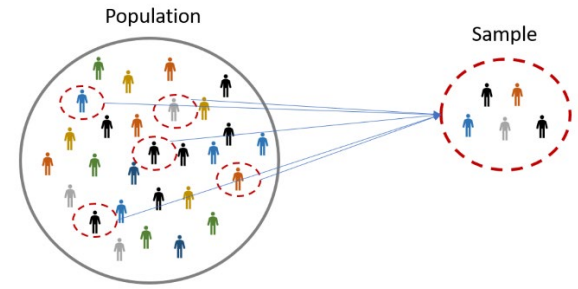
Raking and Weighting

Basics of Raking

Survey data is a collection of information on a sample that is done in a systematic way. The collected data are meant to be representative of a target population, but this does not always happen as intended. Differences can occur due to non-responsive subjects, sampling fluctuations, and survey design issues. One way to improve this relationship is by estimating sampling weights and modifying them using a procedure called raking. Raking adjusts these weights for a survey sample to correspond to the target population. The term raking is in reference to a rake working the soil in multiple directions, alternating until it is smooth. This procedure rakes sampling weights using a set of characteristic variables so that the sample totals agree with the population totals. This agreement is achieved through convergence using an iterative process. [2] Raking will also help reduce sampling and non-response biases.



[10]

Raking Example: Traumatic Brain Injury Model Systems

An applied example using this raking procedure was performed using the Traumatic Brain Injury Model Systems (TBIMS) national database (NDB) [1]. The TBIMS consists of multiple rehabilitation centers that provide data on patients with a moderate-to-severe traumatic brain injury (TBI). The TBIMS-NDB sample intends to make inferences to the national population admitted to inpatient rehabilitation for TBI. Using data provided by The American Medical Rehabilitation Providers Association's eRehabData (eRehab) and the Uniform Data System for Medical Rehabilitation (UDS), the cases in the national TBI population were compared to those in the TBIMS-NDB sample. The TBIMS patients were found to be younger, more likely to be single, and have shorter lengths of stay in rehab than the national TBI population.[3, 9] We then used a raking procedure to estimate sampling weights to allow their sample to be more representative of the national TBI population. An additional example is provided in appendences A, B, and C.

SAS Raking Macro

The raking procedure can be executed through multiple statistical programs, but this paper will focus on the 'rakinge' macro in SAS. This macro uses both a sample data set and the characteristic data from a population. The macro creates weights for the sample using a set of categorical variables found within both data sets. In a two-variable example, the macro arranges the variables into a grid with the categories of one comprising the rows, and the second for the columns. Each cell of this grid represents the initial sampling weights for subjects that fall into the categories of both variables. The macro starts by using the marginal totals for each row and multiplying each cell sampling weight by the ratio of the population totals for that respective category. These new modified cell sampling weights have been raked by the

population row variable. However, these sampling weights may not line up with the column totals. So the macro then uses the marginal totals for each column and multiplies each cell sampling weight by the ratio of the population totals for these column categories. The process continues, alternating between the rows and the columns, until the modified cell sampling weights agree for both variable totals. This continuation works through a specified number of iterations, or until the macro converges. The raking procedure is similar when you add more variables. Each entry sampling weight is raked by each variable until the end totals agree for all variables included in the macro. [4, 6]

Using the same two-variable example, the macro's algorithm constructs a grid with J rows and K columns. Each cell value has a sampling weight that is designated as w_{jk} for cell (j,k). The total sample weights for each row and column are then represented as w_{j+} and w_{+k}, respectively. The population totals for the rows and columns will be denoted with T_{j+} and T_{+k}. For each step, the modified sampling weight for cell (j,k) will be known as m_{jk}. One iteration for this algorithm will include a single rake across the rows and columns. A single iteration looks like [4, 6]:

$$m_{jk}{}^{(0)} = w_{jk} \qquad (j = 1,2,\dots,J;\ k = 1,2,\dots,K)$$

$$m_{jk}{}^{(1)} = w_{jk}{}^{(0)} * (T_{j+}/m_{j+}{}^{(0)}) \qquad \text{(For each } k \text{ within each } j\text{)}$$

$$m_{jk}{}^{(2)} = w_{jk}{}^{(1)} * (T_{+k}/m_{+k}{}^{(1)}) \qquad \text{(For each } j \text{ within each } k\text{)}$$

With each iterative step labeled by *s*, the overall iterative process looks like [4, 6]:

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

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

TBIMS-NDB Example

This raking procedure can be illustrated using TBIMS data. From 2001 to 2007, the TBIMS-NDB had a sample size of 857 patients and the comparison national TBI population consisted of 20,145 patients. Both data sets shared 10 demographic characteristics. Using sex and race/ethnicity, we can demonstrate the two-variable example. Sex was categorized into males and females and race/ethnicity was grouped into white, black, Hispanic, and other. The table below illustrates the grid with rows consisting of the sex categories and the columns consisting of the race/ethnicity categories. Each cell represents the TBIMS sample size for each group. [3, 9]

		Race / Ethnicity			
		White	Black	Hispanic	Other
Sex	Male	429	113	72	28
	Female	165	25	20	5

The first step of this raking procedure is to use the ratio of the population totals for the males and females and multiply each cell for the respective rows. If these new modified sample totals don't agree with the population ratios for each race/ethnicity categories, then the second step is to multiply each race/ethnicity column by the ratio of population totals for those respective categories. This process continues until the

modified sample totals in every cell agree with the population ratios for both variables. The table below displays the sex and race/ethnicity distributions for the national TBI population.

National TBI Population

	%
Gender	
Male	64.37%
Female	35.63%
Race/Ethnicity	
White	78.97%
Black	8.72%
Hispanic	7.31%
Other	5.00%

Using the SAS Raking Macro

The 'rakinge' SAS macro was provided by Izrael, et al. and was used for the raking procedure. This macro was made available from SAS SUGI 25. [4, 7] Using the macro required inputting the appropriate set of data and variables to the macro options. Both the sample data set and the population characteristic data will be needed. The set of variables that will be used in the raking process must be categorical and formatted in the same way for both data sets. If there are subjects with missing values for these raking variable in the sample data set, then an imputation method should be used to get complete data. One potential next step is to subset both data sets by time or location. This will help account for any temporal or spatial variation between the sample and population. With the TBIMS example, both data sets were subset by year of admission to rehabilitation.

For the population characteristic data, frequency tables need to be created for each raking variable. These frequency tables must include a column of category names and a column of percentages. The column of category names should be numbered instead of using the actual group names and the column of percentages should be inputted as a decimal. For example, 23% will be 0.23. It is also important that the column title for the percentages to be named 'PERCENT' in order for the SAS macro to run properly. Once all the data management is complete, the SAS macro options can be filled out.

The required SAS options include *inds*, *outds*, *inwt*, *outwt*, *freqlist*, *varlist*, *numvar*, *cntotal*, *trmprec*, *numiter*, and *prdiag*. *Inds* is the name of the input data set or the sample data set. *Outds* is the name of the output data set produced from the macro. *Inwt* is the variable name of the initial sampling weights to be adjusted. If there are no initial sampling weights, then input the number 1. *Outwt* is the variable name of the adjusted sampling weights computed from the macro. *Freqlist* is the list of all the frequency table names created for each raking variable. *Varlist* is the list of column titles for the column of category names for each frequency table listed in the *freqlist* option. *Numvar* is the number of raking variables. This number should be equal to the number of frequency tables. *Cntotal* is the control total or the sample size of the population data set being weighted to. This number should be multiplied by 100 since the raking variable category percentages are decimals. So if the sample size is 234, the *cntotal* will be 23400. *Trmprec* is the numeric tolerance level for the algorithm to reach convergence. *Numiter* is the maximum number of iterations for the macro to run. *Prdiag* is a yes/no statement for whether the macro prints out a diagnostic report. [4, 7]

Raking and Weighting

```
%rakinge(inds = ,          /* input data set */
         outds = ,         /* output data set */
         inwt = ,          /* weight being adjusted, if there is no weight, 1 is assigned */
         freqlist = ,      /* list of data sets with marginal control totals or percent */
         outwt = ,         /* resulting weight */
         byvar = ,         /* BY variable */
         varlist = ,       /* list of raking variables */
         numvar = ,        /* number of raking variables */
         cntotal = ,       /* general control total */
         trmprec = 1,      /* termination criterion, 1 default */
         trmpct = ,        /* termination based on marginal percent */
         numiter = 50,     /* number of iterations, default 50 */
         prdiag = Y);      /* print detailed diagnostics */
```

Since this raking macro is an iterative process, convergence issues can arise. This could be due to small sample sizes or the complexity of the raking variables. When the input data set has less than 200 subjects, this may result in longer convergence times. If the data has been subset, this issue can potentially be resolved by merging data sets together. Using the TBIMS example, consecutive admission year subsets can be merged together. Raking variables with a large number of categories can also potentially lead to longer convergence times. Collapsing categories should reduce computation time. Some other solutions are to increase the maximum number of iterations for the macro or to increase the tolerance level.

Once the macro has successfully converged, sampling weights will have been created for each sample subject in the output data set. If data subsets were created, then the SAS macro should be run for all data subsets. Once all subsets have successfully converged, they can be combined into a single data set containing all sampling weights.

Adjusting Sampling Weights for Non-Response

For some analyses, the sampling weights need to be adjusted for non-response. Non-response occurs when subjects have missing values for a certain outcome potentially due to sampling issues, health issues, or other systematic factors. These non-responders create bias within the sampling weights. The newly adjusted sampling weights help reduce this bias and allow the analytic sample to be more representative of the target population. This adjustment to the sampling weights is done using the same SAS macro with some additional data management steps. The first step is to create a binary response variable for the outcome of interest that indicates which subjects possess the outcome and which do not. If there are missing values for any model covariates or the original raking variables, then the data can be subset down to only include subjects with complete data, or an imputation method can be used to ensure complete data. The option to use only subjects with complete data can, however, result in additional bias. Next, a logistic model is run using the binary response variable created earlier as the outcome to determine the propensity of a subject responding to the outcome of interest. This logistic model includes all the raking variables and analytic model covariates. The resulting propensity scores from this logistic model are then binned into quintiles and a new categorical variable is created. Finally, a frequency table is created for this categorical propensity score variable, adjusting for the initial sampling weights. If the analytic data has been subset, then this propensity score model and resulting frequency table must be created for each subset.

Raking and Weighting

After the propensity score analysis, the data will be subset by the binary responder variable. The SAS macro will be run using only the data set for the responders as the input data set. The frequency table created for the categorical propensity score variable should be included in the *freqlist* and *varlist,* along with the other raking variables. The *inwt* should be the initial sampling weights created previously. After the macro converges, the newly-outputted data set should be merged with the original sample data set that includes both responders and non-responders for a complete analytic data set. Every responder in this data set will have two weights, their initial sampling weight and their new weight adjusted for non-response, while non-responders will only have their initial sampling weight. Similarly to before, if the analytic data set is subset by time or location, then this process needs to be run for each subset and combined into one final complete data set.

Weight Trimming

Weight trimming is a technique that can be used to truncate high value sampling weights. The goal is to reduce some of the variability and bias created by these large sampling weights. Although there is no set rule for where to trim the sampling weights, some common suggestions include five times the mean sampling weight or the median sampling weight plus six times the interquartile range of the sampling weights. Weight trimming is not a required step for the raking process but a valuable analytic tool. [8]

Conclusion

The procedure of raking and weighting survey data has been shown to improve the sample's overall representativeness to its target population. Utilizing the iterative process provided by the 'rakinge' SAS macro, along with adjusting for non-response and weight trimming, helps researchers control for potential sampling and systematic issues that can arise during the survey data collection process.

References:

1. Traumatic Brain Injury Model Systems National Data and Statistical Center, "TBINDSC", https://www.tbindsc.org/.
2. Brick, J. M., Montaquila, J., & Roth, S. (2003). Identifying problems with raking estimators. In *annual meeting of the American Statistical Association, San Francisco, CA*.
3. Corrigan, J. D., Cuthbert, J. P., Whiteneck, G. G., Dijkers, M. P., Coronado, V., Heinemann, A. W., ... & Graham, J. E. (2012). Representativeness of the traumatic brain injury model systems national database. *The Journal of head trauma rehabilitation*, *27*(6), 391.
4. Izrael, D., Hoaglin, D. C., & Battaglia, M. P. (2000, April). A SAS macro for balancing a weighted sample. In *Proceedings of the twenty-fifth annual SAS users group international conference* (pp. 9-12). Cary (NC): SAS Institute.
5. Liu, H., Cella, D., Gershon, R., Shen, J., Morales, L. S., Riley, W., & Hays, R. D. (2010). Representativeness of the patient-reported outcomes measurement information system internet panel. Journal of clinical epidemiology, 63(11), 1169-1178.
6. Battaglia, M. P., Izrael, D., Hoaglin, D. C., & Frankel, M. R. (2004). Tips and tricks for raking survey data (aka sample balancing). Abt Associates, 1, 4740-4744.

Raking and Weighting

7. Izrael, D., Hoaglin, D. C., & Battaglia, M. P. (2004, May). To rake or not to rake is not the question anymore with the enhanced raking macro. In Proceedings of the Twenty-Ninth Annual SAS Users Group International Conference.
8. Battaglia, M. P., Izrael, D., Hoaglin, D. C., & Frankel, M. R. (2009). Practical considerations in raking survey data. *Survey Practice*, *2*(5), 1-10.
9. Cuthbert, J. P., Corrigan, J. D., Whiteneck, G. G., Harrison-Felix, C., Graham, J. E., Bell, J. M., & Coronado, V. G. (2012). Extension of the representativeness of the traumatic brain injury model systems national database: 2001 to 2010. *The Journal of head trauma rehabilitation*, *27*(6), E15.
10. Population and Sample. https://s3-eu-west-1.amazonaws.com/blog.omniconvert.com-media/blog/wp-content/uploads/2019/10/21150245/sample-size-definition.png

Raking and Weighting

Appendix A: Example Code and Simulated Data

In order to help guide readers through the use of this raking and weighting procedure, example SAS code was provided to illustrate the use the 'rakinge' SAS macro. A sample data set and a national data set were simulated for this example code using R statistical software and provided in Excel CSV (comma separated values) files. Using these simulated data sets, an individual will be able to use the SAS raking macro to create sampling weights and adjust those sampling weight for non-response.

File Name	File Type	Description
SAS Raking and Weighting Macro Example.sas	SAS file	Example SAS code
Sample and National Simulated Data.R	R script	Simulated R code
Sample.csv	Excel CSV	Sample Excel CSV file
National.csv	Excel CSV	National Excel CSV file
rakinge.sas	SAS macro	SAS Raking Macro

The simulated sample data set has a sample size of 1,000 subjects where 500 are designated by year one and 500 are year two. Four demographic variables (Age, Sex, Race/Ethnicity, and Length of Rehabilitation Stay (LOS Rehab)) were randomly generated to be used in the raking procedure. Age was assumed to be normally distributed, and values were generated for each subject using a mean of 40 years and a standard deviation (SD) of 15. Age was then categorized into seven groups (<29, 30-39, 40-49, 50-59, 60-69, 70-79, >79). Sex was randomly assigned using a binomial distribution with a probability of 50%. Race/Ethnicity was categorized into white, black, and other and randomly assigned using 65%, 30%, and 5% probabilities for each category respectively. LOS Rehab was also assumed to be normally distributed and values were generated using a mean of 20 days and a SD of 10. LOS Rehab was then categorized into four groups (<9, 10-19, 20-29, >29). A binary outcome variable was also generated using a probability of 75% for a subject responding to the outcome.

The simulated national data set has a sample size of 100,000 subjects with 50,000 in year one and the rest in year two. The same four demographic variables that were simulated in the sample data set were also randomly generated for each subject in the national data set using the same distributions. Age was generated using a mean of 55 years, a SD of 15, and then categorized into the same seven age groups. Sex was assigned using probability of 65% for males. Race/Ethnicity was assigned using probabilities of 75% for white, 15% for black, and 10% for others. LOS Rehab was generated using a mean of 10 days, a SD of 10, and then categorized into the same four LOS groups.

The changes in the means and probabilities between the two simulated data sets are designed to illustrate that the subjects in the national data are older on average, more likely male, more likely white in race, and spent less time in rehab than the subjects in the sample data. The SAS raking macro will generate weights for the sample subjects to account for these changes.

Raking and Weighting

Appendix B: Sample and National Data Variable Tables

Frequency counts and percentages for all variables within the sample and national simulated data sets (Sample.csv, National.csv) are provided below.

	Sample		National	
	N	**%**	***N***	**%**
Age Category				
< 29 years	217	27.7%	4200	4.2%
30 – 39 years	249	24.9%	10192	10.2%
40 – 49 years	255	25.5%	20117	20.1%
50 – 59 years	172	17.2%	26115	26.1%
60 – 69 years	80	8.0%	21872	21.9%
70 – 79 years	21	2.1%	11996	12.0%
> 79 years	6	0.6%	5508	5.51%
Sex				
Male	490	49.0%	65018	65.0%
Female	510	51.0%	34982	35.0%
Race / Ethnicity				
White	664	66.4%	74875	74.9%
Black	280	28.0%	15008	15.0%
Other	56	5.6%	10117	10.1%
Length of Rehabilitation Stay				
< 9 days	132	13.2%	45949	46.0%
10 – 19 days	313	31.3%	35704	35.7%
20 – 29 days	372	37.2%	15494	15.5%
> 29 days	183	18.3%	2853	2.9%
Outcome Response				
Responder	759	75.9%	-	-
Non-Responder	241	24.1%	-	-

Raking and Weighting

Appendix C: Sampling Weight Results

The resulting sampling weights and adjusted sampling weights created from the example SAS code (using the sample and national CSV files) are summarized below. Means, SDs, and the distribution of the weights are provided for the sampling weights and adjusted sampling weights for when the full sample data set is raked as a whole, and also for when the sample data set is raked separately by year 1 and year 2 data. A few large weights exist (>20) that will more than likely need to be trimmed.

Full Sample Data

	N	Mean (SD)	Min	25th Percentile	Median	75th Percentile	Max
Sampling Weights	1000	1.00 (2.26)	0.01	0.13	0.36	0.94	37.79
Adjusted Sampling Weights	759	1.32 (3.11)	0.02	0.19	0.51	1.31	49.65

Sample Data by Year 1 and Year 2

	N	Mean (SD)	Min	25th Percentile	Median	75th Percentile	Max
Sampling Weights	1000	1.00 (2.16)	0.00	0.09	0.28	0.94	25.89
Adjusted Sampling Weights	759	1.32 (2.93)	0.00	0.14	0.43	1.26	36.39

SD: Standard Deviation
Min: Minimum
Max: Maximum