.

geographic homogeneity varies greatly with geographic scale. For example, a typical 1-in-100 sample includes one household approximately every fifth manuscript census page; if we divide that sample into 100 subsample replicates, however, cases occur only once every 500 pages and thus high heterogeneity. This difference in geographic scale could have significant implications for variance.

Validation

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. Fortunately, a new source can provide near-perfect estimates for the 1880 census. The 1880 Population Database provides individual-level data on the entire population of 50 million Americans, assembled with the help of 11 million hours of volunteer effort by members of the Church of Jesus Christ of Latter-Day Saints (Goeken et al. 2003). This remarkable database provides an ideal laboratory for the evaluation of sample designs. The database allows us to simulate any sample design precisely, and by repeatedly drawing samples from the full 1880 census we can develop highly accurate variance estimates for the IPUMS sample design.[4] We can then compare

[4] The 1880 replicate reconstruction from the full 1880 census universe was completed using the household selection rules for the IPUMS 1880 1% sample (Ruggles and Menard 1995). In order to replicate the IPUMS sample design we used the SAS procedure, PROC SURVEYSELECT to randomly select one person per microfiche page. This process was carried out for 100 replicate 1% samples, and the sampling rules determined the eligibility of the household or individual for inclusion in the final sample. The sampling rules selected households or families if the randomly selected sample line was the household or family head and selected individuals for group quarters larger than 31 persons. If the randomly selected sample line did not meet the sampling rule criteria, the sample line was disregarded. We used PROC SURVEYSELECT to select 100, 1% samples following the IPUMS sampling rules from

these estimates with results based on both the IPUMS subsample replicates and Taylor series linearization to assess the reliability of both methods.

RESULTS

Table 1 compares alternate methods for estimating standard errors of selected variables in the 1880 census. The first two columns are based on sample replication of the entire 1880 population. We drew 100 independent 1% samples from the complete-count database, mimicking the sample design used to create the historical samples. The standard errors derived from these replicates are unbiased estimates of the standard error that would be expected in a 1% sample.

The last three columns present the ratio of the standard error using the full 1880 census subsample replication to the standard error calculated from the one-percent IPUMS sample for 1880 using three methods: subsample replicate, Taylor series linearization with pseudo-strata, and simple random sample assumptions. We regard the full 1880 census subsample replicates as a gold standard, so the ideal ratio would be 1.0. Ratios under 1.0 represent underestimated standard errors, and ratios over 1.0 represent overestimated standard errors.

The top row of Table 1 shows that in 1880 the average age was estimated to be 24.2 years with a full census replication standard error estimate of 0.03. The ratio of the 1% IPUMS sample replicate to the full census replication was 1.0. The ratio for the Taylor series was 1.1, and the

---

the full 1880 microdata census file (Ruggles and Menard 1995). The full 1880 census 1% sample replication has variance estimates that are on a different scale than the replications of the 1% IPUMS file which are based on a sample of only .1% of the records. As a result the full 1880 census 1% replications results are divided by the square root of 100 (i.e., 10) because the standard errors from the full 1880 census replication are based on samples that are 100 times larger than those from the 1% IPUMS.

simple random sample was 0.9. These estimates are all quite close to one another, suggesting that for this variable the particular method of standard error estimation does not matter much. This is not surprising, since age is not highly correlated within clusters (i.e., households) or by strata (i.e., geographic strata).

The poorest performing estimates in Table 1 were for the non-white and non-relative simple random sample estimates, with a ratio of 0.5. When using statistics that assume a simple random sample, which is the default in most statistical packages, the standard error estimates would be about half as large as the standard errors that take the sample design into account. This is the result of household clustering; these characteristics were both highly correlated within households.

For a few characteristics, the estimation methods overstate true standard error. This is most noticeable for the characteristics Male and Socioeconomic Index. These variables vary greatly across geographic areas in 1880, so the implicit stratification significantly reduces standard errors. The comparatively high ratios shown for these variables in Table 1 suggest that none of the estimation methods fully captures the dampening effects of implicit stratification on standard errors.

On the whole, however, the Taylor series and IPUMS 1% subsample replicate estimates both performed well; the ratios for only a few characteristics deviate substantially from 1.0, and the average deviation is negligible. By contrast, assuming a simple random sample leads to substantial underestimates of standard errors for several characteristics that are highly correlated within households.

When we turn to other historical samples from the IPUMS—1850, 1860, 1870, 1900, 1910, 1920, and 1930—we no longer have the ability to have the gold standard of the full census

replication, since only the IPUMS sample data exist for those years. Table 2 therefore varies in structure from Table 1. The first column contains the population parameter estimate from the IPUMS sample, and the second column contains the standard error estimates assuming the data were collected as a simple random sample. The remaining two columns present the Taylor series and subsample replicate estimates as ratios to simple random sample estimates. A high ratio indicates that the standard error estimation method yields larger standard errors than would be obtained from a simple random sample of the same size.

Several characteristics have consistently high ratios across all census years. In addition to non-white and non-relative mentioned in the discussion of 1880, we examine two variables with extremely high ratios: urban residence and farm residence. These are, in fact, household-level variables, not person-level variables, but we have rectangularized the file and added the household-level characteristics to each person record to demonstrate the effects of clustering. Therefore, these two characteristics are identical for every individual in the household. With this perfect correlation, standard errors based on a simple random sample assumption are severely underestimated.

DISCUSSION

Our validation analysis using the 1880 Population Database shows that both the Taylor series and the subsample replicate method compared favorably to the full census replication estimates of the standard error. We had been concerned about the impact of not being able to control for the implicit stratification for the Taylor series estimates, but our pseudo-strata variable tracks the full 1880 sample replication with only minor deviations. We were also had concerns that the subsample replicate estimates could be biased if the importance of geographic homogeneity varies with geographic scale. The analysis demonstrates that the IPUMS replicate

estimates were not severely biased by differences in geographic scale. Based on the 1880 analysis both types of standard error estimates that attempt to adjust for complex sample design work reasonably well to estimate standard errors. This is crucial, since we do not have the same gold standard to evaluate the other years of IPUMS data. Because the IPUMS samples for 1850 through 1930 are all drawn in a similar fashion, we believe it is reasonable to infer that methods which work in 1880 will also work for the entire period, but this cannot be tested at this time.

The subsample replicate and Taylor series produce very similar results for 1850-1870 and 1900-1930, and although we cannot compare these estimates to full census replication as we did in 1880, we infer that both methods produce reasonable standard error estimates in these years as well, since the IPUMS samples during this period were all drawn in a similar fashion as the 1880 sample. We did not find that the Taylor series estimates for the 1850-1930 period substantially overstated standard errors because they cannot explicitly include the effects of implicit stratification; the pseudo-strata variable we created for performing Taylor series worked to incorporate the implicit stratification in the sample design.

We have now added the pseudo-strata variable (PSTRAT) to the publicly available IPUMS data files for the period 1850-1930. This will make it easy for researchers to create Taylor series standard error estimates using the major statistical packages (SPSS, SAS, and STATA). These Taylor series estimates in the statistical programs are coupled with procedures for regression, cross-tabulation and univariate analyses.

The results presented here demonstrate that in certain expected circumstances treating historical IPUMS data as simple random samples will yield underestimates of standard errors, and this could cause researchers to draw unwarranted statistical conclusions. The results also show, however, that for characteristics that are not highly correlated within clusters (i.e.,

households) —such as sex, socioeconomic index, or labor-force participation—a simple random sample assumption can provide a reasonable estimate of standard errors

Many analyses of IPUMS data do not pose standard error estimation problems. Recent IPUMS-based publications have focused, for example, on elderly persons residing with their adult children (Ruggles 2007), mothers of young children (Short, Goldscheider, and Torr 2006), men aged 20-39 (Rosenfeld 2006), and married couples in which the wife is aged 18 to 40 (Schwartz and Mare 2005). In each of these cases, the researchers examined a population subgroup that typically appears just once per household. For example, most households contain no more than one intergenerational coresident group, one mother of small children, one young adult man, or one young married couple. In such cases there is little or no clustering, and thus little reduction in statistical power. It follows that for most analyses, assuming a simple random sample will yield acceptable estimates of standard errors.

There are some situations, however, in which the likelihood of large clustering effects is greater. Analyses of historical school attendance pose risk, since if one child in a family attended school, the odds were high that all the school-age children were in school. Many schooling analyses, however, subdivide the schoolchildren by age and sex; such studies avoid clustering effects, because a given household is unlikely to have multiple children of a particular age and sex. The worst clustering arises with analysis of those population characteristics—such as poverty or urban residence —that almost by definition apply to entire households or families. Even with these topics, however, the clustering problem evaporates if the unit of analysis does not usually occur more than once per household. Thus, for example, studies of the poverty status of families, householders, or mothers would be virtually unaffected by clustering.

In the end, researchers must evaluate their research designs and judge whether they pose a potential risk of clustering that might lead to underestimated standard errors. Where a significant risk exists, we recommend that data users make use of the new strata and cluster variables on the IPUMS web site to produce Taylor series standard errors estimates using the statistical package of their choice. This methodology can be used both for calculating percentages and means (as it was in this paper) and to calculate regression models (e.g., ordinary least squares and logistic regression). These procedures will allow researchers to take advantage of implicit geographic stratification while also paying attention to the clustering of people and their characteristics within sampled households. Although the subsample replicate estimates were also found to produce reasonable standard error estimates and are available in the IPUMS data files, these estimates are harder to obtain as each analysis needs to be run 100 times and standard errors must be calculated from the resulting sampling distribution.

REFERENCES

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.

Goeken, R., C. Nguyen, S. Ruggles, and W.L. Sargent. 2003. "The 1880 United States Population Database." *Historical Methods* 36(4): 27-34.

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. Honolulu, HI.

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

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.

_____. 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.

Rosenfeld, M.J. 2006. "Young Adulthood as a Factor in Social Change in the United States." *Population and Development Review* 43: 617-629.

Ruggles, S. and R.R. Menard. 1995. *Public Use Microdata Sample of the 1880 United States Census of Population: User's Guide and Technical Documentation.* Minneapolis: Social History Research Laboratory, pp. 4-7.

Ruggles, S. 2007. "The Decline of Intergenerational Coresidence in the United States." *American Sociological Review* (forthcoming).

Ruggles, S., M. Sobek, T. Alexander, C.A. Fitch, R. Goeken, P.K. Hall, M. King, and C. Ronnander. 2004. *Integrated Public Use Microdata Series: Version 3.0* [Machine-readable database]. Minneapolis, MN: Minnesota Population Center [producer and distributor].

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.

Schwartz, C.R., and Robert D. Mare. 2005. "Trends in Educational Assortative Mating, 1940-2003." *Demography* 42: 621-646.

Short S.E., Fr.K. Goldscheider and B.M. Torr. 2006. "Less Help for Mother: The Decline in Coresidential Female Support for the Mothers of Young Children, 1880-2000." *Demography* 43 (4): 617-629.

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.

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*. New York: Springer-Verlag.

Table 1. Standard Error Computations Comparing Replicate Estimates from the Complete 1880 Census to Estimates Derived from Sample Data using Alternative Methods

Selected Person Characteristics	Parameter Estimate From Entire 1880 Census	Replicate Variance Estimates Drawn from Entire 1880 Census *	Ratio of Estimates Using the IPUMS 1880 One-percent Sample to Replicate Estimates from Entire 1880 Census: Subsample Replicate Method	Taylor Series Linearization with pseudo-strata	Simple Random Sample
Age (mean)	24.2	0.03	1.0	1.1	0.9
Male (percent)	50.9	0.05	1.2	1.1	1.3
Married (percent)	34.9	0.06	0.9	1.0	1.1
Nonwhite (percent)	13.4	0.10	1.0	0.9	0.5
Foreign Born (percent)	13.6	0.07	1.0	0.9	0.7
Socioeconomic Index (mean)	6.8	0.02	1.3	1.2	1.0
Other Relative (percent)	5.3	0.04	1.1	1.2	0.9
Non-Relative (percent)	9.7	0.08	0.9	0.8	0.5
Average			1.0	1.0	0.9

Source: 1880 Full Census and the 1880 IPUMS 1% Sample

Table 2. Comparison of standard error estimation techniques: IPUMS samples, 1850-1930

	Selected Person Characteristics	IPUMS Estimate	IPUMS Simple Random Sample	Ratio of Estimates to Sample to Simple Random Sample from: IPUMS Taylor Series	Ratio of Estimates to Sample to Simple Random Sample from: Replicate Estimates from IPUMS Samples
1850					
	Age (mean)	23.1	0.10	1.3	1.4
	Male	51.2	0.10	0.9	0.9
	Nonwhite	2.2	0.07	2.2	2.1
	Foreign Born	14.5	0.15	1.8	1.9
	Socio-Economic Index (mean)	5.3	0.03	1.2	1.2
	Child Under Age 5	17.6	0.09	1.0	1.0
	Enrolled In School	20.7	0.12	1.4	1.3
	Labor Force Participant	89.5	0.14	1.2	1.1
	Urban residence	19.9	0.19	1.6	2.1
	Farm residence	52.9	0.27	2.3	2.4
1860					
	Age	23.4	0.09	1.5	1.5
	Male	51.2	0.08	0.8	0.9
	Nonwhite	2.0	0.07	2.2	2.6
	Foreign Born	15.5	0.12	1.6	1.7
	Socio-Economic Index	6.0	0.03	1.1	1.2
	Child Under Age 5	18.1	0.08	1.1	1.1
	Enrolled In School	20.6	0.10	1.3	1.3
	Labor Force Participant	52.5	0.12	0.9	0.9
	Urban residence	22.2	0.15	1.4	1.9
	Farm residence	47.6	0.25	2.2	2.6
1870					
	Age	23.9	0.07	1.5	1.5
	Male	50.4	0.07	0.8	0.8
	Nonwhite	12.9	0.10	2.0	1.9
	Foreign Born	14.4	0.08	1.5	1.3
	Socio-Economic Index	6.0	0.03	1.1	1.3
	Child Under Age 5	16.8	0.06	1.0	1.0
	Enrolled In School	17.0	0.08	1.3	1.3
	Labor Force Participant	52.8	0.10	0.8	0.9
	Urban residence	25.2	0.12	1.5	1.7
	Farm residence	41.3	0.19	2.2	2.4
1880					
	Age	24.1	0.03	1.1	1.2
	Male	50.9	0.07	0.8	0.9
	Married	35.1	0.08	0.9	1.1
	Nonwhite	13.5	0.12	2.0	2.4
	Foreign Born	13.4	0.07	1.4	1.5
	Socio-Economic Index	6.8	0.02	1.1	1.1
	Other Relative Householder	5.7	0.05	1.4	1.6
	Nonrelative Householder	8.4	0.07	1.7	1.7
	Child Under Age 5	16.6	0.06	1.1	1.1
	Enrolled In School	17.9	0.08	1.3	1.5
	Labor Force Participant	55.0	0.08	0.8	0.9
	Urban residence	28.9	0.13	1.5	2.0
	Farm residence	43.3	0.15	2.2	2.1

Table 2. (continued)

Selected Person Characteristics	IPUMS Estimate	IPUMS Simple Random Sample	Ratio of Estimates to Sample to Simple Random Sample from: IPUMS Taylor Series	Ratio of Estimates to Sample to Simple Random Sample from: Replicate Estimates from IPUMS Samples
1900				
Age	25.8	0.02	1.2	1.1
Male	51.0	0.05	0.9	0.8
Married	36.6	0.06	1.0	1.0
Nonwhite	11.8	0.08	2.0	2.2
Foreign Born	13.7	0.06	1.4	1.5
Socio-Economic Index	8.5	0.02	1.1	1.1
Other Relative Householder	6.1	0.04	1.4	1.5
Nonrelative Householder	8.5	0.06	1.8	1.8
Child Under Age 5	14.4	0.04	1.1	1.0
Enrolled In School	16.9	0.06	1.3	1.3
Labor Force Participant	56.7	0.06	0.8	0.8
Urban residence	39.7	0.09	1.5	1.6
Farm residence	38.3	0.11	2.0	2.0
1910				
Age	26.7	0.02	1.2	1.2
Male	51.4	0.05	0.9	0.9
Married	38.8	0.06	1.0	1.1
Nonwhite	11.1	0.07	2.0	2.2
Foreign Born	14.8	0.06	1.5	1.6
Socio-Economic Index	10.3	0.02	1.1	1.1
Other Relative Householder	6.3	0.04	1.4	1.4
Nonrelative Householder	8.8	0.05	1.8	1.8
Child Under Age 5	13.7	0.04	1.1	1.0
Enrolled In School	22.3	0.05	1.2	1.2
Labor Force Participant	59.4	0.05	0.9	0.8
Urban residence	45.0	0.09	1.5	1.7
Farm residence	32.5	0.10	2.0	2.0
1920				
Age	27.5	0.02	1.3	1.3
Male	51.1	0.04	0.8	0.8
Married	40.9	0.04	1.0	0.9
Nonwhite	10.5	0.06	2.0	2.1
Foreign Born	13.4	0.05	1.4	1.4
Socio-Economic Index	10.7	0.02	1.1	1.0
Other Relative Householder	6.4	0.04	1.4	1.5
Nonrelative Householder	7.2	0.05	1.8	1.8
Child Under Age 5	13.2	0.04	1.1	1.2
Enrolled In School	20.8	0.05	1.2	1.2
Labor Force Participant	57.5	0.05	0.8	0.8
Urban residence	51.2	0.09	1.4	1.8
Farm residence	30.2	0.09	1.9	2.0

Table 2. (continued)

Selected Person Characteristics	IPUMS Estimate	IPUMS Simple Random Sample	Ratio of Estimates to Sample to Simple Random Sample from: IPUMS Taylor Series	Ratio of Estimates to Sample to Simple Random Sample from: Replicate Estimates from IPUMS Samples
1930				
Age	28.8	0.03	1.3	1.2
Male	50.6	0.05	0.8	0.9
Married	42.9	0.06	1.0	1.0
Nonwhite	10.1	0.09	2.0	2.2
Foreign Born	11.6	0.05	1.3	1.3
Socio-Economic Index	11.4	0.03	1.1	1.2
Other Relative Householder	6.8	0.05	1.4	1.5
Nonrelative Householder	6.6	0.06	1.7	1.8
Child Under Age 5	11.5	0.05	1.1	1.2
Enrolled In School	22.7	0.07	1.2	1.3
Labor Force Participant	55.9	0.06	0.8	0.8
Urban residence	55.4	0.10	1.4	1.6
Farm residence	24.8	0.13	1.9	2.3
Average			1.4	1.4

Source: 1850, 1860, 1870, 1880,1900, 1910, 1920 and 1930 IPUMS Samples