Methods for Producing Consistent Control Totals for Benchmarking in Survey Sampling

Ismael Flores Cervantes[1]
[1]Westat, 1600 Research Blvd, Rockville, MD 20850

Abstract

Estimates from probability samples may not match known totals of populations due to sampling variation, undercoverage, or nonresponse. In this situation, special weighting adjustments such as raking or poststratification are implemented so the sums of weights match known totals. This process improves the face-validity of the survey because the total estimates reproduce the "true" or accepted total population generally produced by official statistical agencies. The totals, which are used as benchmarks in weighting, are sometimes generated combining different sources and in some situations they do not match the eligible population of the survey. In this paper we take a heuristic approach and propose methods for computing control totals that are consistent and represent the population of interest while maintaining the face validity of the estimates.

Key Words: Benchmarking, Control totals, Iterative Proportional Fitting

1. Benchmarking of Probability Sampling Weights

In the creation of weights for some surveys, there is the need for the sum of adjusted weights to match known totals available from official sources (Deming and Stephan, 1940). In practice, this type of weighting adjustment is generally done through poststratification or raking (Kalton and Flores Cervantes, 2003). In a more formal definition, poststratification and raking are members of a family of estimation procedures known as *sample weight calibration* or *calibration weighting* initially described by Deville and Särndal (1992); and Deville, Särndal, and Sautory (1993). There are several definitions of calibration but one commonality is that these methods use auxiliary information to adjust probability-sampling weights constrained by the calibration equation defined as

$$\sum_s w_k \mathbf{x}_k = \mathbf{t}_x, \tag{1}$$

where w_k is the calibrated weight, $\mathbf{x}_k$ is the auxiliary variable vector of a characteristic of the sampled element k, and $\mathbf{t}_x$ is the population total of $\mathbf{x}_k$ defined as $\mathbf{t}_x = \sum_U \mathbf{x}_k$. In addition to providing face validity to the sum of sampling weights to known population totals (i.e., consistency with external sources), calibrations methods are used to reduce the variance of estimates, and to adjust for nonresponse (Särndal and Lundström, 2005).

There are numerous papers that deal with the left hand of the calibration equation (1). In most of the literature, it is assumed that the control totals are available and ready to use. In practice, this is not the case and these totals have to be estimated. In this paper we describe several methods for producing these consistent control totals. We describe

several approaches and their statistical properties. Some of these methods have been used since 2001 in a telephone survey.

2. Control Totals

The principal sources of control totals are statistical federal or state government agencies. In the US, the Census Bureau disseminates population data by demographic characteristics from programs such as the decennial Census or continued surveys like the American Community Survey. In order to facilitate the use of the data and to address concerns related to the confidentiality of individuals in the population, the data is released in the form of contingency tables. The population totals are cross-classified by few demographic characteristics in these tables. There is partial overlap among the contingency tables; and, there are not tables with a full demographic profile with the combination of all demographic characteristics. Table 2 illustrates this situation. Table 2 lists some tables from the 2010 U.S. Census of Population and Housing Summary File 1 (U.S. Census Bureau, 2011). In this case, no table contains population totals for all combinations of categories by sex, race, ethnicity, age, and group quarter type.

Table 1: Example of tables available in from the 2010 Census of Population available from the Census Bureau

Table Numbers	*Description*
PCT012rxxx	Population subjects by sex by age (103 levels) by race/ethnicity (White alone, Black or African American alone, American Indian and Alaska Native alone, Asian alone, Native Hawaiian and other Pacific Islander alone, some other race alone, two or more races, Hispanic or Latino, White alone, not Hispanic or Latino)
P04300xx	Group quarters population by sex by age (under 18, 18-64, 65 and over) by group quarters type
P01200xx	Population subjects by sex by age (23 groups)
PCO00100xx	Group quarters population by sex by age (18 groups)
PCT020r0xx	Group quarters population by group quarters type by race and ethnicity (White alone, Black or African American alone, American Indian and Alaska Native alone, Asian alone, Native Hawaiian and other Pacific Islander alone, some other race alone, two or more races, Hispanic or Latino, White alone, not Hispanic or Latino)
PCT022r0xx	Group quarters population by sex by quarters type for the population 18 years and over by race and ethnicity (White alone, Black or African American alone, American Indian and Alaska Native alone, Asian alone, Native Hawaiian and other Pacific Islander alone, some other race alone, two or more races, Hispanic or Latino, White alone, not Hispanic or Latino)

The second difficulty is that in some cases the available data do not match the eligible population in the survey. For example, most telephone surveys collect data on the non-institutionalized population. In contrast, the available tables with the most detailed age groups do not provide separate totals this population. In this case, estimates for this population need to be estimated and subtracted before these totals can be used as benchmarks in weighting. Another issue is the different definitions for population characteristics between the published tables and collected data in the survey. For example, the population projections from the California Department of Finance (California Department of Finance, 2013a) use the U.S. Office of Management and

Budget (OMB) race definition with only 5 race groups and excludes the race category "Other" (Office of Management and Budget, 2003). In most surveys, the data collected includes this group as a separate category. If these totals are used, adjustments to either the control totals or to the collected data are required to match these definitions. The last issue is that the totals become outdated due to changes in the population. Although revisions are made after the first release; the updated totals are not generally produced at the same level of detail as in the first release. Therefore adjustments are needed to update the older and more detailed tables to reflect the latest totals.

The motivation for developing methods for the creation of consistent control totals is to address the problems described above. Such methods are most useful in surveys with a large number of raking dimensions and with dimensions that incorporate multiple demographic variables.[1] Furthermore, these methods should be implemented in a timely fashion during the production period for the creation of the weights. This calls for procedures that do not require specialized software.

The methods described in this paper have been used since 2001 in the California Health Interview Survey (CHIS) (California Health Interview Survey, 2011). CHIS is a large telephone survey of non-institutionalized civilian residents of California. CHIS is designed to provide population-based estimates for most California counties and all major ethnic groups, including several ethnic subgroups. In the last step of weighting, the CHIS weights are raked to control totals. Due to the need to produce estimates at these geographic and demographic levels, the more than 10 raking dimensions with close to 1,000 cells are used. The raking dimensions are defined by geographic area, age groups, gender, race, ethnicity, household tenure, number of adults in household and they combine information from the Decennial Census, American Community Survey, and the California Department of Finance (DOF). Since the demographic tables from the Census Bureau include the population in group quarters, one of the methods described in this paper was used to produce control totals that exclude this population but matching at the same time any equivalent total (i.e., without group quarters) by any demographic characteristics reported in any other table. Depending of the year in the CHIS cycle, another method was used to update control totals to reflect changes in the population through the years. A similar method was used to create a bridge between the different race/ethnicity definitions used in the DOF population projections and the Census Bureau tables. The latest implementation of these methods in CHIS 2011-2012 is described in the methodology reports found in the CHIS website at http://healthpolicy.ucla.edu/chis/design/Pages/methodology.aspx.

3. Computing Control Totals

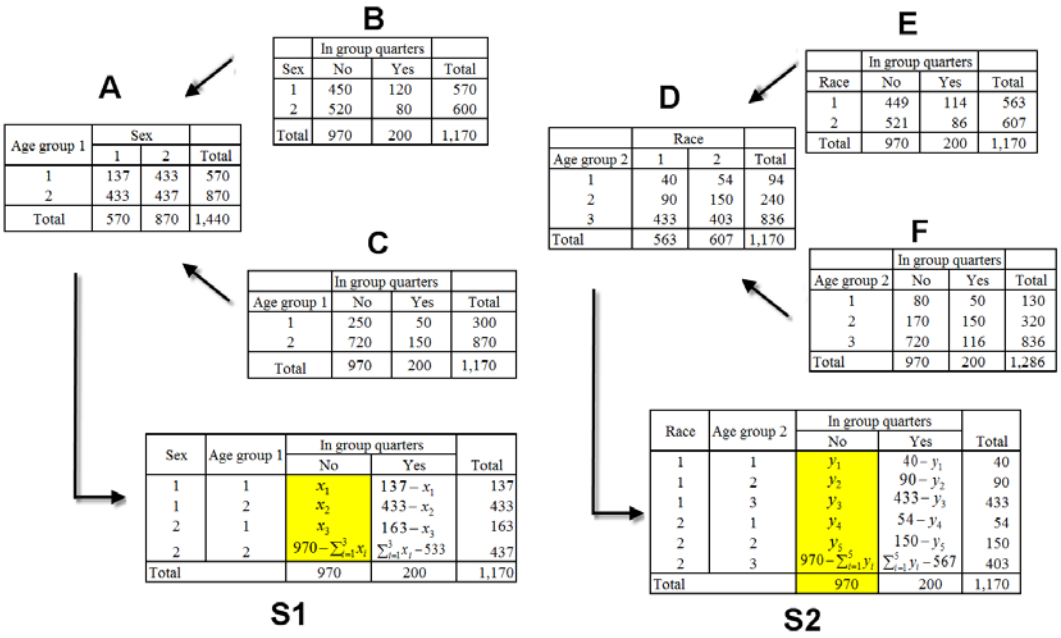
There are two problems to address when computing the total from contingency tables from external sources. The first problem is to generate consistent totals in a single contingency table using auxiliary tables where the demographic categorical variables partially overlap. The second problem is to generate totals in more than one contingency table with totals consistent across all tables using multiple auxiliary tables where there

[1]If the weighting adjustment uses poststratification or raking with few cells, a more naïve approach may be more appropriate.

categorical variables do not necessarily overlap. The following sections describe a heuristic approach to address these situations.

3.1 Computing Consistent Totals for One Contingency Table

To illustrate the first problem, consider three published tables (A, B, and C) shown in Figure 1a. We are interested in creating a raking dimension that excludes the population in group quarters by sex by age group 1. However, the only available information consists of tables A (sex by age group 1), B (group quarters by sex) and C (group quarters by age group 1). In this example, the auxiliary tables are A, B, and C. The categorical variable sex overlaps tables A and B; while the variable age group 1 overlaps tables A and C. The goal is to create the contingency Table S1 with the estimated totals represented by $\mathbf{x} = (x_1, x_2, x_3)^t$ that are consistent with the totals in auxiliary tables A, B, and C. Note that unless there is any other source of data, we cannot compare the estimated totals to the true population totals. Therefore, it is important to understand the statistical properties of these estimated totals and the implicit assumptions that depend on the method used to create them.



(a) (b)

Figure 1: Separating the population in group quarters using auxiliary tables.

3.1.1 Linear programing solution

Estimating **x** is not a new problem and there are several methods that yield a solution. One method is linear programming (LP) where the best outcome is determined (i.e., minimizing a linear cost function) for a list of requirements described as linear relationships. Mathematically, the problem can be expressed as minimize *the objective function* $D = \mathbf{c}^t\mathbf{x}$, subject to the constrains $\mathbf{Ax} = \mathbf{b}$ and $\mathbf{x} \geq 0$, where **c** is a vector of known coefficients, and **A** is the matrix of coefficients used to express the constraints between the totals **x** and the margins of the auxiliary tables **b**. In this example, the variables **A**, **x**, and **b** are defined in Figure 2.

$$\mathbf{A} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix}, \qquad \mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}, \qquad \mathbf{b} = \begin{pmatrix} 459 \\ 250 \end{pmatrix}, \qquad \mathbf{c} = \begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix}$$

Figure 2: Matrix representation of relationships among tables S1, A, B, and C.

Since we are estimating the number of persons, we can add a constrain for the totals to be integers (i.e., $x_i \in Z$), which makes this an Integer Linear Programming (ILP) problem. The main drawback of ILP is the use of heuristic approaches that make it very difficult to implement without specialized software. In cases where the number of cells is very large as in the CHIS dimensions, a solution may not found or the found the solution is not optimal. Removing the integer requirement improves the situation but it requires the use of simplices (i.e., simplex method). This is a geometric approach and an optimal solution may not be found too. However, the main problem with these methods is that there is not a clear understanding of the statistical properties of the estimated totals generated using this method.

3.1.2 Least norm solution

Following a more algebraic approach, we remove the integer and positive constraints and solve the equation $\mathbf{Ax} = \mathbf{b}$. Since $\mathbf{A}$ is not a square matrix, it does not have full rank and its inverse matrix $\mathbf{A}^{-1}$ does not exist. In this case, the system is undetermined with fewer equations than unknown variables (m rows < n columns) with an infinite number of solutions. A way to overcome this is to compute the generalized pseudo inverse matrix or the Moore–Penrose pseudo-inverse as $\mathbf{A}^{+} = (\mathbf{A}^t\mathbf{A})^{-1}$ and use it to find the solution as $\mathbf{x}_0 = \mathbf{A}^{+}\mathbf{b}$. The solution $\mathbf{x}_0$ minimizes the squared Euclidean distance or norm $D = \|\mathbf{x}\|^2$ where the Euclidian distance is defined as $\|\mathbf{x}\| = \sqrt{\sum_n x_i^2}$. This result is known as the *least norm solution*. The implementation requires inverting the matrix $\mathbf{A}^t\mathbf{A}$. There are several drawbacks in this approach. First, the optimal solution may include negative values. In CHIS, this may be acceptable because negative counts are likely to be computed for the population in group quarters (i.e., they represent a small proportion of the total population) and these negative totals are excluded when computing the control totals. The second problem is that the solution is influenced by the larger counts in the distance function D. This is easier to address by modifying the distance function and scaling the contribution of each x_i as $D_{x^2} = \sum_i (x_i - E_{i\in rc})^2 / E_{i\in rc}$ where $E_{i\in rc} = N_r N_c / N$, N_r and N_c are the row and column margins associated to cell x_i, and N is the population grand total. This distance function is known as the *Chi-squared distance*. The optimal solution allocates the counts as if they are generated independently from any column and rows interaction. Notice that since minimizing D does not imply that $D = 0$, the optimal solution does not necessary produces independent counts. On the other hand, it is possible that the optimal solution meets this condition and the generated counts reflect this model of independence implicitly imposed. Although the linear approach is an improvement over the previous method, this approach requires specialized software such SAS IML to compute the inverse of large and sparse matrices.

As mentioned before, the optimal solution may include negative totals that are not acceptable in some situations. The negative total problem can be solved by using nonlinear distance functions where the optimal solution is forced to be non-negative. The disadvantage of the use of non-linear distance functions is that more complex mathematical methods are needed to find the optimal solution.

3.1.3 Nonlinear distance solutions

In the next approach, we start by re-parameterizing $x_i' = e^{x_i}$ and minimizing the linear distance function defined before. Since the domain of the exponential function is only defined for values greater than 0, the optimal total will be greater than 0. An additional parameterization of $\mathbf{x}$ defining $x_i' = e^{r_j + c_k} = e^{r_j} e^{c_k}$ enables us to better understand the statistical properties of the totals in the combined contingency table under this parameterization. Table S1 in Figure1a is represented as a contingency table in Table 2.

Table 2: Re-parameterizing of the contingency tables by gender and group quarters

			In group quarters		
			Yes	*No*	
			Column effect		
Sex	*Age group 1*	*Row effect*	e^{c_1}	e^{c_2}	*Total*
1	1	e^{r_1}	$e^{r_1+c_1}$	$e^{r_1+c_2}$	137
1	1	e^{r_2}	$e^{r_2+c_1}$	$e^{r_2+c_2}$	433
2	2	e^{r_3}	$e^{r_3+c_1}$	$e^{r_3+c_2}$	163
2	2	e^{r_4}	$e^{r_4+c_1}$	$e^{r_4+c_2}$	437
Total			970	200	1,170

Rather than computing the total e^{x_i} in a cell, we compute it as the product of the effects of row r_j and column c_k subject to the same margin constraints. This re-parameterization matches the log-linear model that is more appropriate for contingency tables with positive counts. The log-linear model is $\ln(F_{ij}) = \mu + \lambda_i^A + \lambda_j^B + \lambda_{ij}^{AB}$, where $\ln(F_{ij})$ is natural logarithm of the expected cell frequency in cell ij, μ is the overall mean of the natural logarithm of the expected frequencies, λ_i^A and λ_i^B are the main effects of variables A and B respectively and λ_{ij}^{AB} is the interaction term between the variables A and B. Notice that this parameterization matches the unsaturated model where the interaction term $\lambda_{ij}^{AB} = 0$. Iterative proportional fitting (IPF) is an algorithm that can be used to decompose the matrix of the table in Figure 2 as the outer product $\mathbf{t} = \mathbf{r} \otimes \mathbf{c}$ of two vectors $\mathbf{r} = (e^{r_1}, \ldots, e^{r_R})^t$ and $\mathbf{c} = (e^{c_1}, \ldots, e^{c_C})^t$ maintaining the margins fixed (Stephan, 1942). IPF is an iterative algorithm and it only requires the computation of sums, multiplication and divisions. Although this parameterization provides a better statistical model and a method to solve it, it does not tell us much about the distance function being minimized.

Before studying the distance function for this re-parameterization, we generalize the optimal solution for any distance function[2]. In order to minimize a distance function $D(\mathbf{x})$ with the constraints $\mathbf{Ax} = \mathbf{b}$, we solve the equation $\nabla L(\mathbf{x}, \boldsymbol{\lambda}) = 0$ where $L(\mathbf{x}, \boldsymbol{\lambda}) = D(\mathbf{x}) - \boldsymbol{\lambda}^t(\mathbf{Ax} - \mathbf{b})$, and $\boldsymbol{\lambda}$ is the vector with the Lagrange multipliers. This is equivalent to solving the equation $\mathbf{d}(\mathbf{x}) - \boldsymbol{\lambda}^t\mathbf{A} = \mathbf{0}$ where $\mathbf{d}(\mathbf{x}) = \nabla D(\mathbf{x})$. The general solution is $\mathbf{x} = \mathbf{d}^{-1}(\boldsymbol{\lambda}^t\mathbf{A})$. One nonlinear distant function of interest is the *Kullback–Leibler distance* defined as $D_{KL}(\mathbf{x}) = \sum_i (x_i \ln(x_i) - x_i)$. The distance $D_{KL}(\mathbf{x})$ has desirable properties, such as $d_{KL}(x_i) = \ln(x_i)$ and the inverse is $d_{KL}^{-1}(x_i) = e^{x_i}$ that matches our parameterization. The distance $D_{KL}(\mathbf{x})$ guarantees positive counts but finding the optimal solution is numerically intensive and requires computing Lagrange multipliers and numerically solving a nonlinear system of equations using Newton's Method or any other similar algorithm.

[2] We consider only those distance functions that are twice differentiable and defined at 0.

A close look at the similarities between this approach for generating cells totals in the contingency tables and those used in generalized raking procedures or calibration (Deville and Särndal, 1992; and Deville, Särndal, and Satutory, 1993) provides a key finding that simplifies the computation of the optimal solution[3]. In calibration, there is also an optimization problem where a distance function which involves sampling weights is minimized subject to a set of constrains represented by the calibration equation (1). If we adjust our problem so it resembles a calibration problem, then we can use the same methodology to find the optimal solution.

These authors recognized that the optimal solution for the distance for Case 2 (also called multiplicative method or raking ratio method) in Table 1 in the 1992 paper can be also found using Iterative Proportional Fitting (IPF) instead of the minimization approach[4]. Since Case 2 uses the same distance function $D_{KL}(\mathbf{x})$ (except for the scaling factors), we can treat our problem as a raking problem and apply the same methods used in raking to find a solution. Notice that although the mathematical methods to generate totals contingency tables and raking sampling weights are the same, the theoretical motivation is different. We also need to bear in mind the implicit assumptions from using the distance function $D_{KL}(\mathbf{x})$. As long as these assumptions hold, we can use raking methods to generate these totals.

In our case, since the contingency tables do not have values similar to sampling weights, we can create seeds with any arbitrary value greater than 0 and treat them as weights.[5] We also need to convert the auxiliary tables as control totals files for the raking procedure. There is also available software for raking weights which includes the command *rake* in package *survey* in R (Lumley, 2012), the SAS macro *CALMAR*[6] from the Institut National de la Statistique et des Études Économiques, (Statutory, 1993), and SAS macro *rake_and_trim* from Abt Associates (Izrael, Hoaglin, and Battaglia, 2000). For the CHIS problem, we used Westat proprietary SAS macro *FS_RAKE*. However, the raking implementation based on the IPF algorithm is very easy to implement in any language.

3.2 More Than Two Contingency Tables

In this section we expand the problem to generate consistent totals in more than one contingency table using multiple auxiliary tables where the categorical variables in the auxiliary tables do not necessarily overlap. As an example, consider the contingency Table S2 containing the totals to generate represented by $y = (y_1, y_2, y_3, y_4, y_5)^t$ and auxiliary tables D, E and F shown in Figure 1b. We are interested in control totals that exclude the population in group quarters by race by age group 2. That is, we need to

[3] There are other similarities not discussed here. For example, Case 1 or linear method matches the least norm solution or the Chi-squared distance with the appropriate scaling factors. As a result, the same software used to create calibrated weights based on the linear model can be used to generate the totals based on linear distance functions.

[4] The formal proof that these procedures are equivalent requires to show that the iterative processes $e^{x_{ij}^{n+1}} = e^{x_{ij}^{n}}\left(\sum_R b_i / \sum_R e^{x_{ij}^{n}}\right)$ and $e^{(\mathbf{x}^{n+1},\boldsymbol{\lambda}^{n+1})} = e^{(\mathbf{x}^{n},\boldsymbol{\lambda}^{n})} e^{-k\left(\nabla^2 L(\mathbf{x}^{n},\boldsymbol{\lambda}^{n})^{-1}\nabla L(\mathbf{x}^{n},\boldsymbol{\lambda}^{n})\right)}$ converge to the same minimum $\mathbf{x_0}$.

[5] The distance function $D_{KL}(\mathbf{x})$ involves $\ln(x_i)$ which is only defined for $x_i > 0$.

[6] *CALMAR* also produces other calibrated weights with different distance functions that can be used to generate totals in contingency tables with other distance functions.

create a contingency Table S2 with totals by race by age group 2 and by group quarters using the auxiliary tables D (race by age group 2), E (race by group quarters) and F (age group 2, by group quarters). At the same time, we also want to produce the totals in table S1 from Figure 1a in such a way that the totals in the cells are both consistent between Tables S1 and S2 (i.e., appropriate combinations of $x_1, x_2,$ and x_3 are consistent with appropriate combinations of $y_1, y_2, y_3, y_4,$ and y_5) and consistent across all auxiliary tables (e.g., Tables A, B, C, D, and E). This situation can be extended to include more contingency tables, for example, Table S3 with totals by age by sex by ethnicity and by group quarters, Table S4 with totals by a specific group such as White non-Hispanic by age by group quarters, and so on with their corresponding auxiliary tables (e.g., Tables F, G, etc.). The goal is still the same, which is to generate contingency tables with consistent totals across all tables. Notice that the categorical variable for age (age group 2) in the contingency Table S1 in Figure 1b does not match the levels of age (age group 1) in Table S1 in Figure 1a (otherwise, this example becomes the same described in the previous section). Furthermore, additional auxiliary tables can contain totals that are nested within other variables (i.e., totals at the state level, totals at the county level, or totals at any sub-county area level). The number of contingency tables with consistent totals to generate depends on the number of auxiliary tables available and the raking dimensions we want to create.

The method to produce consistent totals across multiple contingency tables requires "interweaving" all the levels of the categorical levels so the cell totals are generated at the same time. The levels are interleaved in a pseudo-population file with all possible combinations of these levels. Since we are using raking software, we also need convert the auxiliary tables to files containing control totals and rake a sampling weight in the pseudo-population file. The following steps describe the details of this process:[7]

A. Determine the raking dimensions to be used in raking the sampling weights. The raking dimensions used in weighting determine the totals to estimate. For example, if any of the raking dimensions does not use race, then there is no need to compute totals by race. There is also no need to process any auxiliary table by race. Notice that different raking dimensions may have common variables but different number of levels. This situation arises frequently with a categorical variable such as age that has many levels. For example, one raking dimension at the state level may have age groups with 12 levels while another raking dimension defined at the county level may have age groups with 3 levels. Notice when creating raking dimensions, it is not necessary for these levels to be nested.

B. Determine the categorical variables and levels needed to create the raking dimensions. Once the raking dimensions have been decided, the next step is to select the categorical variables and levels needed to create these raking dimensions. We also need to verify there are auxiliary tables with the same categorical variables and levels. Auxiliary tables with more detailed levels are acceptable because they can be summarized by combining the levels. If there are no auxiliary tables with at least the level of detail required by the raking dimension, then the raking dimension need to be redefined or dropped from the adjustment. Special care is needed for the following situations:

[7] These steps should be modified depending on the raking software.

- Different definitions. Make sure that the variables have the same definition in the survey and auxiliary tables. If they do not match these totals should not be used in this process.
- Same variables with different levels. Variables such as age groups that are defined grouping several age ranges should be identified as separate variables; for example, age group 1 and age group 2 in Figures 1a and 1b are not the same categorical variables.
- Irrelevant levels: It is recommended to only use the levels that appear in the raking dimensions. For categorical variables with many levels such as age, it is not recommended to generate totals at a detailed level such as single years, presuming that they can be combined and summarized after generating the contingency table. The raking algorithm is not likely to converge with variables with too many levels.

C. Convert the auxiliary tables from a row-column format to files with the tables in a list format. Each record in the file represents one combination of the cross-tabulation of the row and columns in the table. For example, the file with the list format of Table A in Figure 1a should look like this

AGE_GRP_1	*GRP_Q*	*TOTAL*
1	1	50
1	2	250
2	1	150
2	2	720

Create variables and assign a numeric level for each of the categorical variables identified in Step B. The variable names and levels should be the same across all tables that have the same variables. In this example the variables are AGE_GRP_1 GRP_Q with values 1 and 2. Special care is needed for the following cases:

- Incomplete auxiliary tables. The auxiliary table should be complete before it can be used to estimate the total. However, some auxiliary tables contain totals for one portion of the population. In this case, use the information of the other auxiliary tables to obtain the complement of the population. In most cases, it is derived by subtraction between a complete and the incomplete tables. For example, there is an auxiliary table that includes the population in group quarters by age group. Using the auxiliary table with the total population by age group, compute the total population not in group quarters. Append these totals to create a complete table.

  In another example, suppose there is a raking variable that is defined at the county level and it includes only some counties. In this situation, auxiliary tables for all counties can be appended to create the file. However, the geographic variable that indicates the county, GEO_COUNTY should take a common value such as 0 for all counties that are not used in the raking dimensions. Another approach is to create an incomplete file that includes the totals for all the counties of interest. Then we obtain the totals for the missing counties by subtraction from the auxiliary table that contains the totals at the state level. In this case, the totals for the missing counties should also have the variable GEO_COUNTY.

- Inconsistent auxiliary tables. Sometimes the auxiliary tables will not add to the same total despite representing the whole population. This may be the result of rounding errors or deliberate suppression implemented to reduce disclosure risk. In other cases, the auxiliary tables come from different sources. In this case, adjust the tables (i.e., use poststratification) to the total population. If the table has total by categories, then adjust these separately. Verify that totals from auxiliary variables are consistent for the whole population and for any categorical variables common across the tables. Inconsistent totals are the main reason why the raking process fails. Inconsistent totals for nested common variables and with different levels are very difficult to spot.

D. Summarize the files to create files with margins. Before summarizing the files created in Step C, create the combined variable with values that represent the combination of the categorical levels. For example, for Table B in Figure 1a, the combined variable AGE_GRP_1_GRP_Q should look like this

AGE_GRP_1_GRP_Q	*AGE_GRP_1*	*GRP_Q*	*TOTAL*
11	1	1	50
12	1	2	250
21	2	1	150
22	2	2	720

It is recommended to name the combined variables in a way that reflects the names of the categorical variable being combined. In this example, the combined variable is created as AGE_GRP_1_GRP_Q= AGE_GRP1|| GRP_Q where || is the append operation in SAS.

Proceed to summarize the files by the combined variables and compute new totals by adding the population counts for all records with unique values of the combined variable. Then remove any records in the summarized files where the new total is 0. Although the algorithm can handle 0 margins, the solution requires large negative values of x_i for e^{x_i} to become very small (it never reaches 0). As a check, compute the sums of the new totals in the summarized files. This should match the total population from the auxiliary tables. These files contain the margins for the raking software.

E. Create a pseudo population file. The pseudo population file will contain one record per level from the cross-tabulation of all the categorical levels in all auxiliary tables selected in Step B. Each record should also have the variable name and level. Using the variables and levels create the same combined variables created in Step D. Create an ID variable that enumerates all levels. For example for the tables in Figures 1 and 3, the pseudo population file should look like this (only some combinations are listed here):

ID	*AGE_GRP_1*	*AGE_GRP_2*	*SEX*	*RACE*	*GRP_Q*	*AGE_GRP_1_SEX*
1	1	1	1	1	1	11
2	1	1	1	2	2	11
3	2	2	2	2	1	22
4	2	3	2	1	2	22

ID	*AGE_GRP_1_GRP_Q*	*SEX_GRP_Q*	*RACE_AGE_GRP_2*	*RACE_GRP_Q*	*AGE_GRP_2_GR_Q*
1	11	11	11	11	11
2	12	12	21	22	12
3	21	21	22	21	21
4	22	22	13	12	32

In this file, the combined variables are AGE_GRP_1_SEX (Table A), AGE_GRP_1_GRP_Q (Table B), SEX_GRP_Q (Table C), RACE_AGE_GRP_2 (Table D), RACE_GRP_Q (Table E) and AGE_GRP_2_GR_Q (Table F).

Remove the records in the pseudo population file for levels of the combined variables without margins in the summarized files in Step D. In this process we do not rake records to zero margins.

Special care is needed for the nested categorical variables. For example, consider the levels of AGE_GRP_1 from Figure 1 and AGE_GRP_2 in Figure 3. The levels of AGE_GRP_1 are 1 (0-17), and 2 (18 year or older). The levels of AGE_GRP_2 are 1 (0-17), 2 (17 to 64 years old), and 3 (65 years or older). The only valid combinations of (AGE_GRP_1, AGE_GRP_2) are (1, 1), (2, 2), and (2, 3). As a check, produce frequency tables for each of these variables to make sure that all levels represented are the only valid combinations of nested tables.

F. Assign a seed for all records in the pseudo population file. Any values different than 0 will do it. This variable plays the role of the sampling weight in the raking software.
G. Use any available software and rake the seed using the combined variables as the raking dimensions and the files created in Step D as the margins. In some cases, it may converge to a solution very slowly. If the procedure does not converge, rake a limited number of dimensions to identify the problematic dimension. Possible solutions are to redefine the raking dimension by reducing the number of levels and repeat the previous steps incorporating the reduced dimension. An alternative is to separate the raking dimensions in two or more groups and derive them separately. This was done in CHIS for the control totals that include Asian groups which were created separately from the control totals defined by age, sex, race, and ethnicity.

Once the seeds have been raked, remove the population in group quarters by excluding the records where GRP_Q=2. Then produce totals by summarizing this file by the appropriate categorical variables. Notice that in this case, the file created in Step G contains an artificial or synthetic population. This synthetic population can be summarized using the “raked seed” by any of the categorical variables to compute totals. Since these total come from a single file, these totals are consistent. Notice that this methodology can be used to create complex synthetic populations for applications other than control totals for raking.

3.3 Removing the Independence Requirement in the Contingency Tables

As described in previous sections, when the totals in the contingency tables cells are created using any of these optimization methods, there are implicit assumptions and statistical models induced in these totals. Specifically, when constant cells are raked to the margins, the generated totals follow a model where there are row and column multiplicative effects without interactions. In other words, the cross-classified variables are independent. This assumption may not hold for some populations.

A way to remove the independence condition is to use different initial seed values. The justification is found when we examine a modified version of the Kullback–Leibler distance function that incorporates these values. The distance function becomes $D_{KL}(\mathbf{x}, \mathbf{s}) = \sum_i (x_i \ln(x_i/s_i) - x_i + s_i)$ where s_i is the seed associated to the cell with total x_i. Instead of minimizing the distance between vectors $\mathbf{x} = (x_1, \dots, x_n)^t$ and $\mathbf{1}_n = (1, \dots, 1)^t$ (i.e. constant seeds factor out and they become a constant in the distance function) the procedure minimizes the distance between and $\mathrm{x} = (x_1, \dots, x_n)^t$ and $s = (s_1, \dots, s_n)^t$. If $\min\{D_{KL}(\mathbf{x}, \mathbf{s})\} = 0$ then the raked weight matches exactly the distribution of the variable seeds (including the relationship and interactions between column and rows).

A source of the values for the seeds can be any of the public use micro data files such as the American Community Survey's public use microdata files. The seeds can be generated by computing totals using person weights. We need totals for all levels in the pseudo population file created in Step E in the previous section. Potential problems are empty cells, but they can be avoided by accumulate several public use files (it does not matter if the totals do not match the population total because the procedure only needs the distribution of the seeds). Another option is to assign a small value such as 1 to those cells with positive margins but no seed from the public use files.

4. Dealing Other Issues with Control Totals

4.1 Updating Control Totals

Raking can be used to update control totals. For example, it is common for agencies to release updated population totals during the intra-decennial census years. However, these releases are generally are not at the same detailed levels as those population totals immediately released after the census. These decennial counts can be updated to reflect the more update totals by raking the older counts to the new counts. The raking ratios computed as $\tau_c = \sum_c T_{I,i} / \sum_c T_{0,i}$ where $T_{I,i}$ is the total after raking, and $T_{0,i}$ is the total before raking can be seen as growth factors between periods 0 to I for the population in cell c.[8] The updated raking dimensions are then created summarizing the raked totals. In this case, the implicit assumption is that growth of the population in the cells used in raking from period 0 to I is constant. This assumption is more likely to hold if the time between periods is short. This assumption does not hold when there is differential growth within the cell. The same method can be implemented in case the totals T_0 include the population in group quarters. That is the totals are first raked to the totals in period I and then the population in group quarters are removed from the file after raking.

[8] A formal definition of this growth function and the description of its properties are beyond the scope of this paper.

4.2 Changing Definitions in Population Characteristics

As mentioned in Section 2, external sources of control totals may not use the same definitions for population characteristics. For example, data available from the Census Bureau produced 2010 totals by 6 race categories that include a race group called "other". In contrast, the California Department of Finance produces population projections based on the OMB race definition which excludes the "other" group for years 2010 and later. The Census Bureau released the 2010 Modified Race Data Summary File (MR) that contained totals based on the OMB race definition (U.S. Census Bureau, 2012). The MR file is the used as a bridge between these definitions. First, we rake the totals $T_{\overline{OMB},c}$ from the 2010 non-OMB file (which is created using the auxiliary tables and contain the most detailed levels of all categorical variables used in raking dimensions) to the 2010 MR file totals $T_{OMB,c}$ using common cells defined in both files.[9] The raking factors are defined as $\tau_c = T_{OMB,c}/T_{\overline{OMB},c}$ can be seen as a set of transformation factors that convert non-OMB totals to OMB totals.[10] Since the raked file or synthetic population contains the most detailed levels of all categorical variables of the auxiliary files, then this file can be used to create the raking dimensions for a population that reflect the OMB race definition. The implicit assumption is that there is a linear relationship between the two populations within the cell. This assumption is more likely to hold if the cells are small.

4.3 Example

The control totals for CHIS 2011-2012 were derived using the methods described in previous sections. Figure 3 shows the flow chart of how these methods were used and the files created after each step. The labels in the files in the figure describe the year of the data (2010 or 2012), the race definition used in the totals (OBM or non-OBM), and if the file includes the population in group quarters (GQ and non-GQ).

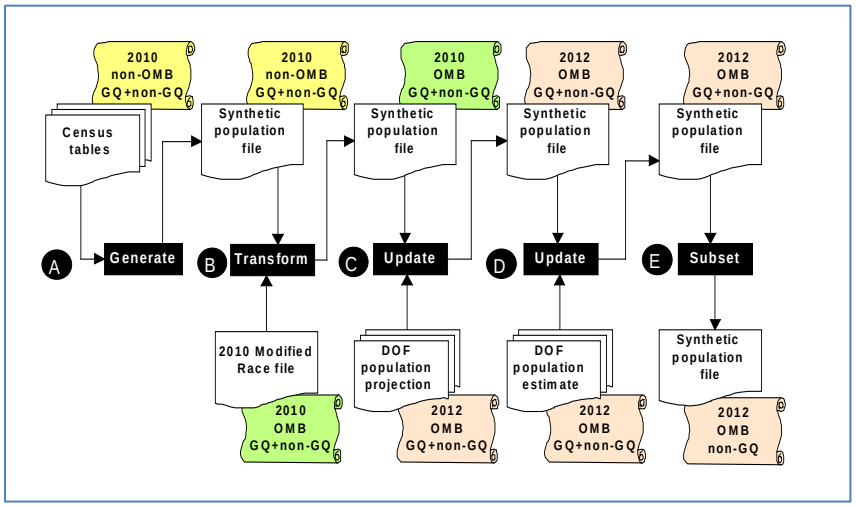


Figure 3: Separating the population in group quarters using tables D, E, and F.

[9] The raking process is more complex than what is described here because the totals that include the race category "Other" do not exist in the file with the OMB definition. That is, they would be raked to a zero total. Although raking can sometimes handle this situation, it is not recommended because the algorithm may not converge. In practice these cells are set to and then remove from file before raking.

[10] A formal definition of the transformation function is beyond the scope of this paper.

There are 6 steps in creation of the control totals:

A. Using the auxiliary tables with data from the 2010 Decennial Census (SF1 files) and the method described in Section 2.2 we create a synthetic population file with the 2010 totals based on the non-OMB definition. This file includes the population living in households and in group quarters (GQ and non-GQ).
B. Using the 2010 Modified Race File and the method described in Section 3.2, transform the synthetic population in Step A to a file with totals based on the OMB definition.
C. Using the 2012 California DOF population projections (California Department of Finance, 2013a) and method described in Section 3.1, update the synthetic population file from Step B to reflect the 2012 population. Notice that the file created in Step B and the MR files have the same OMB definitions, so these raking factors are population growth factors for the period 2010 to 2012.
D. Using the 2012 California DOF population estimates update the synthetic population file from Step C. The California DOF releases more accurate population totals than those from the population projections. These are called population estimates (California Department of Finance, 2013b). These estimates are only available at the county level, and they are used to adjust the synthetic population file from D using poststratification (using raking software with one dimension). In this way, the synthetic population file reflects the latest population counts from the DOF.
E. Subset the synthetic population from Step D to remove the population group quarters. The final synthetic population file contains the 2012 population total, is based on an OMB race definition and excludes the population in group quarters.

The synthetic population file from Step E can be summarized by any of the categorical variables form the auxiliary tables in order to create the raking dimensions. Additional for information and the description of control total for Asian ethnic subgroups in CHIS 2011-2012 will be described in the CHIS 2011-2010 methodology reports located at the CHIS website at http://healthpolicy.ucla.edu/chis/design/Pages/methodology.aspx. It is recommended to verify that the assumptions made when these methods are applied hold before these totals are used.

5. Summary

This paper describes several issues that arise in the creation of totals before they can be used as benchmarks in weighting adjustments. The main problems are limited utility of data, outdated data, use of inconsistent definitions, and difficulties in incorporating more up-to-date releases. We presented several methods to address these problems. The preferred method is based on iterative proportional fitting which is the same procedure used to rake sampling weights in calibration. This approach is simple to implement and does not require specialized software. The implicit model for the produced totals is known, and has good statistical properties. The method can be modified so more than one contingency table is created, and it expands the current methodology limited to 2x2 tables. This is done by interleaving the categorical levels in one single file. The method produces a synthetic population file with totals for many auxiliary tables. Since the control totals are produced summarizing a single file, these control totals are consistent. As in any other statistical procedure, these totals reflect implicit and explicit assumptions made during the estimation process. These assumptions should be examined to ensure

sure they are reasonable before the totals are used. Further research is needed to address problems arising from departures from the assumptions. The same methodology can be used to generate complex synthetic populations for other applications.

Acknowledgements

I would like to thank Dr. J. Michael Brick for his support and comments on this work.

References

California Department of Finance (2013a). Population Projections, 2010-2060. Retrieved September 9, 2013 from http://www.dof.ca.gov/research/demographic/reports/projections/view.php.

California Department of Finance (2013b). E-1 Population Estimates for Cities, Counties, and the State — January 1, 2012 and 2013. Retrieved September 9, 2013 from http://www.dof.ca.gov/research/demographic/reports/estimates/e-1/view.php.

California Health Interview Survey (2011). CHIS 2009 Methodology Series: Report 5 – Weighting and Variance Estimation Methodology Series: Los Angeles, California: UCLA Center for Health Policy Research. Retrieved from http:// http://healthpolicy.ucla.edu/chis/design/Pages/methodology.aspx.

Deming, W. E.; and Stephan, F. F. (1940). On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals are Known. *Annals of Mathematical Statistics*. 11 (4): 427–444.

Deville, J.-C., Särndal, C. E., and Sautory O. (1993), Generalized Raking Procedures in Survey Sampling. *Journal of the American Statistical Association*, Vol. 88, No. 423, 1013-1020.

Deville, J.-C., and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

Izrael, D., Hoaglin, D., and Battaglia, M. (2000). A SAS Macro for Balancing a Weighted Sample. *The Twenty-Fifth Annual SAS User's Group International Conference*, Cary, NC: SAS Institute, Inc.

Kalton, G., and Flores Cervantes I. (2003). Weighting Methods. *Journal of Official Statistics*. 19:81-97.

Lumley, T. (2012) survey: analysis of complex survey samples. R package version 3.28-2.

Office of Management and Budget (2003). Revisions to the Standards for the Classification of Federal Data on Race and Ethnicity. Retrieved from http:// http://www.whitehouse.gov/omb/fedreg_1997standards/.

Särndal, C-E, and Lundström, S. (2005). *Estimation in Surveys with Nonresponse*. New York: Wiley.

Sautory, O. (1993). A macro Calmar. Redressement d'un échantillon par calage sur marges. L Document No. F9310, Institut National de la Statistique et des Études Économiques, Paris.

Stephan, F. F. (1942). Iterative method of adjusting frequency tables when expected margins are known. *Annals of Mathematical Statistics* 13 (2): 166–178.

U.S. Census Bureau. (2011). 2010 Census Summary File 1. 2010 Census of population and Housing, Retrieved September 9, 2013 from www.census.gov/prod/cen2010/doc/sf1.pdf.

U.S. Census Bureau. (2012). 2010 Modified Race Data Summary File. Retrieved September 9, 2013 from http://www.census.gov/popest/research/modified.html