Raking: An Important and Often Overlooked Survey Analysis Tool

LCDR Lew Anderson and Dr. Ronald D. Fricker, Jr.

INTRODUCTION

Within the U.S. military's global fight against indirect and asymmetric warfare, surveys have become an important tool for military planners and operations research (OR) analysts. Surveys help commanders and policy-makers gain valuable insight into an organization's performance and/or a population's attitudes and opinions toward various groups.

When the final product, often in the form of a huge data set with thousands of observations (respondents) and hundreds of variables (questions), arrives in a military analyst's inbox, it can be a daunting task to sift through and properly analyze the data. Furthermore, while the goal of every survey analysis is to estimate something about the population from a sample – where this inference depends on the sample being representative of the population – what if the sample demographics do not match known population demographics? Can an analyst honestly conclude that any results obtained from this survey data are representative of the entire population?

Raking (otherwise known as iterative proportional fitting, sample-balancing, or raking ratio estimation) is a method for adjusting the sampling weights of the sample data based on known population characteristics. By adjusting these weights, the survey sample is essentially forced to resemble the population, therefore making inference to the entire population, not just the sample, possible.

Correctly applied, raking can be a valuable survey analysis tool, allowing analysts to make the appropriate inference from sample to population. Since survey analysis is becoming such an important part of military OR, but tends not to be taught in OR curricula, this paper is intended to provide some insight into using the raking technique from a military OR analyst's point-of-view.

MODEL-BASED VS. DESIGN-BASED INFERENCE

There are two approaches for conducting statistical inference from a sample to an entire population. The most widely-used approach is *model-based inference*, where generalization from sample to population is based on a mathematical probability model. For example, the hypothesis tests taught in introductory statistics courses, such as t-tests and analysis of variance (ANOVA), are model-based; they assume the data, or the statistic of interest, follow a particular probability distribution and, typically, that the data are independent and identically distributed.

In contrast, *design-based inference*, which is used in survey analysis, follows from the idea that the data (people in this case) are selected according to a known and carefully controlled sampling scheme, thus allowing appropriate inference to the population. This is useful and important because it is often not reasonable to assume that survey responses are identically distributed across various population subsets, particularly when the data is collected according to a complex sampling methodology, which is almost always the case in any large, real-world survey.

COMPLEX SURVEY SAMPLING

To truly appreciate the utility of raking, one must first understand a bit about survey sampling, particularly complex sampling. In an ideal surveying world, data would be collected via simple random sampling (SRS) where model-based inferential analysis would be perfectly appropriate. However, real-world constraints, driven by numerous things like a limited survey budget and time, require the use of complex sampling. In using these sampling techniques, analytical complications arise due to the unequal selection probabilities of survey respondents, the occasional lack of independence between respondents, and the fact that in any real-world survey the population is finite.[1]

One commonly used complex sampling technique is stratified sampling, which involves sampling by population strata – uniquely identifiable, non-overlapping groups within the population – that can result in respondents having unequal selection probabilities. For example, important subgroups within the population, such as high-poverty neighborhoods or ethnic minorities, may be over-sampled. These subgroups then have a higher probability of being selected to take the survey and thus would be over represented in the sample.

A popular complex sampling technique used in face-to-face interviews is cluster sampling in which respondents are sampled in groups. With this technique, it is more efficient to transport an interviewer to one location where he or she interviews a number of individuals located in close geographic proximity. However, because of this geographic proximity, cluster sampling can, and often does, result in correlation between observations within groups. That is, people who live or work near each other tend to be more alike, and thus answer survey questions more similarly, than people chosen at random throughout the population.

Furthermore, in most large-scale, real-world surveys, samples are drawn using both stratification and clustering. For example, a population is first split into strata from which either single-stage or multi-stage cluster sampling is then conducted.

Each complex survey sampling design has its own advantages (and disadvantages) and, typically, the sampling approach for any given survey is uniquely designed to meet the objectives and constraints of that survey effort. For example, as compared to a SRS, a stratified sample can produce smaller margins of error. Additionally, if estimates of population parameters are desired for subgroups of the population, then these subgroups can be identified as strata, drawing down the cost of the survey (Lohr 1999).

The key take-away is that simple model-based inference is usually the wrong way to analyze survey data arising from complex sampling. For example, except in special circumstances, stratification will result in a sample that is not representative of the population, in the sense that the various subgroups will be over- or under-represented, and without accounting for this in the analysis, the resulting point estimates will be incorrect. Similarly, except for special circumstances, the use of cluster sampling will result in cor-

[1] The details of complex survey sampling fill books. Interested readers should consult one or more of the standard texts, including Lohr (1999), Cochran (1977), and Kish (1995) to name a few.

related survey responses that, if not appropriately accounted for in the analysis, will result in incorrectly estimated margins of error (and thus incorrect confidence intervals and hypothesis tests).

ANALYZING COMPLEX SURVEY DATA

Regardless of the sampling method used, a survey sample should be drawn from the population probabilistically, and the design-based sampling methodology must meet two specific properties:

1. Every individual (i) in the population must have a non-zero probability (p_i) of ending up in the sample. With SRS and some special complex sample designs, these probabilities are equal, but in general, they could be different for every sampled individual.
2. The probability (p_i) must be known for every individual in the sample.

Since the sample is unlikely to be representative of the entire population, *sampling weights* are used to account for these design-based inequalities. That is, each respondent in the final sample is given a sampling weight (w_i) that is the inverse of the sampling probability: $w_i = 1/p_i$. These weights, perhaps further adjusted to account for other aspects of the survey process (e.g., unequal response rates) are then used in the *Horvitz-Thompson estimator* to do the appropriate inference from sample to population.

For example, if one is interested in using survey data to estimate the (unobserved) population mean response to question j (μ_j) then the Horvitz-Thompson estimator for a sample size n is:

$$\hat{\mu}_j = \frac{\sum_{i=1}^{n} w_i x_{ij}}{\sum_{i=1}^{n} w_i}.$$

Intuitively, the weights are the number of individuals in the population that the respondent represents, where those with a small selection probability represent more of the population and those with larger selection probabilities represent fewer people. In the extreme case of $p_i = 1$, person i has a weight of $w_i = 1$ and thus only represents him or herself.

It is important to note that these weights and the Horvitz-Thompson estimator only provide for correct point estimation. Also required, though beyond the scope of this article, is the correct estimation of standard errors.[2]

A couple take-aways for the OR analyst faced with analyzing a large-scale survey. First, if the data were collected by a commercial company under contract, then the company should provide sampling weights in the analysis file. They should also provide documentation about how the weights were calculated.[3] Second, the weights must be used in the survey analysis, and this will likely require special software designed for survey analysis, particularly in order to calculate the correct standard errors. Third, and most

[2] For those interested, see Lohr (1999) for an introduction to estimating standard errors for complex sample designs.

[3] This presumes that the original contract required the survey company to provide this information – as it should. The contract should also specify that the survey company develop a sampling plan prior to fielding the survey and provide it to the Government for review and approval.

importantly, *any* survey using complex sampling needs sampling weights (provided or calculated) in order to conduct the correct analysis.

But what if, even with the provided weights, the survey estimates of the known population characteristics are off? Or, perhaps even worse, what if the weights are not provided? These are serious but real issues, both of which have been observed with surveys conducted under DoD contract. Fortunately there exist several established techniques, such as raking, to estimate or modify the sampling weights in order to ensure the survey estimates match known population characteristics.

THE RAKING ALGORITHM

Raking is a model-based approach using known population totals (usually from a census) that adjusts the sampling weights so the marginal values of a table sum to those known totals. The actual algorithm involves repeatedly estimating weights across each set of variables in turn until the weights converge and stop changing. Essentially, raking forces the survey totals to match the known population totals by assigning a weight to each respondent.

The most logical and popular survey variables to use with raking are demographics (i.e., gender, age, religion, etc.) for two reasons. First, population totals for demographics are often available, for example, from census data, and second, survey responses to these types of variables typically have a low non-response rate.

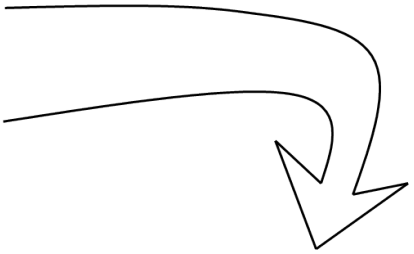
In order to converge, the population totals for each raking variable need to be the same. For example, one raking variable may be pulled from a 2008 census and another from a 2012 census. Since it is unlikely the 2008 population equals the 2012 population, it is necessary to determine and fix a population total and then ratio-adjust any variables that do not sum to the fixed population total.

Careful consideration also needs to be taken to ensure the raking variables in the survey sample and the known population are measured by the same criteria. For example, the survey sample may measure socio-economic status by tangible items around the respondent's household, such as the number of TV sets, whereas the known (census) socio-economic status comes directly from the respondent's household income.

Figure 1 is a hypothetical example of the raking algorithm applied to two demographic variables each with two categories, `Gender` (Female or Male) and `Locale` (Rural or Urban). Notice the survey totals for each category do not match the census totals for the population, which is the case for most complex surveys. Using raking, weights (shown in the lower right table of Figure 1 in parenthesis) are calculated so the weighted survey totals match the census totals. For example, since urban locations were over-sampled, individuals from these locations are given a weight smaller than one, while individuals in the rural locations, that were under-sampled, are given weights greater than one. Similarly, females were over-sampled, so their weights are smaller than their male counterparts who were under-sampled.

As shown, Figure 1 is a fairly easy example when there are only two variables with two categories, but now imagine raking on four to six variables with multiple categories. Thankfully there is software available that does this procedure in a matter of seconds.

	Female	Male	Survey Totals	Census Totals
Rural	20	17	37	57
Urban	39	24	63	43
Survey Totals	59	41		⇑
Census Totals	49	51	⇐	Target



	Female	Male	Survey Totals	Census Totals
Rural	20(1.2978722)	17(1.8262812)	57	57
Urban	39(0.5908348)	24(0.8313842)	43	43
Survey Totals	49	51		⇑
Census Totals	49	51	⇐	Target

Figure 1: Marginal distributions with survey totals being raked to match the census totals by giving each respondent a sampling weight. For example, each female from a rural locale has a sampling weight of 1.2978722.

AN ILLUSTRATIVE (BUT ACTUAL) APPLICATION

In the following example, data were obtained over a year and a half from a survey that was repeatedly fielded in the same region every six months. The purpose of the survey was to gain insight into population attitudes and opinions about various government, military organizations, and VEOs, as well as to assess the population's perceived quality of life and views of various programs. Overall, 500 respondents were sampled in each survey "wave" and respondents answered 144 questions, of which 15 were related to demographics.

For this illustrative analysis, data from two survey waves fielded before a particular event were combined (consisting of 1000 respondents – Survey I) and compared with the data from a survey wave fielded after the event (consisting of 500 respondents – Survey II). The goal was to assess whether and how population attitudes and opinions changed as a result of the event.

The survey was contracted out to a commercial entity and the following summarizes the information provided about the sampling scheme used to collect the data:

- The data was collected using complex survey sampling (multi-stage clustering);
- Sampling weights were provided by the survey company, but seem to inappropriately capture the variation in sampling probabilities (i.e., they seem too uniform);
- Certain geographic locations within the area of interest were purposely avoided; and,
- Adjacent locations from previous survey waves were used in subsequent survey waves.

There are some serious sampling issues mentioned here, particularly the last two. Since certain locations were purposely eliminated from the sampling scheme, the probability of selecting individuals in those locations is zero. And, since some locations were purposively selected (i.e., with probability 1), not every stage of the sampling process was random. Both these issues make traditional design-based survey analysis via the Horvitz-Thompson estimator problematic and, as shown below, they contributed to making the resulting samples unrepresentative of the population. Furthermore, the weights provided by the company did not correct this unrepresentativeness.

To demonstrate these issues, Table 1 shows chi-squared test *p*-values comparing the distribution of demographic characteristics between Survey I and Survey II data. The column labeled "Unweighted" is the results from chi-squared tests conducted on the raw data, while the "Weighted" column are the results appropriately using the survey company's weights in the chi-squared calculation (see Lohr, 1999, for discussion).

Table 1 shows that there are statistical differences at the 0.05 significance level for socio-economic status and personal identity between Survey I and II. Had the sampling been consistently done survey-to-survey, it's very likely (though not guaranteed) that the sample demographics would be stable between surveys. That is, there would be no statistically significant differences. More importantly, if the sampling weights were correctly calculated, they should have corrected the demographic differences between surveys as well. What Table I shows is that neither the raw data nor the (contractor) weighted data in at least one of the waves (if not all of them) are not representative in at least these two demographic dimensions.

Demographics	Unweighted pValue	Weighted pValue
Age	0.12110	0.11096
Locale	0.60073	0.61613
Socio-economic	0.00294	0.00288
Gender	1.00000	1.00000
Age Group	0.80292	0.79262
Marital Status	0.25264	0.24203
Education Level	0.79494	0.78393
Work Status	0.97935	0.98152
Occupation	0.31626	0.32301
Home Owner	0.07428	0.07456
Income	0.25629	0.24393
Religion	0.07076	0.06933
Ethnicity	0.63281	0.63348
Personal Identity	0.01127	0.01162

Table 1: Statistical differences between surveys using the chi-square test at the 0.05 significance level.

APPLYING THE RAKING TECHNIQUE

The first step to raking is finding known population totals. In this case, the known population totals came from a 2010 census, where five census demographics, shown in Table 2, were category for category comparable to and consistent with the survey demographics.

The under-sampled variable categories are highlighted in blue on the left side of table of Figure 2, and, as expected, show dramatic differences between the survey sample data and the census data. Here, the survey data under-represent single individuals, who are under 30 years old, in households with incomes over 10,000 a month, and from an urban locale. Similarly, the survey data over-represent older, married, poorer, and rural residents.

	Survey I	Survey II	Census
Locale			
Rural	0.54	0.56	0.47
Urban	0.46	0.44	0.53
Age Group			
Under 30	0.26	0.29	0.37
30-39	0.26	0.25	0.24
40-49	0.22	0.22	0.18
50-59	0.14	0.12	0.12
Over 60	0.12	0.12	0.09
Marital Status			
Married	0.78	0.75	0.51
Single	0.17	0.18	0.45
Widow	0.05	0.07	0.04
Ethnicity			
A	0.20	0.19	0.07
B	0.13	0.15	0.28
C	0.06	0.05	0.16
D	0.48	0.53	0.03
E	0.15	0.12	0.46
Income			
Below 4000	0.27	0.26	0.07
4001-5000	0.19	0.17	0.11
5001-10000	0.38	0.34	0.30
10001-20000	0.12	0.13	0.37
Over 20000	0.05	0.10	0.15

	Survey I	Survey II	Data
Locale			
Rural	0.49	0.44	0.47
Urban	0.51	0.56	0.53
Age Group			
Under 30	0.34	0.35	0.37
30-39	0.26	0.25	0.24
40-49	0.19	0.17	0.18
50-59	0.12	0.13	0.12
Over 60	0.09	0.09	0.09
Marital Status			
Married	0.48	0.51	0.51
Single	0.47	0.45	0.45
Widow	0.05	0.04	0.04
Ethnicity			
Ethnicity A	0.08	0.07	0.07
Ethnicity B	0.27	0.29	0.28
Ethnicity C	0.16	0.15	0.16
Ethnicity D	0.04	0.05	0.03
Ethnicity E	0.47	0.46	0.46
Income			
Below 4,000	0.07	0.07	0.07
4,001-5,000	0.11	0.10	0.11
5,001-10,000	0.31	0.30	0.30
10,001-20,000	0.37	0.37	0.37
Over 20,000	0.15	0.14	0.15

Figure 2: Unweighted Survey I and II demographics on the left and weighted demographics, after the implementation of the raking algorithm, are on the right. In the left table, the blue bars highlight survey demographics that significantly deviate from the population demographics. Observe how, after raking, the weighted survey sample demographics in the right table closely resemble the census demographics.

Using the `survey` package in `R` with the `calibrate` function, the census totals for each variable were used to calculate weights, where the procedure is performed separately for Surveys I and II. The result, as seen in the table on the right side of Figure 2, is the weighted data now match one another and the census data quite closely across each variable category. Likewise, performing the chi-square test on the demographics now produces no statistically significant differences between Surveys I and II. The result is that the weighted data are more representative of the population.

A key take-away is that, in some instances, the weighted survey data (with raked sampling weights) change the analysis results. For example, as shown in Figure 3, the raw (i.e., unweighted) data for the question “How do you rate your life now as compared to 12 months ago?” is statistically different be-

tween Surveys I and II, with almost 1 in 10 people shifting their view toward the "Worse Now" category. This seems to show that a not insubstantial portion of the population believe their lives are getting worse. However, the weighted data show no statistically significant difference between surveys, and people, for the most part, have the same view of their lives now as compared to 12 months ago.

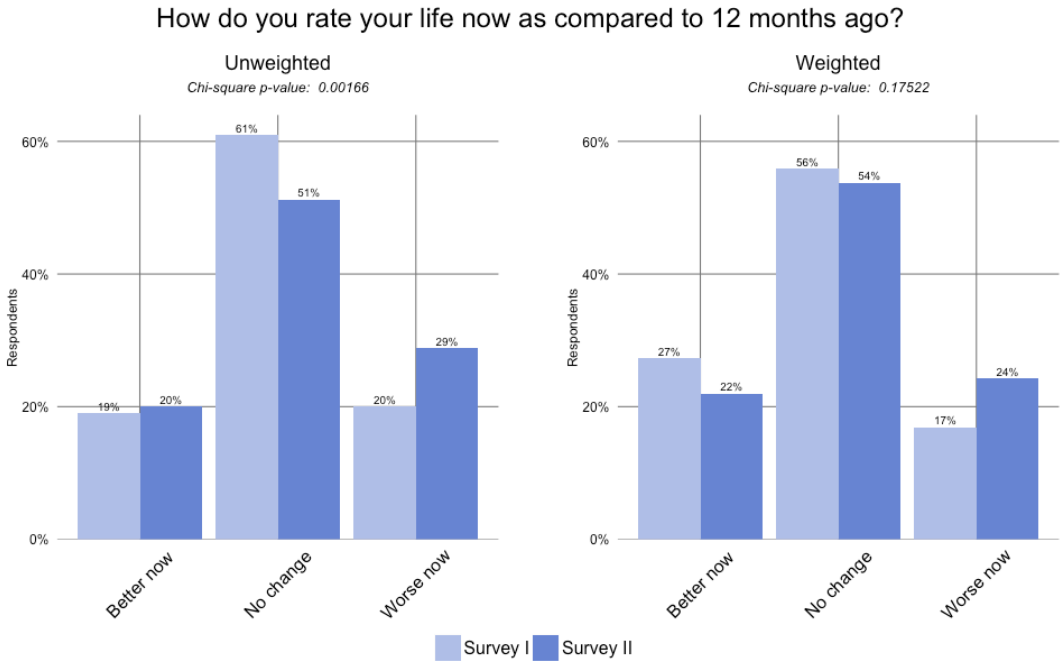


Figure 3: The unweighted data seem to show that respondents are reporting their lives are worse now compared to a year ago (left plot). However, the difference is not statistically significant for the weighted data (right plot).

To give some insight into how the sampling weights adjust the responses, this question is broken down in Figure 4 by the income demographic. Here, note that those with lower incomes are over-represented in the sample and are thus disproportionately affecting the results compared to the population. The weighting corrects for this imbalance in the sample.

Now, it may be that those with lower incomes *do* rate their lives as worse now than a year ago. And that may be a critical issue to address. But when doing the correct inference from sample to population, using the raked weights, there is no significant change *over the entire population*.

Furthermore, although not shown here, there are instances in this data where the weighted sampling design does not change the analytical outcome for other survey questions and the same conclusion would be drawn regardless of whether the sampling weights were used or not. Even so, the percentages of each category within the question will be different when using sampling weights in the analysis.

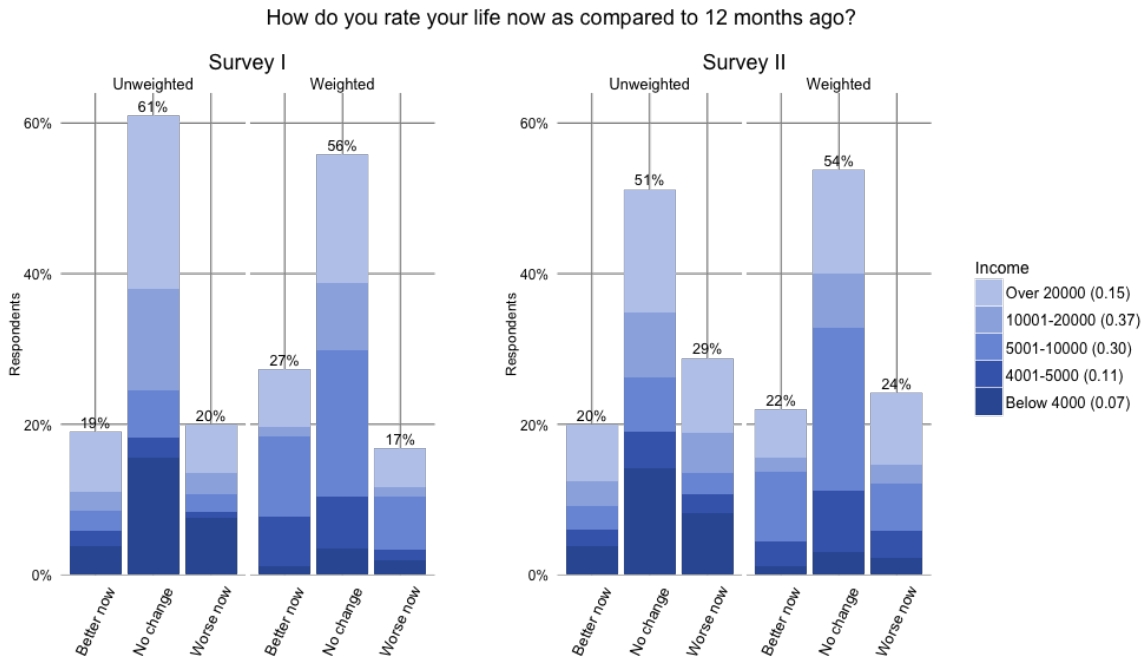


Figure 4: This stacked bar plot shows how respondents belonging to each household income category answered the question, "How do you rate your life now as compared to 12 months ago?"

SUMMARY AND CONCLUSIONS

Most surveys collect data using complex sampling designs and that design must be appropriately accounted for in any analysis of the data. Weights, which are one important part of a correct analysis, may not be provided with the survey data given to an analyst and, even if they are, they may not have been calculated correctly or the weighted data may not match known population quantities. In such cases, raking is a very useful tool with which the military OR analyst can calculate improved weights. For an in-depth discussion of the raking algorithm and its other applications, see Bishop et al. (1975).

REFERENCES

Lohr SL (1999) Sampling: Design and Analysis, Second Edition (Duxbury Press, Pacific Grove, California).

Lumley T (2010) Complex Surveys: A Guide to Analysis Using R (Wiley, New York, New York).

Cochran WT (1977) Sampling Techniques, Third Edition (Wiley, New York, New York).

DeBell M, Krosnick JA (2009) Computing weights for American national election study survey data. Retrieved March 26, 2014, www.electionstudies.org/resources/papers/nes012427.pdf.

Kish L (1995) Survey Sampling (Wiley, New York, New York).

Bishop Y, Fienberg S, Holland P (1975) Discrete Multivariate Analysis: Theory and Practice (MIT Press, Cambridge, Massachusetts).