Estimation for Person-pair Drug-related Characteristics in the Presence of Pair Multiplicities and Extreme Sampling Weights

James R. Chromy and Avinash C. Singh
A.C. Singh, Statistics Research Division, RTI International, Research Triangle Park, NC 27709, asingh@rti.org

Abstract

For the National Household Survey on Drug Abuse, Brewer's method is adapted to select 0, 1, or 2 persons at the second phase from the dwelling unit (DU) selected for screening at the first phase. The pair-level (2 persons per DU) sample forms a subsample of the larger person-level (1 or 2 persons per DU) sample. For studying drug-related behavioral relationships among members of the same household, pair data is required because the outcome variable is generally other-member specific, i.e., it depends on the other member selected from the household. However, the parameter of interest is generally at the person-level and not at the pair-level. For example, in the parent-child pairs, one may be interested in the proportion of the pairs in which parent reports talking to child about drugs when child used drug in the past year, but the target population may consist of only children, and not all pairs. In estimation for pair domains, two major problems arise: one is that of multiplicities because for a given domain defined by the pair relationship, several pairs in the household could be associated with the same person. In this case, it may be desirable to use the average measure of behavior relationship for each member, and this gives rise to multiplicities. Thus the design weights need to be divided by the person-level multiplicity factors for each domain of interest. Therefore, multiplicity factors need to be produced along with the final set of calibrated weights. The other problem is that of extreme weights that may arise due to small selection probabilities for certain pair-age groups. This may lead to unstable estimates for which we propose a Hajek-type modification. This modification essentially entails calibration (like poststratification) to controls for the number of persons in households belonging to each domain of interest; these controls can be obtained from the larger sample of singles and pairs, i.e., one or two persons selected from DUs. However, the multiplicity factor, being domain-specific, renders the calibration adjustment factor domain-specific. This raises the question of finding one set of calibration weights for use with all domains or outcome variables. We propose a way out by performing a multivariate calibration with respect to a key set of pair domains. This type of poststratification is then followed by a repeat poststratification to further control the extreme weights by imposing separate bound restrictions on the initially identified extreme weights.

Key words: Multiplicity factors; Extreme weights; Hajek-type Modification; Calibration.

1. Introduction and Motivation

Traditionally, most household surveys have been designed either to measure characteristics of the entire household or to focus on a randomly selected respondent from among those determined to be eligible for the survey. Selecting more than one person from the same household was considered ill-advised since persons from the same household tended to repeat the same general information characteristic of the entire household. Selecting only one person per household totally avoided the clustering effect on the variance. The "one person per household" sampling approach, however, precludes the opportunity to gather information about the relationships among household members. In this paper, we examine the richer analytic capability of a survey designed to assure a positive pairwise probability of selection among all eligible household members in each sample household. Achieving positive probabilities for all pairs within sampled households permits unbiased estimation of the within dwelling unit component of variance. It also facilitates the study of the relationships of social behaviors among members of the same household besides providing efficient data collection. This paper focuses on the second objective, the study of behavioral relationships among persons residing in the same household.

The National Household Survey on Drug Abuse (NHSDA) samples were selected so that any two survey-eligible persons residing in the same dwelling unit[1] would have a positive probability of both being selected into the sample. The probability of selecting a pair p of persons (associated with person i in the household) residing in the same dwelling unit h can be represented by

$$\pi_{hip} = \pi_h \pi_{ip|h}$$

[1]The NHSDA target population includes not only the household population, but also members of non-institutional group quarters. The term dwelling unit is used to identify both households (the listing units for the household component) and group quarters units (the listing units for the elgible group quarters usually rooms or beds). The terms household or housing unit and dwelling unit are used interchangeably.

where π_h is the probability of selecting dwelling unit h, and $\pi_{ip|h}$ is the conditional probability of selecting pair p containing person i given that the dwelling unit h has been selected. Since 1999, a modified version of Brewer's method (Brewer, 1963) for selecting probability proportional to size samples of size 2 has been applied to select 0, 1, or 2 persons per household. With pair data, possible domains of interest in practice might be:

Child-Parent pairs:

1. (Parent, Child 12-17) with focus on parent, i.e., the target population consists of parents.
2. (Child 12-17, Parent) with focus on child.
3. (Parent, child 12-14) with focus on parent.
4. (Child 12-14, Parent) with focus on child.
5. (Parent, Child 15-17) with focus on parent.
6. (Child 15-17, parent) with focus on child.
7. (Parent, Child 12-20) with focus on parent.

Sibling pairs:

8. (Child 12-14, Child 15-17) with focus on Child 15-17.
9. (Child 12-17, Young Adult 18-25) with focus on young adult.

For pair data, an important module of the drug questionnaire, termed the Parenting Experience, is administered to the parent only when the child is also selected for the interview. Based on this module, we can use the parent-child pair domain (with child being the focus, i.e., the target population consists of children living with parents) to introduce issues related to the pair data analysis. Table 1shows an example of analysis where it is of interest to estimate the distribution of reported parent-child discussions about drugs by the child's (adolescent's) drug use in the past year.

We now introduce two problems that arise with pair data analysis: one is that of multiplicities, and the other is that of extreme weights. The problem of multiplicities (see Sirken, 1972) arises because for two-parent households with children, each child has two inclusion possibilities (one with each parent) in the population of all parent-child pairs. Thus, if children form the target population, it would be desirable in general to assign one observation per child, and a reasonable way to achieve this would be to take the average of the two responses corresponding to the pairs associated with the two parents. In other words, the response for each child-parent pair from two-parent households is divided by the number of parents-- the multiplicity factor. Note that the multiplicity factor depends on the person and the pair relationship domain. Similarly, if parents form the target population, then the multiplicity factor would be the number of children for each parent in the household. The multiplicity problem does not arise if the person-level outcome variable is not other-member specific (e.g., child reports talking to a nonspecific parent in the case of two-parent households), or if the outcome variable is at the pair-level itself, e.g., child-parent pair drug behavior. Section 2 considers in more details estimation of parameters in the presence of multiplicities.

The other problem of extreme weights with pair data arises because the Brewer's method may give rise to very small pair selection probabilities as older age groups are assigned much lower sampling rates. One can use the analogy with the Basu's elephant example, and the Hajek-type (1971) ratio adjustment to address this problem. This is considered in Section 3. Concluding remarks are presented in Section 4.

Table 1: Parents Reporting Talking about Drugs by Adolescent Drug Use

Parent reports talking to child	Child used drugs ?	
	Yes	No
Yes	(1,1)	(1,2)
No	(2,1)	(2,2)

Table 2: Pairs by Persons per Dwelling Unit

Persons	Pairs	Persons	Pairs
1	0	5	10
2	1	6	15
3	3	7	21
4	6	8	28

2. The Problem of Pair Multiplicities

Before addressing the development of estimators it is useful to (1) define the population of pairs and (2) to define the population parameters for pair data. The number of eligible person pairs per dwelling unit increases as the number of eligible persons increases, but it increases much faster as shown in Table 2. This points out the need for careful interpretation of any data that relates to the population of pairs.

Table 3 below shows examples of the number of pairs of certain specified types by dwelling unit composition. In

this case, the number of population pairs of selected types (parent-child, mother-child, or father-child) depends both on the number (and type) of parents and the number of children. Analysts do not usually wish to draw inferences about the population of all pairs; they may instead wish to draw inference about relationships to other persons with a subpopulation defined in terms of the characteristics of one member of the pair.

Table 3: Parent-Child Pairs by Household Composition

Dwelling Unit Composition		Number of Pairs by Type		
Parent	Child	Parent-child	Mother-child	Father-child
Both	1	2	1	1
	2	4	2	2
	3	6	3	3
Father only	1	1	0	1
	2	2	0	2
	3	3	0	3
Mother only	1	1	1	0
	2	2	2	0
	3	3	3	0

We will illustrate how multiplicities appear in the definitions of parameters and estimates for the example of Table 1. Consider estimation of the total number of children who used drugs in the past year, and with whom parents report talking about drugs. Let $y_{hip}(d)$ = drug related behavior outcome for pair p containing the individual i belonging to domain d in household h. The variable y_{hip} is generally a 0-1 variable. However, it may also be nonbinary for this domain, e.g., if the outcome measure is the number of times the parent had conversation with the child about drug abuse.

Now, for the population of all individuals who belong to the domain d, the total parameter is defined as

$$\tau_y(d) = \sum_{h=1}^{H} \sum_{i=1}^{N_h(d)} \sum_{p=1}^{M_{hi}(d)} \frac{y_{hip}(d)}{M_{hi}(d)}$$

i.e., total of averages over pairs (p) associated with the individual i over all i in domain d and in the household h. Here $M_{hi}(d)$ denotes multiplicity (i.e., the number of pairs associated) for the person i in domain d, and $N_h(d)$ can be thought as the multiplicity count for the household h, i.e., the number of persons in the household that are in domain d.

Similarly, the population mean parameter is defined as

$$\mu_y(d) = \tau_y(d) \Big/ N(d), \quad N(d) = \Sigma_h N_h(d)$$

The Horvitz-Thompson (HT) estimator of $\tau_y(d)$ is

$$\hat{\tau}_y(d) = \Sigma_h \Sigma_i \left(M_{hi}(d)\right)^{-1} \Sigma_p y_{hip}(d) 1_{hip \in s} \Big/ \pi_{hip}.$$

The estimator of the mean is given by

$$\hat{\mu}_y(d) = \hat{\pi}_y(d) \Big/ \hat{N}(d),$$

$$\hat{N}(d) = \Sigma_h \Sigma_i \left(M_{hi}(d)\right)^{-1} \Sigma_p 1_{hip}(d) 1_{hip \in s} \Big/ \pi_{hip},$$

Observe that for the sake of simplicity, the weight adjustments for nonresponse are not shown in the above estimator.

We next consider some alternative target populations for parent-child relationships, as shown in Table 4. It may be noted that there are several options for specifying the target population depending on the outcome being reported. Option 1 (reports for all possible parent-child pairs) defines the population as all possible parent-child pairs where both the parent and the child reside in the same dwelling unit. Option 2 reports mother-child pairs and father-child pairs separately. Option 3 reports on the average relationship with two parents when both are present in the same dwelling unit. In options 2 and 3, the population is defined as children residing in the same dwelling unit as one or both of their study-eligible parents. If one or both parents residing in the dwelling unit are not study eligible (e.g., members of the military), then pairs with those parents must be excluded from our study population.[2] Option 4 provides an example where the

[2]In certain non-institutional group quarters, listing is done by person or bed rather than by family unit even when the family unit resides together. Since the beds or

focus is on the parents and it averages responses over all their children residing with the parent.

Option 2 requires twice as many tables with fewer pairs represented in each table. Note that the population size for option 1 is sum of the population sizes defined in option 2 less an adjustment for double counting. For option 3, children contribute to the population count equally whether they live with one or both parents. Option 4 is similar to option 3 with the focus shifted to parent side of the pairs.

3. The Problem of Extreme Weights

As mentioned in Section 1, Brewer's method may give rise to extreme weights for selected pairs. A modification of the Brewer's method was used for NHSDA which involves the introduction of three dummies. It can be described as follows.

First we assign trial sampling rates to all eligibles based on age and the specified rates. Since Brewer's method is designed to select samples of size 2 only (and the sample size selected is the sum of sampling rates), we must adjust the sampling rates, $\pi_{i|h}$, to add to 2. This is done in one of two ways:

1. If the trial sampling rates add to more than 2, set the Final Sampling Rates by simply scaling back the Trial Sampling Rates to add to 2 and leave the final sampling rates for the three dummies at zero.

2. If the trial sampling rates add to less than two, set the Final Sampling Rates at their Trial Sampling Rate levels (including zeroes for ineligibles), and allocate the difference between 2 and their sum evenly among the dummy persons.

The dummies are needed to give Brewer's algorithm a chance to select 2 records in every household, but sometimes it may select 2 eligible persons, other times it may select 2 dummy persons, and it may also select 1 dummy and 1 eligible person. This trick is needed to produce probability samples at the specified probabilities (sampling rates) and to insure that all pairs have a chance of being selected while controlling both individual sampling rates and the pairwise rates. In every case, it will select 0, 1, or 2 eligible persons. Three dummies (rather than 1 or 2) are required to avoid division by a very small number under some household compositions.

Brewer's method sets the pairwise probabilities at (here $\pi_{ip|h}$ is denoted as $\pi_{ij|h}$ where the pair p contains members i and j.)

$$\pi_{ij|h} = \frac{\pi_{i|h}\pi_{j|h}}{K}\left[\frac{1}{1-\pi_{i|h}} + \frac{1}{1-\pi_{j|h}}\right]$$

By setting K at

$$K = 2 + \Sigma_i \frac{\pi_{i|h}}{1-\pi_{i|h}},$$

we guarantee that the sum of the pairwise probabilities taken over all unique pairs will be exactly one , i.e.

$$\sum_i \sum_{j>1} \pi_{ij|h} = 1.$$

It also guarantees that

$$\sum_{j\neq 1} \pi_{ij|h} = \pi_{i|h}$$

for all values of i.

Randomly selecting the sample of 0, 1, or 2 eligibles then reduces to selecting one pair at random. Note that the pair-wise selection probability could become very small depending on the household composition. For instance, for a household with two persons in the age group 50+, if the selection probability for each member is .04, then the three dummies get the selection probability of .64 each, and the inverse of the selection probability for the pair of persons each aged 50 + becomes as high as 2224. Thus for pair data, the proportion of extreme weights among the initial design weights could be very high resulting in a high unequal weighting effect. This would make the Horvitz-Thompson estimator very imprecise although it remains unbiased.

To overcome the above extreme weight problem, we consider the Hajek-type modification to the Horvitz-Thompson estimator, which is basically a ratio-type poststratified estimator. Hajek (1971) had suggested this estimator in his comment on the problem raised by Basu's elephant fable which we briefly describe here for the sake of historical interest. A circus owner had 50 elephants, and wanted to estimate the total weight to help him make arrangements for shipping . To save time, he only wanted

---

persons are treated as dwelling units in the survey process, the opportunity to define child-parent pairs is lost. If emphasis on pairs is to be continued, we may wish to re-examine the sampling and field procedures used to define and sample dwelling units within group quarters.

to weight Sambo (an average sized elephant), and use 50 times its weight as an estimate. However, the circus statistician, being highly conscious of the optimality and unbiasedness of the HT-estimator, objected about the potential bias of his estimate because of the purposive selection. Instead, he suggested random selection of an elephant with a very high probability of 99/100 for Sambo, and the rest including Jumbo (the biggest in the herd) with probability 1/4900 each. The circus owner was very unhappy with the statistician's response of 100/99 times the Sambo's weight as the estimate if Sambo got selected in this random draw, and was outraged with the response of 4900 times the Jumbo's weight if Jumbo happened to get selected. It was obvious to the owner that this new estimator was extremely poor, although he didn't know anything about its unbiasedness. The story had an unhappy ending with the circus statistician losing his job. To alleviate the instability of the HT-estimator, Hajek suggested to multiply it by 50 divided by inverse of the selection probability, which reduces simply to 50 times the weight of the selected elephant. Clearly this estimator would give a reasonable answer most of the time although it is no longer unbiased. It should be noted that, in practice, to deal with the above problem of instability in estimation due to extreme variability in the population, survey statisticians would prefer a suitable stratification at the design stage followed by poststratification at the estimation stage.

Now for a given domain d, Hajek's estimator is given by

$$\tilde{\tau}_y(d) = \tilde{N}(d)\hat{\mu}_y(d)$$

where $\tilde{N}(d) = \Sigma_h N_h(d) 1_{h\epsilon S} / \pi_h$

is the HT- estimator of the total number of persons in the domain d from the bigger sample of households (i.e., households from which one or two persons were selected). Notice that in the above ratio-type adjustment, the adjustment factor is domain specific. However, in practice, it is desirable to produce a single set of final weights which can be used for all domains. So to make the Hajek-type adjustment on a number of domains simultaneously, one can perform a multivariate calibration which is a type of poststratification with domain-specific controls $\tilde{N}(d)$ corresponding to a selected set of domains. These are in addition to the nondomain-specific controls obtained from the first phase screener data. This idea of multivariate calibration is similar to the one used in multivariate regression composite estimation by Singh, Kennedy, and Wu (2001). After the above calibration for poststratification, the resulting weights may still have extreme weights which may cause some instability. To address this concern, a repeat poststratification step with the same controls but with tighter bounds on the extreme weights identified at the previous poststratification step can be performed as suggested by Folsom and Singh (2000). This repeat poststratification redistributes the total weight such that sample distributions for various demographic domains are preserved.

It may be noted that, in practice, due to item nonresponse about pair relationships, imputation for the person-level $(M_{hi}(d))$, and the household level $(N_h(d))$ multiplicities may be required for certain pairs and households.

4. Summary

It is clear that weights based on pairwise probabilities are required for many drug behavior analyses of the NHSDA data. For this purpose the analyst needs to make some fundamental decisions about defining population parameters when the person has the same relationship (parent of or child of) to more than one person in the household. For the two problems of multiplicities and extreme weights that might arise in pair data analysis, it was shown how the estimator could be adjusted in the presence of multiplicities, and how the weights could be calibrated to alleviate the problem of extreme weights.

References

Brewer, R. K. W. (1963). "A Model of SystematicSampling With Unequal Probabilities." *Australian Journal of Statistics* **5**: 5-13.

Hajek, J. (1971). Comment on D. Basu's paper "An essay on the logical foundations of survey sampling, part one", in *Foundations of Statistical Inference* (eds. Godambe and Sprott), Holt, Rinehart, and Winston, p 236.

Folsom, R.E., and Singh, A.C. (2000). A generalized exponential model for weight calibration for adjustments for extreme weights, nonresponse and poststratification. *ASA Proc. Surv. Res. Meth. Sec.*, 598-603.

Singh, A.C., Kennedy, B., and Wu, S. (2001). Regression composite estimation for the Canadian Labour Force Survey. *Survey Methodology*, 27, 33-44.

Sirken, M.G. (1972). Stratified sample surveys with multiplicity. JASA, 65.

Table 4. Comparison of Alternate Approaches to Defining Population Parameters for Parent-Child Relationships

Feature	Population and reporting strategy			
	Population = pairs	Population = children		Population = parents
	Option 1. Report for all possible parent-child pairs.	*Option 2*. Report mother-child and father-child separately.	*Option 3*. Average parent-child relationships if child resides with both parents.	*Option 4*. Average parent-child relationships if parent resides with children.
Population units	All parent-child pairs where both reside in the same dwelling unit.	A. Mother-child pairs. B. Father-child pairs. (both reside in the same dwelling unit.	Children residing in dwelling units with at least one parent present.	Parents residing in dwelling units with one or more of their own children.
Tables required	1	2	1	1
Population size	Count of children living with mothers + count of children living with fathers - the count of children living with both parents.	A. Count of children living with mothers for table 1. B. Count of children living with fathers for table 2.	Count of children living with one or both parents.	Count of parents living with one or more children.
Population Total parameters	$\sum_{[parent-child\ pairs]} y_{hip}$	A. $\sum_{[mother-child\ pairs]} y_{hip}$ B. $\sum_{[father-child\ pairs]} y_{hip}$	$\sum_{[parent-child\ pairs]} \frac{y_{hip}}{M_{hi}(1)}$	$\sum_{[parent-child\ pairs]} \frac{y_{hip}}{\pi_{hip}}$
Multiplicity factor	None	None	$M_{hi}(1)$, the number of parents residing in the same hth DU with the child.	$M_{hi}(2)$, the number of children residing in the same hth DU with the parent.
Population Total Estimator	$\sum_{[parent-child\ pairs]} \frac{y_{hip}}{\pi_{hip}}$	A. $\sum_{[mother-child\ pairs]} \frac{y_{hip}}{\pi_{hip}}$ B. $\sum_{[father-child\ pairs]} \frac{y_{hip}}{\pi_{hip}}$	$\sum_{[parent-child\ pairs]} \frac{y_{hip}}{M_{hi}(1)\pi_{hip}}$	$\sum_{[parent-child\ pairs]} \frac{y_{hip}}{M_{hi}(2)\pi_{hip}}$