Proceedings of Statistics Canada Symposium 2016
Growth in Statistical Information: Challenges and Benefits

Better Adjusted Weights for Respondents in Skewed Populations

Glen Meeden, Zack Almquist and Charles Geyer[1]

Abstract

In the standard design approach to missing observations, the construction of weight classes and calibration are used to adjust the design weights for the respondents in the sample. Here we use these adjusted weights to define a Dirichlet distribution which can be used to make inferences about the population. Examples show that the resulting procedures have better performance properties than the standard methods when the population is skewed.

1. Introduction

In the design based approach to survey sampling, information about the population is used when selecting the design and after the sample has been observed to adjust the sampling weights. In the formal Bayesian approach to survey sampling, information about the population is incorporated in a prior distribution. After the sample is observed, inferences are based on the posterior distribution of the unobserved units in the population given the values of the observed units in the sample. The posterior distribution does not depend on the sampling design. In large scale surveys Bayes methods have been little used in practice because it is difficult to find sensible prior distributions. In particular, it is not clear how to incorporate into a prior distribution the type of information contained in the design, which is then used in calibration to account for missing observations. An advantage of the Bayesian approach is that one can find point and interval estimators for many population parameters by simulating from the posterior distribution. Here we will argue that one can combine features from both approaches, and this results in improved inferences. Estimates will be based on a posterior distribution, but, paradoxically, one need not specify a prior distribution. Instead, after the sample has been observed, one selects a posterior distribution that depends on the sampling design and all the other information that is available to the statistician.

2. The Horvitz-Thompson Estimator

Consider a finite population of size N, where y is the variable of interest and x is an auxiliary which carries information about y. Given a sampling design of fixed sample size n, let π_i denote the probability that unit i is selected in the sample. If it is assumed that the y_i's are roughly proportional to the x_i's then a popular design is random sampling without replacement, where the probability that unit i is selected is proportional to x_i. In this case $\pi_i = n(x_i / T_x)$, where T_x is the population total of the x values. If $\text{wt}_i = 1/\pi_i$ then this is the weight assigned to unit i. A unit gets a small weight if there are just a few other units in the population with similar x values and a large weight if there are many other units with a similar value of x. Given a unit in the sample its weight represents how many other units in the population that are similar to it.

If s denotes the labels of the units in the sample then $\text{HT} = \sum_{i \in s} \text{wt}_i * y_i$ is the Horvitz-Thompson estimator of the population total. It is easy to check that it is an unbiased estimator and that for any sample s we have $\sum_{i \in s} \text{wt}_i x_i = T_x$. In other words, the HT estimator is calibrated on x (Särndal, 2007). In the design approach to survey

[1] School of Statistics, 313 Ford Hall, 224 Church ST S.E., University of Minnesota, Minneapolis, MN 55455-0460

sampling these weights play an important role. Not only do they define the estimator but are used to get an estimate of variance for the estimator.

If the design is simple random sampling without replacement, then $\text{wt}_i = N/n$ and for every sample $\sum_{i\in s}\text{wt}_i = N$. For most other designs this is not the case. For this reason the HT estimator is not robust against the assumption that $y_i \propto x_i$. If one replaces each y_i with $y_i + \lambda$ for some fixed number λ then the behavior of the HT estimator becomes much less desirable. It still is unbiased but its absolute error can become much larger and its confidence intervals much wider. This happens because the sum of the weights in a sample does not equal the population size. For more discussion of this point see Strief and Meeden (2014). If one renormalizes the weights so that they sum to the population size then they will no longer be calibrated for x. A better thing to do, we believe, is to find weights which are close to the design weights, are calibrated for x and sum to N. This is a quadratic programming problem and there are many computer packages which will find the solution.

In the next section we will discuss this approach in more detail when we have missing observations and compare it to more Bayesian approach.

3. Missing Observations

3.1 The standard approach

The standard approach to observations missing at random is to assume that for each i there is a probability, say ψ_i, that unit i is observed when it is included in the sample. This response probability is assumed to be independent of the sampling design and so the probability that we actually observe y_i in our sample is $\pi_i\psi_i$ which then yields a weight for the unit. Unfortunately the ψ_i's are almost never known. To overcome this lack of knowledge of the ψ_i's the statistician uses the observed values of x to construct weighting adjustment classes with the hope that respondents and nonrespondents in the same class are similar, that is, the ψ_i's within each class are roughly constant. This assumes that the x values are known for each unit in the full sample. Then within each class the total weight of the units in the sample falling in this class is split equally among the respondents in the class. Let s_r be the labels of the respondents in the sample. Given a sample s, for $i \in s_r$ let wt_i be its adjusted weight found by this procedure.

In general, the wt_i's will neither be calibrated nor sum to N. Let $\gamma = \{\gamma_i : i \in s_r\}$ denote a possible set of weights. As we indicated just above we recommend finding a new set of weights, say γ^*, which is a solution to the problem

$$\min_{\gamma} f(\gamma) = \sum_{i\in s_r}(x_i/\text{wt}_i)(\gamma_i - \text{wt}_i)^2$$

subject to the constraints

$$\sum_{i\in s_r} x_i\gamma_i = T_x \quad \text{and} \quad \sum_{i\in s_r}\gamma_i = N,$$

where we assume T_x is known. One may also include the additional constraints that $p_i \le \text{bd}$, $i \in s_r$, where $0 < \text{bd} < 1$ is some number selected by the statistician.

Our choice of the function measuring how far a set of weights is from the wt_i's is a popular one but other common choices will not change the story very much. We let CHT denote the estimator based on the set of weights found as the solution to the above problem.

3.2 A stepwise Bayes approach

Implicit within the standard approach is the assumption that the only possible values for units in the population are those that have appeared in the sample. Given a sample, let $p = \{p_i : i \in s_r\}$, where p_i is the proportion of units in the population that are assumed to be identical to respondent i. We can think of p as an unknown parameter that we wish to estimate. In fact, any set of weights for the respondents can be converted into an estimate of p simply by dividing by the the total sum of the weights. If p is an unknown parameter then given a $0 < \text{bd} < 1$ a natural parameter space for p is the polytope which is the collection of vectors p satisfying

$$\sum_{i \in s_r} p_i = 1 \quad \text{and} \quad \sum_{i \in s_r} x_i p_i = \mu_x \quad \text{and} \quad 0 \le p_i \le \text{bd}, \text{ for all } i \in s_r,$$

where μ_x is the mean of the x values of the population. We denote this polytope by Γ_{bd}. It is usually the case that for the $\gamma*$ defining CHT, γ^* / N will be a point in the relative boundary of Γ_{bd}.

We will take a stepwise Bayesian (STB) approach to the problem of estimating p by defining a "posterior" distribution for p after the sample has been observed. Since we are using the STB approach we do not need to define a single prior distribution on which inferences will be based. An introduction to this approach can be found in Ghosh and Meeden (1997). Because of space limitations, we will omit the STB justification for the methods given here.

Our posterior distribution will also depend on $\text{wt} = \{\text{wt}_i : i \in s_r\}$. We get this posterior in two steps. Let $\hat{w} = \{w_i : i \in s_r\}$, where $w_i = n_r \text{wt}_i / \sum_{j \in s_r} \text{wt}_j$. In the first step we use $\hat{w}$ as the parameter for a Dirichlet distribution which is restricted to the set Γ_{bd}. (We will explain a bit later how we choose bd.) We then use the R (R Core Team, 2016) package `polyapost` (Meeden et al., 2015) to calculate the expectation of p, say $\hat{p}$ under this distribution. Note that $\hat{p}$ is always in the relative interior of Γ_{bd}, in contrast to γ^* / N which is usually on the relative boundary.

Note that $\hat{p}$ depends on the design but also takes into account the information in x. Then $N\hat{p}_i$ is a sensible weight for unit i and $\sum_{i \in s_r} N\hat{p}_i y_i$ is an estimate of the population total of y. We denote this estimator by WD.

We still need a way to evaluate the variability of the estimator WD. Rather than using the distribution which gave us $\hat{p}$ we follow Strief and Meeden (2014) and define a second Dirichlet distribution. The parameter for this distribution is $\alpha = n_r \cdot \hat{p}$. From the stepwise Bayes prospective we can base our inferences for the vector p on the Dirichlet distribution with parameter vector α. Note this distribution is no longer restricted to Γ_{bd} but lives on the full $n_r - 1$ dimensional simplex (of probability vectors whose components are nonnegative and sum to one). Because this posterior allows for vectors of p which only satisfy the constraint on average, this helps to account for the fact that the $\hat{p}_i$'s are estimates whose true values are not known.

Given this posterior, it is straightforward to find the posterior variance of the corresponding estimate of the population total. We will assume it is approximately normal and use the usual normal approximation to get an approximate 95% confidence interval for the population total. We will call this distribution the weighted Dirichlet posterior. For other population parameters of interest one can just simulate from this distribution to find point and interval estimates. We want to emphasize that this posterior is not based on any model assumptions on how x and y are related. It just assumes that units which have similar x values will tend to have similar y values and that the response probability is a "smooth" function of x.

We note that Rao and Wu (2010) also base inferences on a Dirichlet distribution but their justification is different from that given here.

4. Simulation Results

We constructed three populations with $N = 10{,}000$ units each based on the same population of x values. Since we are interested in skewed populations we let the auxiliary variable x be a random sample from a lognormal distribution whose mean and standard deviation of the log are $\sqrt{e}/2$ and $\sqrt{e^2 - 2e}/2$, where e is the base of natural logarithms. The minimum value, the 25%, 50%, and 75% quantiles, and maximum value for this population are 0.09, 1.41, 2.26, 3.60, and 26.73.

In the first population, the y_i's were conditionally independent given the x_i's, and the conditional distribution of y_i given x_i was normal with mean $8x_i^2$ and standard deviation 0.4. The correlation between x and y is 0.825.

In the second population we let the mean function of the y be a function of x that is not the identity function. This mean function, m, was defined as follows,

$$m(x) = \begin{cases} 1000(2-x)^2, & x \le 2 \\ 4(x-2)^2, & x > 2 \end{cases}$$

For this population the correlation between x and y is -0.35.

In the third population the conditional distribution of y_i given x_i was normal with mean $1.5x_i$ and standard deviation 3. The correlation between between x and y is 0.75. In these last two populations the distribution of the y_i's were conditionally independent given the x_i's just like in the first.

Next we need to model the missing observations. To this end we need to define the vector of ψ_i's. We will assume that the units with a large x value will be less likely to respond than units with a small value of x. For convenience the population is labeled so that x is an increasing function of the indices. We begin by considering the vector which is the sequence that goes from 0.4 to 0.2 in 9,999 equal steps. This is a smooth decreasing function of the labels. In practice, we would not expect the response vector to be so smooth. So we added independent random errors from a normal distribution with mean 0 and standard deviation 0.05 to each component. We then applied to the resulting vector the linear function that rescaled its components back to the interval $[0.2, 0.4]$. This was the vector ψ which we used to define the probability that a unit responds in our simulations. Of course, none of the estimators we compute are based on knowing ψ.

We used two different sampling designs in our simulations. The first was simple random sampling. Let v be the vector that goes from 0.3 to 0.8 in 9,999 equal steps. The second design used sampling proportional to v. Since the the x_i's are an increasing function of the labels this will result in more units with larger values of x_i in the sample.

For each of the three populations, we took 500 samples of size $n = 150$ for each of the two designs. In each case the average number of respondents was about 44. For each set of simulations for the CHT estimator and the WD estimator we calculated their average value, their average absolute error, the average length their approximate 95% confidence interval and the frequency of their intervals containing the true population total.

Under simple random sampling the ratios of the average absolute error of the CHT estimator to the average absolute error of the WD estimator for the three populations were 1.24, 1.37 and 0.93. For the second design these ratios were 1.12, 1.27 and 0.93. We have done other simulations which will not be presented here where the populations were less skewed. In these cases we found that the behavior of the two estimators are quite similar except when $y_i \propto x_i$, where the the CHT estimator does slightly better. This suggests that unless there is strong evidence that $y_i \propto x_i$ one should use the WD estimator especially when the population of interest is skewed.

Both estimators were nearly unbiased. For example, in the first two populations both estimators are biased downwards by just over one percent. A natural question is how do the CHT weights differ from the WD weights? The WD estimator gives more weight to the units with the largest and smallest x values and hence correspondingly somewhat less weight for the units in the middle. For the first population when the design was simple random sampling for a

given sample considered the units with the minimum value, 0.25 quartile, median, 0.75 quartile, and maximum value of x. The average weights assigned to these units under the CHT estimator were 181, 202, 217, 249 and 263. Whereas for the WD estimator these average weights were 201, 201,213, 233 and 331.

The frequency of coverage of the WD approximate 95% confidence intervals for the first population was 0.984 and 0.976 for the two designs. For the second population these numbers were 0.948 and 0.896 while for the third they were 0.962 and 0.992.

Our first population is similar to one discussed in Dorfman (1994) and Rao et al. (2003). There it was observed that assuming a linear relationship between y and x when in fact it was quadratic can lead to poor confidence intervals whose actual coverage probabilities are far from their nominal levels. The problem is that one cannot use a quadratic model when the population total of the x_i^2 values are not known. To investigate this further we selected 500 simple random samples of size 44 from the first population where there were no missing observations. Recall that 44 is the average number of responders we saw in our previous simulations. The ratio of the average absolute error for the regression estimator from this set of simulations to the average absolute error of the WD estimator from the first set of simulations was 1.06. The associated 95% confidence intervals for the regression estimator contained the true population total only 67.8% of the time and the ratio of the average length of the two methods was 0.36.

We did a second similar simulation for the second population and the behavior to the regression estimator with no missing observations was similar to the WD estimator with missing observations.

Finally, we did a similar simulation for the third population. Here the ratio of the average absolute error of the regression to that of the average absolute error of the WD estimator in the first set of simulations was 0.88. We also computed the average absolute error of the HT estimator and of its CHT version which ensures that the weights are both calibrated for x and sum to N. The average absolute error of the second was just 1% larger than that of the regression estimator while the average absolute error for the HT estimator was 50% larger then that of the regression estimator. So even though $y_i \propto x_i$ in this population the HT estimator performs poorly because the population is so skewed.

In these simulations, given a sample, we set $\mathrm{bd} = 5/n_r$, where n_r is the number of respondents in the sample. If instead of 5 we had used in the numerator of bd any number ranging from 3 to 9 our results would have not changed much. But setting $\mathrm{bd} = 2/n_r$ would be too small. So we see that our method is quite robust against the choice of bd.

5. Comments

All the simulations done here used the R package `polyapost` which, since version 1.4-2, allows one to compute expectations of any Dirichlet distribution constrained to a polytope (the set satisfying a finite family of linear equality and inequality constraints). This makes it easy for anyone familiar with R to compute the WD estimator. For simplicity we have only considered the situation with just one auxiliary variable but one can incorporate more than one variable in the set of constraints.

The stepwise Bayes approach outlined here is an extension of some of the ideas in Strief and Meeden (2014). Those authors argued that the sampling design did not matter after the data was collected, and they used linear constraints on the auxiliary variables and the uniform distribution over a polytope to come up with a set of weights. Subsequent work has shown that the approach given here, which incorporates the design, yields better results. We have seen that the resulting procedures make no model assumptions about how y and x are related. They combine pre-sample and post-sample information available to the sampler in a coherent and objective manner and can yield procedures with good design properties for skewed populations where standard methods fail. Further investigation needs to be done on the behavior of the approximate confidence intervals of our weighted Dirichlet posterior. Alternatively, one could use our weights in standard design based methods to get an estimate of variance of our point estimator.

References

Dorfman, A. H. (1994). A note on variance estimation for the regression estimator in double sampling. *Journal of the American Statistical Association*, 89:137–140.

Ghosh, M. and Meeden, G. (1997). *Bayesian Methods for Finite Population Sampling*. London: Chapman and Hall.

Meeden, G., Lazar, R., and Geyer, C. (2015). R package polyapost: Simulating from the polya posterior, version 1.4-2. http://CRAN.R-project.org/package=polyapost.

R Core Team (2016). *R: A language and evironment for statistical computing*. R Foundation for Statistical Computing. http://R-project.org.

Rao, J. N. K., Jocelyn, W., and Hidiroglou, M. A. (2003). Confidence interval coverage properties for regression estimators in uni-phase and two-phase sampling. *Journal of Official Statistics*, 19:17–30.

Rao, J. N. K., and Wu, C. (2010). Bayesian pseudo empirical likelihood intervals for complex surveys. *Journal of the Royal Statistical Society, Series B*, 72:533–544.

Särndal, C. (2007). The calibration approach in survey theory and practice. *Survey Methodology*, 33:99–119.

Strief, J., and Meeden, G. (2014). Objective stepwise bayes weights in survey sampling. *Survey Methodology*, 39:1–27.