Survey Methodology
Part E

Fall 2015

Seppo Laaksonen

Weighting Adjustments
due to Unit Non-response
Called also Re-weighting

Content

Survey micro data
Sampling Design
Response Rates
Missingness mechanisms
Auxiliary variables, Covariates
Response propensity modeling
Design weights
Basic weights
Post-stratification calibration
Raking ratio calibration
Generalised regression estimation and calibration
Calmar and Calmar 2
PSW = Propensity score weighting (response propensity modeling weighting)
PSW plus calibration
General conclusion

Weighting and Re-weighting process

It can be considered to cover the following 7 actions:

(i) Sampling design before the fieldwork
(ii) Weights for the gross-sample (*n*) using (i), 'design weights'
(iii) Sampling Design File before and after the fieldwork, this includes auxiliary variables from registers, other administrative sources, also from the fieldwork
(iv) 'Basic weights' for the net sample or for the respondents (*r*), assuming *MARS*
(v) Re-weighting strategies assuming MAR(C): specification, estimation, outputs
(vi) Estimation: point-estimates, variance estimation = sampling variance plus variance due to missingness.
(vii) Critical look at the results including benchmarking these against recent results (how plausible they are?)

Two types of Auxiliary variables (covariates)

(i) Macro or aggregate:

- Known (frame) population statistics by strata or post-strata or calibration margin (benchmarking): e.g. region, gender, age group, industry, number of employees and their share by gender, age group, occupation and education,

(ii) Micro:

- All variables (possibly useful but not known necessarily in advance) that are available both for the respondents and for the non-respondents: e.g. a code for region, area, psu, gender, age or age group, industry, education level, marital status, year of marriage, socio-economic group, dwelling unit size, number of children in a household, type of home, type of living area (grid), number of rooms, mother tongue, citizenship, employment status, living or not in a municipality born, R&D intensity, ownership, ...

Missingness/Response rates and propensity

Usually, it is first computed a response rate and then this is analyzed by categories of auxiliary variables.
Naturally, this gives opportunity to understand non-response and in-eligibility as well, and some ideas for re-weighting can be found.
Currently, the response rates are lower and lower in developed countries (40-60%). In-eligibility rates can also be high if possible to get correctly (10-25%).

Towards weighting

We should always have a valid sampling design, that can be simple or more or less complex. Some examples soon.

Each sampling design is determined for a gross sample. But the data file after the fieldwork is available (for most variables) for a net sample only, that is, for the unit respondents.

The core variables in the sampling design data file include:
- Identity code (both confidential and non-confidential)
- All inclusion probabilities of the sampling design
- Other sampling design variables (stratum, psu, ...)
- Auxiliary variables (above)
- Survey modes (single, mixed. multi)
- Fieldwork outcome
- Technical variables and good meta data

Inclusion probabilities

We should always have a valid sampling design, that can be simple or more or less complex.

Each sampling design is determined for a gross sample. But the data after the fieldwork is available (for most variables) for a net sample only, that is, for the unit respondents.
And we have calculated based on this design

- The design weights for the gross sample

But due to unit nonresponse, we also need the weights for the net sample, i.e., for the respondents.
I call these 'basic weights' or 'base weights' but some use 'design weights' for these as well, but note that this assumes that non-response is ignorable, e.g. within explicit strata but not necessarily within 'cluster *psu*'s'. However, the whole psu rarely is missing, instead units within *psu*'s.

Inclusion probabilities

The sampling file should include one inclusion probability variable at minimum. This is the case in one stage sampling. But the number of the probabilities is growing while more stages are used in the design.

Usually, the inclusion probabilities are independent of each other, that is, the final inclusion probability is the product of all stage probabilities.
There are designs in which case these probabilities are not independent but thus we do not consider these cases in details.
Note however that it is possible that all probabilities are not known for all units, i.e. there may be missingness for all or some nonrespondents.

Towards Re-weighting

As in the previous examples, we can have
- A specific sampling design, simple or more or less complex

And we have calculated based on this design
- The design weights for the gross sample

But due to unit nonresponse, we also need the weights for the net sample, i.e., for the respondents.
I call these 'basic weights' or 'base weights' but some use 'design weights' for these as well, but note that this assumes that non-response is ignorable, e.g. within explicit strata but not necessarily within 'cluster *psu*'s'. However, the whole psu rarely is missing, instead units within *psu*'s.

Inverse probability weighting is used in clinical studies for these weights.

Towards Re-weighting 2

Re-weighting thus starts from the valid basic weights that will be tried to improve so that the estimates will be less biased than the initial ones. Usually, there is not in mind to improve all estimates but some key estimates. The other estimates are often improved at the same time but not maybe all of them.

As already seen, good auxiliary data are necessary to make re-weighting successful. If you have little good auxiliary variables, you cannot do much. So, you have to work for the auxiliary data service hardly during the survey process.

Re-weighting methods

I do not try to explain all possible re-weighting methods since they are too many. Often it is however difficult to recognise what a certain method is about since so many various terms are used. I will not be an exception. My terms are somewhat new for you, I guess, but they are in my opinion quite clear, I hope.

I will concentrate on the two methodologies
Calibration and Propensity weighting (called also response propensity based weighting)
And their combination, or synergic application.
This could be called Joint Propensity and Calibration Weighting (JPCW).

Before that; I briefly describe Post-stratification that is possibly the most common reweighting method.

Post-stratification

Is a basic calibration method that is useful to apply if you have such population level data (macro auxiliary data) that are not yet exploited in the sampling design. This is often the case.

Post-stratification is not, unfortunately, simple still, since it is conditional to the initial sampling design. This means that there may be difficulties to compute appropriate post-stratified weights. A big problem is often that the data is too small in some post-strata. THIS is obviously the main reason why the other calibration methods are developed. We consider them later in this part. First however, we explain how to implement post-stratification, or how to create the post-stratified sampling weights?

Post-stratification 2

If the sampling design is simple random sampling, you can create the post-stratified weights:

- If your data file consists of a categorical variable for which the target population statistics are available.
- Naturally, the number of respondents should be big enough as in ordinary stratification.

The post-stratified weights have the same form as in stratification, that is, (in which *g* means a post-stratum *g=1,..,G.)*

$$w_k = \frac{N_g}{r_g}$$

This method is often used even though it is not known how close to *srs* the sampling is. For example, when obtained by CATI a number of respondents more or less randomly, the weights are calculated assuming that they are selected randomly within post-strata.

Post-stratification 3

I present another case that is maybe most common. Now the sample has been drawn by explicit stratification and with a certain allocation. The strata are symbolised by *h*. When the respondents are known, responding problems are found. For example, if the stratification is regional as it is often, the basic weights adjust for regional representativeness, not for anything else. However, it was found that females participated better than males, and educated people as well. This may lead to post-stratification given that the target population statistics are available at the same categories as in the survey data file. The tabulation of next page illustrates the situation.

Illustrating post-stratification

	Initial stratification = Pre-stratification							
	Region 1		Region 2		Region R			
Post-strata within pre-strata	Little educa-ted	More edu-cated	Males	Fe-males	Little educated males	Little educated females	More educated males and females	

It thus is possible to flexibly create the post-strata within each pre-stratum. Its purpose is either that the response rates vary by these post-strata, or the target is to reduce the sampling error that occurs if post-strata are more homogenous than initial strata. The weights are of the same form as all stratified weights

$$w_k = \frac{N_{hg}}{r_{hg}}$$

The strategy for creating ‘propensity sampling reweights’ is as follows

(i) We have the gross sample design weights that are the inverses of the inclusion probabilities. Explicit stratification is used.

(ii) We assume that the response mechanism within each stratum is ignorable, and hence compute the initial (basic) weights analogously to the weights (i). These are available only for the respondents k, and symbolised by w_k.

(iii) Next we take those initial weights and divide these by the estimated response probabilities (called also response propensities) of each respondent obtained from the probit or logit model, and symbolised by p_k.

(iv) Before going forward, it is good to check that the probabilities p_k are realistic, that is, they are not too small, for instance. All probabilities are below 1, naturally.

The strategy for creating ‘propensity sampling reweights’, continues

(v) Since the sum of the weights (iii) does not match to the known population statistics by strata *h*, they should be calibrated so that the sums are equal to the sums of the initial weights in each stratum. This is made by multiplying the weights (iii) by the ratio

$$q_h = \frac{\sum_h w_k}{\sum_h w_k / p_k}$$

(vi) It is good also to check these weights against basic statistics. If the weights are not plausible, the model should be revised.

Example of the response propensity weighting

This uses the response propensity model of Part D that consists of the four auxiliary variables; Education, Region and the interaction of Gender and Agegroup. The predicted values or response propensities were calculated using this model and then continued to the adjusted weights, RP in the below table. The table also includes the basic weight figures, all these as analysis weights that are not as confidential as proper sampling weights. You see that the variation of the weights increases when adjusting.

Analysis weight	Mean	Minimum	Maximum	Coefficient of Variation	Sum
Basic	1.00	0.95	1.02	3.32	1624
RP	1.00	0.67	1.70	19.46	1624

Calibration by Calmar 2

I first summarize calibration as Särndal (2007, *Survey Methodology 33*, 99-119) presents it:

"The calibration approach to estimation for finite populations consists of
a computation of weights that incorporate specified auxiliary information and are restrained by calibration equation(s),
the use of these weights to compute linearly weighted estimates of totals and other finite population parameters: weight times variable value, summed over a set of observed units,
an objective to obtain nearly design unbiased estimates as long as nonresponse and other non-sampling errors are absent."

Calibration by Calmar 2

Calmar 2 is a new version of the initial Calmar that can be downloaded from INSEE website. Some maybe do not like that the manuals are in French, but it is good for everyone to learn some basics of this language.

Calmar 2 is a SAS macro as the initial Calmar as well. This means that you cannot do your own applications but insert necessary parameter values in the programme only. It was not easy to start to work with it but I found a person (Josiane Guennec) from INSEE who was willing to help us to use the software and the document: Sautory, Olivier ja Le Guennec, Josiane (2005). La macro Calmar 2: Redressement d'un échantillon par calage sur marges. Institut National de la Statistique et des Etudes Economiques Direction Generale.) .

Calibration by Calmar 2

I do not explain a general approach to calibration in details. The basic idea is thus to calibrate the re-weights so that the certain margins (macro auxiliary statistics) are correct. There are a number of strategies to succeed with this target. Usually, the algorithm is such that the distance between the initial (basic) weight and the calibrated weight will be minimized. It is easy to notice that the distance function can be different. Calmar 2 gives opportunity to apply the five alternatives:

- Linear
- Raking ratio that is in fact exponential
- Logit
- Truncated linear
- Sinus hyperbolicus

Calibration by Calmar 2

A user has to choose the starting weight that is ‘basic weight’ in our pure Calmar application. Secondly, he/she have to create a file that includes the margins. The number of margins and their categories are technically limited, but in practice, it is good to be realistic. There are in Calmar 2 also two margin levels possible, such as for individuals and for households, respectively. The third point needed is to choose one of these five methods. The methods give opportunity to put the certain constraints as follows: lower limit and the upper limit of the ratio of ‘calibrated weight/starting weight.’ This may be useful in order to avoid negative weights and other extreme weights. This option is both in method ‘Logit’ and ‘Truncated linear.’ Raking ratio and sinus hyperbolicus (both are exponential based) do not provide negative weights.

I have tested how to get negative weights. For example, this occurred for units, if the number of margin categories was rather high. Also, if too many margins are tried to use, this may happen more often.

Calibration by Calmar 2

There are many nice things in CALMAR 2. For example, it shows what is the distribution of categories of auxiliary variables based on the initial weights and respectively, the true values that should be achieved by calibration.

There is an example about this on next page.

CALMAR 2 thus gives such types of figures for all margins, and it is principly possible to calculate an overall summary of these differences.

CALMAR 2 example, one sample out of 100 simulations

COMPARAISON ENTRE LES MARGES TIRÉES DE L'ÉCHANTILLON (AVEC LA PONDÉRATION INITIALE) ET LES MARGES DANS LA POPULATION (MARGES DU CALAGE)					
VARIABLE	MODALITÉ	MARGE ÉCHANTILLON	MARGE POPULATION	POURCENTAGE ÉCHANTILLON	POURCENTAGE POPULATION
AGEG	1	22699.73	21595	13.05	12.41
	2	52728.41	42980	30.31	24.70
	3	47224.04	47565	27.14	27.34
	4	42531.51	49630	24.45	28.53
	5	8801.31	12215	5.06	7.02

VARIABLE	MODALITÉ	MARGE ÉCHANTILLON	MARGE POPULATION	POURCENTAGE ÉCHANTILLON	POURCENTAGE POPULATION
GENDER	1	79940.68	85575	45.95	49.19
	2	94044.32	88410	54.05	50.81

VARIABLE	MODALITÉ	MARGE ÉCHANTILLON	MARGE POPULATION	POURCENTAGE ÉCHANTILLON	POURCENTAGE POPULATION
REGION = Stratum	PKT	38710.00	38710	22.25	22.25
	Maas	75845.00	75845	43.59	43.59
	KaupEt	32550.00	32550	18.71	18.71
	KaupPo	26880.00	26880	15.45	15.45

Based on Initial weights

True

Sample and 'true' statistics are not equal by age group and gender but

they are equal by region since the region is a stratification variable as well.

Calibration by Calmar 2
And Simulation data

My examples are based on the simulated data set that is created much from the Finnish European Security Survey 2010. Thus, we have created the population based on its sample. This data set was selected since its auxiliary data pattern is good. I already have shown its micro auxiliary variables. I selected three macro auxiliary variables only although there could be more possibilities in the data set. But this selection is such that is often applied in practice. It should be noted that these macro variables or calibration margins should be true values and in real life, there not so many opportunities for these. My margin are as already observed in the CALMAR 2 output: Gender, Age Group and Region.

The last variable is also used in sampling design, since we wanted to use a fairly common design, i.e. Stratified Random Sampling, but so that allocation by strata is not equal or proportional that is realistic in real-life. NOTE that this same margin is needed to include in CALMAR as well. Otherwise, the true margin values is not guaranteed for Region or Stratum.

Simulations

The simulated data set, naturally, is not a extended copy of the initial data. Random numbers is several steps make this data set unit. Since we have a number of estimate to test, we can get varying results, thus not the one single conclusion.

So far, we have created 150 simulations that seem to give very stable results. If the results are too close to each other, we interpret these so that these methods are about as good.

But before that I explain how we have applied the response propensity weighting.

My new strategies

(i) Combine Calibration and Propensity Weighting

(ii) Use the design weights in estimating propensities. This was successful in the paper

Now I have no time for the second strategy but it is in some cases definitely useful, if the sampling weights vary substantially.

In the rest I will show how CALMAR 2 has been used after my propensity weighting, thus as an additional stage after (v).

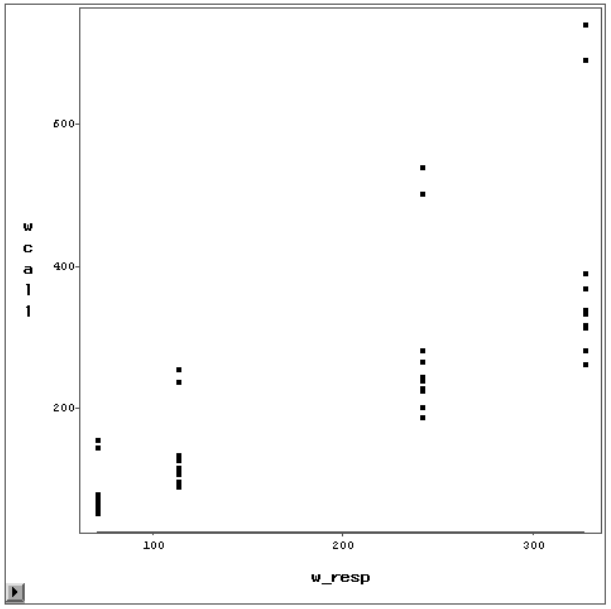
This is thus simply such a strategy that the initial weights in CALMAR 2 are the propensity adjusted weights, instead of the basic weights. You know that there are different calibration strategies as well.

One simulation result for various weights

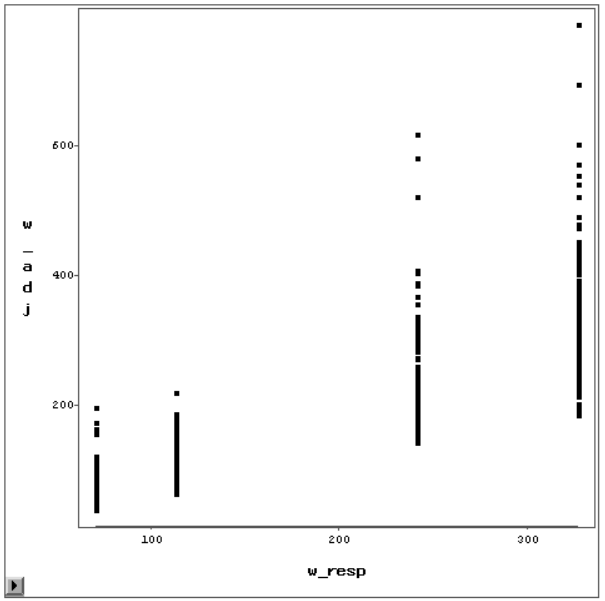
This is just to show how the weights vary

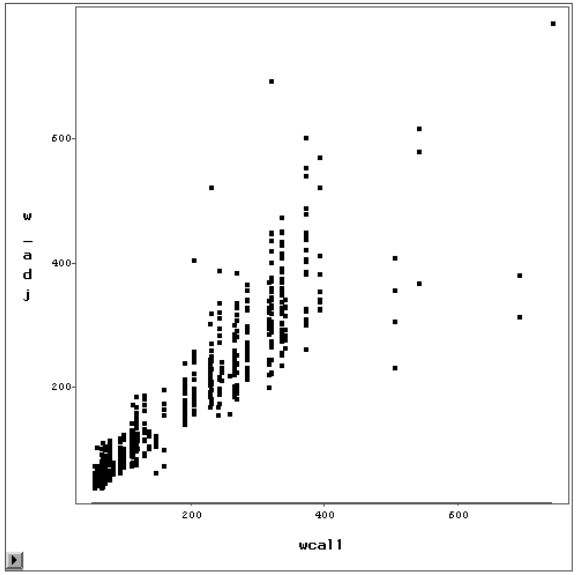
The following page are some comparisons at micro level

Weight	Explanation of weight	n or r	Mean	Coeff of Variation	Mini-mum	Maxi-mum	Sum	
w_sample	Design weight	2200	79.1	62.4	36.2	151.7	173985	
w_resp	Basic weight	1071	162.5	64.2	74.5	334.1	173985	
wcal1	Linear	1071	162.5	66.1	58.4	436.4	173985	
wcal2	Raking ratio	1071	162.5	66.084	59.4	443.2	173985	
wcal3	Logit	1071	162.5	66.086	59.4	443.5	173985	
wcal4	Truncated linear	1071	162.5	66.084	58.4	436.4	173985	
wcal5	Sinus hyperbolicus	1071	162.5	66.088	59.5	440.6	173985	
w_adj	Propensity weight	1071	162.5	71.69	46.1	638.7	173985	
wcal1_2	Prop+Linear	1071	162.5	71.51	42.0	643.8	173985	
wcal2_2	Prop+Raking ratio	1071	162.5	71.52	42.1	643.6	173985	
wcal3_2	Prop+Logit	1071	162.5	71.52	42.1	643.6	173985	
wcal4_2	Prop+Truncated linear	1071	162.5	71.51	42.0	643.8	173985	
wcal5_2	Prop+Sinus hyperbolicus	1071	162.5	71.52	42.1	643.8	173985	

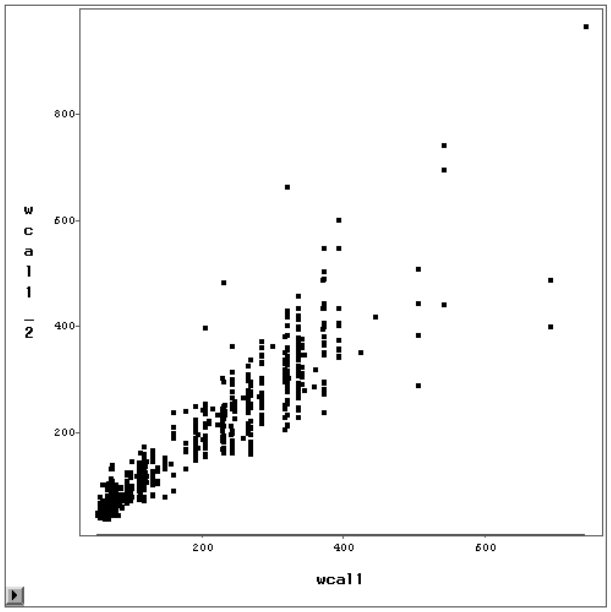


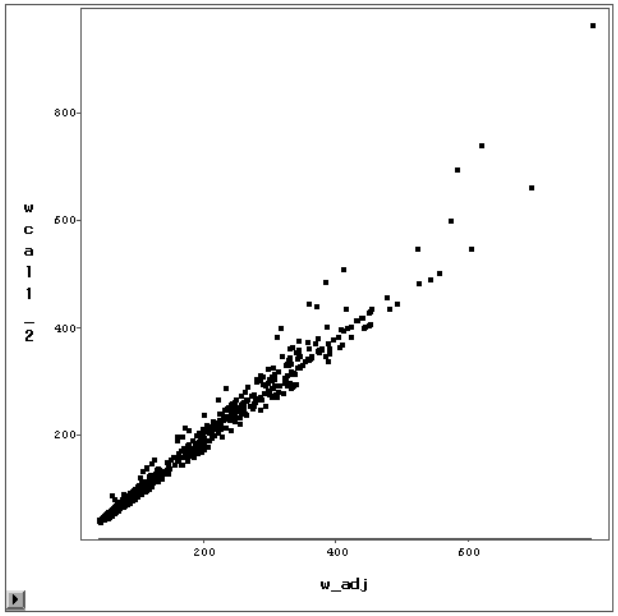
Scatter between basic weight (w_resp=x-axis) and linear calibrated weight (left), and propensity weight (right)



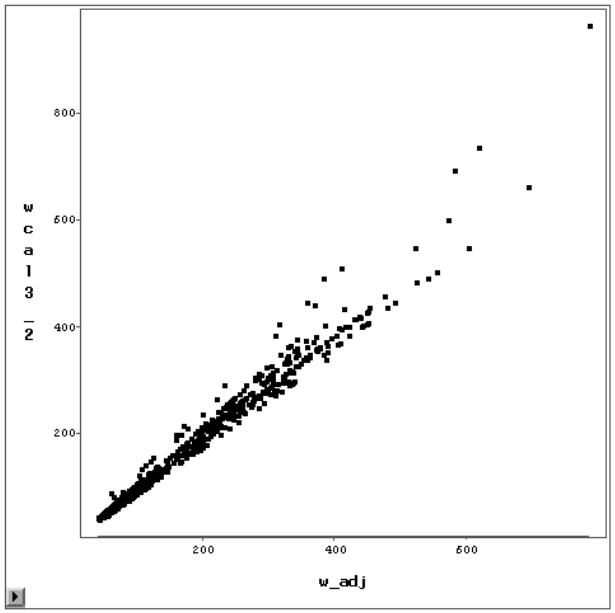


Scatter between linear calibrated weight (wcal=x-axis) and propensity weight (left) and propensity and linear calibrated weight (right)





Scatter between propensity weight (w_adj=x-axis) and linear calibrated weight (left), and logit calibrated weight (right)



Our conclusion is that there are three groups of weights:

- Design weights are not interesting from the estimation point of view but in general and comparing against other weights
- Basic weights are relatively close to the design weights although unit non-response has some influence on them
- Calibrated weights vary very little from each other. This is surprise for us
- Propensity adjusted weights are more varying than calibrated weights as expected
- Joint propensity and calibrated weights are varying little as calibrated weights.

We made an exercise that aims at illustrating, how well weighting works with several different variables and the indicators estimated on those.

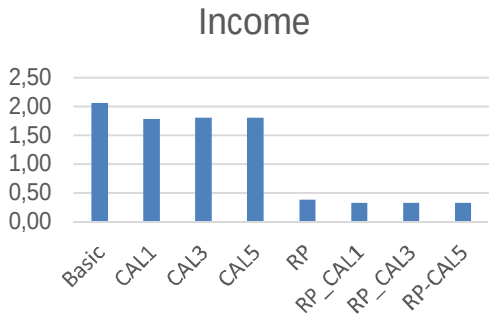
The comparisons are fairly easy to do since we use simulated data that is much on characteristics of real data including its missingness mechanism. We already found from real data that the well adjusted weights do not change dramatically an estimate in all cases. It is possible that the estimates are already fairly correct but it is also possible that our auxiliary variables are not well predictable that thus is often the case.

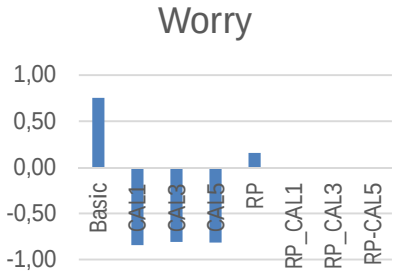
Our simulation exercise illustrates the situation with 8 different weights as described on next page.

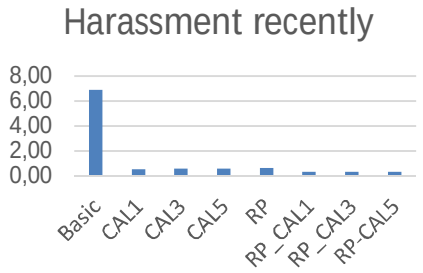
Eight weighting methods in simulations:

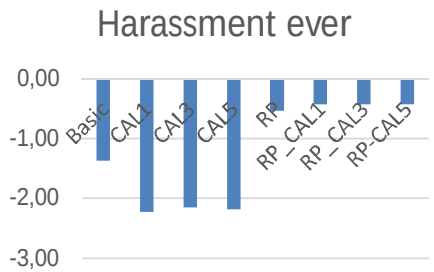
- Basic weights (design weight assuming ignorable nonresponse)
  BASIC in Graph
- Three pure calibration methods with three margins (gender, age group, region) and following distance functions: linear, logistic and sinus hyperbolicus
  CAL1, CAL3 and CAL5 in Graph
- Response propensity weighting with 8 auxiliary variables
  RP in Graph
- The same three calibration methods after the response propensity weighting
  RP_CAL1, RP_CAL3 and RP_CAL5 in Graph

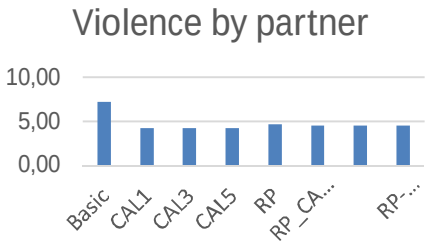
The true value of the graphs is = 0, and the differences are relative to these true values.

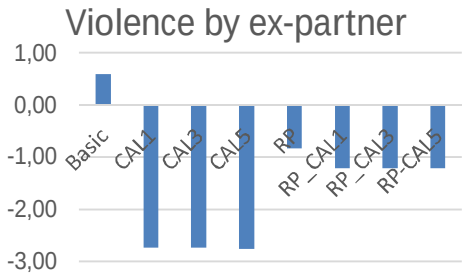


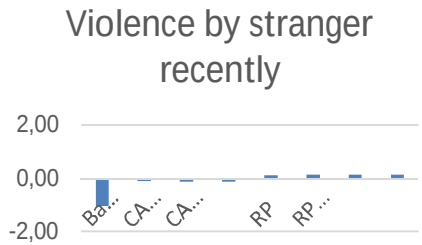


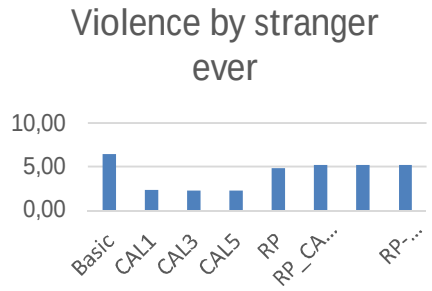












Basic = Basic weight
CAL1-3 = Calibrated weights
RP = Response propensity weight
RP_CAL1-3 = Joint RP+CAL weight

Conclusion of the simulation results

The differences between the three calibration estimates are minor. This is concerned the pure calibration methods CAL1, CAL3 and CAL5 on one hand, and the same methods after the response propensity weighting on the other.

Almost all weights with adjustments improve the estimates to some extent. The study also shows that the combination of the response propensity weighting and calibration is a superior method to pure calibration. Nevertheless, it is not best in each case. It is even so that the basic weights are best in one case (violence by ex-partner). The two reasons behind this are obvious: a small number of respondents and non-good auxiliary variables. Calibration is best only in one case (violence by stranger ever). Surprisingly, the results are for this indicator worsening when calibrating after response propensity weights. We thus see that any weighting method does not work ideally in each case.