Surveying Kindergarten Children in the Absence of a Sampling Frame: a Study in Indirect Sampling

Hans Kiesl
Regensburg University of Applied Sciences, Germany
Email: hans.kiesl@hs-regensburg.de

Abstract

One of the cohorts of the German National Educational Panel Study (NEPS) consists of a sample of kindergarten children. No nationwide frame of kindergartens in Germany was available for sample selection, contrary to the situation for primary schools. Following the works of Deville and Lavallée, we present a solution by indirect sampling, using links between kindergartens and primary schools. We give a short review of the theory of indirect sampling and present some theory for unbiased estimation after an additional subsampling stage. Finally, we describe how this sampling and estimation procedure was applied to construct a kindergarten sample for NEPS. [1]

Keywords: educational survey, generalized weight share method, indirect sampling.

1. Introduction

The National Educational Panel Study (NEPS) is a new longitudinal educational survey in Germany with a quite complex sampling design. Six different samples from various age cohorts are drawn from the population and followed over time. One of the six target populations are children aged 4 (in 2010) who attend a kindergarten (or nursery school). Unfortunately, no complete list of all kindergartens in Germany, which could have been used as a sampling frame, was available. Therefore it was not possible to draw a sample of kindergartens directly. In a situation like this, indirect sampling might be an alternative. Here, a sample s_A is drawn from some other population U_A whose units are linked to the units of the target population U_B; the sample s_B from U_B then consists of all units from U_B that are linked to some unit in s_A. It is possible to construct unbiased estimators from s_B; the basic references for the theory of indirect sampling are Deville and Lavallée (2006) and Lavallée (2007). In our application, a (direct) sample from the population of primary schools in Germany was drawn (for which a sampling frame was available); we then used links between kindergartens and primary schools to end up with an indirect sample of kindergartens. Since it was not feasible for budget reasons to use the complete indirect sample of kindergartens, we had to add another stage of subsampling.

2. Review of Indirect Sampling

The idea underlying indirect sampling is quite simple. Consider the task to estimate the total t_Y of a variable Y in some population U_B. If a sampling frame for U_B were available,

[1]This work was partially funded by the German Research Foundation (DFG), Priority Programme "Education as a lifelong process" (SPP 1646) under the grant KI 1646/1-1.

we would (directly) sample from this frame, using any suitable sampling design with inclusion probabilities $\pi_i^B > 0$ for every $i \in U_B$. To get a design-unbiased estimator for t_Y, we could use the Horvitz-Thompson-estimator (HT-estimator) $\hat{t}_{Y,HT} = \sum_{s_B} \frac{y_i}{\pi_i^B}$, where the summation is over all units in the sample s_B.

In certain applications no sampling frame for U_B is available. We might, however, have a sampling frame for a population U_A, whose elements are somehow "linked" to the elements in the population of U_B. A natural idea is then to draw a sample s_A from U_A (with known inclusion probabilities) and subsequently choose all elements from U_B that are linked to elements in s_A and define them as the sample s_B from U_B (see figure 1). Because of the indirect selection of elements from U_B, this procedure is called indirect sampling. The question remains, whether (and how) unbiased estimation for t_Y from s_B is possible, since the calculation of inclusion probabilities for s_B might be difficult or impossible under this setting.

The theory of indirect sampling (although not yet called that way) started with Ernst (1989) and the references therein in the context of longitudinal household surveys. In several papers (e.g. Deville and Lavallée (2006), Deville and Maumy-Bertrand 2006) the theory was generalized to other situations. Lavallée (2007) is a comprehensive treatment of indirect sampling with theory and applications.

Let U_A and U_B be two populations, and let θ be a non-negative function (the "link function") on $U_A \times U_B$, i.e. for every $a \in U_A$ and every $b \in U_B$ we have $\theta_{ab} \geq 0$. We say that a link exists between $a \in U_A$ and $b \in U_B$, if and only if $\theta_{ab} > 0$. We then call θ_{ab} the "weight" of the link between a and b. (The question how best to define this link function depends on the application at hand. We come back to this later.)

For every $b \in B$, let $\theta_{+b} := \sum_{a \in U_A} \theta_{ab}$ be the sum of the weights of all links from U_A to $b \in B$. We assume that $\theta_{+b} > 0$ for all $b \in B$, i.e. there exists a link to every $b \in B$. (Otherwise unbiased estimation is impossible.)

The key observation is that the total of any variable Y in the population U_B might be written as follows:

$$t_Y = \sum_{b \in U_B} y_b = \sum_{b \in U_B} \left(y_b \cdot \underbrace{\sum_{a \in U_A} \frac{\theta_{ab}}{\theta_{+b}}}_{=1} \right) = \sum_{a \in U_A} \sum_{b \in U_B} \frac{\theta_{ab}}{\theta_{+b}} y_b = \sum_{a \in U_A} \tilde{y}_a = t_{\tilde{Y}},$$

with $\tilde{y}_a := \sum_{b \in U_B} \frac{\theta_{ab}}{\theta_{+b}} y_b$.

Thus, the total of Y in population U_B can be written as the total of the variable $\tilde{Y}$ in population U_A. If it is possible to draw a probability sample from U_A, the HT-estimator may be used for unbiased estimation of $t_{\tilde{Y}}$ and thus also of t_Y.

Let s_A be a sample of elements from U_A, and let π_a be the inclusion probability of $a \in s_A$. The sample s_B is defined as the set of all units in U_B that have a link to some element of s_A; more formally $s_B := \{b \in U_B | \theta_{ab} > 0 \text{ for some } a \in s_A\}$. The HT-estimator for the total of $t_{\tilde{Y}}$ is now also an unbiased estimator for the total of t_Y, thus we call it the indirect sampling estimator $\hat{t}_{Y,\mathrm{IS}}$ for the total of t_Y:

$$\hat{t}_{Y,\mathrm{IS}} := \hat{t}_{\tilde{Y}} = \sum_{a \in s_A} \frac{\tilde{y}_a}{\pi_a} = \sum_{a \in s_A} \sum_{b \in U_B} \frac{\theta_{ab}}{\theta_{+b}} \frac{y_b}{\pi_a} \overset{(*)}{=} \sum_{a \in s_A} \sum_{b \in s_B} \frac{\theta_{ab}}{\theta_{+b}} \frac{y_b}{\pi_a} = \sum_{b \in s_B} w_{b_s} y_b \quad (1)$$

with weights

$$w_{b_s} = \sum_{a \in s_A} \frac{\theta_{ab}}{\pi_a \cdot \theta_{+b}}. \tag{2}$$

(Equality (*) is due to the fact that by definition of s_B we have $\theta_{ab} = 0$ for $a \in s_A$ and $b \notin s_B$.)

The weights w_{b_s} are in general different from the inclusion probabilities π_b^B, and they are sample dependent (therefore the subindex s): the weight of unit $b \in s_B$ depends on which units in U_A that are linked to b are actually in the sample s_A.

3. Subsampling of Indirect Samples

3.1 Subsampling Within Units of s_B

Up to now, we assumed that U_B is the set of final sampling units. Now suppose that the elements of U_B are actually clusters of individual units, i.e. U_B is the set of primary sampling units (but note that the links are still defined between units in U_A and U_B). The value y_b of a cluster $b \in U_B$ is now itself a total of individual values. Let y_{bi} be the value of the variable Y of the i-th element in cluster b, let e_b be the set of all elements in cluster b. Then, $y_b = \sum_{i \in e_b} y_{bi}$. Suppose that we draw only a subsample from e_b; in this case we do not observe y_b, but we can estimate it from the sample.

To be more precise, consider a second stage of sampling, independently within the clusters b of the first stage sample s_B. The sample of cluster b is called s_b. Let $\pi_{i|b}$ be the (conditional) inclusion probability of unit i in cluster b, given that $b \in s_B$. Then y_b can be estimated by $\hat{y}_b = \sum_{s_b} \frac{y_{bi}}{\pi_{i|b}}$, and an estimator for t_Y may be defined as follows:

$$\hat{t}_{Y,\mathrm{IS,sub}} := \sum_{a \in s_A} \sum_{b \in s_B} \frac{\theta_{ab}}{\theta_{+b}} \frac{\hat{y}_b}{\pi_a} = \sum_{a \in s_A} \sum_{b \in s_B} \frac{\theta_{ab}}{\theta_{+b}} \frac{\sum_{i \in s_b} y_{bi}/\pi_{i|b}}{\pi_a} = \sum_{b \in s_B} \sum_{i \in s_b} w_{bi_s} y_{bi}$$

with weights $w_{bi_s} = \sum_{a \in s_A} \frac{\theta_{ab}}{\pi_a \cdot \pi_{i|b} \cdot \theta_{+b}}$.

Lavallée (2007, section 5.1) calls this procedure *two stage indirect sampling* and shows that $\hat{t}_{Y,\mathrm{IS,sub}}$ is unbiased for t_Y. In the next subsection, we will introduce an additional subsampling step that was necessary in our application of kindergarten sampling.

3.2 Subsampling Linked Units for Every $a \in s_A$

The indirect sampling estimator (1) assumes that sample s_B consists of all units of U_B that are linked to the units in the direct sample s_A from population U_A. In some applications, where U_B is much larger than U_A or where the number of links is highly variable among the units in U_A, this might result in a vary large, or at least quite unpredictably sized sample s_B. From a practical point of view, it might be desirable to draw only a subsample of s_B as the final sample from U_B.

Consider first the simple case that U_B consists of final sampling units (we return to the case in which U_B is a population of PSUs below). The direct sampling procedure results in the direct sample s_A of U_A. Consider now for any $a \in U_A$ the set Ω_a of all units in U_B that are linked to a. (Note that U_B is the union of all Ω_a ($a \in U_A$), but apart from the special case of cluster sampling, the Ω_a need not be pairwise disjoint.) The idea is now to independently

draw a subsample Ω_a^{sub} from Ω_a for every $a \in s_A$ (which we view as the second sampling stage). For any $b \in \Omega_a$, let $\pi_{b\in\Omega_a^{\text{sub}}}$ be the (conditional) probability to be in subsample Ω_a^{sub}, given that a is in sample s_A. The final indirect sample s_B^{fin} then consists of the union of all Ω_a^{sub}. Because the subsampling is done independently, some units $b \in U_B$ might appear in different Ω_a^{sub}; this has to be considered when constructing an estimator.

The following estimator is unbiased for the total t_Y in U_B:

$$\hat{t}_{Y,\text{IS,2stage}} := \sum_{b\in s_B^{\text{fin}}} w'_{b_s} y_b \qquad \text{with} \qquad w'_{b_s} = \sum_{a\in s_A} \frac{\theta_{ab} \cdot \mathbb{1}(b \in \Omega_a^{\text{sub}})}{\pi_a \cdot \pi_{b\in\Omega_a^{\text{sub}}} \cdot \theta_{+b}},$$

where $\mathbb{1}(b \in \Omega_a^{\text{sub}})$ is equal to 1, if $b \in \Omega_a^{\text{sub}}$, and 0 otherwise.

3.3 Combining Both Subsampling Steps

Finally, we turn again to the situation where the elements of U_B are actually clusters of individual elements. Suppose that we independently draw subsamples s_b from every cluster b of the final indirect sample s_B^{fin}. Let $\pi_{i|b}$ be the (conditional) inclusion probability of unit i in cluster b, given that $b \in s_B^{\text{fin}}$. Then y_b can be estimated by $\hat{y}_b = \sum_{s_b} \frac{y_{bi}}{\pi_{i|b}}$. Under this three stage procedure, an unbiased estimator for t_Y may be defined as follows:

$$\hat{t}_{Y,\text{IS,3stage}} := \sum_{b\in s_B^{\text{fin}}} \sum_{i\in s_b} w''_{bi_s} y_{bi} \qquad \text{with} \quad w''_{bi_s} = \sum_{a\in s_A} \frac{\theta_{ab} \cdot \mathbb{1}(b \in \Omega_a^{\text{sub}})}{\pi_a \cdot \pi_{b\in\Omega_a^{\text{sub}}} \cdot \pi_{i|b} \cdot \theta_{+b}}. \qquad (3)$$

In practice, nonresponse might occur on every sampling stage. Notice that $\hat{t}_{Y,\text{IS,3stage}}$ in (3) can be easily adapted for this case: $\frac{1}{\pi_a}$ might be replaced by weights adjusted for nonresponse occuring in the direct sampling stage; $\frac{1}{\pi_{b\in\Omega_a^{\text{sub}}}}$ might be replaced by weights adjusted for nonresponse occuring in the subsampling among the Ω_a; $\frac{1}{\pi_{i|b}}$ might be replaced by weights adjusted for nonresponse among the units of s_b. If the nonresponse adjustments result in (approximately) unbiased estimators on every sampling stage, the final estimator will also be approximately unbiased.

4. A New Application: Sampling of Kindergarten Children for NEPS

The German National Educational Panel Study (NEPS) is a new longitudinal educational survey with a very complex design. The main study consists of six different samples representing different age cohorts of the population in Germany at the start of the survey in 2010 (newborns, children aged 4 attending kindergarten, fifth graders, ninth graders, university freshmen, adult population aged 25 and over). Contrary to other educational surveys like PISA or TIMSS, the NEPS is designed as a longitudinal survey with repeated test procedures and questionnaires each year. For an overview of the NEPS, see Blossfeld et al. (2011) or www.neps-data.de/en.

One of the NEPS sample cohorts are children attending kindergarten; to be more specific, the target population was defined to consist of children attending kindergarten in the school year 2010/2011 who were expected to beginn schooling in school year 2012/2013. These children had been around the age of 4 years in 2010/2011, since children in Germany start attending primary school around the age of 6 (depending on their exact date of birth).

Children in Germany are not obliged to go to a kindergarten, but more than 90 % of all children aged 4 attend some kind of kindergarten or pre-school. (Children not attending any kindergarten were not part of the target population.) Unlike in some other countries, kindergartens in Germany are usually completely separated from primary schools. There is also a great variety of types and organizational forms of kindergartens. Due to this diverse nature, there are no complete listings of kindergartens in Germany available for sample selection. On the other hand, a complete sampling frame for primary schools is available. Also, despite the spatial separation, kindergartens might be seen as "linked" to primary schools, since every child that eventually leaves a kindergarten joins a particular primary school. Thus, indirect sampling was used to draw a sample of kindergarten children for NEPS.

Using the notation from the previous sections, let U_A be the population of primary schools, and let U_B be the population of kindergartens. There are two rather obvious ways to define a link function on $U_A \times U_B$:

(i) $\theta_{ab} = 1$ if there was at least one child moving from kindergarten b to school a in a particular reference period; otherwise $\theta_{ab} = 0$. In this case, θ_{+b} is the number of schools that received children from kindergarten b within the reference period.

(ii) $\theta_{ab} =$ number of children having moved from kindergarten b to school a in a particular reference period. In this case, θ_{+b} is the number of children that moved from kindergarten b to any school in U_A within the reference period.

Note that the reference period used above does not have to cover the time of actually drawing the sample. An important condition for unbiased estimation, however, is that every kindergarten that exists at the time of drawing the sample had been linked to at least one primary school in the reference period. This condition is not met, if we have a kindergarten $b \in U_B$ that did not send children to any primary school in the reference period. In reality, such kindergartens might exist (e.g. a newly founded kindergarten with 4 year old kids only), but we regard their number to be negligible.

Both definitions of θ_{ab} above could be used for unbiased estimation. The decision which definition to use in the end depended on practical considerations concerning how easy it was to get the values θ_{ab} and θ_{+b} from the sampled units.

In our application, the first step was to draw a direct sample of primary schools s_A (which was achieved by systematic pps-sampling; see Aßmann et al. (2012) for details). Then, every school $a \in s_A$ was asked to provide the set of linked kindergartens Ω_a and the values of θ_{ab} for every $b \in \Omega_a$. In the case of link function (i) this meant providing the set of kindergartens that sent children to school a in the reference period (which was chosen to be the school year 2009/2010). In the case of link function (ii) this meant for every child that joined school a as a first grader in the reference period to provide the name (and address) of the kindergarten the child had attended before. Pre-tests for the survey had shown that primary schools are usually able to come up with both kinds of information from their files, and this was confirmed in the final survey.

Since the number of kindergartens that a primary school is linked to has a high variance (some sampled schools have more than 20 linked kindergartens), for budget reasons a decision was made not to survey the complete indirect sample s_B but to use some kind of subsampling within Ω_a. s_B^{fin} was therefore drawn by the procedure described in section 3.2. For every $a \in s_A$, between 1 and 4 kindergartens from Ω_a have been selected with probability proportional to the θ_{ab}. The reason for this is that the complete direct sample

of primary schools s_A was used to get a sample of first graders in school year 2012/2013, and it was desired to find in every school $a \in s_A$ at least some children that were in the kindergarten sample of 2010/2011.

In every kindergarten $b \in s_B^{\text{fin}}$, we then had to ask for the value of θ_{+b}. In the case of link function (i) this meant providing the number of primary schools that children who left kindergarten b during the reference period moved to as a first grader. In the case of link function (ii) this simply meant providing the number of children who left kindergarten b during the reference period and who joined any primary school. Pre-tests had shown that the latter information could be given by kindergartens much more reliably. Kindergarten administrators know quite well how many children left, but they usually do not know exactly which primary schools (or how many of them) these children joined. For this reason, it was decided to use link function (ii) to construct the indirect sampling estimator for the kindergarten cohort in NEPS.

In the chosen kindergartens, no subsampling among the children in the desired age range was originally planned. However, due to non-response, the three stage estimator (3) was used for the construction of non-response adjusted design weights. The first wave of the kindergarten sample in NEPS finally consists of 2,996 kindergarten children in 279 kindergartens.

5. Conclusion

Indirect sampling proved to be a feasible way to draw a sample of kindergarten children for the German National Educational Panel Study (NEPS) in the absence of a proper sampling frame. Starting with a sample of primary schools, an indirect sample was generated using links between kindergartens and schools. We have shown that unbiased estimation of population totals is possible, even if for every sampled primary school only a subset of the linked kindergartens can actually be surveyed for budget reasons. We are confident that our experiences with the NEPS may be useful for future applications of indirect sampling.

REFERENCES

Aßmann, C., Koch, S., Steinhauer, H.W., and Zinn, S. (2012): NEPS Starting Cohort 2 Kindergarten (SC2), SUF-Version 1.0.0 Data Manual [Supplement]: Weighting, https://www.neps-data.de/Portals/0/NEPS/Datenzentrum/Forschungsdaten/SC2/1-0-0/SC2_1-0-0_Weighting_EN.pdf

Blossfeld, H.-P., Roßbach, H.-G., and von Maurice, J. (Eds.) (2011): *Education as a Lifelong Process - The German National Educational Panel Study (NEPS)*, Zeitschrift für Erziehungswissenschaft, Special Issue 14, Heidelberg: VS Verlag für Sozialwissenschaften.

Deville, J.-C., and Lavallée, P. (2006): Indirect Sampling: the Foundations of the Generalised Weight Share Method, *Survey Methodology*, 32, 165–176.

Deville, J.-C., and Maumy-Bertrand, M. (2006): Extension of the Indirect Sampling Method and Its Application to Tourism, *Survey Methodology*, 32, 177–185.

Ernst, L. (1989): Weighting issues for longitudinal household and family estimates, in *Panel Surveys*, eds. D. Kasprzyk, G. Duncan, G. Kalton and M.P. Singh, New York: John Wiley and Sons, 139–159.

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

Lavallée, P. (2007): *Indirect Sampling*, New York: Springer.