Indirect Sampling in Case of Asymmetrical Link Structures

Torsten Harms[*]

Abstract

Estimation in case of indirect sampling as introduced by Huang (1984) and Ernst (1989) and developed by Lavalle (1995) Deville and Lavalle (2006) extends the application of sampling theory for settings when the sampling frame and target population are not identical, but rather connected through specified links. This approach has found widespread application, particularly through the generalized weight share method (GWSM), and can be applied generally to sample from networks. In order to ensure unbiased estimates, all inbound links for any sampled unit in the target population must however be known. While this is not a problem in many types of link structures (such as family relationships or household association), it prevents important applications in networks with asymmetrical link structures, such the world-wide web, lose social relationships, or refreshment samples. This paper extends the application of the generalized weight share method (GWSM) to produce design-unbiased estimators even for such situations, where the inbound links for each element are generally unknown. This is achieved by adequately reflecting the link-structure which is partially revealed during the sampling process. An empirical evaluation as well as a comparison with other estimation methods, particularly Monte-Carlo approaches, evaluates the proposed method.

Key Words: Indirect, sampling, linkage, networks

1. Introduction and Motivation

In survey sampling it often occurs that a sample $s \subset U$ is taken from a sampling frame U, however the properties of interest lie within a population U' which is not sampled directly but rather assessed through following links $c_{i,i'}$ that connect the elements $i \in U$ with $U' \in s'$. Let i and i' to be connected if $c_i = 1$ and i and i' unconnected if $c_{i,i'} = 0$. The sampling of $s \subset U$ thus defines a sample $s' \subset U'$ which is given by $\{i' \in U' : \exists i \in s, c_{i,i'} = 1\}$. The reasons for this so-called indirect sampling can be manifold: In many cases, indirect sampling is carried out, simply because no adequate sampling frame (e.g., lists or databases) or established method of approach exists for U'. Another reason might be already available information or samples on U which greatly reduce the cost of fieldwork. And finally, in panel surveys it is often desirable to allow for some sort of linkage between the initially sampled population U and the current population U'. The classic example are newborns that are linked to their initially sampled parents.

Except for the trivial case of a one-on-one linkage between U and U', the estimation of properties in U' poses some methodological challenges. This has been addressed by Sirken (1970) and Huang (1984) who coined the term multiplicity estimates to allow for multiple links between elements of the sampling frame and target population. A unified approach based on the works of Ernst (1989) has been developed by Lavalle (1995) and Deville and Lavalle (2006) under the term Generalized Weight Share Method (GWSM).

Consider the estimation of the population total $T_{y'}$ of the variable $y'_{i'}$ ($i' \in U'$). The general weight share method assigns each element $i' \in s'$ a new weight $w_{i'}$, which can be used to construct a Horvitz-Thompson type estimator $\hat{T}_{y'}^{\mathrm{GWSM}} = \sum_{i' \in s'} y'_{i'} w_{i'}$.

In order to arrive at the weights $w_{i'}$, the initial design weights d_i of $i \in s$ are "shared" according to certain link functions $l_{i,i'}$ in the following way: $w_{i'} := \sum_{i' \in s} l_{i,i'} d_i$.

[*] ISS International Business School; contact: torsten.harms@gmx.com

The link function depends on both, the linkage between the elements $i \in U$ and $i' \in U'$ as well as the desired form of "weight sharing". A detailed description can be found in Lavalle (2007). A commonly used link function is given by the so-called fair share approach with $l_{i,i'} := \frac{c_{i,i'}}{\sum_{j \in s} c_{j,i'}}$. In the case of only one link from $i \in U$ to $i' \in U'$, the link function $l_{i,i'} = 1$. In the case of 2 inbound links from $i, j \in U$ to $i' \in U'$, the link functions are $l_{i,i'} = l_{j,i'} = 0.5$.

The classical Horvitz-Thompson type form of the estimator $\hat{T}_{y'}^{\mathrm{GWSM}}$ and the high flexibility to incorporate any form of linkage and link function make the GWSM very appealing, particularly in large official or social surveys. As laid out by Lavalle (2007), variance estimators can be constructed and an extension to several stages of indirect sampling is straightforward.

In order to provide unbiased estimates, it can be shown that for all $i' \in U'$ the sum of the inbound links $\sum_{i \in U} l_{i,i'}$ must be equal to 1. This requires in particular that all links $c_{i,i'}$ must be known for all $i' \in s' \subset U'$. While this is not a challenge with so-called symmetrical links such as family or household association, there also exist many situations of so-called asymmetrical links, where the inbound links from all elements in the sampling frame are generally not known. Important cases include

- *Refreshment samples:* It is often unknown, if the refreshment sampling frame would generally lead to re-sampling of already sampled persons.
- *Web sampling on the internet:* Where by definition only outbound links from websites are known
- *Sampling in social networks:* Such as snowball sampling or by following asymmetrical links such as friendship

2. Existing research

The classical approach do deal with asymmetrical link functions is to obtain estimates of the number of inbound links. This can be either done by Markov-Chain Monte-Carlo Methods (MCMC), see Thompson (2006), or by using capture-recapture methods such as the so-called respondent-driven sampling (RDS) proposed by Heckathorn (1987).

In the simplest form of the MCMC approach, the links in the network are followed repeatedly and, following a certain "burn-in phase", the frequency of visits of $i' \in U'$ converges to a factor that is proportional to the number of inbound links to i'. There are some challenges associated with this approach, mainly that it basically requires following a large proportion of nodes (which makes it impractical for large population surveys) and that it has limitations in the case of unconnected parts of the graph defined by the links. In the classical case of 2-stage indirect sampling, following the link structure from U to U' terminates after 1 step at the target population U'. This requires probabilistic jumps back to the sampling frame U, which in essence would mean substantial over-sampling of U (and thus of U'). While the efficiency of the MCMC approach decreases substantially in the above described classical case of 2-stage indirect sampling with no outbound links at the second stage, it has gained popularity particularly for sampling in large, very well connected networks.

The other main approach to estimate the number of inbound links via Respondent driven Sampling (RDS) has also been applied successfully in practice. A challenge here might be the administration of the capture-recapture approach through identification cards. However this corresponds largely to the challenge of tracking the link function in the classical GWSM. Generally, however the classical theory to of the GWSM cannot be used in

RDS sampling which would also make it necessary to use different tools/software under this approach.

As a summary on the existing research. it can be said, that working and tested approaches exist for dealing with asymmetrical link functions. However, these approaches are not straightforward extensions of the GWSM and thus generally require a ”system change” for the user.

3. The Estimated Generalised Weight Share Method (EGWSM)

In order to extend the weight share approach for the above-mentioned settings where the inbound links are unknown, we propose a modified weight share approach that estimates the link function $l_{i,i'}$ based on the link structure that is revealed in the sample. Similarly to the general principle in sampling, this is achieved by reweighting the observed links in order to achieve unbiased estimates even for those situation that were not observed.

Given an estimator $\tilde{l}_{i,i'}$ for $l_{i,i'}$, the proposed estimator for the population total of y' of U' would be

$$\hat{T}_{y'}^{\text{EGWSM}} = \sum_{i' \in s'} y'_{i'} \sum_{i \in s} \tilde{l}_{i,i'} d_i$$

Where EGWSM stands for Expected Generalized Weight Share Method. Due to the linear form, it is sufficient ot show for the unbiasedness, that $\mathrm{E}[\tilde{l}_{i,i'}] = l_{i,i'}$.

Let $m_{i'}$ be the set of elements $i \in U$ with links to i' (thus $c_{i,i'} = 1$). In the case of asymmetrical link structures, given a sample $s \subset U$, we can generally only observe a subset of links to $i' \in U$, namely those from the sampled elements $i \in \tilde{m}_{i'} := m_{i'} \cap s$. The estimator for $\tilde{l}_{i,i'}$ must thus be based on the set $\tilde{m}_{i'}$ rather than the unknown $m_{i'}$.

The construction principle for $\tilde{l}_{i,i'}$ can be best described on an example: Consider the elements i, j, k linking to i' and assume a fair share approach is used. Thus we have $m_{i'} = \{i, j, k\}$, and the true (but unknown) $l_{i,i'} = 1/3$.

Now consider the case that of i, j, k, only i is sampled. In this case $\tilde{m}_{i'} = \{i\}$ and we clearly need to set $\tilde{l}_{i,i'} = 1$ since we do not observe any other links and the sum of all inbound links to i' must be equal to one in order to ensure unbiasedness. (Note that the true $l_{i,i'} = 1/3$ thus $\tilde{l}_{i,i'}$ overestimates $l_{i,i'}$ in this case.)

In the case of 2 observed inbound links to i', say $\tilde{m}_{i'} = \{i, j\}$. We need to compensate for the the known overweighting in the cases with $\tilde{m}_{i'} = \{i\}$. Thus we set for this case $\tilde{l}_{i,i'} = 1 - \frac{1}{2}\frac{1}{\pi_{ij}}$ where π_{ij} is the joint selection probability of i and j. Note that the compensation factor $\frac{1}{2}$ is weighted with the inverse of detecting the two inbound links.

Similarly in the case of $\tilde{m}_{i'} = m_{i'}$, that is all units i, j, k are sampled, we have: $\tilde{l}_{i,i'} = 1 - \frac{1}{2}\frac{1}{\pi_{ij}} - \frac{1}{2}\frac{1}{\pi_{ik}} + \frac{1}{3}\frac{1}{\pi_{ijk}}$ where π_{ijk} is the joint selection probability of i, j, k.

Under the fair share approach, the general rule for $\tilde{l}_{i,i'}$ is given by:

$$\tilde{l}_{i,i'} = 1 + \sum_{m \subset \tilde{m}_{i'}} (-1)^{|m|-1} \frac{1}{|m|} \frac{1}{\Pr(m \subset s)}$$

An extension for general link functions can be found via the same principle via induction, but note that not in all cases a closed form of the estimator $\tilde{l}_{i,i'}$ for the link function might be possible.

As can be seen the link function $\tilde{l}_{i,i'}$ now depends on the sample and thus the variance of the estimator is expected to increase, compared to the situation of known links that is typically assumed in indirect sampling. Still, due to the finite size of the population and sample, the variance remains bounded and the estimator is by construction design-unbiased, if the sampling is such, that all inbound links can be detected with one sample (that is $\forall i' \in U' \exists s : m_{i'} \subset s$). Note that the estimator is only unbiased for the estimation of totals not the estimation of population shares as the number of elements in U' is generally unknown.

Due to the construction of our estimator $\hat{T}_{y'}^{\mathrm{EGWSM}}$, we assume the variance to increase compared to the weight share estimator $\hat{T}_{y'}^{\mathrm{GWSM}}$. This is due to fact that the fixed link function $l_{i,i'}$ is replaced by the estimator $\tilde{l}_{i,i'}$. However note that there might be situations, where the inbound links $l_{i,i'}$ are generally unknown for elements i not in s and thus the classical weight-share estimator cannot be constructed. As we can see by the formula for $\tilde{l}_{i,i'}$ the variance increase stays small either in situations where multiple inbound links are the exception (e.g., refreshment samples) and/or in situation where the multiple inclusion probabilities of elements in $m_{i'}$ are rather large. An example for the latter situation might be cluster sampling of families and then following links defined by family relationships.

So far, we have only considered single stage indirect sampling. In practice many indirect sampling schemes consist of more than one stage. Typical examples include panel surveys, where follow-up rules are used to follow participants or household through several panel waves. As already discussed by Lavalle (2007) the generalized weight share method (GWSM) can easily be extended to incorporate multi-level sampling. As our proposed estimator only replaces the known link function $l_{i,i'}$ with an estimator $\tilde{l}_{i,i'}$, this also applies to our proposed method.

The most common approach to deal with multi-stage sampling is to consider the stages sequentially: Based on the first sampling stage from U to U' we can consider the weights $w_{i'} := \sum_{i \in s'} \tilde{l}_{i,i'} d_i$ as "design-weights" for the "sample" $s' \subset U'$. Note that the terms "sample" and "design-weights" are not strictly correct as sampling only occurs in the selection of $s \subset U$. However s' with $w_{i'}$ can be used in the same fashion as a classical direct sample and the Horvitz-Thompson estimator corresponds to our proposed EGWSM estimator and is unbiased. In a subsequent stage following links from U' to U''. This sample s' and the design weights $w_{i'}$ are then used in the same fashion in the equivalent way as s and d_i in the first stage following U to U'. Similarly to the GWSM, also the EGWSM is also unbiased at this second stage, assuming the above mentioned regularity conditions to be met (most notably that for all elements $i'' \in U''$, the sampling design and follow-up rules are as such, that the full set of inbound links $m_{i''}$ can be detected via sampling (thus $\exists s' \subset U' : m_{i''} \subset s$).

4. Empirical Application

In order to empirically assess the performance of our proposed EGWSM estimator, particularly against the GWSM estimator but also other alternatives, a small simulation study was carried out with the following settings:

- *Population:* The populations U and U' are given by the sets $U = \{1, 2, \ldots, 10\}$ and $U' = \{1, 2, \ldots, 17\}$. The variable of interest y' was given by $y'_{i'} = i'$. Thus $T_{y'}$ = 153

- *Link function:* We assumed a links c between U and U' as given by figure 1. For the link function, we used the fair share approach with $l_{i,i'} = \frac{c_{i,i'}}{\sum_{i \in U} c_{i,i'}}$

- *Sampling design:* We used SI sampling with $n = 4$ in U. Note that, given the link structure with a maximum of 4 inbound links in U', this sampling scheme automatically fulfills our requirement, that for any $i' \in U'$, the full set of inbound links from $m_{i'}$ from U to U' can be detected in a single sample.

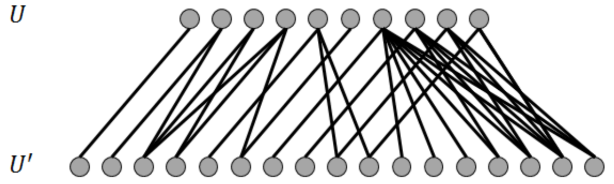

Figure 1: Simulation population solid lines describe existing link

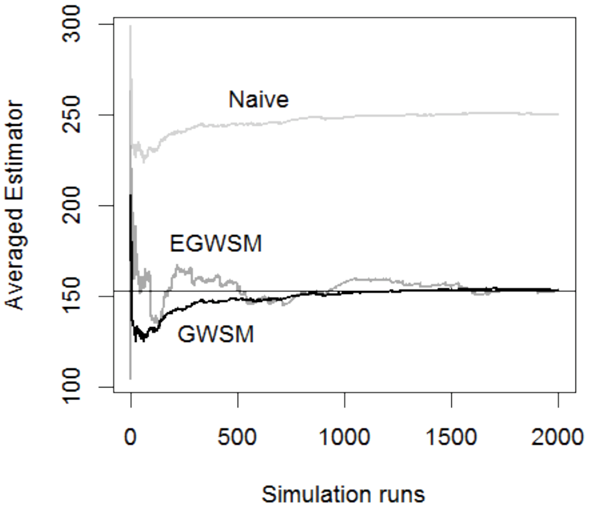

Figure 2: Cumulative convergence of estimators for 2000 simulation runs

In our study the following 3 estimators were considered:

- *Classical fair share estimator:* Given by $\hat{T}_{y'}^{\text{GWSM}} = \sum_{i' \in s'} y'_{i'} \sum_{i \in s} l_{i,i'} d_i$ where the link function reflects the above-described fair-share method.
- *Proposed estimator:* Given by $\hat{T}_{y'}^{\text{GWSM}} = \sum_{i' \in s'} y'_{i'} \sum_{i \in s} \tilde{l}_{i,i'} d_i$. Thus the link function $l_{i,i'}$ is estimated based on the available information
- *Naive estimator:* $\hat{T}_{y'}^{\text{NAIVE}}$ which is similar to the above estimators, except that the link-function is estimated naively by the inverse of the observed (!) inbound links. This estimator is generally biased and corresponds to the Monte-Carlo estimator with no burn-in phase. In our case, as the elements in U' with larger values $y'_{i'}$ tend to have more inbound links, we expect the naive estimator to have a positive bias as the it by construction underestimates the number of inbound links.

We carried out 2000 simulation runs of the above mentioned design. The cumulative convergence of the proposed estimators is shown in figure 2.

As expected, the naive estimator exhibits a positive bias. Both our proposed estimator as well as the classical fair share estimator show are unbiased by construction. Also the effect of the increased variance of our proposed estimator over the classical fair share approach can be clearly seen. Again note that this increased variance comes from estimating the link function. This might be required in situation where the inbound links are not fully

Table 1: Results of simulation study

Estimator	rel. bias	variance	MSE
GWSM	0.002	4271	4269
EGSM	0.002	62409	62378
NAIVE	0.635	4358	13799

known, as described in the beginning of this paper. The empirical relative bias, empirical variance and empirical MSE are given in the table 1. From there we can also see that the variance inflation from the need to estimate the link function is substantial with a variance increase by factor 14. This might by reduced by capping extreme resulting weights at the price of a small bias.

5. Conclusion and Outlook

In this paper, an extension of the generalized weight share method (GWSM) has been proposed which extends the application of the GWSM to settings where not all inbound links to a unit in the target population are known. Applications of the Expected Generalized Weight Share Method (EGWSM) range from refreshment samples, through sampling of hyperlinks on the internet to following asymmetrical links in social networks. An appealing feature of the EGWSM is, that the sole change compared to the GWSM is the replacement of the unknown link function with a design-unbiased estimator, which in turn leads to design-unbiased estimates of the proposed estimator and allows for the use of same or slightly modified analytical tools.

The proposed EGWSM approach might be an alternative to existing MCMC methods or respondent driven sampling, particular for statistical agencies or researchers that are familiar with the the GWSM. While the proposed estimator is design-unbiased, the challenge of increased variance (over the situation of known links) exists. Thus all efforts should be made to find out the link structure before using this approach.

On the outlook, several further research the question on the EGWSM remain open. Most notably, practitioners might likely consider to cut-off extreme resulting weights in order to reduce variance of the price of a negligible bias. In addition, the evaluation of this approach in a real survey as well as an analysis of the performance of the variance estimator. In addition

REFERENCES

Deville, J.-C., Lavalle, P. (2006). Indirect sampling: The Foundation of the Generalized Weight Share Method. *Survey Methodology*, 32-2, p. 165-176.

Ernst, L. (1989). Weighting issues for longitudinal household and family estimates. In *Panel Surveys* Wiley & Sons, New York, p. 135-159.

Heckathorn, D.D. (1987). Respondent-Driven Sampling. *Social Problems*. 44-2, p. 174-199

Huang, H. (1984). Obtaining Cross-Sectional Estimates from a Longitudinal Survey. *Proceedings of the American Statistical Association*. p 670-675.

Lavalle, P. (1995). Cross-sectional Weighting of Longitudinal Surveys of Individuals and Households Using the Weight Share Method. *Survey Methodology*, 21-1, p. 25-32.

Lavalle (2007). *Indirect Sampling*. Springer, New York.

Sirken, M.G. (1970). Household Surveys with Multiplicity. *Journal of the American Statistical Association*. 65-329, p. 224-227.

Thompson, S.K. (2006). Adaptive Websampling. *Biometrics*, 62-4, p. 1224-1234.