A better rim weighting algorithm

Article *in* International Journal of Market Research · August 2016

DOI: 10.2501/IJMR-2016-036

1 author:

Michael Baxter
Kantar Group

Volume 58 | Issue 4 2016

International Journal of Market Research

MRS Evidence Matters™

International Journal of Market Research

Contents

FORUM

A better rim weighting algorithm

Michael Baxter
Kantar Media

This paper proposes the asymmetric rim weighting algorithm as an alternative to rim weighting (also called raking). The latter is currently a popular method for grossing up the results of a sample survey, but asymmetric rim weighting produces results that are more efficient and have fewer high weights, with little or no increase in processing time.

Introduction

Results from sample surveys are typically grossed up to give data for the total population. The simplest procedure is that if there are n respondents and N people (or n responding households and N households in total) in the population, all results are multiplied by a grossing factor N/n. However, more sophisticated grossing procedures are available that can reduce bias caused by unrepresentative sampling and non-response bias (Brick & Kalton 1996). A sample may come from a one-off survey, or a panel that is monitored regularly. The grossed data may be presented as absolute numbers, such as the number of people born in Lithuania and working in the United Kingdom, or the number of people who watched a particular television programme. Alternatively, they may be given as a percentage of the total, such as the percentage of people who intend to vote for each political party in a forthcoming election.

Received (in revised form): 9 December 2015

DOI: 10.2501/IJMR-2016-036

Unrepresentative sample

Ideally, the sample should closely represent the population being surveyed, in that it should include the same proportions of people of each significant characteristic (for example, age and sex) as there are in the population as a whole. A well-designed random sample survey, such as those discussed by Moser and Kalton (1971), should achieve this approximately, but it is unlikely to be achieved perfectly. Further, the survey may be deliberately unrepresentative; it may over-sample some groups so that robust data on these groups can be produced. For example, in the United Kingdom Family Resources Survey, people in Scotland are over-sampled so that data for Scotland will be based on a sufficiently large sample to give reliable results (Bowditch & Rusgys 2010).

Non-response bias

Even if the sample is well chosen, there is usually non-response bias, since people with certain characteristics may be harder to contact or less likely to respond than other people; for example, many surveys have problems with contacting people in single-person households. Even when someone is contacted, he or she may refuse, or provide a response that is unusable. As uncontactable or refusing people may differ on average from other people, neglect of this will bias the results.

A solution?

One popular method that can reduce both of these problems is to gross up the survey results to a series of rims of known totals (rim weighting or raking) (see Deming & Stephan 1940). Appendix 1 shows the rims used when grossing up a recent household survey in Switzerland, using data produced by the Swiss Federal Statistical Office. If this procedure is used, different respondents will get different weights. As a result, the weighted results from the survey will have a demographic or regional profile that, for the variables in the rims, is close to that of the actual population. This should improve all of the data that can be derived from the survey, although it cannot guarantee the removal of all biases. For example, the first rim in Appendix 1 has 25 cells, containing the populations of each of the 25 elementary zones. If the weights of all the individuals in any zone are added up, the total will equal the total population of that zone. The third rim has 54 cells (nine age groups multiplied by two sexes multiplied by three linguistic regions). Again, if the weights of all the individuals in any age/sex/linguistic region group are added up, the total will equal the total population of that group.

There may be bias due to interactions between two variables, or due to variables not included in the weighting. Bias due to interactions can be removed by interlocking two variables (for example, having a rim of household size by region rather than separate rims for household size and region), but this can lead to rims with very large numbers of cells.

A cell is a particular value of the variable used in a rim. Thus, with household size, there might be four cells, corresponding to households of sizes 1, 2, 3 and 4+. If there are ten regions, there would be ten cells, one for each region. Thus a rim of household size by region would need four times ten, or 40, cells. Unless the sample is very large, the number of households in some cells may be quite small.

There are other weighting methods – see, for example, Sharot (1986), and Kalton and Flores-Cervantes (2003). However, this paper concentrates on rim weighting, which is current best practice.

The current algorithm for rim weighting

Rim weighting is an iterative procedure; its mode of operation is outlined in, for example, Sharot (1986) and a detailed example is given below.

Illustration of rim weighting

This is an iterative process. If the universes are specified correctly, it will normally converge, although this is not guaranteed (Ireland & Kullback 1968).

Step 1
Assign each respondent an initial weight, known as the *pre-weight*; this is usually 1 for all respondents.

Step 2
Consider the first rim. Suppose, for example, that it is by area and the country is divided into five areas. Respondents are classified by which area they are in (see Table 1 for an example).

Each respondent is assigned a weight so that the sum of the weights of all respondents in a cell equals the universe for that cell. At the first stage, if all the pre-weights were 1, this is just the universe divided by the number of respondents; the results are given in the last column.

Table 1 Position after step 2

Area	Number of respondents	Universe	Weight
1	37	127	3.4324
2	24	95	3.9583
3	32	87	2.7188
4	29	101	3.4828
5	39	122	3.1282
Total	161	532	

If there are pre-weights other than 1, the procedure in step 3 would be followed.

Step 3

Consider the second rim. Suppose, for example, that it has three cells: adult male, adult female, child. Respondents are classified into these three cells. Since respondents now no longer all have the same weight, we have to consider the sum of their weights. The results might be as shown in Table 2.

It is now necessary to make the sum of weights in each cell in this rim equal the universe for that cell. Thus each weight in a cell is multiplied by the weight adjustment for that cell, which is the universe divided by the sum of weights. The weight for an adult male in area 1 is now:

$$0.9509 \times 3.4324 = 3.2637$$

The weight for a child in area 2 is now:

$$1.0645 \times 3.9583 = 4.2135$$

There are thus now $5 \times 3 = 15$ possible different weights (assuming that there are respondents in each age–sex group in all areas).

Table 2 Position after step 3

Group	Number of respondents	Sum of weights	Universe	Weight adjustment
Adult males	68	224.0095	213	0.9509
Adult females	68	226.2588	232	1.0254
Children	25	81.7317	87	1.0645
Total	161		532	

Step 4
Repeat step 3 for each rim in turn. By this point, two respondents will usually have the same weight only if their demographic characteristics on all of the rims are the same.

Step 5
Check for convergence. This means that, for each rim, the sum of weights in each cell is calculated and compared to the universe. If they are all approximately equal, as specified by the convergence criterion, the process has converged and the program stops. The usual convergence criterion is that the absolute difference between the sum of weights and the universe for each cell in each rim is less than 0.5; this means that, when rounded to the nearest integer, the sum will equal the universe. Another possibility is to require the ratio of the sum of weights to the universe to be close to 1, say between 0.999 and 1.001. This is not recommended; such a criterion would be excessively tight for small universes and too loose for large ones.

If the weighting has not converged, the results will not be usable because it means that weighted data will not equal the specified rim totals. Thus, using the example in Appendix 1, the estimated population for some of the elementary zones would not necessarily equal their actual population.

Step 6
If the process has not converged, then the weights that were calculated at the end of step 4 are set as pre-weights and steps 1 to 5 are repeated until there is convergence. A limit on the number of iterations should be set, and if the procedure has not converged after this number, it should halt as a failure. It is also possible to halt the procedure as a failure if, after any iteration, all of the sums are further from the universes than after the previous iteration so the process is not converging.

The efficiency of a grossed-up survey

The grossing up procedure assigns a weight to each respondent. Those with characteristics that are over-represented among respondents compared with the population as a whole will have weights lower than average, and those with characteristics that are under-represented among respondents compared with the population as a whole will have weights higher than average. (Of course, this just refers to those characteristics for which reliable independent data can be found to allow a rim to be calculated.) An estimate for any characteristic measured by the survey can then be found by summing the individual weights of all respondents who have that characteristic, so that over-represented people will get less weight and under-represented ones will get more, reducing the effects of disproportionality as far as possible. For example, suppose there were a survey of a population of 1,000, sample size 12, which measured whether someone had fair or dark hair. If the respondents had the characteristics and weights in Table 3, the best estimate would be that there were 420 fair-haired and 580 dark-haired people, whereas the unweighted estimate would be that there are 500 of each.

Table 3 Getting unbiased estimates from weighted results

Hair	Weight	Hair	Weight	Overall total
Fair	85	Dark	144	
Fair	73	Dark	94	
Fair	81	Dark	114	
Fair	66	Dark	75	
Fair	54	Dark	66	
Fair	61	Dark	87	
Total	420		580	1,000

This reduction of bias comes with a penalty. The sampling error will be larger than if the results were not weighted. However, the benefits of reducing the bias should outweigh the increase in sampling error. (This is often expressed by saying that an analysis should minimise the mean square error, defined as bias squared plus variance of sampling error.) This increase is often quantified by defining an effective size of the sample, which is less than its actual size. If individual i has weight w_i, the effective size n_{eff} is:

$$n_{eff} = \frac{\left(\sum w_i\right)^2}{\sum w_i^2}$$

See Conway (1982) and Kish (1992) for further details of effective size. In a perfectly balanced random survey, the sampling error is proportional to the square root of the reciprocal of the sample size. When calculating the sampling error of a weighted survey, the actual size must be replaced by the effective size. Since this is smaller than the actual size, the sampling error is larger than in a balanced survey of the same size.

The efficiency E is the effective size expressed as a percentage of the actual size, or:

$$E = \frac{n_{eff}}{n} \times 100\% = \frac{\left(\sum w_i\right)^2}{n\sum w_i^2} \times 100\%$$

The reciprocal of the efficiency (ignoring the factor of 100) is sometimes called the *design effect* or *deff*, though strictly speaking this refers only to the effect of deliberate disproportionality in selecting the sample, ignoring the correction for response bias. A better term is *weighting effect* or *weff*; Kalton and Flores-Cervantes (2003) use the term *variance inflation factor*, *F*. The standard deviation of the sampling error is increased by a factor equal to the reciprocal of the square root of the efficiency, so if the efficiency is 64%, the standard deviation of the sampling error increases by a factor of $1/\sqrt{0.64} = 1.25$. With the data in Table 3, the effective sample size is about 11.08, so the efficiency is 92.4% and the standard deviation of the sampling error increases by a factor of 1.04.

Another way to express the efficiency is that if the weights have mean μ and variance σ^2 the efficiency is:

$$E = \frac{1}{\left(1 + \frac{\sigma^2}{\mu^2}\right)} \times 100\%$$

Since μ is necessarily equal to the simple grossing factor N/n, it is fixed given the sample size. Thus the efficiency increases if the variance or standard deviation of the weights decreases, and vice versa. If the sample is very unbalanced, the variance of the weights will be large. Thus the efficiency will be low, so the effective sample size will be much less than the actual size and the sampling error will be considerably larger than for a perfectly balanced sample of the same size.

Weight capping

When implementing rim weighting, a modification called weight capping, or weight trimming, is sometimes used (see Potter 1988, 1990, 1993). After each loop of the algorithm, any weights exceeding a given cap are reduced to equal this cap. The cap is usually a multiple of the pre-weight (see the description above of how rim weighting works) or the average weight, say three times. It may happen that due to the cap the weighting fails to converge; in that case, the weighting procedure must be repeated with a higher cap. The rule might be that caps of, say, 3.5 and four times the pre-weight are tried, and if there is still no convergence the run is abandoned. It is also possible to have a second, lower cap (sometimes called a floor) so that all weights below this cap are increased, but this is less common.

Capping increases the efficiency by reducing the range and hence the standard deviation of the weights. This will reduce the sampling error of the results. However, there is a penalty in that the bias of the results may increase, because people with the highest weights often have similar characteristics. For example, if the highest weights are all associated with households consisting of just one person aged 20–29, and all these weights are capped, then the weighted results for such households will be too low. Since the sum of the weights must equal the total universe, it cannot change, so the uncapped weights will increase, meaning that the weighted results for other households will tend to be too high. Thus capping may increase the mean squared error (bias squared plus variance) of the results.

If the sample is well balanced, capping should be unnecessary because there will not be any very high weights. If it is a panel rather than a one-off survey, capping may be only a temporary expedient until panel balances can be rectified.

Asymmetric rim weighting

The purpose of this article is to introduce asymmetric rim weighting as an alternative procedure to the conventional one. While it is equally effective at removing bias, the weights give higher efficiencies, and the need for weight capping is eliminated or reduced.

Consider the step of adjusting to a given rim. Let the weights be w_1 to w_n, $\Sigma w_i = W$, and the required rim total be T. The standard procedure for rim weighting is to multiply each weight by the same factor, T/W; the total will now be T as required, and the efficiency is unchanged since the mean μ and standard deviation σ are both multiplied by T/W so their ratio is unchanged. This method may be called R for ratio adjustment.

An alternative procedure would be to add $(T - W)/n$ to each weight. Again the total will now be T as required. This method may be called E for equal adjustment. The variance of the weights is unchanged. If $T > W$, the mean weight increases by $(T - W)/n$. Thus σ^2/μ^2 decreases and the efficiency increases, so in this case the procedure is better than R.

However, if $T < W$ the mean weight decreases hence so does the efficiency. Further, if any weight is less than the amount of the adjustment, it would become negative, which is not permissible. Thus in this case the procedure is worse than R.

If all of the weights are initially equal (either 1 or some other constant pre-weight), as is usually the case, then on the first step the results of E and R will be identical; within each cell, all of the weights will become T/n.

Thus asymmetric rim weighting may be defined as follows. Let the weights be w_1 to w_n, $\Sigma w_i = W$, and the required rim total be T:

- if $T < W$, decrease the weights by multiplying each of them by T/W
- if $T = W$, do nothing
- if $T > W$, increase the weights by adding $(T - W)/n$ to each of them.

An example is given in Table 4. The given set of 17 weights (total 142.74) is to be adjusted to a universe of 217.

The adjustment for the R method is to multiply each weight by 217/142.74 = 1.520, while for the E method it is to add (217 – 142.74)/17 = 4.37 to each weight.

Table 4 Comparison of the two methods of adjustment

	Before adjustment	After *R* adjustment	After *E* adjustment
	5.27	8.01	9.64
	5.62	8.54	9.99
	5.01	7.62	9.38
	9.36	14.23	13.73
	3.46	5.26	7.83
	9.73	14.79	14.10
	5.18	7.87	9.55
	10.95	16.65	15.32
	9.09	13.82	13.46
	12.98	19.73	17.35
	5.37	8.16	9.74
	16.04	24.38	20.41
	5.16	7.84	9.53
	8.82	13.41	13.19
	14.38	21.86	18.75
	7.99	12.15	12.36
	8.33	12.66	12.70
Total	142.74	217.00	217.00

The weights after E adjustment have a smaller range and a lower standard deviation than those after R adjustment (7.83–20.41 and 3.61 for E compared with 5.26–24.38 and 5.49 for R). If these are the final weights, the efficiency for E is 93.0% and for R it is 85.2%.

In practice, there are several rims, and the system must be iterated to convergence. At any step, T may be greater or less than W and the appropriate method must be chosen. Within a given iteration, some rims will need E and others will need R; the selection may be different at the next iteration. At convergence, all the weights will sum to the correct rim totals, so will be as unbiased as those from the current method. However, it is not obvious that the simple argument above about efficiency will work

Table 5 Results from eleven data sets

Efficiencies

Data set	Current	Asymmetric	Change
1	71.6	72.1	+0.5
2	69.5	70.1	+0.6
3	69.6	70.2	+0.6
4	74.4	74.8	+0.4
5	68.4	69.3	+0.9
6	75.0	75.4	+0.4
7	74.7	75.0	+0.3
8	66.2	67.6	+1.4
9	77.2	77.5	+0.3
10	79.7	79.9	+0.2
11	78.4	78.7	+0.3

Minimum and maximum weights

	Current		Asymmetric	
Data set	Min.	Max.	Min.	Max.
1	0.0604	3.64	0.0356	2.43
2	0.1055	7.89	0.0752	6.31
3	0.0867	4.09	0.0473	3.02
4	0.1155	3.47	0.0812	2.66
5	0.1041	6.20	0.0441	4.30
6	0.1384	4.72	0.0879	3.54
7	0.1220	4.09	0.0782	3.18
8	0.0101	4.68	0.0002	3.49
9	0.1008	4.99	0.0862	3.93
10	0.1362	3.33	0.1239	2.70
11	0.1267	3.89	0.1074	3.03

in this more complex case. It must be assessed empirically whether the proposed method converges and produces more efficient weights. Results from applying this method to 11 data sets are given in Table 5. These are all household surveys carried out in Switzerland. Data set 1 included 22,500 individuals and the others included around 10,000 individuals each. There were four weighting rims (see Appendix 1).

Thus, in all 11 cases, the new method converges. It always produces a higher efficiency than the current method, although the gain is small. Both the minimum and maximum weight always decrease. However, the mean weight (which equals the total universe divided by the number of respondents) is necessarily unchanged, so some weights must increase.

Table 5 Results from eleven data sets (continued)

Percentiles of the weights

	Current			Asymmetric		
Data set	**10th**	**Median**	**90th**	**10th**	**Median**	**90th**
1	0.1223	0.3229	0.6042	0.1194	0.3207	0.6201
2	0.2166	0.5889	1.1466	0.2171	0.5843	1.1778
3	0.2151	0.5207	1.1257	0.2111	0.5113	1.1633
4	0.2311	0.5568	1.1217	0.2268	0.5448	1.1479
5	0.3036	0.7790	1.6686	0.2908	0.7822	1.7378
6	0.3066	0.7697	1.4842	0.3007	0.7690	1.5217
7	0.2340	0.7416	1.4069	0.2377	0.7330	1.4411
8	0.0725	0.8295	1.5353	0.0658	0.8481	1.5732
9	0.2133	0.9673	1.6078	0.2131	0.9671	1.6227
10	0.2401	0.8478	1.4267	0.2438	0.8472	1.4531
11	0.2330	0.8270	1.4236	0.2320	0.8246	1.4483

Number of iterations needed

Data set	Current	Asymmetric	Change
1	18	23	+5
2	17	17	0
3	17	20	+3
4	18	20	+2
5	20	22	+2
6	18	19	+1
7	19	21	+2
8	23	24	+1
9	89	86	–3
10	17	17	0
11	19	23	+4

The 10th percentile and the median of the weights usually, but not always, decrease slightly. The 90th percentile is substantially less than the maximum weight, suggesting that there are comparatively few weights that are much larger than average. This percentile always increases, so the increase in weights must be largely in the weights that are rather larger than average but not the very largest ones.

The number of iterations generally increases slightly, but not by much. Further, the time per iteration may decrease, as addition is faster than multiplication. The run time is longest for the ninth set, and is about nine seconds for both methods on a PC with a processor running at 2.60 GHz. For the other sets, which required far fewer iterations, the run times are much shorter. There is no evidence that the run times of the two methods differ significantly.

Reducing the maximum weight usually reduces the need for weight capping, if this is used, and so reduce any bias due to the weight capping. For example, for data set 6, there are 9,334 weights, average 0.8462. Of these, 2,421 are higher under the new method, but the maximum decreases from 4.72 to 3.54. If the cap is 3.385, i.e. four times the average weight, 11 weights would need to be reduced with the current method but only one with the asymmetric method. On the other hand, the minimum weight is also reduced. Thus if there is also a floor, there will probably be more small weights that need to be increased. However, floors are less common than caps.

Even higher efficiencies can sometimes be obtained by requiring the procedure to use the E method at every step. However, it often happens with this method that some of the weights are negative so the results would not be acceptable. With both the current and the proposed method, negative weights can never occur.

Conclusion

From the evidence presented in this paper, the proposed new method seems to be generally superior to the current rim weighting method, and would improve the accuracy and robustness of market research findings, hence leading to better outcomes for users of the research. It is likely to be particularly advantageous if weight capping is being implemented. It should thus replace the current method, especially if weights are capped.

A spreadsheet that can implement both methods is available from the author.

Appendix 1: Rims used to weight a Swiss survey

1. Population in each of 25 elementary zones.
2. Population by size of township: >500,000; 200,000–499,999; 100,000–199,999; <100,000; isolated village; rural.
3. Nine age groups (0–2, 3–14, 15–19, 20–29, 30–39, 40–49, 50–59, 60–69, 70+) by sex and by linguistic region (German, French, Italian).
4. Housewife_husband/Non-housewife_husband by household size (1, 2, 3, 4, 5+) by linguistic region.

Acknowledgements

The author would like to thank Les Taylor for helpful discussions, Greg Fulford for coding the two rim weighting algorithms, and the referees for their valuable feedback.

References

Bowditch, E. & Rusgys, G. (2010) *Family Resources Survey Annual Technical Report: 2008–09*. Available online at: www.ons.gov.uk/ons/rel/frs/family-resources-survey---technical-report/2008-09/technical-report.pdf (accessed 24 May 2016).

Brick, J.M. & Kalton, G. (1996) Handling missing data in survey research. *Statistical Methods in Medical Research*, 5, pp. 215–238.

Conway, S. (1982) The weighting game. *Market Research Society Conference Papers*, pp. 193–207.

Deming, W.E. & 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, pp. 427–444.

Ireland, C.T. & Kullback, S. (1968) Contingency tables with given marginals. *Biometrika*, 55, pp. 179–188.

Kalton, G. & Flores-Cervantes, I. (2003) Weighting methods. *Journal of Official Statistics*, 19, 2, pp. 81–97.

Kish, L. (1992) Weighting for unequal P_i. *Journal of Official Statistics*, 8, pp. 183–200.

Moser, C. & Kalton, G. (1971) *Survey Methods in Social Investigation*. London: Heinemann Educational.

Potter, F.J. (1988) Survey of procedures to control extreme sampling weights. *Proceedings of the American Statistical Association, Section on Survey Research Methods*, pp. 453–458.

Potter, F.J. (1990) A study of procedures to identify and trim extreme sampling weights. *Proceedings of the American Statistical Association, Section on Survey Research Methods*, pp. 225–230.

Potter, F.J. (1993) The effect of weight trimming on nonlinear survey estimates. *Proceedings of the American Statistical Association, Section on Survey Research Methods*, pp. 758–763.

Sharot, T. (1986) Weighting survey results. *Journal of the Market Research Society*, 28, 3, pp. 269–284.

About the author

Michael Baxter is the Chief Statistician in the Audience Intelligence division of Kantar Media. He is a former Head of Methodology at the United Kingdom Central Statistical Office and a former chairman of the Official Statistics Section of the Royal Statistical Society. His work has appeared in journals ranging from *Metrika* to *Economic Trends*.

Address correspondence to: Michael Baxter, Kantar Media, Kantar House, Westgate, London W5 1UA, UK.

Email: michael.baxter@kantarmedia.com