Compositional Model Inference

J. Michael Brick[1]
[1]Westat, 1600 Research Boulevard, Rockville, MD 20850

Abstract

The ability to conduct surveys using opt-in Web respondents has raised concerns about whether these samples are valid. Probability sampling theory is not applicable because the units are not subject to being sampled with a known and non-zero probability of selection. Frameworks have been proposed for Web opt-in surveys, but these generally have features that are not well suited for general-purpose surveys. This paper proposes a model-based framework for making inferences from non-probability samples that we refer to as a compositional approach. The paper outlines the assumptions required for making inferences from these types of samples, and suggests some evaluation measures for assessing the assumptions.

Keywords: non-probability, survey, model diagnostics

1. Introduction

Non-probability sampling (NPS), specifically methods such as opt-in web panels and other methods of recruiting respondents without a population frame, make inference using the design-based theory of probability sampling (PS) impossible. Nonetheless, many NPS surveys instead use some form of weighting the sample observations that mimic methods used in PS. One form adjusts the sample observation to population controls like poststratification (Stephan and McCarthy, 1958, pp. 148-149). Propensity score adjustment (PSA), where the propensity score is estimated to be proportional to the inverse of the conditional probability of sample being observed or selected, is another common adjustment method (Taylor, 2000; Terhanian and Bremer, 2000). The foundations for these approaches in NPS are not well defined, and this issue hinders investigations of the validity of the assumptions and the potential effects on the quality of NPS estimates.

The goal of this article is to describe a structure that could be used with current methods for weighting NPS and might also be applicable for PS that have substantial nonresponse or incomplete coverage. We refer to this as a "compositional approach" because it, like most NPS schemes, uses weighting based on auxiliary data on the composition of the population to make estimates of finite population totals. The weighting methods are not new (Lee and Valliant, 2009), but defining the inferential basis allows for investigations of the quality of estimates.

2. Poststratification and Inverse Probability Weighting

Poststratification and PSA have different origins and bases. Poststratification creates weights so that the weighted sums of the sample observations match known population totals. Let *s* denote the set of responding units. The poststratified weights for an equal probability sample, and as applied in most NPS, are $w_{ps,g} = N_g n_g^{-1}$ for all $k \in s_g$, where *g* denotes one of the *G* non-overlapping and exhaustive groups or poststrata, N_g is the number of population units, and n_g is the number of responding units in *g*. Raking or iterative proportional fitting is a multidimensional extension that controls the weighted sample sums so that they match on more than one dimension simultaneously (Brackstone and Rao, 1979).

When poststratification is used for NPS, the implicit outcome model is $E_O y_k = \mu + \alpha_g = \mu_g$ for all $k \in s_g$, *g*=1,…,*G* (Wu and Sitter, 2001). Under this model, the estimated total

$$\hat{y}_{ps} = \sum\nolimits_{k \in s} w_{ps,k} y_k = \sum\nolimits_{k \in s_g} N_g \bar{y}_g \qquad (1)$$

is unbiased for the population total, *Y*. With two or more dimensions, the model assumes no interactions and is $E_O y_k = \mu + \alpha_g + \beta_h$, for all $k \in s_g \cap s_h$, where α_g and β_h are the main effect of two dimensions with *h*=1,…*H* ($\sum\nolimits_{k \in s_g} \alpha_k = \sum\nolimits_{k \in s_h} \beta_k = 0$).

PSA was developed for causal analysis of observational studies to reduce selection bias due to differences between treated and control or untreated cases (Rosenbaum and Rubin, 1983). In this setting, the propensity is the conditional probability of being sampled for treatment given a set of covariates. PSA is also used in PS to reduce nonresponse bias. Inverse probability or propensity weighting (IPW) has also been used in observational studies to reduce selection bias (Robins, Rotnitzky, and Zhao, 1994). Lee (2006) and Lee and Valliant (2009) describe IPW for NPS that relies on a reference sample derived from a PS or census. We follow their approach.

As in casual analysis, we assume there are two potential outcomes that are response to the NPS: Web sample (denoted r_1) and response to the reference sample (r_0). The propensity is the conditional probability of observation *k* is from the NPS given a set of covariates, **Z** that are measured in both samples, $\pi_{Z_k} = \Pr\left(R_k = 1 \middle| \mathbf{Z}\right)$ where $R_k = 1$ if *k* is from the NPS and $R_k = 0$ if it is from the reference sample. The inverse of the estimated propensity for a particular observation from a NPS are used like a Horvitz-Thompson (1952) weight to produce estimates.

Lee (2006) estimates the propensity by combining the observations from the NPS and PS, and running a logistic regression model with the common variables to predict response from the NPS. The combined sample is then categorized into classes or cells and the “propensity” is estimated computed as the estimated proportion of reference sample in each cell, $\hat{\pi}_{Z_k} = \frac{\hat{N}_{c,ps}}{N_{ps}} \frac{n_{nps}}{n_{c,nps}}$ for all $k \in s_g$ where the first term is the estimated proportion in class *c* from the PS and the second factor is the inverse of the observed proportion in class *c* from the NPS (n_{nps} is the total NPS sample and $n_{c,nps}$ is the number in class *c*). The inverse of the estimated propensity for a cell scaled to the population total is the IPW weight for observations from that cell. Estimates from the NPS produced with $w_{ipw,k} = \hat{\pi}_{\mathbf{Z}_k}^{-1} \frac{N}{n_{nps}}$.

To estimate the propensity for NPS, the standard assumptions described by Rosenbaum and Rubin (1983) apply. Specifically, a key assumption is strong ignorablity which includes the condition $0 < \pi_{\mathbf{Z}_k} \le 1$ for all *k* and $(r_0, r_1) \perp R \mid \mathbf{Z}$). These assumptions imply $E_M(R_k \mid \mathbf{Z}) = \pi_{\mathbf{Z}_k}$. In addition, for each class *c* the PS sample must be large enough so that $\frac{\hat{N}_{c,ps}}{N_{ps}}$ is estimated reliably and the NPS must be large enough for estimating $\frac{n_{nps}}{n_{c,nps}}$ reliably.

The IPW estimator of the population total is

$$\hat{y}_{ipw} = \sum\nolimits_{k \in s} w_{ipw,k} y_k . \tag{2}$$

This estimator is unbiased and consistent when the propensities are accurately estimated and the model assumptions hold.

2.1 Compositional Model

The compositional approach is related to both calibration weighting (Deville, Särndal, and Sautory, 1993) and augmented inverse probability weighting (AIPW) (Rotnitzky, Robins, and Scharfstein, 1998). Like doubly robust estimators, the estimator is approximately unbiased and consistent if either the outcome or missingness model holds. It is possible neither model may be appropriate for some variables in NPS, but the estimator may still reduce biases by accounting for more auxiliary data than is possible under either model alone.

The missingness and poststratification outcome models are:

$$E_M(R_k \mid \mathbf{Z}) = \pi_{\mathbf{Z}_k} \text{, and} \tag{3}$$

$$E_O y_k = \mu + \alpha_g + \beta_h , \tag{4}$$

where the two-dimensional outcome model is given for ease of presentation. We assume that raking dimensions are constructed so that raking converges.

The compositional approach first estimates a propensity and that is converted to a weight as described above, $w_{ipw,k} = e_{\mathbf{Z}_k}^{-1}$ and these weights are scaled to the known population total, $\sum_{k\in s} w_{ipw,k} = N$ for convenience. The second step adjusts the IPW weights to known control totals of the population, in addition to *N*. The properties of these weights, w_k, are:

1) $w_k > 0 \;\forall\; k \in s$.
2) $\sum_{k\in s} w_k \boldsymbol{\delta}_k = \mathbf{N}$, where $\boldsymbol{\delta}_k$ is a *p*x1 vector of 0 and 1's indicating group membership, and $\mathbf{N}$ is a *p*x1 vector of population counts of units in the groups.
3) Estimates of totals are linear in the weights, $\hat{y}_{com} = \sum_{k\in s} w_k y_k$. Other estimators considered are functions of weighted totals.
4) The estimates are approximately unbiased and consistent if either model *M* or *O* holds.

The first three properties follow directly from the construction of the weights. Suppose the outcome model for poststratification holds, then

$$E_O\hat{y}_{com} = \sum_g \sum_{k\in s_g} N_g \frac{w_{ipw,k}}{\sum_{k\in s_g} w_{ipw,k}} E_O y_k = \sum_g N_g(\mu+\alpha_g) = Y \ .$$

Consistency when either model holds follows from the WLLN.

For the one-dimensional case, the outcome model is $E_O y_k = \mu_g$ for all $k \in s_g$; the missingness model is $E_M(R_k \,|\mathbf{Z}_k) = \pi_{\mathbf{Z}_k}$. The compositional restriction involves $\boldsymbol{\delta}_k$, a *G*x1 vector where all entries are 0 except it is 1when unit *k* is in the poststratum or group and $\mathbf{N} = (N_1, N_2, ...N_G)'$ is the *G*x1 vector of the number of population units in the postrata ($\sum_{g=1}^{G} N_g = N$). The poststratified weight is $w_k = N_g \hat{N}_{ipw,g}^{-1} w_{ipw,k}$ for unit *k* in poststratum *g*, where $\hat{N}_{ipw,g}^{-1} = \sum_{k\in s_g} w_{ipw,k}$. The estimated compositional total,

$$\hat{y}_{com} = \sum_g \ \sum_{k\in s_g} \frac{N_g}{\hat{N}_{ipw,g}} w_{ipw,k} y_k = \sum_g \frac{N_g}{\hat{N}_{ipw,g}} \tilde{y}_g \, , \tag{5}$$

where $\tilde{y}_g = \sum_{k\in s_g} w_{ipw,k} y_k$.

For the two-dimensional case, the outcome model is $E_O y_k = \mu + \alpha_g + \beta_h$ for all $k \in s_g$ and $k \in s_h$. Now $\boldsymbol{\delta}_k$ is a vector where the first *G* entries are 0 except it is 1when unit *k* is in the *g*th group of dimension one and the last *H* entries are constructed the same way for the dimension two variable, with *p*=*G*+*H*. The vector of population totals is $\mathbf{N} = (N_{g1}, N_{g2}, ...N_{gG}, N_{h1}, N_{h2}, ...N_{hH})'$ and $\sum_{i=1}^{G} N_{gi} = \sum_{i=1}^{H} N_{hi} = N$. The compositional weights are the raked IPW weights, which can be written as the IPW weights multiplied by row and column factors ($w_k = \hat{\alpha}_g \hat{\beta}_h w_{ipw,k}$, $\forall\, k \in s_g \cap s_h$) . See Deville, Särndal, and Sautory (1993). The estimator is

$$\hat{y}_{com} = \sum\nolimits_g \sum\nolimits_h \ \sum\nolimits_{k \in s_g \cap s_h} w_{ipw,k} \hat{\alpha}_g \hat{\beta}_h y_k = \sum\nolimits_g \sum\nolimits_h \hat{\alpha}_g \hat{\beta}_h \tilde{y}_{gh} \ , \qquad (6)$$

where $\tilde{y}_{gh} = \sum\nolimits_{k \in s_g \cap s_h} w_{ipw,k} y_k$.

While similarities of the compositional approach to calibration are obvious, the relationship to AIPW is also strong. The fitted value for observation $k \in s_g$ under the poststratification model is $\bar{y}_{ipw,g} = \hat{N}_{ipw,g}^{-1} \tilde{y}_g$, and the compositional estimator can be written as

$$\hat{y}_{com} = \sum\nolimits_{k \in s} w_k y_k = \sum\nolimits_g \sum\nolimits_{k \in s_g} w_k (y_k - \bar{y}_{ipw,g}) + \sum\nolimits_g \sum\nolimits_{k \in s_g} w_k \bar{y}_{ipw,g} \ . \qquad (7)$$

This is a weighted combination of the residual, $(y_k - \bar{y}_{ipw,g})$, and the fitted estimate. Thus, the compositional estimator corresponds to the construction of the AIPW estimator as a weighted sum of residuals and fitted values (Robins, Sued, Lei-Gomez, and Rotnitzky, 2007).

2.2 Variance Structure

The approach thus far is a mean model, and to complete the model specification, a variance structure is needed. We assume a population structure with clustering generates the data. The specifics depend on the data collection mechanism. Assume clusters exist and that observations within a cluster may be correlated, but those in different clusters are independent. Redefine the sample observations as y_{ck} for observation *k* in cluster *c* for *c*=1,…,*C*. If both the number of clusters and the number of observations per cluster are relatively large but the overall proportion of the population observed is small, an ultimate cluster variance estimation technique such as used in PS could be applied. We employ resampling methods to capture the variability associated with the multiple steps.

With *C* clusters, define the jackknife variance estimator of the population total using standard delete-one cluster jackknife. The replicate weights are created as follows:

1) Replicate IPW weights are computed by estimating propensity weights *C* times, where the sample observations from the *r*th cluster are dropped in the logistic regression for *r*=1,…,*C*. The *r*th replicate IPW weight for sample unit *k* is the inverse of the propensity score scaled to *N*, say $w_{ipw,k}^{(r)}$ if $k \notin r$ and $w_{ipw,k}^{(r)} = 0$ if $k \in r$.
2) These weights are poststratified or raked to match the compositional population totals. For example, with poststratification the *r*th replicate weight is $w_k^{(r)} = \dfrac{N_g}{\hat{N}_{ipw,g}^{(r)}} w_{ipw,k}^{(r)}$ for $k \in s_g$ where $\hat{N}_{ipw,g}^{(r)} = \sum\nolimits_{k \in s_g} w_{ipw,k}^{(r)}$.

3) Replicate estimates are

$$\hat{y}_{com}^{(r)} = \sum_{k \in s} w_k^{(r)} y_k \ , \tag{8}$$

and the estimator of the variance of $\hat{y}_{com}$ is

$$v(\hat{y}_{com}) = \frac{C-1}{C} \sum_{r=1}^{C} \left(\hat{y}_{com}^{(r)} - \hat{y}_{com} \right)^2 \ . \tag{9}$$

With large samples, confidence intervals are based on normal theory approximations. The 95 percent confidence interval is estimated by $\hat{y}_{com} \pm 2\sqrt{v(\hat{y}_{com})}$. The compositional estimator is unbiased and consistent when the model assumptions hold.

3. Case Study

A collaboration between Pew Research Center, SurveyMonkey, and Westat led to the development and implementation of a set of coordinated surveys in 2014. Our primary interest here is the online NPS survey administered to the SurveyMonkey Audience panel. Since 2011, more than four million people joined the SurveyMonkey panel, with the only incentive being a 50-cent contribution to a charity of their choice for each survey taken. The survey was conducted from September-October 2014 with 5,301 adult panelists living in the United States.

The reference sample was a PS of adults conducted by Westat at about the same time using mail and an address-based sample. The survey contained many of the same items in the SurveyMonkey instrument. The survey had 2,668 completed interviews and an American Association for Public Opinion Research response rate (RR3) of 29 percent. The mail weights were the product of base weights (inverse of sampling the address and one adult per household) and raked to totals of the adult population using seven raking dimensions based on American Community Survey and National Health Interview Survey estimates.

Four sets of weights were computed for the Web survey using the methods described earlier. The first set of weights raked the unweighted responses to the same seven dimensions used for the mail survey and is denoted as *Raking*. The second set of weights are called *IPW-L* weights, with propensities estimated from a logistic regression that selected significant variables from all the substantive common variables in the mail and Web surveys. Variables used in the raking were excluded. The variables were recoded into dichotomies and four main effects and three interactions were included in the final model. Propensities for the combined set were partitioned into quartiles to create cells for weighting. No raking adjustment was made for the IPW-L weights.

The other two sets of weights followed the two-step compositional approach, with IPW and raking adjustments. The first of these, called *Comp-L*, raked the IPW-L weights to the seven dimensions from the mail survey. The other computed the IPW by forming 16 cells based on 4 binary variables (trust people, trust media, trust public schools, use Internet daily) identified by the nearest neighbor option in the MatchIt software in R (Ho, Imai, King, and Stuart 2007). These weights were raked to the same seven dimensions. This set of weights is called *Comp-N*.

A jackknife, delete-1 cluster approach is used to compute the variances of the NPS estimates. Clusters are those responses from the same metropolitan statistical area (MSA). Responses from MSAs with less than two respondents are randomly distributed to the clusters. A total of 100 clusters were created, each containing between 14 and 378 responses.

3.1 Comparison of NPS Web and Mail Survey Estimates

We begin by comparing the 34 substantive estimates from the NPS to those from the PS mail survey. Since the mail estimates may be biased due to nonresponse and other errors, the differences are not bias estimates.

The estimates from the mail and Web surveys do not differ dramatically, and there are no major differences between the weighting methods. The median absolute difference between the Web and mail survey were 2.9, 2.2, 2.9, and 2.7 percentage points for the 4 weighting methods respectively. The estimates from the two compositional weights are a bit closer to the regression line as evidenced by the standard deviations of the differences (4.5, 2.7, 2.4, and 2.1 percentage points, respectively).

3.2 Assessment of Model Assumptions

Since all persons do not access the Internet either at home or in other places, the strong ignorability assumption ($0 < \pi_{\mathbf{z}_k} \leq 1$ for all *k*) is violated, at least partially. The strong ignorability assumption is often evaluated using data from the study by examining "common support" assumption. The distributions of the estimated propensities for the mail and Web surveys when the propensities are estimated by logistic regression were compared, and there was limited overlap in the distributions. The separation is troubling especially because the additional separation due to households without access to the Internet being excluded from the NPS is not even considered in this examination. The median propensity for the Web survey is above the third quartile of the mail survey. For the IPW estimates of Comp-N, we examined the magnitude of the IPW adjustments in each of the 16 matching cells. The relative magnitude showed considerable variation, with 8 large factors for respondents who do not use the Internet daily. This finding is consistent with the separation seen for the IPW-L.

Next, define the relative effect of poststratification on the bias as

$$RE(y) = 100\left(\frac{\sum_g N_g \hat{N}_{ipw,g}^{-1} \tilde{y}_g - \sum_g \tilde{y}_g}{\sum_g \tilde{y}_g}\right). \tag{10}$$

Large values of *RE* do not necessarily imply reduced bias, although we would hope the adjustment would reduce bias. For most items RE(y) is small and only exceeds 10 or 20 percent for a handful of items, suggesting raking did not substantially reduce biases.

Even if raking has little effect, the NPS estimates are unbiased if the outcome model holds. With poststratification, the outcome model implies that estimates of a domain within a poststratification cell have the same mean. This assumption can be studied by looking at standardized residuals estimated as

$$\hat{\tau}_{gd} = (\overline{y}_{gd} - \overline{y}_g) / \sqrt{\hat{\text{v}}(\overline{y}_{gd} - \overline{y}_g)} \ , \tag{11}$$

where $\overline{y}_{gd}$ is the estimated mean in domain *d* and poststratum *g* and $\hat{\text{v}}(\overline{y}_{gd} - \overline{y}_g)$ is the variance of the difference estimated using the replication scheme. It is reasonable to assume the standardized residuals are approximately normally distributed even though the outcome model does not assume normality.

A Q-Q plot of 2,105 standardized residuals for the responses to ratings of general health showed many large residuals. While the distribution of the residuals is not a direct assessment of the common mean assumption, it raises many concerns.

Variances of the Web estimates of Comp-L and Comp-N were computed. The median design effect (deff) for the Comp-L estimates was 14.9 with a mean of 48.3. These are extraordinarily large. For Comp-N, the median deff was 5.5 and mean is 6.5. The effective sample size is starkly lower than if the computations were based on variation in the weights.

Discussion

The compositional model provides a framework for some of the more popular weighting approaches used in NPS. Even though the evaluations showed there were problems with the model assumptions, the NPS estimates themselves were similar to those from the mail PS. For many uses, the estimates from the low-cost Web survey are likely to be fit-for-use. Even when this is the case, the estimates are much less precise than might be indicated by the sample size.

References

Brackstone, G.J., and Rao, J.N.K. (1979). An investigation of raking ratio estimators. *Sankhya*, 97-114.

Deville, J.C., Särndal, C.E., and Sautory, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, *88*(423), 1013-1020.

Ho, D., Imai,K., King, G., and Stuart, E. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis* 15, 199-236. http://gking.harvard.edu/files/abs/matchp-abs.shtml.

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

Lee, S. (2006). Propensity score adjustment as a weighting scheme for volunteer panel web surveys. *Journal of Official Statistics*, *22*, 329-349.

Lee, S., and Valliant, R. (2009). Estimation for volunteer panel web surveys using propensity score adjustment and calibration adjustment. *Sociological Methods & Research,* 37, 319-343.

Robins, J.M., Rotnitzky, A., and Zhao, L. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association,* 89, 846-866.

Robins, J., Sued, M., Lei-Gomez, Q., and Rotnitzky, A. (2007). Comment: Performance of double-robust estimators when “inverse probability" weights are highly variable. *Statistical Science*, 22, 544-559.

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

Rotnitzky, A., Robins, J.M., and Scharfstein, D.O. (1998). Semiparametric regression for repeated outcomes with nonignorable nonresponse. *Journal of the American Statistical Association*, *93*(444), 1321-1339.

Stephan, F.F., and McCarthy, P.J. (1958). *Sampling opinions: An analysis of survey procedure.* New York: JohnWiley & Sons.

Taylor, H. (2000). Does Internet research work? Comparing online survey result with telephone survey. *International Journal of Market Research*, 42, 58-63.

Terhanian, G., and Bremer, J. (2000). *Confronting the selection-bias and learning effects problems associated with Internet research.* Research paper. Harris Interactive.

Wu, C., and Sitter, R.R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, *96*, 185-193.