Estimating Birth Counts for Small Geographical Domains Used for Control Totals in Raking Adjustment

Amang Sukasih[1], Donsig Jang[1], Barbara Lepidus Carlson[1], Mary Edith Bozylinsky[2]
[1]Mathematica Policy Research. Inc., 600 Maryland Ave., SW, Suite 550, Washington, D.C. 20024
[2]EMMES Corporation, 401 North Washington St., Suite 700, Rockville, MD 20850 USA
(ASukasih@Mathematica-MPR.com)

Abstract

This paper presents an application of raking adjustment in the Fragile Families Survey (FFS). Raking is an iterative post-stratification method that is usually used to control the distribution of the sample so that it is consistent with some known population distributions or totals by key variables. The problem in this work, however, is that domains available for control totals are not exactly matched to domains of interest for analysis. In this paper, we explain how weighting, raking, and creating synthetic estimates of total births for small domains of interest were done in the FFS using available CDC birth record data.

Keywords: Weighting, Estimation; Natality data; City of birth; CDC; Fragile Families.

1. Introduction

Raking is an iterative post-stratification method that is usually used to control the distribution of the sample so that it is consistent with some known population distributions or totals by key variables. This technique is often used for survey weighting so that weighted sums of sample units are in accordance with control totals by key characteristics. Such control totals are usually obtained from the sample frame or external data such as census counts, administrative data, projections or other survey data. However, domains available for control totals are sometimes not exactly matched to domains of interest for analysis. For example, natality data from the Centers for Disease Control and Prevention (CDC) is often used as a good source for population birth totals in the U.S. Data on place of birth (state and county) are available, from which the total number of live births by state or county can be produced. Unfortunately, the same variable for smaller geographical units like city is not available, yet often some policy research studies need such birth totals by city of birth. (City of residence for the birth mother is available on the CDC natality data, but not city where birth occurred.) No other data sources on births by city of occurrence were available. The Fragile Families and Child Wellbeing Study, or Fragile Families Survey (FFS), is one example of a study where births were sampled independently within hospitals within selected cities, and city-level analysis was of primary interest. It was desirable to rake by city-level control totals of births by race/ethnicity, age, education attainment, and marital status in this study.

This paper will present estimation techniques of total number of live births by city used for raking using available CDC birth record data. The empirical example will be limited to only weighting for data from the baseline survey.

2. FFS Data

FFS is a longitudinal survey that targets a population of unmarried parents through a sample of births from unmarried couples. The objective of the Fragile Families study is to be able answer the following questions:

(1) What are the conditions and capabilities of unmarried parents, especially fathers?

(2) What is the nature of the relationships between unmarried parents?

(3) How do children born into these families fare?

(4) How do policies and environmental conditions affect families and children?

To do so, the study follows a cohort of nearly 4,700 children born in large U.S. cities between 1998 and 2000, and collects data from both mothers and fathers at birth (baseline), and when children are ages one, three and five, plus in-home assessments of children and their home environments at ages three and five.

The study selected births through a three-stages sampling: (1) first-stage: sampled large cities, (2) second-stage: sampled hospitals within city, and (3) third-stage: sampled births within hospital. At the first stage, the 77 large cities were stratified into nine strata based on welfare generosity, the strength of welfare support, and the strength of the local labor market. Then a national sample (16 cities out of 77 large cities) was selected through PPS. One city is treated as a non-respondent. At the second stage, selection of hospitals depended on the city. In Oakland, Austin, Newark, Richmond, and Corpus Christi, all hospitals were selected. In New York and Chicago, a random sample of hospitals from the frame of hospitals with 1,000+ non-marital births per

year was selected. In other cities, the hospitals were first ordered based on the number of non-marital births (the 1996/1997 data), and then hospitals were selected in order starting from hospital with the most non-marital births until 75% of non-marital births in the city were covered. At the third stage, births were selected within hospitals; however, the births were not sampled from a list frame. Both married and un-married births were sampled until reached preset quota. The births were sampled by including births during a period of time (either predetermined start and end dates, or predetermined start date and the end date was determined by when reached quota).

When weighted, the data represent the non-marital (and marital) births in large U.S. cities (i.e., the 77 cities with 200,000+ population). For more information on the Fragile Families study including the list of the 77 cities, the reader can visit www.fragilefamilies.princeton.edu .

3. Weighting in FFS

In this study, the weighting process was done in five steps: (1) computed base weight, (2) adjusted for nonresponse, (3) raked/calibrated to known population, (4) trimmed weights, and (5) re-raked. The computation of the base weight took into account unequal selection probabilities due to unequal sampling rates across cities or groups in the population and nonrespondents in the sample. However, the weighted estimates produced using these weights could drastically over- or under-estimate the true number in the Fragile Families (FF) population because of certain sampling features. For example, sampling the births during a period of time and then annualizing through weighting do not guarantee a proper representation of births throughout the year. Thus, the distributions of characteristics in the sample may fail to mimic those in the population.

The estimates' accuracy may be improved if we know the population's distribution and are able to adjust the individual survey weights to such known distribution. A commonly used adjustment method is *raking*. The basic nonrespondent-adjusted sampling weights are used as an input for the raking process. Raking adjusts these weights by aligning the total sum of the weights for selected variables, which are considered as risk factors in the study.

For this study, external information of the population is available for the raking process. Natality data for the years 1998, 1999, and 2000 are available from Centers of Disease Control and Prevention (CDC). These data contain important characteristics such as mother's marital status, mother's race/ethnicity, mother's age, mother's education, etc. Using these known population data, the FF survey weights can be adjusted by aligning the sample distribution (weighted) to the population distribution based on selected variables. In turn, the adjusted weights would properly represent births in the population. For example, the weighted estimate of total non-marital births for an overall domain (city-level or national) should approximately equal the total of non-marital births in the Fragile Families population.

In addition to demographic variables, the CDC natality file has variables that represent geographical location for both the birth's occurrence and the mother's residency. This information may be used to match an individual CDC birth record to a geographical area. Each record can be identified by city-, county-, and state-codes for the mother's residence. Unfortunately, city-codes are not available for the birth's place of occurrence. Nevertheless, we use the available population information for the raking process. Discussion on how we utilize this information and overcome the problem of incomplete information for the birth's occurrence is provided in the following sections.

The output of the raking process is the final survey weights that have been "raked" and "trimmed" (and re-raked after trimming). Raking is done to produce three sets of weights: individual city-level weights, national-level weights with all cities included in the data, and national-level weights where Austin is excluded from the data. The details of the process are explained below.

4. Estimating Totals for Raking

Raking is a method of adjusting the weights to ensure that the weighted counts of the sample are consistent with the known counts of the population. This is a post-stratification technique done within each raking cell constructed by raking variables along with constraints that the sum of weights should match the known population totals for each level of all raking variables. In this study, the variables used for the raking process are given in Table 1. Even though the adjustment was done within individual raking cells, the raking process only requires known marginal population totals rather than totals for individual cells.

Raking is an iterative process in which adjustments are made to scale the weights to the known marginal population totals for each raking variable. In each step, the weights are "raked" so that the weighted counts equal the population totals for each level of a particular raking variable. After each step, however, the weighted counts and the population counts may not be equal for the levels of other variables, so the adjustment process is repeated in a cycle until the differences between the weights in the previous iteration and the current iteration converge to a

predetermined value. MPR implemented the raking algorithm using a SAS macro.

Table 1. List of Raking Variables

Variable name	Description	Levels
Marital Status	Mother's Marital Status	2
Education	Mother's Education Level	5
Race/Ethnicity	Mother's Race/Ethnicity	4
Age	Mother's Age	7

For the FF baseline survey, raking adjusts the weights attached to individual sampled births, so that the sums of these weights match population counts. The goal is to produce three sets of weights—individual city-level, national-level with all cities, and national-level excluding Austin—so we need population counts for the 77 large cities. As mentioned in the previous section, city information is not available in CDC data for the birth's occurrence; thus, population birth counts are not available at the city-level. Population birth counts *are* available for the county-level, however, and other useful information from the CDC natality file is city and county information for the mother's residency. We estimate the total number of births for the city-level using the information available from the sample data and from the CDC natality file as follows.

Even though city information is not available for the birth's occurrence, certain cities may use the county's information in place of the missing city information. Two types of cities qualify: 1) cities that have identical boundaries as their county and/or 2) cities that contain all the hospitals for their county. For these cities, we may use county birth totals as city birth totals. Table 2 presents a list of the cities from the FF samples that are type(s) 1 and/or 2.

Table 2. List of FF Cities of Type(s) 1 and/or 2

FF city	Type
New York	1,2
Norfolk	1,2
Baltimore	1,2
Philadelphia	1,2
Richmond	1,2
Corpus Christi	2
San Antonio	2

For the remaining FF cities, we estimated the number of births for each city using both residence and birth's occurrence information. The county-level (population) birth count for a particular county was partitioned into three parts:

A = total number of births given by mothers living in the FF city,

B = total number of births given by mothers living in a city other than the FF city (but within the same county of occurrence as a FF city),

C = total number of births given by mothers living in a county other than the county in which they gave birth

The estimate of the FF city-level birth count is computed as

$$D = rA + sB + tC \qquad (1)$$

where

D = estimate of total number of births in the FF city,

r = proportion of births occurring in the FF city given by mothers living in the FF city,

s = proportion of births occurring in the FF city given by mothers living in the non-FF city within the county,

t = proportion of births occurring in the FF city given by mothers living outside the county.

The value of r is assumed to be large ($0.9 < r < 1$), since it is reasonable to assume that in general mothers who live in a particular city give birth in a hospital within the same city, especially for large cities such as FF cities.

Intuitively, s should be larger than t. In this case, however, there is no compelling reason to treat non-FF cities differently regardless of their locality within or outside an FF county; therefore, the values of s and t are assumed to be equal. Under this assumption, the above equation can be simplified into

$$D = rA + u(B + C) \qquad (2)$$

where

u = proportion of births occurred in the FF city given by mothers lived in the non-FF city within or outside the county.

Now, for a particular FF city, the proportion of births where the mother lives in the same FF city is

$$P = \frac{rA}{rA + u(B + C)}. \qquad (3)$$

This value can be estimated from the sample when u, whether the mother lives in the city of interview, is

available from the survey. Suppose p denotes such a proportion from the sample data. Hence, the estimate of u can be obtained by replacing P with p in (3) and solving it for u as follows:

$$u = \frac{\left(\frac{1}{p} - 1\right) rA}{B + C}. \qquad (4)$$

In this study, we assumed r = 0.9, and u was computed based on the values of A, B, and C obtained from CDC data. The value of p was obtained from the survey data as given in Table 3.

Table 3. Estimate of Proportion of Births given by Mothers Living in the Same City by FF City (p)

FF city	Proportion of mothers living in the city of interview
Austin	0.90
Boston	0.79
Chicago	0.85
Detroit	0.86
Indianapolis	0.78
Jacksonville	0.89
Milwaukee	0.83
Nashville	0.56
Newark	0.64
Oakland	0.95
Pittsburgh	0.80
San Jose	0.72
Toledo	0.65

Source: FF survey data

5. Alternative Estimates of Totals for Raking

The above estimator of control totals utilized available birth data from CDC (total births by county of occurrence and total births by city of residency) adjusted the estimates for under- and over-coverage, and also used data on total hospitals within city/county. One, however, may compute the estimates of total births by city of occurrence differently. The following are three alternative options explored:

Option 1: Use total births by city of mother's residency

This option is a quick method to estimate the total births by city. However, with this option the estimator fails to include birth occurrences in a FF city where the mother lived outside the city. This results in under coverage of the population. On the other hand, the estimator includes some births that occurred outside the FF city; that results in over coverage of the population.

Option 2: Use total births by city of mother's residency, and adjusted for over coverage

This option utilizes information from the CDC data to get the proportion of births where county of residence equals county of occurrence, and uses this proportion to adjust for the over coverage found in Option 1. With this option, the total number of births within the city of occurrence is estimated as:

$$D = P_1 \times D_R \qquad (5)$$

where

D = estimate of total number of births within city of occurrence,

D_R = total number of births within city of residence,

P_1 = proportion of births where county of residence equals county of occurrence (from CDC population data); that is,

$$P_1 = \frac{\text{births within county to mothers living in that county}}{\text{births within county}}$$

Notice that the proportion is computed at the county level; however, the proportion is needed at the city level (which is unfortunately not available from CDC data).

Option 3: Use total births by county of occurrence, and adjust for coverage of city level

Another alternative is to use total births within county of occurrence, and then adjust the coverage to the city level by multiplying this number by a proportion of total births within city of residence to total births within county of residence. From CDC population data, this proportion can be computed based on place of mother's residency. With this option, the total number of births within the city of occurrence is estimated as:

$$D = P_2 \times D_O \qquad (6)$$

where

D = estimate of total number of births within city of occurrence,

D_O = total number of births within county of occurrence,

P_2 = proportion of births within city of residence to births within county of residence from CDC population data); that is,

$$P_2 = \frac{\text{birth counts within city (of residence)}}{\text{birth counts within county (of residence)}}.$$

Note that the proportion is computed based on the place of mother's residency; however, the proportion is actually needed based on place of birth occurrence (which is unfortunately unavailable for city level).

Appendix Table 1 provides a comparison between the estimates of characteristics of non-marital births where the control totals for raking were computed based on Option 1 and those computed based on synthetic method as described in Section 4.

6. Baseline Survey Analysis Weights

6.1 City-level Weights

The birth weights for individual FF cities were developed to provide users of the survey data with final survey weights for analyses of individual cities. Using the methods explained in Section 4, these weights have been adjusted/raked to be consistent with total population counts of US births for individual cities based on CDC data. The following years of CDC natality files have been used in raking the weights for individual FF cities:

1998 CDC data:	Austin, Oakland
1999 CDC data:	Baltimore, Philadelphia, Richmond, Detroit, Newark
2000 CDC data:	Corpus Christi, Indianapolis, Boston, Nashville, Jacksonville, San Antonio, New York, Norfolk, Chicago, Milwaukee, Pittsburgh, San Jose, Toledo

Appendix Table 2 provides a summary of the city-level weights for the 20 FF cities in the sample. Note that not all 20 cities were selected probabilistically. Only the 15 cities that were selected randomly are included in the national weights.

6.2 National-level Weights

The national-level weight is the final survey weight attached to individual births for analyses that pool all records (15 cities part of the national sample) within the sample. The analysis will generalize to births that occurred in the 77 large cities defined as the FF population. The weights were developed based on city-level weights (computed in the earlier steps), which were in turn raked to total (population) birth counts in the 77 cities based on CDC data.

We produced two sets of national-level weights: one was computed based on all 15 cities in the national sample with all 77 cities as the population being targeted (`finalntlwt`), and the other one was computed based on only 14 cities (Austin is excluded) in the sample with all 77 cities as the population being targeted (`afinalntlwt`)[1].

Table Appendix 2 presents some statistics that were estimated using `afinalntlwt`.

References

Deming, WE and Stephan, FF (1940), "On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals are Known." Annals of Mathematical Statistics, 11, 427-444.

Lohr, Sharon L. (1999), "Sampling: design and analysis", Duxbury Press, North Scituate, MA.

Oh, H.L. and Scheuren, F. J. (1983), "Weighting Adjustment for Unit Nonresponse," pp. 143-184 in Incomplete Data in Sample Surveys (Volume 2), (Eds. W.G. Madow, I. Olkin and D.B. Rubin), New York: Academic Press.

Potter, Frank J. (1990), "A Study of Procedures to Identify and trim Extreme Sampling Weights", ASA Proceedings of the Section on Survey Research Methods, American Statistical Association (Alexandria, VA), pages 225-230.

Potter, Frank J. (1993), "The Effect of Weight Trimming on Nonlinear Survey Estimates", ASA Proceedings on the section on Survey Research methods, American Statistical Association (Alexandria, VA), pages 758-763.

Vu, Thu (2003), "Methodology Report on Sample Weights Construction for the Fragile Families Baseline Survey." Center for Health and Wellbeing, Princeton University, Princeton, NJ.

Reichman, N.E., Teitler, J.O., Garfinkel, I., and McLanahan, S.S. (2001), "Fragile Families: Sample and Design." Children and Youth Services Review, Vol. 23, Nos. 4/5, pp. 303-326.

[1] National analyses using afinalntlwt should include only samples in the 14 cities, while analyses using finalntlwt should include all records in the sample (15 cities).

Appendix Table 1. Characteristics of non-marital birth (percentage) based on different control totals

Characteristic	Previously published	Synthetic estimate of control total	Un-weighted estimate	Option 1 (based on city level residency)
Substance use during pregnancy				
Any alcohol use	10	10	10	10
Any drug use	6	5	5	5
Any cigarette use	23	23	23	22
Low birth weight baby	10	12	11	12
Health insurance				
Medicaid	71	75	74	75
Private	23	17	21	17
Other	6	8	5	8
Enough time in hospital	81	82	80	82
Initiation of prenatal care				
1st trimester	77	78	78	78
2nd trimester	18	16	18	16
3rd trimester	3	3	3	3
No prenatal care	2	2	2	2
Baby's living arrangement				
Mother and father	49	49	48	48
Mother only	30	29	30	30
Mother and others	21	22	22	22
Total respondents	2,659	2,360	2,360	2,360
Estimate of total	NA	450,491	NA	404,108
Deff (weight)	NA	3.65	NA	3.63

Appendix Table 2. Summary of city-level survey weights, by FF city

FF city	Sample size	Min	Max	Mean	Median	Sum	Deff*
Oakland	330	1.5	125.7	17.3	11.1	5692.4	2.4
Austin	326	6.7	185.2	32.1	14.7	10460.5	2.3
Baltimore	338	13.0	175.9	44.2	33.8	14949.0	1.4
Detroit	327	13.7	126.3	41.0	36.5	13398.8	1.2
Newark	342	7.0	48.5	13.0	11.2	4452.7	1.2
Philadelphia	337	13.4	345.2	69.2	51.9	23327.0	1.5
Richmond	327	2.4	36.2	11.8	10.3	3844.0	1.3
Corpus Christi	331	1.1	91.2	15.9	7.6	5250.0	2.2
Indianapolis	325	7.3	273.8	40.6	25.7	13196.5	1.9
Milwaukee	348	6.0	179.0	29.1	23.3	10131.8	1.4
New York	297	36.7	2936.8	391.7	261.0	116325.0	2.0
San Jose	326	3.9	1040.6	48.7	13.8	15880.2	6.0
Boston	99	13.9	375.3	72.5	40.8	7181.1	1.9
Nashville	102	1.0	801.4	79.9	41.5	8148.3	3.0
Chicago	134	11.5	4143.7	357.4	169.1	47893.3	3.0
Jacksonville	100	18.4	466.9	102.7	62.4	10267.6	1.9
Toledo	101	9.5	224.1	48.7	36.5	4921.0	1.9
San Antonio	100	26.2	1552.7	240.4	115.1	24038.0	2.5
Pittsburgh	100	5.1	396.6	36.9	26.9	3686.0	3.1
Norfolk	99	12.0	156.9	42.4	30.3	4195.0	1.5

*In this table the design effect measures extra variability added in the variance of statistics due to unequal weights