Kolenikov, Stas. 2016. "Post-Stratification or Non-Response Adjustment?" *Survey Practice* 9 (3). https://doi.org/10.29115/SP-2016-0014.

ARTICLES

Post-stratification or non-response adjustment?

Stas Kolenikov*

Keywords: calibration, weighting, raking, non-response adjustment

https://doi.org/10.29115/SP-2016-0014

Survey Practice

Vol. 9, Issue 3, 2016

This paper considers the conceptual similarities and differences in weight adjustment steps known as nonresponse adjustment, post-stratification, and calibration. The distinction is based on the information requirements, i.e., whether the data necessary for the specific type of weight adjustment exist at the population level, at the level of the original sample that includes both respondents and nonrespondents, and/or at the respondent level. An illustrative example is provided where the different weights are constructed.

ELEVATOR VERSION

Post-stratification means that the weights are adjusted so that the weighted totals within mutually exclusive cells equal the known population totals. This term is misused most of the time, as in practice it is often attributed to weighting methods for which only margins of a multivariate contingency table are known, but not the higher order cells. These methods should be referred to as calibration, and the specific raking algorithm is usually used. While calibrated weights are based on achieving the alignment between the sample and the known population figures, non-response adjusted weights are based on achieving the alignment between the responding sample and the original sample.

A ROUND OF GOLF VERSION FOLLOWS

Before we move on to discuss the different types of weighting adjustments, let us check our gear to make sure we are on the same foot. The target population of a survey is a (hopefully, well-defined) group of social, economic, etc. units, such as persons, households, or businesses. It can be large (e.g., the decennial U.S. Census has a target population of everybody in the country) to pretty narrow. (National Immunization Survey has the target population of households with children 19–35 months of age.) The frame of the survey is a technical method by which the units from the target population are found and enrolled into the sample. The frame can be explicit, as is the case with list samples – e.g., schools in a district with known contact information; or implicit, as is the case with random digit dialing (RDD) samples, where no full list of phone numbers in the United States may exist, but the ways to generate phone numbers that are likely to lead to a real person are known (Brick and Tucker 2007). In some cases, an explicit frame may have to be

* Abt SRBI

constructed by the survey organization – e.g., by listing housing units in area face-to-face samples. The original sample (which is a somewhat nonstandard term) is the list of units to be approached with the survey request. This list must contain the unit identifier and contact information (which is often one and the same – e.g., the phone number in RDD surveys and the household address in area and mail surveys) and may contain additional information (e.g., a sample of hospital patients is likely taken from the hospital records and may have variables such as age, gender, length of stay and diagnosis). The responding sample (also a somewhat non-standard term) is the final list of units that are considered completed interviews, i.e., have a final disposition code of 1.1 (AAPOR 2015).

Now that we checked our clubs, let us move to the course. I usually do not like the term "post-stratification." It has a very specific technical meaning (Holt and Smith 1979), which is to say that you adjusted the weights so that the weighted totals agree with the known counts or proportions of nonoverlapping, mutually exclusive cells. (I prefer to think in terms of totals rather than proportions; as the sampling theory books explain (Thompson 1997), totals work better for certain technical reasons.) It is a close relative of stratification, and in some cases, the standard errors formulae that you can use for stratification also work for post-stratification. (Although these days, all these formulas are hidden behind the software packages interfaces.) Stratification works by splitting the population into non-overlapping groups (strata), with the purpose of drawing samples independently between strata. Stratification can utilize a single variable (example: area, address-based sampling [ABS] or RDD designs in which you can only stratify by geography, somewhat approximately on RDD) or several variables (example: establishment surveys, in which the strata are usually defined as a cross-classification of industry, defined by the first digits of the North American Industry Classification System (NAICS), geography, and categories of the establishment size, measured by revenue or employment).

The crucial feature of stratification is that it relies on the frame data available before sampling. So when somebody says that his or her general population survey was stratified on race, age, gender, education, and income, then the literal meaning is that they knew these characteristics for every unit on the frame before sampling. In the United States this is impossible to achieve in general population surveys, because no general population frame, like RDD or ABS, has this sort of data. (Although in some European countries that have population registers, this is entirely plausible; also, list samples of specialized populations, like beneficiaries in medical programs or students in a university, may have this sort of demographic variables available through the administrative channels, so the samples can be properly stratified on them.) If this was not the case, then the survey methods report should have stated something different – e.g., that the sample was post-stratified on race, age, gender, education, and income.

Unlike stratification, post-stratification relies on the data obtained in the survey itself that were not available before sampling, and adjusts the weights so that the totals in each group are equal to the known population totals. It still needs the post-stratification cells to be mutually exclusive and cover the whole population. The post-stratified weight for unit i in group g is:

$$w_{gi}^{\text{PS}} = w_{gi}^{\text{SRC}} \frac{\text{Population total for group g}}{\Sigma_{\text{responding } k \in g} w_{gk}^{SRC}} \quad (1)$$

The superscript PS stands for post-stratified; SRC, for source (which could be the base weight of the survey, or frame-integrated weight if multiple frames are being used, non-response weights as explained below, or some other intermediate weight that serves as the input to post-stratification). All weights within a group are increased proportionately so that the sum of post-stratified weights equals the known population total. If as a result of the random sampling error and/or non-response, a group becomes relatively underrepresented compared to the population, post-stratification aligns the representation of that group to match that of the population.

Like in stratification, the post-stratification cells are mutually exclusive, and cover the whole population. So when somebody says that his or her sample was post-stratified on race, age, gender, education and income (we already learned that we cannot stratify an RDD sample on these variables), it technically means that he or she obtained a five-way table of population counts and adjusted the weights in each cell. While I could buy somebody doing a two-way cross-classification of demographics, I think anything beyond that is a big, BIG stretch for most types of our data unless the sample sizes run well into thousands and tens of thousands.

So when somebody said that his or her weight was post-stratified by race, age, gender, education, and income, he or she probably mean that his or her weight was calibrated on these variables. Calibration (Deville and Sarndal 1992) means that the weights were made to agree with the known population totals for each margin, i.e., the weighted totals in the groups defined by race are equal to the known population totals; the weighted totals in the groups defined by gender are equal to the known population totals; etc. However, it is not guaranteed that the totals agree in finer cells, such as race-by-age or education-by-gender-by-age. They are likely to be close, but calibration does not attempt to perfectly align them with the population totals unless these interactions are themselves entered explicitly as calibration targets. The reasons why these interactions may not be practical are that these totals may not be known, to begin with; and that the sample sizes in the higher-order crossed cells may be small leading to unstable weights.

Specific implementations of calibration vary. European agencies like to use linear calibration, in which the adjustment factor is a linear combination of the calibration variables. (With a stretch of technical accuracy, we can say that the linearly calibrated weights are obtained as predictions in a linear regression where instead of equating the normal equations to zero, you equate them to the negative of discrepancy between the original weighted total for that variable and the known population total.) Most of the U.S. agencies rely on a form of iterative proportional fitting, or raking, in which each of the margins is adjusted at a time in turn. First, you adjust the margin of race, so that each of the weighted totals of race categories aligns with the known population total. (This is precisely post-stratification on race). Then you post-stratify on age, then on gender, then on education, then on income. By the time you are done with income, in all likelihood, your weighted totals on race disagree with the population figures, so you cycle back and post-stratify on race again, and then keep cycling until you get close enough on all variables. Implementations in SAS Institute (Battaglia, Hoaglin, and Frankel 2009) and Stata (Kolenikov 2014) are available online. Both papers describe practical aspects of implementation like stopping criteria and trimming.

In view of calibration/raking, post-stratification is simply calibration with one margin (although it may be a complicated margin with two- or three-way interactions). I implement the true post-stratification simply by running my Stata program with one single target.

This has been a long enough post already, and yet I have not touched non-response adjustments (NRA). Non-response is whatever stands between you receiving your original sample (general population samples from the vendor such as Marketing System Group (MSG) or Survey Sampling International (SSI) for RDD or ABS samples or list samples for specialized clients that have the full population somewhere in their systems, e.g., a list of patients of a hospital) and you closing the study with your final responding sample that is only a fraction of the original sample. Non-response is a complicated confluence of noncontacts (your mail was never opened by the intended recipient who just tossed the envelope; your calls on the phone always went to the answering machine), refusals (the intended respondent explicitly told you that he or she is not interested in continuing the survey), unknown eligibility (it is unclear from the message on the phone if the phone number is active), etc.

If the original sample came with auxiliary information, this information can be used to scale the responding sample back to resemble the original sample more closely:

$$w_i^{\mathrm{NRA}} = w_i^{\mathrm{SRC}} f_i^{\mathrm{NRA}} \qquad (2)$$

With an abuse of notation, SRC subscript is recycled to denote the use of weights that serve as input to non-response adjustments (base weights, frame integrated weights, etc.), NRA stands for non-response adjustment, and f_i^{NRA}

is the non-response adjustment factor. Since you know the relation between your original sample and the population, provided by the base weights, the non-response adjusted weights will align your sample with the population. List samples from client populations often have the variables necessary for such adjustments. For example, in surveys of university alumni, the information about the year of the degree and the major is typically available, and often date of birth, gender, and race/ethnicity can be found as well. Even if the data on the population counts are not provided by the client university, and all you receive is the sample of the alumni names, contact information, and the minimal demographics such as above, you can still try to adjust the responding sample back to the original sample.

There are several methods of non-response adjustment. Let me highlight two of them.

Non-response adjustment can be carried out within cells defined by the existing sample information on both respondents and non-respondents, very much like in post-stratification. You can break the full sample into the cells defined by cross-classification of cohorts and majors, and define non-response adjustment factor as the inverse of the plain or weighted response rate:

$$f_{gi}^{\text{NRA}} = \frac{\text{\# originally sampled units in cell g}}{\text{\# respondents in cell g}} \tag{3}$$

or

$$f_{gi}^{\text{NRA}} = \frac{\Sigma_{\text{originally sampled unit } k \in g} w_{gk}^{\text{SRC}}}{\Sigma_{\text{respondent } k \in g} w_{gk}^{\text{SRC}}} \tag{4}$$

Another popular method of non-response adjustment is by modeling response propensity (Little 1986). Defining the dependent variable as 0 for non-response and 1 for response, you can run a logistic (or probit, although it is less frequently used) regression model using the auxiliary variables available on the full sample (graduation year, major, demographics) as the explanatory variables. Then you can define the non-response adjustment factor as the inverse of the predicted response propensity $f_i^{\text{NRA}} = 1 + \exp(-x_i'\hat{\beta})$ where $\hat{\beta}$ are coefficient estimates from the logistic regression, or as the inverse of the mean response rate within a group of units with similar response propensities. A common approach is to break the sample into five groups by the response propensity, either as groups of equal size, in terms of number of responding units or sampled units, or as intervals of equal length; and then using these groups as non-response adjustment cells to be used in expression (3) or (4).

Table 1 A hypothetical sample from alumni population.

Cohort	Degree	Gender	Population count	Original sample	Respondents
2007	Bachelor	Male		480	153
2007	Bachelor	Female		520	261
2007	Bachelor	Total	10,000	1,000	414
2007	Graduate	Male		230	120
2007	Graduate	Female		220	120
2007	Graduate	Total	3,000	500	240
2012	Bachelor	Male		460	207
2012	Bachelor	Female		540	315
2012	Bachelor	Total	12,000	1,000	522
2012	Graduate	Male		250	151
2012	Graduate	Female		250	173
2012	Graduate	Total	3,500	500	324

A word of caution is applicable to all of post-stratification, calibration, and cell non-response adjustment: make sure that each of the cells has at least 50 respondents. Otherwise, you may be adjusting on something too noisy, and the weights may blow up leading to undesirably high design effects.

Non-response adjustments and calibration are not mutually exclusive, and when the data allow, can and should be used together. There is evidence (Krueger and West 2014) that performing just the calibration/raking is insufficient without properly accounting for non-response processes.

Let me provide an illustrative example. Let us say that we are taking a sample of a university alumni. The alumni records provide breakdown by cohort and degree, and the sample that is drawn from these records additionally has the alumni gender. An informative response was simulated to produce higher response rates for females, for holders of graduate degrees, and for more recent graduates, reflecting the known demographic trends for non-response, and a likely better contact information for recent graduates. The counts are given in Table 1.

Let me construct four different weights, each of which can be considered reasonable, and in some circumstances, the only feasible.

1. Weight 1, the non-response adjusted weight: a logistic regression with main effects of cohort, degree and gender was fit to the data. No attempt to model interactions was made, although the regression fit poorly as shown by an "almost significant" p-value of the Hosmer-Lemeshow (Hosmer, Lemeshow, and Sturdivant 2013) goodness-of-fit $\chi 2$ statistic (p=0.056). The NRA weight was produced as the base weight divided by the estimated response propensity. No aggregation of respondents was done, and response propensities were used as is. This would be the only feasible weight if the population totals in column 4 of Table 1 were not known.

2. Weight 2, the post-stratified weight: post-stratification (1) was based on the four cells of cohort and degree. This would be the only feasible weight if the original sample did not have any additional information on top of the existing population information (although in this case, gender is additionally available), and population counts in all cells are known.
3. Weight 3, raked to margins only: this weight is constructed using the raking algorithm, with population targets being the 22,000 Bachelor vs. 6,500 graduate degrees, and 13,000 vs. 15,500 graduates in a cohort. This would be the only feasible weight if these counts were known, but not the counts of cohort-by-degree cells.
4. Weight 4, non-response-adjusted, post-stratified weight: the non-response adjusted Weights 1 were taken and further post-stratified in the same way as Weights 2 were. This weight uses as much information about both the population and the response process as possible, and thus is likely to be the most accurate one.

Table 2 reports the values and the summaries of the weights, and Table 3 provides the estimated totals. Note that the true population counts are only known for the total rows of gender. All other total entries are statistical estimates.

Table 2 Summaries of weights.

Cohort	Degree	Gender	Base weight	Base weight, scaled†	Estimated response propensity	NRA Weight 1	PS Weight 2††	Raked Weight 3§	NRA+PS†† Weight 4
2007	BA	Male	10	20.0	0.347	28.80	24.15	24.04	28.99
2007	BA	Female	10	20.0	0.472	21.18	24.15	24.04	21.32
2007	Grad	Male	6.67	13.3	0.474	14.05	12.5	12.69	13.99
2007	Grad	Female	6.67	13.3	0.603	11.06	12.5	12.69	11.01
2012	BA	Male	12	23.9	0.454	26.43	22.99	23.08	26.53
2012	BA	Female	12	23.9	0.583	20.58	22.99	23.08	20.66
2012	Grad	Male	7	14.0	0.585	11.96	10.80	10.66	11.87
2012	Grad	Female	7	14.0	0.704	9.95	10.80	10.66	9.87
Unequal weighting			1.054	1.054		1.112	1.095	1.095	1.115
DEFF=1+CV^2									

†With the scaling correction for the overall response rate.

††Post-stratified to the four cells of degree by cohort.

§Raked to the margins of degree and cohort only.

Table 3 Estimated totals.

Cohort	Degree	Gender	Base weight†	NRA weight	PS weight	Raked weight	NRA+PS	True population
2007	BA	Male	3055.3	4406.1	3695.7	3678.6	4435.8	
2007	BA	Female	5212.0	5527.1	6304.3	6275.2	5564.2	
2007	BA	Total	8267.2	9933.2	10000.0	9953.8	10000.0	10,000
2007	Grad	Male	1597.5	1686.2	1500.0	1523.1	1678.9	
2007	Grad	Female	1597.5	1326.9	1500.0	1523.1	1321.1	
2007	Grad	Total	3195.1	3013.0	3000.0	3046.2	3000.0	3,000
2012	BA	Male	4960.3	5470.6	4758.6	4776.9	5492.2	
2012	BA	Female	7548.3	6482.1	7241.4	7269.3	6507.8	
2012	BA	Total	12508.7	11952.7	12000.0	12046.2	12000.0	12,000
2012	Grad	Male	2110.7	1805.9	1631.2	1609.6	1792.1	
2012	Grad	Female	2418.3	1721.1	1868.8	1844.2	1707.9	
2012	Grad	Total	4529.0	3527.0	3500.0	3453.8	3500.0	3,500

†With the scaling correction for the overall response rate.

The base weights reflect the stratification by cohort and degree used in the sampling plan. They are identical across gender since gender was not used at the sampling stage. To produce entries in Table 3, the base weights were scaled up to sum up to the population size. The total estimates reflect the non-response biases in the sample: the graduates from the later cohort and the graduate degree holders are clearly over-represented when compared to the known totals. The base weights have the lowest degree of variability, which translates to the lowest apparent unequal weighting design effect of 1.054.

NRA Weight 1 combines the base weights and the response propensities. Since the latter varied across the eight cells with known demographics, the resulting weights demonstrate these differences, as well. However, since these weights do not make any attempts to equate the totals with the population figures, the latter are off in the total rows. As these weights incorporate both the four different level of the base weights and eight cell-specific non-response adjustment factors, they demonstrate a higher variability and a higher apparent DEFF of 1.112.

Post-stratified Weight 2 explicitly aligns the weighted totals with those of the population, and hence reproduced them exactly. These weights removed the main effects of cohort and degree in the non-response process, adjusting up the representation of the earlier cohorts and BA graduates. However, this weight is agnostic to gender, which is a covariate of non-response, and we see in Table 2 that post-stratified weights do not differ by gender. Hence non-response associated with gender (after accounting for cohort and degree) remains in the sample estimates. Unequal weighting effects are between those of the base weights and the non-response adjusted weights: post-stratification effectively uses only four non-response effective adjustment factors, compared to eight incorporated into the NRA Weight 2.

While post-stratified Weight 2 effectively uses four known totals (in each cohort-by-gender cell), raked Weight 3 only uses three known totals (overall total, total for cohort 2006, total for BA; the totals for cohort 2012 and graduate degrees can be obtained as the balance from the overall total). It does not use information on gender, producing identical weights across gender in all cohort-by-degree cells in Table 2, just as the base weights and post-stratified weights did. The raked weights failed to reproduce the (otherwise known) totals in the cohort-by-degree cells. In other aspects, these weights seemed to be closer to the post-stratified Weight 2 than to any other weight, although this is only an artifact of the non-response simulation model that did not feature very strong interaction.

The non-response adjusted, post-stratified Weight 4 utilized both non-response adjustment and post-stratification steps. While Weight 2 took the base weights as the input weights to post-stratification, Weight 4 took the non-response adjusted Weight 1 as the input. It shows some face validity in both demonstrating variation across gender, and matching the known totals for cohort-by-degree cells. While this weight arguably removes the greatest fraction of non-response bias compared to other weights, it correspondingly demonstrates the greatest unequal weighting effect: the trade-off between robustness and efficiency is a very typical one. Weight 4 starts off with four levels of base weight, uses the eight cell-specific factors incorporated into Weight 1 to bring the responding sample back to the original sample, and then further incorporated four additional cell-specific cohort-by-degree factors used in post-stratification.

Based on both methodological considerations and evidence presented, the most accurate weight, in the sense of having the potential to produce the estimates with the least amount of non-response bias, is the weight that is based on the combination of the explicit non-response adjustment and post-stratification to the most detailed known figures (cells of degree and cohort).

SUMMARY

A stylized flow of the survey data from frame to sample to the final data set, as well as the steps that the survey statistician can undertake to balance the sample back to the population, are represented on Figure 1 below.

- Do not say “post-stratification”/“post-stratified weight” unless the weight adjustment cells were mutually exclusive. Say “calibration”/“calibrated weight” or “raking”/“raked weight” instead, especially when unsure.
- Calibration is applicable when you have population totals. For general population surveys, they typically come from the ACS (demographics; http://www.census.gov/acs/www/), NHIS (phone

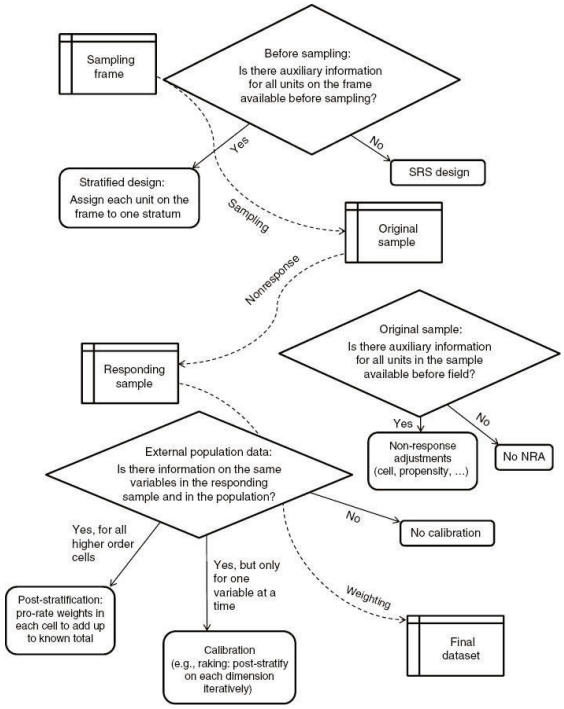

Figure 1 Information requirements for the different weight adjustments.

use; http://www.cdc.gov/nchs/nhis.htm) or CPS (detailed labor force participation and economics characteristics; http://www.bls.gov/cps/).

- Non-response adjustments are applicable when your sample comes with auxiliary variables available for both respondents and non-respondents. Then you can adjust your responding sample to make it closer to the original sample on these variables.
- Non-response adjustment and calibration are not mutually exclusive. If the existing frame, sample and population data permit, both can be used to enhance the sample quality: a non-response adjustment can be made to align the responding sample with the original sample, and calibration can further be applied to align the resulting sample with the population.

Non-response adjustments and calibration are two of potentially many steps in creating survey weights. This short tutorial provides but a cursory look at these two steps. Other components of weights may include frame integration, eligibility adjustments, multiplicity adjustments, etc. For more discussion, see Kalton and Flores Cervantes (1998), Valliant et al. (2013) and Lavallee and Beaumont (2015).

REFERENCES

AAPOR. 2015. *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys*. 8th ed. Deerfield, IL: American Association for Public Opinion Research.

Battaglia, M.P., D.C. Hoaglin, and M.R. Frankel. 2009. "Practical Considerations in Raking Survey Data." *Survey Practice* 2 (5): 176.

Brick, J., and C. Tucker. 2007. "Mitofsky-Waksberg: Learning from the Past." *Public Opinion Quarterly* 71 (5): 703–16.

Deville, J.C., and C.E. Sarndal. 1992. "Calibration Estimators in Survey Sampling." *Journal of the American Statistical Association* 87 (418): 376–82.

Holt, D., and T. Smith. 1979. "Post Stratification." *Journal of the Royal Statistical Society* 142 (1): 33–46.

Hosmer, D.W., S. Lemeshow, and R.X. Sturdivant. 2013. *Applied Logistic Regression*. 3rd ed. Hoboken, NJ: Wiley.

Kalton, G., and I. Flores Cervantes. 1998. "Weighting Methods." In *New Methods for Survey Research*, edited by A. Westlake, J. Martin, M. Rigg, and C.J. Skinner. Southampton, UK: Association for Survey Computing.

Kolenikov, S. 2014. "Calibrating Survey Data Using Iterative Proportional Fitting (Raking)." *The Stata Journal* 14 (1): 22–59.

Krueger, B.S., and B.T. West. 2014. "Assessing the Potential of Paradata and Other Auxiliary Data for Nonresponse Adjustments." *Public Opinion Quarterly* 78 (4): 795–831.

Lavallee, P., and J.-F. Beaumont. 2015. "Why We Should Put Some Weight on Weights." *Survey Insights: Methods from the Field, Weighting: Practical Issues and 'How to' Approach*. http://surveyinsights.org/?p=6255.

Little, R.J. 1986. "Survey Nonresponse Adjustments for Estimates of Means." *International Statistical Review* 54 (2): 139–57.

Thompson, M. 1997. *Theory of Sample Surveys*. New York: Chapman & Hall.

Valliant, R., J.A. Dever, and F. Kreuter. 2013. *Practical Tools for Designing and Weighting Survey Samples*. New York: Springer.