Improving Population Estimates Through Raking by Joint Distribution of Multiple Grouping Variables

Yilan Huang

Joint work with Dr. Honghu Liu

Department of Biostatistics
University of California, Los Angeles

yilanh19@ucla.edu

JSM 2024, Portland, Oregon

Table of Contents

Motivation

- Weighting adjustments are commonly applied in surveys to compensate for nonresponse and noncoverage.
- Typically, each observation is assigned a sampling weight that is equal to the reciprocal of the probability of selection.
- The sampling weights can be further adjusted to compensate for the fact that persons with certain characteristics are not as likely to respond to the survey.
- Raking is a post-stratification procedure to match marginal distributions of a survey sample to known population margins on a specified set of variables.

Motivation

- Raking offers a distinct advantage in that only the marginal control totals of each auxiliary variable are needed.
- When the raking margins are highly correlated, the raking process may take a large number of iterations to converge.
- One possible solution is to combine multiple variables to form a single dimension, when the joint distribution with respect to the auxiliary variables is available.
- The aim of our study is to compare the performance of the raking algorithm when using a single variable versus using multiple grouping variables for raking margins.

Basic Algorithm

- In a cross-classification that has J rows and K columns, we denote the sum of individual weights $w_i (i = 1, ..., n_{jk})$ in cell (j, k) by w_{jk}.
- The initial row totals and column totals of the sample weights are w_{j+} and w_{+k} respectively.
- Similarly, we denote the corresponding population control totals by T_{j+} and T_{+k}.

K columns

J rows		Female	Male	
	African American			
	Asian	(j,k) cell		w_{j+}
	Hispanic			
	White			
	Other			
		w_{+k}		

Figure: Example of the basic raking algorithm.

Basic Algorithm

- The iterative raking algorithm produces modified weights.
- In the two-variable cross-classification, we use $m_{jk}^{(1)}$ for the sum of the modified weights in cell (j, k) at the end of step 1.
- If we begin by matching the control totals for the rows, T_{j+}, the initial steps of the algorithm are

$$
\begin{aligned}
m_{jk}^{(0)} &= w_{jk} && (j = 1, ..., J; k = 1, ..., K) \\
m_{jk}^{(1)} &= m_{jk}^{(0)}(T_{j+}/m_{j+}^{(0)}) && (\text{for each } k \text{ within each } j) \\
m_{jk}^{(2)} &= m_{jk}^{(1)}(T_{+k}/m_{+k}^{(1)}) && (\text{for each } j \text{ within each } k)
\end{aligned}
$$

- The adjustment factors are actually applied to the individual weights in the cell (j, k).

Basic Algorithm

- In the iterative process, an iteration rakes both rows and columns. For iteration $s(s = 0, 1, ...)$, we may write

$$m_{jk}^{(2s+1)} = m_{jk}^{(2s)}(T_{j+}/m_{j+}^{(2s)})$$
$$m_{jk}^{(2s+2)} = m_{jk}^{(2s+1)}(T_{+k}/m_{+k}^{(2s+1)})$$

- The raking algorithm proceeds by proportionately scaling the m_{jk} such that the relations are satisfied in turn

$$\sum_{k}^{K} m_{jk} = m_{j+} = T_{j+}, \sum_{j}^{J} m_{jk} = m_{+k} = T_{+k}$$

- The process terminates either after a fixed number of iterations or when each marginal total of the raked weights is within a specified tolerance of the corresponding population control total.

Comparison between Raking Estimators

- We derive estimators with the two raking approaches separately, for a simple case of a 2 x 2 cross-classification.
- Scenario 1: we rake on the two auxiliary variables sequentially.
- We denote $m_{jk}^{(2)}$ as the sum of the modified weights in cell (j, k) at the end of step 2:

$$m_{11}^{(2)} = (m_{11}^{(0)} \times \frac{T_{1+}}{m_{1+}^{(0)}}) \times \frac{T_{+1}}{(m_{11}^{(0)} \times \frac{T_{1+}}{m_{1+}^{(0)}}) + (m_{21}^{(0)} \times \frac{T_{2+}}{m_{2+}^{(0)}})} \quad (1)$$

- Scenario 2: assuming that the population totals for the cross-classification T_{jk} are known, we combine the two auxiliary variables to form a single margin.
- We denote $u_{jk}^{(2)}$ as the sum of the modified weights in cell (j, k) at the end of step 2.

Comparison between Raking Estimators

- Assume that for a general survey characteristic Y_{jki}, we are interested in estimating the population total Y with the statistic

$$\tilde{Y} = \sum_{j}^{J} \sum_{k}^{K} \frac{m_{jk}}{n_{jk}} (\sum_{i}^{n_{jk}} y_{jki}) \tag{2}$$

- Let $\bar{y}_{jk}$ be the sample mean for cell (j, k). We have

$$\tilde{Y}_2 - \tilde{Y}_1 = \sum_{j=1}^{2} \sum_{k=1}^{2} \{ \frac{u_{jk}^{(2)}}{n_{jk}} (\sum_{i}^{n_{jk}} y_{jki}) - \frac{m_{jk}^{(2)}}{n_{jk}} (\sum_{i}^{n_{jk}} y_{jki}) \} \tag{3}$$

$$= \sum_{j=1}^{2} \sum_{k=1}^{2} (T_{jk} - m_{jk}^{(2)}) \bar{y}_{jk} \tag{4}$$

- The difference between the two raking estimators depends on population totals, initial sampling weights, and the individual-level survey outcome.

Properties Conditioning on Sample

- Assume that for a general survey characteristic Y_{jki}, in a simple random sampling setting, we are interested in estimating the population total Y with the statistic

$$\tilde{Y} = \sum_{j}^{J} \sum_{k}^{K} \frac{m_{jk}}{n_{jk}} (\sum_{i}^{n_{jk}} y_{jki}) \tag{5}$$

- Let $\bar{Y}_{jk}$ be the population mean for cell (j, k). Given sample counts $\underline{n} = (n_{11}, n_{12}, ..., n_{JK})$, the conditional expected value of $\tilde{Y}$ is

$$E(\tilde{Y}|\underline{n}) = \sum_{j}^{J} \sum_{k}^{K} m_{jk} \bar{Y}_{jk} = Y + \sum_{j}^{J} \sum_{k}^{K} (\bar{Y}_{jk} - \bar{Y})(m_{jk} - N_{jk}). \tag{6}$$

Properties Conditioning on Sample

- Applying the equation

$$\bar{Y}_{jk} - \bar{Y} = (\bar{Y}_j - \bar{Y}) + (\bar{Y}_k - \bar{Y}) + (\bar{Y}_{jk} - \bar{Y}_j - \bar{Y}_k + \bar{Y}), \quad (7)$$

  the conditional bias given $\underline{n}$ can be written as

$$Bias(\tilde{Y}|\underline{n}) = \sum_j^J (\bar{Y}_j - \bar{Y})(m_{j+} - T_{j+}) + \sum_k^K (\bar{Y}_k - \bar{Y})(m_{+k} - T_{+k}) + \sum_j^J \sum_k^K (\bar{Y}_{jk} - \bar{Y}_j - \bar{Y}_k + \bar{Y})(m_{jk} - T_{jk}).$$

- When the iterative raking algorithm converges, the first two terms of the conditional bias should be approximately zero, leaving the third to be the driving term.
- For the third term to be zero, it is sufficient that the Y_{jk} be such that there is no interaction.

Comparison through LA County Smile Survey

- *Los Angeles County Smile Survey* screened kindergarten and third-grade children at a representative sample of public elementary schools during the 2018–2019 and 2019–2020 school years.
- The oral health measures collected included the number of teeth with untreated decay, the number of teeth with treated decay, the status of each permanent first molar, and the urgency of need for dental care.
- In this empirical study, we focus on data collected from the 4,546 third-grade children.
- Sampling weights are determined by the inverse probability of selection, based on the sampling design.

Comparison through LA County Smile Survey

- Raking is used to generate weights to ensure that the survey totals matched the known population totals in terms of the school district (LAUSD vs. non-LAUSD), gender, race/ethnicity, and socioeconomic status (SES).
- The selection criteria for these raking variables include that:
  - control totals of these variables add up to the same population total
  - usually no missing categories
  - the raking variables are associated with oral health outcomes, participant coverage, and response rates
- We compare weights and estimations across three weighting schemes:
  - Sampling weights
  - Raking 1: LAUSD + SES + Race/Ethnicity + Gender
  - Raking 2: LAUSD x SES x Race/Ethnicity + Gender

Comparison through LA County Smile Survey

Table 1. Compare outcomes using sampling weights and raking weights

Oral health related outcomes	Weighting method	Mean or %	Standard error
Untreated Decay (%)	Sampling weights	20.889	1.200
	Raking 1	20.493	1.042
	Raking 2	20.337	1.033
Untreated Decay (number)	Sampling weights	0.4217	0.0287
	Raking 1	0.4099	0.0236
	Raking 2	0.4062	0.0233
Caries Experience (%)	Sampling weights	66.391	2.021
	Raking 1	64.637	1.121
	Raking 2	64.417	1.119
Caries Experience (number)	Sampling weights	3.1487	0.1424
	Raking 1	3.0365	0.0819
	Raking 2	3.0224	0.0796
Needs Urgent Care (%)	Sampling weights	20.880	0.286
	Raking 1	20.356	0.256
	Raking 2	19.979	0.253

Comparison through Simulation Study

- In the case study analysis, it is not possible to evaluate bias due to the lack of a gold standard or any known truth about the outcome measures of interest in the population.
- The simulation study aims to measure the magnitude of differences in terms of empirical bias, standard error, and execution time.
- We evaluate the estimates for population prevalences and means for oral health outcome variables.
- We assume that the outcome variable model contains three main effect covariates: school district, SES (yes and no) and race/ethnicity (Hispanic and non-Hispanic).
- All statistically significant two-way interaction terms between the three main effect variables are included in the outcome model.

Comparison through Simulation Study

Table 2. Empirical results for the three weighting schemes over repeated sampling (10,000 simulation samples)

Oral health related outcomes	Population or weighting method	Mean or %	Standard error	Time (secs)	Relative bias ($x10^{-2}$)	Relative SE ($x10^{-3}$)
Caries Experience (%)	Population	65.279				
	Sampling weights	66.00122	0.3325516		1.10636	5.0943
	Raking 1	65.28097	0.1603526	0.0390043	0.00302	2.4564
	Raking 2	65.28055	0.1590280	0.0142241	0.00237	2.4361
Caries Experience (number)	Population	2.52016				
	Sampling weights	2.539091	0.0135787		0.75118	5.3880
	Raking 1	2.520230	0.0113744	0.0434305	0.00278	4.5134
	Raking 2	2.520206	0.0112566	0.0157441	0.00183	4.4666

Summary

- We compared the performance of the raking algorithm when using a single variable versus using multiple grouping variables for raking margins.
- The raking algorithm has the potential to reduce both average bias and average standard error in estimates compared with sampling weights.
- Furthermore, raking through cross-classification, especially when there is an interaction effect between auxiliary variables, can improve the performance of an estimator, based on joint distributions available at the population level.

References

- Kalton G, Flores-Cervantes I. Weighting methods. Journal of official statistics. 2003 Jun 1;19(2):81.
- Bell BA, Onwuegbuzie AJ, Ferron JM, Jiao QG, Hibbard ST, Kromrey JD. Use of design effects and sample weights in complex health survey data: a review of published articles using data from 3 commonly used adolescent health surveys. American journal of public health. 2012 Jul;102(7):1399-405.
- Battaglia MP, Izrael D, Hoaglin DC, Frankel MR: Practical considerations in raking survey data. Survey Practice 2009, 2(5):1-10.
- Battaglia MP, Izrael D, Hoaglin DC, Frankel MR. Tips and tricks for raking survey data (aka sample balancing). Abt Associates. 2004 May 11;1:4740-4.
- Brick JM, Montaquila J, Roth S. Identifying problems with raking estimators. Inannual meeting of the American Statistical Association, San Francisco, CA 2003.
- Oh HL, Scheuren FJ. Modified raking ratio estimation. Survey Methodology. 1987 Dec;13(2):209-19.

References

- Lumley T. Package 'survey'. Accessed on 2024-04-08. Available from: https://cran.r-project.org/web/packages/survey/index.html.
- Chen TC, Clark J, Riddles MK, Mohadjer LK, Fakhouri THI. National Health and Nutrition Examination Survey, 2015-2018: Sample design and estimation procedures. National Center for Health Statistics. Vital Health Stat 2(184).
- Oral Health | CDC; 2024. Accessed on 2024-04-08. Available from: https://www.cdc.gov/oralhealth/index.html.
- Oral Health Surveillance Report, 2019; 2021. Accessed on 2024-04-08. Available from: https://www.cdc.gov/oralhealth/publications/OHSR-2019-index.html.
- Smile Survey 2020: The Oral Health of Los Angeles County's Children. Oral Health Program, Los Angeles County Department of Public Health; 2020.

Thank You!

Comparison through Simulation Study

The implementation of the simulation is based on the framework and involves the following steps:

1. Generate an artificial finite population of size N that contains 8 subpopulations defined by the categories of the three auxiliary variables (2 x 2 x 2). The subpopulation size N_{ij} is determined based on the population distribution of LA County third-grade students.
2. Generate the value for the outcome variable Y. The outcome model parameters are determined based on the LA County Smile Survey third-grade data.
3. Select a stratified random sample of size n from the population, with school districts as the strata.
4. Conduct survey weighting using sampling weights, raking with single-variable margins, and raking with joint distribution (school district x SES x race/ethnicity), respectively. Obtain the estimates for the outcome variables and then compare the empirical results.

Comparison through Simulation Study

- The model for the outcome variable Y is specified as follows. For a binary response:

$$P(Y_{ijk} = 1) = \text{logit}^{-1}(\mu + \alpha_i + \beta_j + \gamma_s + \alpha_i\beta_j + \alpha_i\gamma_s + \beta_j\gamma_s + \epsilon_{ijk})$$
$$i = 1, 2; j = 1, 2; s = 1, 2; k = 1, ..., N_{ij}$$

- where N_{ij} is the population size in cell (i, j) for the survey, and $\epsilon_{ijk} \sim N(0, \sigma^2)$.
- In the model, α measures the main effect of SES, β measures the main effect of race/ethnicity, and γ measures the main effect of school district.
- For a count response, we employ a Poisson regression model with a similar covariate structure.