Weighting in Modern Surveys

Shiro Kuriwaki*

POLISCI 450D October 27, 2021

> "Survey weighting is a mess."
>
> Gelman (2005, *Statistical Science*)

Motivation

Most social science PhD students will use surveys for their own research. Modern surveys are increasingly cheap to collect, but perhaps increasingly unrepresentative. Therefore common challenges that students need to address with their survey data are often:

- "Isn't your survey data unrepresentative of the population you care about?"
- "What happens to your findings among subset S of your survey?"

Both questions require basic intuition about survey weights to answer.

Illustrative Data (for Problem Set)

Since the 1950s, the German government conducts a 1% sample of its residents. The U.S. Census Bureau also conducts a similar survey called the American Community Survey. Such surveys are the basis of many research articles, and the basis of even more government and private industry analytics.

Say we want to make inferences about Native Born Germans in former East Germany (GDR) states. You run an online Qualtrics survey of this population. Because it is online, you have no control over any strata before running the survey. Rivers (2007) was one of the first papers that showed how to match and re-weight such opt-in online surveys.

In Canvas, download the rds file `samp_GDR.rds`, which is a survey of $n = 1{,}000$ of this population. It has the following variables:

*Postdoctoral Fellow. www.shirokuriwaki.com

- `Y` is a binary outcome variable of interest (e.g. Yes to some public opinion question). It is a synthetic variable I made up; the content is not important here.
- `educ` Education, which has three levels. "Abitur" are those who pass the final exam at the end of high school aiming for higher education (like the A-levels in the UK). "Vocational" are vocational schools in the former GDR. The rest are coded as "No HighEd". The dummy variables associated with this levels are `abitur` and `poly`.

Also download the rds file `pop_GDR.rds` (N = 100,000). This is the population this poll was sampled from, which is called a *frame* in polling. In modern surveys, we do not have actual frames available, but we will use it here to explore the validity of weighting methods.

- In this data, `D` is an indicator for Response / Selection / Survey Inclusion. It is recorded in the frame. It is 1 if the item in the frame ended up in the survey, 0 otherwise.
- Usually, the outcome `Y` is not in the population frame. (The whole point of a poll is that you can only observe the outcome you are interested in the sample). But we leave it in here for later validation.

(To avoid releasing proprietary data, this is quasi-synthetic of the actual data I obtained.)

The poll is a *biased* sample of the population, so the goal is to make good (i.e. low bias, low variance) inferences about the population. Load the data and answer the following questions.

1 Post-stratification weighting

The educational breakdown in the population and survey is as follows:

Population

Education	Fraction
Not High-Ed	0.50
Abitur	0.15
Vocational	0.35

Survey

Education	Fraction
Not High-Ed	0.33
Abitur	0.25
Vocational	0.42

Question 1.1. *What are post-stratification weights that will make the weighted proportion of the education in the survey equal to the target distribution in the population?*

It is often convenient to re-scale the weights by a scalar so that they are mean 1. We call this **normalized** weights w_i for individual i.

Question 1.2. *What transformation would you apply to your weights to normalize them?*

Question 1.3. *Compute the sample proportion of Y, which we'll call $\widehat{\mu}$. How would you compute the weighted proportion in the sample with weights w_i? Is this weighted estimated guaranteed to be an unbiased estimator of the population mean of Y, which we'll call μ?*

Next, suppose we think that the outcome should vary by geography too. However, we are worried that the non-representativeness of the survey with respect to education will also show up within geographies. You tabulate the data and find this frequency table.

State	Education	n	frac
Brandenburg	Not HighEd	77	0.08
Brandenburg	Abitur	45	0.04
Brandenburg	Vocational	63	0.06
Mecklenburg-Vorpommern	Not HighEd	51	0.05
Mecklenburg-Vorpommern	Abitur	22	0.02
Mecklenburg-Vorpommern	Vocational	51	0.05
Saxony	Not HighEd	109	0.11
Saxony	Abitur	93	0.09
Saxony	Vocational	142	0.14
Saxony-Anhalt	Not HighEd	58	0.06
Saxony-Anhalt	Abitur	42	0.04
Saxony-Anhalt	Vocational	86	0.09
Thuringia	Not HighEd	70	0.07
Thuringia	Abitur	25	0.03
Thuringia	Vocational	66	0.07

Question 1.4. *How would you post-stratify on both education and state? Verbally describe the procedure. You will implement it in R in the problem set.*

Question 1.5. *Your friend is interested in another outcome. He takes the weighted mean of Q with those weights. Does this estimator has the same guarantees of representativeness as $\widehat{\mu}_w$?*

Question 1.6. *Your friend is interested in estimating the average of Y among non-married men in East Germany. She subsets the survey data to non-married men, takes the associated weights you computed, and produces weighted mean of the sample. What guarantees does this estimate have?*

In the rest of this document, we'll think about the limitations of this post-stratification weighting approach. That is, if you wanted to post-stratify on as many variables as possible, what new problems do you think this will create for you? All this will naturally motivate *Multilevel Regression and* Poststratification (MRP), which we will discuss next week.

2 Standard Errors

> "It is not always clear how to use weights in estimating anything more complicated than a simple mean or ratios, and standard errors are tricky even with simple weighted means."
>
> Gelman (2007)

Weights are great at reducing bias but there is a bias-variance tradeoff. Let's first start with standard errors with no weights are no necessarily, i.e. when the sample is a SRS:

Question 2.1. *Show that when the outcome of interest is binary, a simple random sample (SRS) of size n has a (95%) margin of error of roughly $1/\sqrt{n}$. Interpret this quantity with proper units.*

Recall the standard error is the standard deviation of your estimator (i.e. the standard deviation of its sampling distribution). Standard Errors changes with weights. A classic formulation for computing standard errors with weights is the design effect formulation by Leslie Kish.

With normalized weights w_i, the **design effect** in practice computed as

$$\mathrm{D}_{\text{eff}} = \frac{\frac{1}{n}\sum_{i=1}^{n} w_i^2}{(\frac{1}{n}\sum_{i=1}^{n} w_i)^2} \tag{2.1}$$

The definition of a design effect is the standard error of the weighted estimator relative to the standard error of the (unweighted) simple random sample version of that estimator.

Question 2.2. *Show that D_{eff} is a function of the variance of normalized weights. Is it an increasing or decreasing function of the variance of weights?*

Kish's **effective sample size** n_{eff} is defined as the SRS sample size that would be needed to achieve the same variance as the weighted estimator. Therefore, it can be computed as:

$$n_{\text{eff}} = \frac{n}{\mathrm{D}_{\text{eff}}} \tag{2.2}$$

Question 2.3. *Show that the standard error of a weighted proportion $\widehat{\mu}_w$ is an increasing function of the variance in weights.*

This relationship between weights, variance, and effective sample size has clear implications for bias-variance tradeoffs and reducing total mean square error.

3 The Balance Test Fallacy and Calibration Methods

> "Contrary to what is assumed by many theoretical statisticians, survey weights are not in general equal to inverse probabilities of selection"
>
> Gelman (2007)

Question 3.1. *In the population frame, use the variable* `D` *(selection) to estimate the propensity score. What is the Horvitz-Thompson weighted estimator?*

Inverse probability weighting (IPW) works well in theory. Remember the propensity score theory (Rosenbaum and Rubin 1983): If you control for the correct propensity score model in a biased sample, then your estimate of the population mean is unbiased (i.e., asymptotically).

But propensity scores must be estimated from the data, and this is hard (Kang and Schafer 2007). Thus the balance test is hard to get right, and is also misleading: aggressive weighting will lead to false negatives (Imai, King, and Stuart, *JRSS A* 2008). And small propensity scores "explode" the variance: a probability of 0.01 implies a weight of 100. The idea of *calibration* is to avoid modeling the propensity score altogether, and instead just balance the covariates.

Imai and Ratkovic (*JRSS B*, 2014) point out that a logit propensity score model's MLE is identical to ("dual" to) the covariate balancing constraint. That is, define the propensity score model as a logit:

$$\pi_\beta(X_i) = \frac{\exp X_i'\beta}{1 + \exp X_i'\beta} \tag{3.1}$$

The likelihood of all the data is the Bernoulli form

$$\pi_\beta(X_i)^{D_i}(1 - \pi_\beta(X_i))^{1-D_i} \tag{3.2}$$

for all observations i. To make things easier, as usual in MLE, we take the log:

$$\sum_i D_i \log\{\pi_\beta(X_i)\} + (1 - D_i)\log(1 - \pi_\beta(X_i)) \tag{3.3}$$

To find the parameter β that maximizes this likelihood (MLE), we solve for the score constraint: i.e. the first derivative of the log likelihood with respect to β.

Question 3.2. *Show that the MLE of the propensity score model is identical to the covariate balancing condition.*

Question 3.3. *Using the* `CBPS` *package (Imai et al.), compute Covariate Balancing Score Weights from the population and survey.*

Hainmueller (*Political Analysis*, 2012) formulated the entropy balancing score:

$$\{w_i^{\texttt{ebal}}\}_{i=1}^{N} = \arg\min_{w} \sum_{i:D_i=0} w_i \log(w_i/w_i^0), \tag{3.4}$$

$$\text{subject to } \frac{1}{N}\sum_{i:D_i=0} w_i f(X_i) = \frac{1}{n}\sum_{i:D_i=1} w_i f(X_i) \tag{3.5}$$

where w_i^0 are merely base weights as starting values (e.g. all 1s), and $f(X_i)$ are various moments of the covariates (e.g. means). This also tries to finds low-variance weights that match a constraint. The idea derives from **rake weighting** in surveys:

> "In particular, the [entropy balancing method] heavily borrows from the survey literature that contains several reweighting schemes which are used to adjust sampling weights so that sample totals match population totals known from auxiliary data (see Sarndal and Lundstrom 2006 for a recent review and earlier work by Deming and Stephan 1940, Ireland and Kullback 1968, Oh and Scheuren 1978, and Zaslavsky 1988 who proposed a similar log-linear reweighting scheme to adjust for undercount in census data)." (Hainmueller 2012, p.27)

Balancing weights are an active area of statistical research. See for example work by Zubizarreta et al. (2017)

Question 3.4. *Using the* `ebal` *package (Hainmueller), compute the same sort of weights with entropy balancing propensity scores.*

Question 3.5. *In Question 1.4, we presumed you had access to the joint distribution of state and education. Suppose that the German government does not give you that joint distribution, but it* does *tell you the breakdown of education in all of the population and populations of each state. What can you do to try and post-stratify the data to the joint distribution of state and education? To validate your weights, you can use any number of raking functions in R:* `survey::rake`, `iterake`, *or* `autumn`.