Part IV

WEIGHTING ADJUSTMENTS

Introduction to Part IV

Jelke Bethlehem[a] and Mario Callegaro[b]

[a]*Statistics Netherlands, The Netherlands*

[b]*Google UK*

IV.1 Panel problems

It is the objective of most online panels to collect and publish reliable and accurate statistical information about specific populations. If the fundamental principles of probability sampling are applied, unbiased estimates of population characteristics can be computed, and also a margin of error for these estimates can be determined.

Computing estimates and margins of error for non-probability online panels are still much-debated topics. The Office of Management and Budget (2006) defines *estimation error* for nonprobability samples as "the difference between a survey estimate and the true value of the parameter in the target population" (p. 31). For estimates for nonprobability online panels, a *credibility interval*, rather than a margin of error, is a popular metrics used. Online electoral polling is a good example of its use. "The credibility interval reflects the statistical uncertainty generated by a statistical model that relies on Bayesian statistical theory" (AAPOR, 2012, p. 1).

In daily practice, ensuring that results from a panel represent the target population is often not an easy task. There are always phenomena affecting the representativity of the outcomes of online panels. One such problem is *undercoverage*, a phenomenon in which not all members of the target population are represented in the sampling frame. This can happen, for example, in the recruitment phase of an online panel for the general population if every member of the population does not have access to the Internet. (See the Introduction to Part I of this volume.)

Another problem is *nonresponse*. This phenomenon occurs when members of the target population who have been selected for the sample do not provide the required information. Nonresponse can have several causes, the most common of which are refusal, non-contact, and inability. (See the Introduction to Part II in this volume.)

Online Panel Research: A Data Quality Perspective, First Edition.
Edited by Mario Callegaro, Reg Baker, Jelke Bethlehem, Anja S. Göritz, Jon A. Krosnick and Paul J. Lavrakas.

Companion website: www.wiley.com/go/online_panel

Nonresponse can occur in the recruitment phase. It can also occur in the waves of a longitudinal study panel, in which case it usually has a monotone pattern: the group of respondents decreases with each subsequent wave. Once individuals stop responding, they are lost to the panel. This type of nonresponse is usually called *attrition*.

In the case of a cross-sectional study panel, there can also be nonresponse in both phases. Nonresponse in the second phase (for specific surveys) may have a monotone pattern (attrition), but nonresponse may also be in reaction to a specific topic of one survey. Panel members who do not like a particular topic may decide not to participate in that survey but may respond to a subsequent survey.

A third problem affecting the representativity of online panels is *self-selection* (Public Works and Government Services Canada, 2008). If panel recruitment is not based on probability sampling but is voluntary, that is, it is left to individuals themselves to become a member of the panel, the researcher has no control over the composition of the panel. As a result, the panel will typically consist only of persons who like to do surveys and/or are interested in the topics of these surveys. Therefore, self-selection panels often suffer from a substantial lack of representativity.

The problem with nonresponse, undercoverage, and self-selection is that panel members are usually different from those not in the panel. Consequently, the panel is not representative of the population, which in turn prohibits valid statistical inference. Wrong conclusions about the population would be drawn from the panel. To avoid this, the results must be corrected for this lack of representativity. Weighting adjustments make up a family of commonly used techniques designed to correct for this problem. A short overview will be given here.

IV.2 Weighting adjustments

Weighting adjustments attempt to improve the accuracy of survey estimates by using auxiliary information. *Auxiliary information* is defined as a set of variables that have been previously measured in a survey and for which information on their population distribution (or complete sample distribution) is available.

By comparing the response distribution of an auxiliary variable to its population (or complete sample) distribution, it can be determined if the sample is representative of the population (with respect to this variable). If this distribution differs considerably, one must conclude that the sample lacks representativity. To correct this, adjustment weights are computed. Weights are assigned to records of the respondents. Estimates of population characteristics are then computed by using weighted instead of unweighted values. Weighting adjustments are often used to correct surveys that are affected by nonresponse. An overview of weighting adjustments can be found in Bethlehem and Biffignandi (2012) and Särndal and Lundström (2005).

We will explore various correction techniques in the rest of this chapter. Our focus is on weighting adjustments for cross-sectional study panels, but note that these techniques are equally applicable for longitudinal study panels.

Because nonresponse can occur during recruitment phase and during the subsequent surveys of the cross-sectional study panel, it would imply that two corrections are required. A possible first approach could be to ignore the two phases of nonresponse. Weights would then be obtained by directly aligning response distributions for auxiliary variables with their population distributions. However, this is not the most effective way to conduct adjustment weighting. Weighting in two steps is preferred. In the first place, recruitment nonresponse may be a different phenomenon than survey nonresponse; therefore, it may require a different

model containing different variables. In the second place, there are a lot more auxiliary variables available to correct for the survey nonresponse. For many online panels, new members take a profile survey in which they answer basic demographic questions. All these variables can be used to weight the survey data. In contrast, there are often fewer auxiliary variables available for weighting adjustments in the recruitment phase.

To summarize, weighting adjustments for an online panel is a two-step process as follows:

1. Compute weights for all panel members in such a way that the panel is representative with respect to the target population.
2. For each survey, compute weights in such a way that the survey is representative with respect to the panel.

The final weights are obtained by multiplying the recruitment weights by the survey weights.

IV.3 Effective weighting

Auxiliary variables are a vital ingredient of weighting techniques, but not every auxiliary variable is effective in terms of weighting. The set of auxiliary variables used for weighting should satisfy two conditions:

1. The auxiliary variables must be able to completely explain the response behavior of the individuals.
2. The auxiliary variables must be able to completely explain the target variables of the survey.

Auxiliary variables for the recruitment phase are often scarce; therefore, it will not be easy to find auxiliary variables that satisfy both conditions. Moreover, due to the multi-purpose nature of many online panels, it will not be clear in advance what the target variables will be.

If recruitment is based on probability sampling, the main representativity problem will be caused by nonresponse. The target population will usually be well-defined, and there will be a corresponding sampling frame, which may contain auxiliary variables. For example, members for the LISS panel (Scherpenzeel, 2008) were selected by means of a random sample from the population register. Consequently, for all individuals in the sample (whether they responded or not) a set of demographic variables were available. Moreover, the Statistics Netherlands had population distributions for many more auxiliary variables for the target population of this panel. So there were a lot of opportunities for weighting adjustments.

If panel recruitment is based on self-selection, weighting techniques can be applied as well. There may, however, be a problem with the definition of the target population. Due to self-selection, it is not always clear what the target population is. Consequently, individuals who choose to join the panel may not belong to the intended target population. If the actual target population is not clear, it will not be possible to find the proper population distributions of auxiliary variables. Moreover, for self-selection panels, there is no sampling frame; therefore, this source of auxiliary variables is not available. For these reasons, it may be difficult to make effective weighting adjustments.

Baker et al. (2010) describe weighting adjustments for online panels in more detail in an AAPOR report. Their conclusion is that researchers should avoid self-selection online panels when one of the research objectives is to accurately estimate population values. The

report states that there is "no generally accepted theoretical basis from which to claim that survey results using samples from non-probability online panels are projectable to the general population."

In a recent AAPOR report on non-probability sampling, the authors discuss the performance of weighting adjustment for self-selection panels (see Baker et al., 2013) and conclude that "adjustments seems to reduce to some extent, but do not by any means eliminate coverage, nonresponse, and selection bias inherent to opt-in panels."

The next section describes a number of weighting adjustment techniques. For the sake of convenience, only weighting adjustments for the recruitment phase are described. Note that the weighting adjustments for each specific survey are similar. To keep things simple, it is assumed the all individuals in the target population have access to the Internet; therefore, there are no under coverage effects. We also assume that a simple random sample has been selected from the population.

IV.4 Weighting adjustment techniques

IV.4.1 Post-stratification

Post-stratification is a well-known and often used weighting technique (see Cochran, 1977, or Bethlehem, 2002). To perform post-stratification adjustments, categorical auxiliary variables are needed. By crossing these variables, the population and sample are divided into a number of non-overlapping strata (subpopulations).

All elements in one stratum are assigned the same weight, and this weight is equal to the population proportion in that stratum divided by the sample proportion in that stratum. Suppose that crossing the stratification variables produces L strata. The number of population elements in stratum h is denoted by N_h, for $h = 1, 2, \ldots, L$. Hence, the population size (N) is equal to $N_1 + N_2 + \ldots + N_L$. The weight w_k for an element k in stratum h is now defined by

$$w_k = \frac{N_h/N}{n_h/n}, \tag{IV.1}$$

where n_h is the number of respondents in stratum h, and n is the sample size. If the values of the weights are taken into account, the result is the post-stratification estimator

$$\bar{y}_{ps} = \frac{1}{N}\sum_{h=1}^{L} N_h \bar{y}_h \tag{IV.2}$$

where $\bar{y}_h$ is the response mean in stratum h. So, the post-stratification estimator is equal to a weighted sum of response means in the strata.

It can be shown that the bias of weighted estimates is small if there is a strong relationship between the target variable and the stratification variables. The variation in the values of the target should manifest itself between strata but not within strata. In other words, strata should be homogeneous with respect to the target variables. In nonresponse correction terminology, this type of missing data comes down to the missing at random (MAR) assumption.

The bias of the estimator will also be small if the variation of the response probabilities is small within strata. This implies that there must be strong relationships between the auxiliary variables and the response probability.

In conclusion, application of post-stratification will successfully reduce the bias of the estimator if proper auxiliary variables can be found. Such variables should satisfy the following four conditions:

1. They have to be measured in the survey.
2. Their population distribution must be known.
3. They must be strongly correlated with all target variables.
4. They must be strongly correlated with the response behavior.

Unfortunately, such variables are not very often available, or there is only a weak correlation. A reference survey may be an option in this situation.

IV.5 Generalized regression estimation

Post-stratification is a rather simple and straightforward weighting technique. More advanced weighting adjustment techniques are described in Bethlehem (2002) and Särndal & Lundström (2005). One such technique is *generalized regression estimation*, also known as *linear weighting*.

Generalized regression estimation assumes there is a set of auxiliary variables $X_1, X_2, \ldots X_p$ that can be used to predict the values of a target variable Y. The generalized regression estimator is defined by

$$\overline{y}_{GR} = \overline{y} + (\overline{X} - \overline{x})'b, \tag{IV.3}$$

where $\overline{y}$ is the sample mean of the target variable. $\overline{X}$ is the vector of population means of the auxiliary variables, and $\overline{x}$ is the vector of sample means of the auxiliary variables. Furthermore, b is the (estimated) vector of regression coefficients. This estimator reduces the bias if the underlying regression model fits the data well.

Post-stratification is a special case of generalized regression estimation. If the stratification is represented by a set of dummy variables, where each dummy variable denotes a specific stratum, expression IV.3 reduces to expression (Iv.2).

By rewriting expression (Iv.3), it can be shown that generalized regression estimation is a form of weighting adjustment (see, for example, Bethlehem & Biffignandi, 2012). The value of a weight for a specific respondent is determined by the values of the corresponding auxiliary variables.

Generalized regression estimation can be applied in situations other than post-stratification. For example, post-stratification by age, class, and sex requires the population distribution of the crossing of age, class, by sex to be known. If just the marginal population distributions of age, class, and sex separately are known, post-stratification cannot be applied. Only one variable can be used. However, generalized regression estimation makes it possible to specify a regression model that contains both marginal distributions. In this way, more information is used, and this will generally lead to better estimates.

Generalized regression estimation has the disadvantage that some correction weights may turn out to be negative. Such weights are not wrong but simply a consequence of the underlying theory. Usually, negative weights indicate that the regression model does not fit the data too well. Some analysis packages are able to work with weights, but they do not accept negative weights. This may be a reason not to apply generalized regression estimation.

It should be noted that generalized regression estimation will only be effective in substantially reducing the bias if the MAR assumption applies to the set of auxiliary variables used.

IV.6 Raking ratio estimation

Correction weights produced by generalized regression estimation are the sum of a number of weight coefficients. It is also possible to compute correction weights in a different way, namely, as the product of a number of weight factors. This weighting technique is usually called *raking ratio estimation, iterative proportional fitting*, or *multiplicative weighting*.

Multiplicative weighting can be applied in the same situations as generalized regression estimation as long as only qualitative auxiliary variables are used. Correction weights are the result of an iterative procedure. They are the product of factors contributed by all cross-classifications (of stratification variables). To compute weight factors, the following process has to be carried out:

1. Introduce a weight factor for each stratum in each cross-classification term. Set the initial values of all factors to 1.
2. Adjust the weight factors for the first cross-classification term so that the weighted sample becomes representative with respect to the auxiliary variables included in this cross-classification.
3. Adjust the weight factors for the next cross-classification term so that the weighted sample is representative for the variables involved. Generally, this will distort representativeness with respect to the other cross-classification terms in the model.
4. Repeat this adjustment process until all cross-classification terms have been dealt with.
5. Repeat steps 2, 3, and 4 until the weight factors no longer change.

The advantage of using multiplicative weighting is that computed weights are always positive. The disadvantage is that there is no clear model underlying the approach. Moreover, there is no simple and straightforward way to compute standard errors of weighted estimates. In contrast, a generalized regression estimation is based on a regression model that allows for computing standard errors.

IV.7 Weighting adjustment with a reference survey

In the previous section, it was shown that correction techniques are effective provided that auxiliary variables have a strong correlation with the target variables of the survey and with the response behavior. If such variables are not available, one might consider conducting a *reference survey*. A reference survey is based on a probability sample and a data collection mode that leads to high response rates and little bias, e.g., CAPI (computer-assisted personal interviewing) with laptops or CATI (computer-assisted telephone interviewing). CAPI and CATI surveys tend to have high response rates. They can be used to produce accurate estimates of population distributions of auxiliary variables. These estimated distributions can be used as benchmarks in weighting adjustment techniques.

The reference survey approach has been used by several market research organizations (see Börsch-Supan et al., 2004 and Duffy et al., 2005) to reduce the bias caused by respondents' self-selection process.

An interesting aspect of the reference survey approach is that any variable can be used for adjustment weighting as long as it is measured both in the reference survey and in the online panel. For example, some market research organizations use "webographics" or "psychographic" variables that divide the population into "mentality groups." (See Schonlau et al. (2004) for more details about the use of such variables.)

It should be noted that use of estimated population distribution will increase the variance of the estimators. The increase in variance depends on the sample size of the reference survey: the smaller the sample size the larger the variance. Therefore, using a reference survey may reduce the bias at the cost of increasing the variance.

IV.8 Propensity weighting

Propensity weighting is used by several market research organizations to correct for a possible bias in their web surveys. Examples can be found in Börsch-Supan et al. (2004) and Duffy et al. (2005). The original idea behind propensity weighting goes back to Rosenbaum and Rubin (1983, 1984), who developed a technique for comparing two populations. The technique is used to attempt to make the two populations comparable by simultaneously controlling for all variables that were thought to explain the differences. In the case of an online panel, there are also two populations: those who participate in the online panel (if asked), and those who do not participate.

Propensity scores are obtained by modeling a variable that indicates whether or not someone participates in the survey. Usually a logistic regression model is used where the indicator variable is the dependent variable and attitudinal variables are the explanatory variables. These attitudinal variables are assumed to explain why someone participates or not. Fitting the logistic regression model involves estimating the probability (propensity score) of participating, given the values of the explanatory variables.

The propensity score $\rho(X)$ is the conditional probability that a person with observed characteristics X responds, i.e.,

$$\rho(X) = P(r = 1|X)$$

It is assumed that within the strata defined by the values of the observed characteristics X_k, all persons have the same response probability. This is the MAR assumption. The propensity score is often modeled using a logit model:

$$\log\left(\frac{\rho\left(X_k\right)}{1-\rho(X_k)}\right) = \alpha + \beta' X_k$$

Once response propensities have been estimated, they can be used to reduce a possible response bias. There are two general approaches: *response propensity weighting* and *response propensity stratification*.

Response propensity weighting is based on the principle of Horvitz and Thompson (1952) that an unbiased estimator always can be constructed if the selection probabilities are known. In the case of nonresponse, selection depends on both the sample selection mechanism and the response mechanism. The idea is to adapt the Horvitz–Thompson estimator by including the (estimated) response probabilities.

There are more advanced estimators than the Horvitz–Thompson estimator. One example is the generalized regression estimator. This estimator also can be improved by including response propensities. (For more details, see Bethlehem, Cobben, & Schouten, 2011.)

Response propensity stratification takes advantage of the fact that estimates will not be biased if all response probabilities are equal. In this case, selection problems will only lead to fewer observations, but the composition of the sample is not affected. The goal is to divide the sample in strata in such a way that all elements within a stratum have (approximately) the same response probabilities. Consequently, unbiased estimates can be computed within strata. Next, stratum estimates are combined into a population estimate.

IV.9 The chapters in Part IV

In Chapter 12, Stephanie Steinmetz, Kea Tijdens, Annamaria Bianchi, and Silvia Biffignandi explore the possibility of improving a self-selection online panel by applying weighting adjustments. They use a sample from the WageIndicator survey for their exploration. This is a continuous self-selection web survey that is conducted in 75 different countries with the objective of collecting labor-related information. For the analysis in this chapter, only the Dutch version of the survey is used.

Steinmetz et al. compare the results of their survey with the results of a different survey taken from the LISS panel. This is an online panel whose sample was selected from the Dutch population register. They use the panels in two ways. First, they use the LISS panel as a benchmark for assessing the quality of the WageIndicator survey. In their comparison of estimates for each survey, they conclude that the estimates for the WageIndicator survey are substantially biased. They note also that LISS panel estimates are not unbiased. Second, they use the LISS panel as a source of auxiliary variables for weighting adjustments. In fact, the LISS panel is used as a reference survey. The authors analyze the effects of various forms of propensity weighting.

Steinmetz et al. demonstrate that propensity weighting can help to reduce the bias, but they also conclude that the effect is rather limited and depends on the type of propensity weighting that is applied.

In Chapter 13, Weiyu Zhang takes a completely different approach to correction. Instead of applying some kind of weighting adjustment technique, she attempts to solve the nonresponse problem by imputing the values of the variables for the missing persons. Her focus is on nonresponse in specific surveys taken from an online panel. This panel already contains many survey variables for both respondents and nonrespondents, which can be used as auxiliary variables in a correction procedure. Zhang investigates an approach using a regression model where the missing values of the target variables for the nonrespondents are estimated by means of imputation.

After imputation, estimates based on the complete (imputed) sample can be compared with estimates based on just the respondents. It becomes clear that indeed uncorrected estimates may be biased. The author warns that often the explanatory power of the regression models is low, which may affect the accuracy of the imputed variables.

References

AAPOR. (2012, October 8). *AAPOR Statement*: Understanding a “credibility interval” and how it differs from the “margin of sampling error” in a public opinion poll. Retrieved July 1, 2013 from: http://www.aapor.org/Understanding_a_credibility_interval_and_how_it_differs_from_the_margin_of_sampling_error_in_a_publi.htm.

Baker, R., Blumberg, S.J., Brick, J.M., Couper, M.P., Courtright, M., Dennis, et al., (2010). Research Synthesis: AAPOR Report on Online Panels. *Public Opinion Quarterly*, *74*, 711–781.

Baker, R., Brick, J.M., Bates, N.A., Battaglia, M., Couper, M.P., Dever, J.A., Gile, K.J. & Tourangeau, R. (2013). Report on the AAPOR Task Force on Non-probability Sampling.

Bethlehem, J.G. (2002). Weighting nonresponse adjustments based on auxiliary information. In R. M., Groves, D. A., Dillman, J. L. Eltinge, & Little, R. J .A. (Eds.), *Survey nonresponse* (pp. 275–288), New York : John Wiley & Sons.

Bethlehem, J. G., & Biffignandi, S. (2012), *Handbook of web surveys*. Hoboken, NJ : John Wiley & Sons.

Bethlehem, J. G., Cobben, F., & Schouten, B. (2011), *Handbook of nonresponse in household surveys*. Hoboken, NJ: John Wiley & Sons.

Börsch-Supan, A., Elsner, D., Faßbender, H., Kiefer, R., McFadden, D., & Winter, J. (2004), *Correcting the participation bias in an online survey*. Report, University of Munich, Munich, Germany.

Cochran, W. G. (1977). *Sampling techniques* (3rd ed). New York: John Wiley & Sons.

Duffy, B, Smith, K., Terhanian, G., & Bremer, J (2005), Comparing data from online and face-to-face surveys. *International Journal of Market Research*, *47*, 615–639.

Horvitz, D. G., & Thompson, D. J. (1952), A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, *47*, 663–685.

Office of Management and Budget. (2006). *Standards and guidelines for statistical surveys*. Retrieved July 1, 2013 from: http://www.whitehouse.gov/sites/default/files/omb/inforeg/statpolicy/standards_stat_surveys .pdf.

Public Works and Government Services Canada. (2008). *The advisory panel on online public opinion survey quality: Final report June 4, 2008*. Ottawa: Public Works and Government Services Canada. Retrieved July 1, 2013 from: http://www.tpsgc-pwgsc.gc.ca/rop-por/rapports-reports/comiteenligne-panelonline/tdm-toc-eng.html.

Rosenbaum, P. R., & Rubin, D. B. (1983), The central role of the propensity score in observational studies for causal effects, *Biometrika*, *70*, 41–55.

Rosenbaum, P., & Rubin, D. (1984). Reducing bias in observational studies using sub classification on the propensity score. *Journal of the American Statistical Association*, *79*, 516–524.

Särndal, C. E., & Lundström, S. (2005). *Estimation in surveys with nonresponse*. Chichester: John Wiley & Sons.

Scherpenzeel, A. (2008), An online panel as a platform for multi-disciplinary research. In I. Stoop, & M. Wittenberg (Eds.), *Access panels and online research, panacea or pitfall?* (pp.101–106) Amsterdam: Aksant.

Schonlau, M., Zapert, K., Payne Simon, L., Haynes Sanstad, K., Marcus, S., Adams, J. Kan, H., Turber, R. & Berry, S. (2004), A comparison between responses from propensity-weighted web survey and an identical RDD survey. *Social Science Computer Review*, *22*, 128–138.