Target Estimation and Adjustment Weighting

Introduction

This Element provides a concise overview of the use of adjustment weights to analyze unrepresentative survey samples. Such unrepresentativeness can arise from the process by which subjects are sampled from the population (e.g., if nonprobability sampling is used) or in the process by which survey responses are obtained from sampled subjects (e.g., if responses are nonrandomly missing). If related to outcome variables of interest, nonrandom sampling and/or nonresponse can bias estimators of population quantities. Almost all surveys, whether historical or contemporary, are at least somewhat vulnerable to such biases.

Adjustment weighting is a simple yet flexible method of addressing sampling and nonresponse bias. It entails assigning each sampled unit an adjustment weight, which is then incorporated into estimators. An advantage of weighting over alternative methods of adjustment, such as multiple imputation, is that it does not require an explicit parametric model for each outcome variable. Rather, adjustment weighting typically involves simple modifications of nonparametric design-based estimators for probability samples, which weight units by the inverse of their probability of being selected under the sampling design.

This Element focuses on a framework for adjustment weighting known as calibration, which subsumes such commonly used methods as poststratification and raking. In this framework, a sample is “calibrated” to a set of population targets derived from auxiliary information (e.g., census data). Calibration ameliorates nonresponse bias to the extent that the variables that define these targets predict units’ response probabilities and outcome values.

A distinguishing feature of this work is that we give equal weight (no pun intended) to two basic steps in the workflow of weighting-based survey inference: the estimation of population targets and the estimation of adjustment weights. The first step (target estimation), though typically ignored by texts on survey weighting, precedes the second step (weight estimation) temporally and can exceed it in complexity and difficulty. Auxiliary information often consists of partial, noisy, and internally inconsistent population estimates, and deriving a single set of population targets from this information is often far from straightforward. Like weight estimation, which requires good working models of the nonresponse mechanism and of the outcome of interest, target estimation implicitly depends on a measurement model relating the auxiliary information to the true population distribution.

In explaining the workflow of weighting-based survey inference, we employ a mix of theoretical discussion and empirical illustration. Section 1 sets the scene by emphasizing the ubiquity of sampling and nonresponse bias and
2 Quantitative and Computational Methods for the Social Sciences

explaining the problems this poses for classical design-based inference. It then
describes in general terms how such biases can be ameliorated by incorporat-
ing auxiliary information into design-based estimators, and it outlines a general
workflow for doing so. Section 2 delves into the specifics of adjustment weight-
ing. It focuses on weight estimation, temporarily assuming that the auxiliary
information is free from measurement error. It shows how calibration sub-
sumes many commonly used weighting techniques and then discusses criteria
and procedures for the critical task of selecting population targets. Relaxing the
assumption of error-free auxiliary information, Section 3 turns to the typically
neglected task of estimating population distributions and deriving population
targets from them. Section 4 illustrates the tasks of calibration and target esti-
mation in detail, using an application to a survey conducted before the 2016 US
presidential election. Section 5 applies similar techniques to a more complex
historical application: quota-sampled public opinion polls conducted between
1936 and 1952. Section 6 discusses methodological extensions and concludes.
Key terms and abbreviations are defined in a glossary at the end of the Element.

To make it easier for readers to use the methods we describe, we provide
illustrative code implementing them in the open-source statistical software R (R
Core Team 2018). Each section ends with an appendix containing code snippets
related to the topics discussed therein. In addition, all of the code in this Element
can be accessed and run reproducibly on Code Ocean (https://codeocean.
com/). There is a separate “capsule” for each section, all of which have the tag
caughey-et-al-weighting-element. A direct link to each capsule can be
found in the Example Code subsection of the corresponding section.

1 The Problem of Unrepresentative Survey Samples

1.1 Survey Sampling: From Quotas to Probability and Back

Again

Opinion polling as we now know it originated in the mid-1930s with George
Gallup’s, Elmo Roper’s, and Archibald Crossley’s pioneering surveys of the
American public (Converse 1987, 87). Unlike straw polls such as those con-
ducted by Literary Digest magazine, which solicited survey responses from
telephone directories and other class-biased lists, Gallup and his fellow poll-
sters consciously constructed samples that were relatively small but observably
representative of the population of interest (in Gallup’s case, the US electorate).
They did so using the technique of quota sampling, in which interviewers
were sent to purposively selected locations and instructed to interview specified
proportions of subjects in distinct demographic categories (Berinsky 2006).
Target Estimation and Adjustment Weighting

The accuracy of quota sampling was validated when these pollsters correctly predicted the outcome of the 1936 presidential election, a success which contrasted markedly with the failure of the much larger Literary Digest poll. The latter’s poor performance was due to two compounding factors: the sampling bias caused by its unrepresentative sampling frame and the nonresponse bias caused by differential response rates, both of which skewed the Literary Digest’s sample in a Republican direction (Squire 1988). By forcing survey samples to match the target population in specified respects, quota sampling substantially reduced the scope for such biases.

Barely a decade later, however, quota sampling experienced its own embarrassing failure when Gallup and other pollsters mispredicted the 1948 presidential election.1 Partly in response to this debacle, US survey organizations transitioned to a new procedure for selecting respondents: probability sampling.2 Instead of constructing samples that match the target population in observable respects, probability sampling entails selecting interview subjects from the sampling frame at random according to known probabilities. By the 1950s, most commercial pollsters had adopted probability sampling, as had new academic survey organizations, such as the University of Michigan’s Survey Research Center (SRC).

Because almost all surveys continued to rely on in-person interviews, early probability samples were based on area sampling. By the 1970s, however, many commercial polling organizations had transitioned to telephone samples selected with random digit dialing (RDD). Though issues of coverage error remained, probability sampling all but eliminated sampling bias in surveys. Moreover, survey response rates remained high, ranging from about 50% for the typical consumer telephone poll to more than 70% for academic surveys such as the American National Election Studies (ANES) and 95% for the best government surveys (Wiseman and McDonald 1979; Luevano 1994; Dixon and Tucker 2010, 597).

Since the 1980s, however, response rates for both in-person and telephone surveys have fallen dramatically, in the United States as well as in other countries (Leeuw and de Heer 2002). The response rate for the SRC’s RDD-sampled

1 According to a postmortem of election polling in 1948, this prediction failure actually had less to do with the deficiencies of quota sampling per se than with late opinion movement after the last polls were conducted (Mosteller et al. 1949).
2 By the mid-1930s, Jerzy Neyman and other statisticians had laid the theoretical basis for probability sampling, and by the end of that decade area sampling methods were well established in the US Census Bureau and other government agencies. The practical merits of probability versus purposive sampling, however, continued to be debated for at least another decade (Berinsky 2006, 501–502).
Survey of Consumer Attitudes, for example, fell from about 70% in 1981 to just above 50% in 2006 (Dixon and Tucker 2010, 597). Face-to-face ANES surveys have followed a similar trajectory (Hillygus 2016). Declines among commercial telephone surveys have been even more precipitous. Response rates for Gallup surveys had dropped to 28% by 1997 and to 7% by 2017 (Marken 2018). Contemporary response rates for probability-based internet panels, which were developed in the 1990s, are if anything lower.

Given the theoretical potential for even a small amount of nonresponse to bias design-based survey estimators (Cochran 1977, 363; but see Groves 2006), these declines have seriously concerned pollsters and survey researchers. One reaction has been to rely more heavily on statistical methods for addressing nonresponse bias, including weighting (Brick and Montaquila 2009) and, less commonly, multiple imputation (Peytchev 2012). A more radical response has been to abandon probability sampling entirely and return to quota sampling, particularly for opt-in online panels (Ansolabehere and Rivers 2013) but also for telephone surveys (Moy 2015). Weighting, imputation, and quota sampling all require **auxiliary information** on the characteristics of the target population, which is available in increasing abundance (if not necessarily quality) from consumer databases, administrative records, and other sources. The importance of such adjustments was dramatically illustrated in 2016, when the undersampling of low-education white men contributed to election surveys’ underestimation of Donald Trump’s vote share in several key states (Kennedy et al. 2018).

Survey sampling has thus in a sense come full circle. Early opinion polls relied on quota sampling mainly because, in an age without widespread telephone access, it was the most cost-effective means of constructing approximately representative national samples. Despite lacking a firm basis in probability theory, quota samples seemed to work well enough in practice, at least by the metric of predicting election outcomes. The embarrassment of the 1948 election, in conjunction with the development of area-sampling methods, persuaded most American pollsters to switch to probability sampling. The advent of near-universal telephone access and development of RDD sampling, which was both probability-based and inexpensive, inaugurated what might be considered the golden age of survey sampling; but, like all golden ages, it was only temporary. By the twenty-first century, all but the highest-quality (and most-expensive) surveys suffered from response rates so low as to call into question the utility of solely design-based survey inference. Today, as in the 1930s, nearly all opinion surveys rely on purposive selection or adjustment (e.g., weighting) of their samples to render them observably representative of the population of interest.
Target Estimation and Adjustment Weighting

1.2 Survey Inference with Unrepresentative Samples

As the foregoing overview suggests, for a fairly brief period it was plausible to draw population inferences from surveys based on their sampling design alone. When feasible, design-based statistical inference is straightforward and appealing, for it does not rely on hard-to-validate assumptions about the data-generating process for the survey outcome of interest. Nor does design-based inference require any auxiliary data beyond the sample itself (though incorporating auxiliary information can often increase estimators’ precision). Rather, population means and other parameters can be consistently estimated based solely on knowledge of each population unit’s sampling probability \( \pi_i \).

1.2.1 Design-Based Inference without Auxiliary Information

When the sampling model is known by design and all sampled individuals provide valid responses, the “workflow” of inference – the sequence of steps leading from population to sample to estimate – can proceed unproblematically along the bottom of Figure 1.1, without recourse to data beyond that contained in the sample itself. The circle in the lower left corner in this figure represents a target population \( U \) of \( N \) units indexed by \( i \). In this population, the outcome of interest \( y \) and auxiliary variables \( x \) have the joint distribution \( f_U(y, x) \), and the target of inference is some parameter \( \theta_y \) of \( f_U(y, x) \). To draw inferences about \( \theta_y \), we rely on a sample \( S \) of size \( n_S \), of which a subset \( R \) of \( n_R \) responders (the respondent set) provide nonmissing responses with joint distribution \( g_R(y, x) \).

In a probability sample with full response, the sampling model that generated the observed data is known by design, and thus parameter estimation and inference can be based solely on that design.

In the case of a simple random sample (SRS), where units’ sampling probabilities \( \pi_i \) are equal and independent across units, the sample average \( \bar{y}_S = \frac{1}{n_S} \sum_{i \in S} y_i \) is an unbiased estimator of the population mean \( \mu_y \). For more complex sampling designs in which the \( \pi_i \), though known, are unequal and possibly correlated, unbiased estimation is provided by the Horvitz-Thompson (HT) estimator,

\[
\hat{\mu}_y^{HT} = \frac{\sum_{i \in S} d_i y_i}{E(\sum_{i \in S} d_i)},
\]

(1.1)

In this Element we consider only unit nonresponse, ignoring the possibility that a sampled unit might provide responses to some questions but not others (item nonresponse). For a useful discussion of methods for addressing item nonresponse, see Särndal and Lundstrom (2005, chap. 12).
6 Quantitative and Computational Methods for the Social Sciences

**Figure 1.1** The workflow of survey inference. Observed quantities are enclosed in rectangles and unobserved in circles. If the sampling model is fixed by design (i.e., the inclusion probabilities \( \pi_i \) are known) and nonresponse is absent (i.e., \( \rho_i = 1 \forall i \)), parameter estimation can proceed based on inverse-probability weights \( d_i = 1/\pi_i \). In the face of unrepresentative sampling and/or nonresponse, however, weights must be calculated based on the population targets \( \tilde{T}_x \) and on assumptions about \( x \)'s relationships with \( \rho \) and \( y \). The targets \( \tilde{T}_x \) themselves must be estimated based on auxiliary information \( \tilde{I}_x = \{\tilde{I}_{x1}, \ldots, \tilde{I}_{xM}\} \) and a measurement model relating \( f_\pi(x) \) and \( \tilde{I}_x \).

where \( d_i = \pi_i^{-1} \) is \( i \)'s inverse probability or **design weight** and \( E(\sum_{i \in S} d_i) \) is the expected weighted sample size (Horvitz and Thompson 1952). Owing to the HT estimator’s high variance, in practice it is often preferable to use the ratio or **Hájek estimator** of the mean,

\[
\hat{\mu}_y^H = \frac{\sum_{i \in S} d_i y_i}{\sum_{i \in S} d_i}, \tag{1.2}
\]

which substitutes the realized weighted sample size for the expected (Hájek 1958). Given probability sampling and full response, the Hájek estimator is consistent for \( \mu_y \) and approximately unbiased, and its sampling variance is generally substantially smaller than that of \( \hat{\mu}_y^{HT} \) (Miratrix et al. 2018, 279; Aronow and Miller 2019, 228). Its approximate variance can be estimated as

\[
\hat{\text{var}}(\hat{\mu}_y^H) = \left( \sum_{i \in S} d_i \right)^{-2} \sum_{i \in S} \sum_{j \in S} \frac{\pi_i - \pi_i\pi_j}{\pi_i\pi_j} (y_i - \hat{\mu}_y^H)(y_j - \hat{\mu}_y^H) \tag{1.3}
\]

4 Lumley (2010, 85) calls the ratio estimator “approximately unbiased” because its bias is much smaller than its standard error (proportional to 1/n rather than 1/\( \sqrt{n} \)).
where \( \pi_{ij} \) is the probability that both units \( i \) and \( j \) are sampled (Miratrix et al. 2018, appendix C). (For R code implementing the Hájek ratio estimator, see Listing 1.2.)

Alternatively, the sampling distribution of almost all common survey statistics, including \( \hat{\mu}_y \), can be estimated via the \textbf{bootstrap}. A general procedure for bootstrapping complex sampling designs is to repeat the following \( B \) times (e.g., \( B = 999 \)):

1. Take a with-replacement sample of size \( n_S \) from \( S \), respecting the original sampling design (e.g., sampling each unit with probability proportional to \( \pi_i \)).
2. Using this bootstrap sample \( S^{(b)} \), calculate the estimate \( \hat{\theta}^{(b)}_y \).

The resulting collection of bootstrap estimates \( \{\hat{\theta}^{(1)}_y, \ldots, \hat{\theta}^{(B)}_y\} \) approximates the sampling distribution of \( \hat{\theta} \). The standard deviation of the bootstrap distribution provides a consistent estimate of the standard error of the sampling distribution. Confidence intervals can also be estimated with the \( [(\alpha/2 \times B)]^{th} \) and \( [(1 - \alpha)/2 \times B)]^{th} \) largest values of the bootstrap distribution, where \( 1 - \alpha \) is the level of the confidence interval (for the many variations and subtleties of the bootstrap, see Davison and Hinkley 1997). The bootstrap is thus a valuable alternative when analytical formulas for sampling distributions are unknown or unreliable. (For R code implementing the bootstrap, see Listings 1.1 and 1.2.)

Unfortunately, most real-world surveys do not approximate the ideal conditions required for design-based inference. Even high-quality probability-sampled surveys, in which \( \pi_i \) is known for each unit, usually do not obtain valid responses from every sampled individual. In such cases, the respondent set \( R \) is a proper subset of the sample \( S \). Moreover, units’ ex ante probability of responding if sampled (their \textbf{response probability} \( \rho_i \)) is typically both unknown and heterogeneous across units. Under such conditions, the consistency of design-based estimators is no longer guaranteed. In an SRS with nonresponse, for example, the bias of the unweighted sample mean is

\[
E(\bar{y}) - \mu_y = R_{\rho y} \sigma_\rho \sigma_y / \bar{\rho},
\]

(1.4)

where \( R_{\rho y} \) is the population correlation between \( \rho \) and \( y \), \( \sigma_\rho \) and \( \sigma_y \) are the population standard deviations of \( \rho \) and \( y \), and \( \bar{\rho} \) is the population average of \( \rho \) (Bethlehem, Cobben, and Schouten 2011, 249). The bias of the HT estimator has a similar form that also depends on the association between \( \rho \) and \( y \) (Bethlehem 1988). In short, unless the outcome of interest is independent of population units’ probabilities of being sampled and responding, purely design-based estimators will in general be biased and inconsistent.
1.2.2 Adjustment Weighting to Address Sampling and Nonresponse Bias

The most common method for addressing bias due to unrepresentative sampling and nonresponse is adjustment weighting, which includes common techniques such as poststratification and raking as well as the more general framework of calibration (Deville and Särndal 1992). Unlike design weights, which can be derived from the sampling design before the survey has been conducted, adjustment weights must be calculated afterwards based on the data actually obtained. Adjustment weights are, however, analogous to design weights in that they can be incorporated into estimators of population parameters. The weighting estimator for the mean, for example, has the same form as the Hájek estimator in (1.2) but with the adjustment weights \( \tilde{w}_i \) substituted for the design-based ones \( d_i \):

\[
\hat{\mu}_W = \frac{\sum_{i \in R} \tilde{w}_i y_i}{\sum_{i \in R} \tilde{w}_i}, \tag{1.5}
\]

Unlike \( d_i \), \( \tilde{w}_i \) is a random quantity whose value must be calculated from auxiliary data beyond that contained in the sample itself. Specifically, calculating \( \tilde{w} \) requires population targets \( \tilde{T}_x = \{\tilde{T}_{x1}, \ldots, \tilde{T}_{xG}\} \) for certain auxiliary variables \( x \) measured in the survey. Suppose, for example, that we wished to weight an SRS of adults to match the proportion of men and women in the population (a simple example of poststratification). In this case, two-category gender would be the only auxiliary variable, and the population targets would consist of estimates of the proportion of men and women in the population, \( \tilde{T}_x = (\tilde{P}_{\text{men}}, \tilde{P}_{\text{women}}) \), derived from auxiliary information \( \tilde{I}_x \) external to the sample itself. Each male respondent would be assigned the weight

\[
\tilde{w}_i = \frac{\tilde{P}_{\text{men}}}{p_{\text{men}}}, \tag{1.6}
\]

where \( p_{\text{men}} \) is the proportion male among respondents. Weights for women would be defined analogously. Because each gender’s weight would be proportional to its underrepresentation in the respondent set relative to the population, this adjustment would ensure that the weighted respondent set matched the gender breakdown in the population.

Weighting eliminates nonresponse bias if it renders the response probability \( \rho \) and the outcome \( y \) totally uncorrelated, as is the case in an SRS with full response. In principle, this condition can be satisfied if, conditional on the auxiliary variables \( x \), survey responses are either (1) missing at random (MAR) or (2) independent and identically distributed (IID). For example, in the case of poststratification, bias is eliminated if either (1) \( \rho \) is constant within cells or
Target Estimation and Adjustment Weighting

(2) $y$ is IID within cells. In practice, adjustment weighting is unlikely to exactly satisfy either of these conditions, but substantial reductions in nonresponse bias are nevertheless possible if $x$ strongly predicts both $\rho$ and $y$.

Because the beneficial effects of weighting can depend heavily on which auxiliary variables are available and on the specific outcome of interest, applied methodological texts typically recommend that auxiliary variables be selected with great care (e.g., Särndal and Lundstrom 2005; Lumley 2010; Bethlehem, Cobben, and Schouten 2011). Given that there may be multiple outcomes of interest in a given survey, it can even be desirable to use different sets of weights for different analyses of the same survey data. Few texts, however, provide much concrete guidance on how exactly to select auxiliary variables. Perhaps for this reason, the overwhelming majority of applied survey researchers seem instead to rely on only a single set of weights – often ones provided by the original creators of the dataset they are analyzing. Such reliance on a single set of (design) weights may be reasonable when analyzing probability samples with minimal nonresponse, but it is less tenable if units’ probability of being sampled or of responding are not known ex ante.

1.2.3 Constructing Population Targets from Auxiliary Data

While the selection of auxiliary variables may be neglected, survey methodologists (not to mention practitioners) have given even less attention to the construction of the population targets used to create the weights in the first place.\textsuperscript{5} In poststratification, for example, it is almost universally assumed that the cell population proportions $\tilde{P}_c$ are known exactly rather than estimated. More generally, survey researchers generally presume that they possess a direct measure of $f_U(x)$, the joint distribution of auxiliary variables in the population. This presumption is frequently unjustified. Rather, the auxiliary information $\tilde{l}_x$ to which researchers have access often consists of $M$ disparate,

\textsuperscript{5} We are not aware of any textbook on survey weighting that gives more than cursory attention to target estimation, though Valliant, Dever, and Kreuter (2018) do offer a brief discussion. The following quotation, from Gelman (2007, 155), while unusual for explicitly acknowledging that population targets must be estimated, nevertheless indicates the general neglect of this problem:

In some cases the cell populations are unknown and must be estimated. For example, [in] the Current Population Survey . . . the counts are too sparse to directly estimate deep interactions (e.g., the proportion who are white females, 30–45, married, with less than a high school education, etc.) . . . For this paper, we shall ignore this difficulty and treat the [cell counts] as known.
noisy, and possibly inconsistent data sources, \( \{\hat{T}_1, \ldots, \hat{T}_M\} \), from which estimated population targets \( \hat{T}_x \) must be derived (Deville 2000; Caughey and Wang 2019).

To illustrate, consider the relatively favorable scenario where a government census has collected data on all possible auxiliary variables. In the United States, privacy concerns preclude the release of the full individual-level census files for many decades after the Census is conducted. Fortunately, individual-level census data are available in the form of 1% or 5% microsamples, though to avoid identifying individual respondents the microdata sometimes mask certain variables, such as urban residence (see, e.g., Ruggles et al. 2017). However, the US Census Bureau often separately reports the aggregate distribution of the masked variables, sometimes cross-classified with other demographic or geographic factors such as race or state. The Census also conducts special surveys, such as the Current Population Study (CPS), that contain more or less accurate estimates of other variables, such as self-reported voter turnout. As a final complication, these various data sources are typically available at irregular intervals – once per decade in the case of the full US Census, annually in the case of the CPS, and sometimes in between in the case of aggregate data such as urban residence.

Population estimates derived from such disparate sources of auxiliary information are subject to several sources of error and other complications. The first is simply random sampling error. The confidence interval for a percentage estimated from the 60,000-observation CPS, for example, can have a width of nearly a percentage point.\(^6\) Even the sampling variability of census microsamples, which typically contain at least 1 million observations, can be nontrivial for some purposes, such as estimation at the level of states or other subpopulations. Only aggregate data that summarize the entire universe of cases, such as census reports on urban population by state, are entirely free of sampling error.

A second, potentially more serious source of error stems from systematic mismeasurement (or differential measurement) of auxiliary variables in the population or the sample. Not only does self-reported turnout, for instance, tend to be exaggerated by survey respondents but the magnitude of overreporting varies systematically across population groups (Ansolabehere and Hersh 2012). It may therefore be problematic to treat the CPS as an unbiased estimate of the voting population.\(^7\) To take an example we consider in more detail

\(^6\) For example, the width of the 95% confidence interval for an estimated percentage of 50% is \( \sqrt{0.5 \times (1 - 0.5)/60,000 \times 1.96 \times 2} = 0.8\% \).

\(^7\) Although the CPS is a very high-quality survey, scholars have shown that its estimates of turnout and other quantities still suffer from nontrivial bias and have devised weights to address this problem (e.g., Hur and Achen 2013; McDonald 2019).