

Triad Sampling in Household Surveys

Jeremy Aldworth and James R. Chromy

Statistics Research Division, RTI International, Research Triangle Park, NC 27709

Key Words: Brewer's Method, Sampford's Method, Household Rosters, Person Selection, Unequal Probability Sampling

Introduction

The National Survey on Drug Use and Health (NSDUH) is currently designed to produce data about both individuals and pairs residing in the same dwelling unit (DU). There is now interest in extending the sampling mechanism to produce data about triads (i.e., groups of three) within the same DU. Triads of particular interest consist of two parents and a child, or one parent and two children.

Computer-assisted dwelling unit screening provides the mechanism for targeted sampling of individuals, pairs, and even triads of individuals within a DU. Selection algorithms can be programmed into the screening instrument to implement unequal probability sample selection. Chromy and Penne (2002) showed how an adaptation of Brewer's (1963, 1974) method for samples of size two was used to select samples of 0, 1, or 2 persons from DUs containing at least one eligible¹ person. They also developed a second modification to control the number of pairs selected. In this paper, Chromy and Penne's adaptation of Brewer's method and their second modification to control the number of persons selected from a DU is extended to the case of sampling triads within DUs containing at least one eligible person.

Many surveys limit person selection so that only one person is selected per DU. Selection schemes can be designed to select eligible persons with equal probability or to oversample certain domains (e.g., domains defined by age, race, or gender). Because the number of eligible persons per DU may vary substantially, the overall design-based weight may vary considerably among sample DUs.

If the selection procedure is designed so that any three persons in the same DU always have a positive probability of all three being selected, then it is also possible to support an analysis of triad data and to study the relationships among person triads residing in the same DU. Of particular interest are triads consisting of two parents and a child, or one parent and two children. Note that within any selected triad, there are three combinations of selected pairs, so relationships among pairs can also

be studied. Allowing selection of up to three persons per DU may also reduce the number of DUs that must be surveyed and, as a result, reduce overall survey costs. Possible negative impacts of sampling triads of persons from the same DU include possible reductions in response rates, an increased design effect due to greater clustering within a DU, and possible response biases.

The NSDUH is currently designed to select up to two persons in the same DU. In the 2002 survey about 180,000 DUs were screened for eligibility and a roster of eligible persons aged 12 or older was obtained. Target sample sizes were specified for each age group (5) and state (51), giving a total target sample size of about 80,000 persons. The highest sampling rates were applied to persons aged 12 to 17 and persons aged 18 to 25. To achieve near equal probability sampling (*Epssem*²) within state-age groups, within-DU samples of 0, 1, or 2 persons were allowed. The sample design required that every eligible person must have a positive probability of selection and every within-DU person-pair must also have a positive probability of selection. The screening interview was conducted using a hand-held computer. After the interviewer had enumerated and recorded all eligible persons at a DU, the computer was programmed to select the sample of 0, 1, or 2 persons.

In this paper, the pair sampling procedure is extended so that person triads in the same DU can be selected. This includes a second modification along the lines of Chromy and Penne (2002) in order to increase the number of pair and triad selections and to reduce the number of single-person selections. To illustrate this triad sample selection process, simulations were applied to several years of NSDUH survey data, and for several values of the tuning parameter governing the numbers of single, pair, and triad selections. In addition, empirical response rates were obtained from some of the same data for the purpose of seeing how response rates may be affected by choosing more than one person from a DU.

Basic Triad Sampling Algorithm

Sampford (1967) formulated a sample selection method to select a sample of size n from a population (or stratum) of size $N > n$, and where the first-order inclusion probabilities are proportional to some size measure. This method is now adapted to the problem

¹The NSDUH target population includes all civilians aged 12 or older not residing in institutions. Eligible dwelling units include both housing units and rooms or persons within non-institutional group quarters.

²Kish (1963, p 21) uses this term to describe any selection method for which population elements have equal probabilities of selection.

of selecting 0, 1, 2, or 3 persons from every DU with at least one eligible person. Within each state, target sampling rates can be set for different age groups, and target selection probabilities can then be set within DUs to achieve an Epssem design within those state-age groups.

Define the target selection probability for person i in DU h as P_{hi} . Then, to insure that all triads have a positive probability of selection, all person probabilities have to be strictly less than 1; and arbitrarily, the maximum P_{hi} is set to 0.99.

In Sampford's (unadapted) method of sampling triads, the sum of first-order inclusion probabilities is always equal to $n = 3$. However, since a DU does not contain a population (or stratum), it is unlikely that $S_h \equiv \sum_i P_{hi} = 3$. If S_h is strictly greater than 3, a multiplicative scaling factor, $F = 3/S_h$, is applied to all the target selection probabilities so that they are scaled down to sum to exactly 3. If S_h is strictly less than 3, it may not be possible to simply scale the selection probabilities up, since a scaled-up probability may be greater than or equal to 1. This problem is remedied by creating 4 dummy persons and distributing the remaining size measure, $(3 - S_h)$, to them equally. Then, by including the four dummy persons, the sum of the adjusted person probabilities is exactly 3.

Sampford's method can then be applied to select a person triad using the triad selection formula

$$P_{hijk} = \frac{P_{hi}P_{hj}P_{hk}}{K} \times \left[\frac{1}{(1-P_{hi})(1-P_{hj})} + \frac{1}{(1-P_{hi})(1-P_{hk})} + \frac{1}{(1-P_{hj})(1-P_{hk})} \right],$$

where

$$K = 3 + 2 \sum_i \frac{P_{hi}}{(1-P_{hi})} + \sum_i \sum_{i < j} \frac{P_{hi}P_{hj}}{(1-P_{hi})(1-P_{hj})},$$

and P_{hijk} is the probability of selecting persons i, j , and k from DU h .

If the selected triad consists of three real persons (no dummy persons), then all three persons are selected. If the selected triad consists of two real persons and one dummy person, then a pair is selected. If the selected triad consists of one real person and two dummy persons, then one person is selected. If the selected pair consists of three dummy persons, no one is selected from that DU. Within any selected triad there are three combinations of selected pairs.

The probability of selecting a pair within a DU containing $r \geq 3$ persons (real or dummy) is given by the following formula:

$$P_{hmn} = \sum_{k \neq m, n} P_{hmnk},$$

where P_{hmn} is the probability of selecting persons m and n from DU h .

A more computationally tractable form of this equation is

$$P_{hmn} = \frac{1}{K} \left\{ \frac{P_{hm}P_{hn}}{(1-P_{hm})(1-P_{hn})} \left[\sum_i \frac{P_{hi}(3-P_{hi}-P_{hm}-P_{hn})}{(1-P_{hi})} - \frac{P_{hm}(3-2P_{hm}-P_{hn})}{(1-P_{hm})} - \frac{P_{hn}(3-P_{hm}-2P_{hn})}{(1-P_{hn})} \right] \right\},$$

where K is defined as above. And, as discussed above in the case of triads, if the selected pair consists of two real persons then a pair is selected.

Modification to Increase Number of Triads

Initial simulations of this method on NSDUH data demonstrated that the method worked as predicted, but for certain age-group combinations too few triads were selected. As a consequence, a simple modification of the method (based on Chromy and Penne, 2002) was applied to increase the number of triads selected for those age-group combinations, while maintaining the Epssem person selections within state-age groups. This modification can only be implemented when the sum of person probabilities S_h is less than 3. The general approach is to increase the number of times 2 or 3 persons are selected, but to maintain the person selection probabilities. To accomplish this, we maintained the relative sizes of the selection probabilities within all DUs, and scaled the individual person probabilities to sum to 3 or to as close to 3 as possible without violating some other condition required for probability sampling of persons and triads. This reduced to computing a scaling factor (≥ 1) for DUs where the initial sum of target person selection probabilities is less than 3. Then, based on a preliminary random number, the computer does not select any persons from the DU or it applies the basic sampling algorithm with the scaled up probabilities of selection.

Operationally, when the initial sum of person probabilities S_h is less than 3, the scaling factor is computed as

$$F_s = \min \left\{ \frac{T(\lambda)}{S_h}, \frac{0.99}{\max\{P_{hi}\}} \right\}$$

where

$$T(\lambda) = S_h + \lambda(3 - S_h); \quad 0 \leq \lambda \leq 1.$$

This ensures that no person selection probability is adjusted to be greater than 0.99; if this is not a limitation and $\lambda = 1$, then the sum of person selection probabilities is set to exactly 3. Based on a

preliminary uniform (0,1) random number, R_1 , the basic sampling algorithm is then applied with the scaled up person selection probabilities if $R_1 \leq 1/F_S$. Otherwise, no persons are selected from the DU. The tuning parameter λ provides a measure of control over the size of the scaling factor. Thus, if $\lambda = 0$, then $F_S = 1$, and no change in the basic algorithm occurs; if $0 < \lambda \leq 1$, then $F_S \geq 1$. If $F_S > 1$, then triad and pair selection probabilities are increased, while the person selection probabilities are kept the same. Triad and pair selection probabilities are increased since the method results in fewer instances of selecting exactly one (real) person and more instances of selecting 0, 2, or 3 persons. Subject to the upper limit, $0.99/\max\{P_{hi}\}$, the scaling factor increases as S_h , the sum of person probabilities in a DU, decreases. And since the person probabilities differ by age group, the scaling factor is going to be influenced not only by the number of eligible persons in a DU, but also how those persons are categorized by age group.

Simulation Study of Modified Algorithm

The 2000, 2001, and 2002 survey DU rosters and target selection probabilities were used to simulate the impact of applying the modified triad sampling algorithm for different values of λ .

Specifically, for each year in question and for each value of $\lambda = 0.00, 0.25, 0.50, 0.75$, and 1.00 , the numbers of persons, pairs, and triads selected were computed from a single simulation of the survey data, and the results are given in Table 1a. The P_{hi} values used in the simulations were the actual state-age group probabilities that were used in the surveys. Note that since the results were obtained from only one simulated outcome in each case, there will be some sampling variability in this simulation exercise.

Table 1a shows that for each year, the total number of persons selected does not vary much with λ , and this minor variation can be attributed to simulation error. On the other hand, the total number of pairs and triads increases fairly markedly with λ , as expected. The increase in the number of pairs is not linear; the largest incremental increase occurs from $\lambda = 0$ to 0.25 , and then from $\lambda \geq 0.75$ there is little or no increase. Maximum increases in the number of selected pairs climb to around 10,000 extra pairs (roughly a 30% increase). The increase in the number of triads approximates linearity better. Maximum increases in the number of selected triads are relatively smaller at around 1,000 extra triads (roughly a 15% increase).

Table 1a. Numbers of persons, pairs, and triads selected by year and λ , computed from one simulated outcome in each case.

Year	Selection	λ				
		0.00	0.25	0.50	0.75	1.00
2000	Persons	97,783	97,829	97,495	97,910	97,246
	Pairs	33,582	38,342	42,133	44,314	44,100
	Triads	5,560	5,851	6,058	6,451	6,589
2001	Persons	95,880	95,674	95,741	95,859	95,578
	Pairs	35,245	39,456	43,123	45,088	45,048
	Triads	6,167	6,425	6,558	6,947	7,026
2002	Persons	86,374	86,473	86,450	86,016	86,251
	Pairs	32,428	36,411	39,891	41,109	41,560
	Triads	5,782	6,073	6,319	6,549	6,740

Compare the 2000 results of Table 1a with those of Chromy and Penne (2002) given in Table 1b, in which the pair sampling algorithm was applied to the 2000 survey data for the same values of λ . The triad sampling algorithm yields a larger number of persons and pairs selected. Averaging over the five levels of λ , the pair sampling algorithm yields 92,644 persons,

but the triad sampling algorithm yields 97,653 persons from the same DU data. This suggests that approximately 94.9% of those DUs would have been required to yield the desired sample size if the triad sampling algorithm had been used in the 2000 survey. This gives an indication of the potential cost savings due to triad sampling over pair sampling.

Table 1b. Numbers of persons and pairs by λ , computed from one simulated outcome of the 2000 survey data in each case (Chromy and Penne, 2002).

Year	Selection	λ				
		0.00	0.25	0.50	0.75	1.00
2000	Persons	92,942	92,795	92,650	92,495	92,339
	Pairs	22,849	25,146	27,444	29,738	32,031

Since the scaling factor typically differs considerably by age group within states, the number of persons, pairs, and triads selected is also broken down by age group. For simplicity, the simulated results of the 2002 NSDUH will only be displayed.

The total number of persons selected, and the percentage of those selected in which only one person per DU was selected, is categorized by age group in Table 2a.

Table 2a. Total number (N) of persons selected, and percentage of those persons in which exactly one person per DU was selected, by age group and λ (NSDUH 2002).

Age Group of Person	λ									
	0.00		0.25		0.50		0.75		1.00	
	N	%	N	%	N	%	N	%	N	%
12-17	29,241	31	29,252	30	29,271	30	29,217	30	29,279	30
18-25	29,173	39	29,241	33	29,119	32	29,126	32	28,979	32
26-34	8,050	72	7,887	50	7,877	27	7,773	20	7,947	19
35-49	12,357	52	12,526	37	12,516	21	12,520	15	12,603	16
50+	7,553	83	7,567	62	7,667	34	7,380	23	7,443	22
All ages	86,374		86,473		86,450		86,016		86,251	

Within each age group, the total number of persons selected remains more or less constant as a function of λ , reflecting the fact that the person probabilities are maintained, give or take a small amount of sampling variability. However, the percentage of those persons who come from DUs in which they were the only one selected changes with age group and λ . For 12 to 17 year olds, the percentage of those who come from DUs in which only one person was selected is constant around 30%. For 18 to 25 year olds, the percentage decreases from 39% to 33% at $\lambda = 0.25$, and thereafter remains constant around 32%. For the 26 or older age groups, there are huge percentage decreases. For example, in

the 50 or older age group, the percentage decreases from 83% at $\lambda = 0$ to 22% at $\lambda = 1$.

The total number of pairs selected is categorized by age group pairs in Table 2b. If the youngest person in a selected pair is 12 to 17 years old, then the total number of pairs selected is nearly constant. If the youngest person in a selected pair is 18 to 25 years old, then the total number of pairs selected increases a little, in the range of 11.9% to 32.7% for $\lambda \geq 0.75$. If the youngest person in a selected pair is at least 26 years old, then the total number of pairs selected may increase hugely, as Table 2b shows.

Table 2b. Total number of pairs selected, by age groups and λ (NSDUH 2002).

Age Groups of Persons in Pair	λ				
	0.00	0.25	0.50	0.75	1.00
12-17, 12-17	7,821	7,881	7,951	7,889	7,967
12-17, 18-25	5,626	5,640	5,577	5,596	5,566
12-17, 26-34	1,047	1,039	1,097	1,073	1,056
12-17, 35-49	5,652	5,814	5,819	5,843	5,847
12-17, 50+	676	712	684	665	690
18-25, 18-25	7,295	7,973	8,111	8,260	8,165
18-25, 26-34	974	1,191	1,229	1,254	1,266
18-25, 35-49	1,837	2,145	2,290	2,179	2,339
18-25, 50+	624	791	774	811	828
26-34, 26-34	231	782	1,491	1,676	1,831
26-34, 35-49	155	445	765	923	980
26-34, 50+	64	173	365	395	430
35-49, 35-49	273	851	1,529	1,874	1,878
35-49, 50+	70	325	652	764	722
50+, 50+	83	649	1,557	1,907	1,995
All ages	32,428	36,411	39,891	41,109	41,560

The total number of triads selected is categorized by age group triads in Table 2c. If the youngest person in a selected triad is 12 to 17 years old, then λ has little change on the number of triads selected. If the youngest person is 18 to 25 years old, then there is some evidence of increased numbers of triads (up to 150 in some cases). If the youngest person in a selected triad is at least 26 years old, then the number of triads increases from zero in all cases for $\lambda = 0$, to a number in the range of 23 to 58 selections for $\lambda = 1$.

Table 2c also displays the number of triads selected within specific age group combinations that might be of interest. For example, consider triads consisting of two parents and a child, or one parent and two children. Suppose the children are 12 to 17 years old, and the parents are at least 35 years old. Then the number of triads selected for two children and one parent younger than 50 is much greater than if that parent were 50 or older, or if one child and two parents were selected. In these cases, λ will have little effect on the number of triads selected, since the youngest person in the triad is 12 to 17 years old.

Table 2c. Total number of triads selected, by age groups and λ (NSDUH 2002).

Age Groups of Persons in Triad	λ				
	0.00	0.25	0.50	0.75	1.00
12-17, 12-17, 12-17	712	694	716	692	720
12-17, 12-17, 18-25	828	837	848	859	836
12-17, 12-17, 26-34	189	181	208	192	184
12-17, 12-17, 35-49	1,289	1,355	1,359	1,401	1,397
12-17, 12-17, 50+	112	127	117	111	123
12-17, 18-25, 18-25	548	546	504	521	498
12-17, 18-25, 26-34	56	66	65	65	54
12-17, 18-25, 35-49	633	667	660	628	663
12-17, 18-25, 50+	89	80	99	94	105
12-17, 26-34, 26-34	14	15	18	14	15
12-17, 26-34, 35-49	12	22	15	14	19
12-17, 26-34, 50+	6	4	2	6	2
12-17, 35-49, 35-49	92	85	77	85	77
12-17, 35-49, 50+	12	17	16	15	14
12-17, 50+, 50+	0	2	4	4	4
18-25, 18-25, 18-25	751	784	838	898	894
18-25, 18-25, 26-34	96	104	130	152	154
18-25, 18-25, 35-49	206	271	313	293	352
18-25, 18-25, 50+	76	112	110	118	125
18-25, 26-34, 26-34	10	13	16	15	15
18-25, 26-34, 35-49	13	15	13	19	13
18-25, 26-34, 50+	13	15	13	15	9
18-25, 35-49, 35-49	16	27	44	25	26
18-25, 35-49, 50+	5	10	9	8	17
18-25, 50+, 50+	4	8	7	8	11
26-34, 26-34, 26-34	0	1	13	26	48
26-34, 26-34, 35-49	0	4	9	17	47
26-34, 26-34, 50+	0	0	16	36	49
26-34, 35-49, 35-49	0	1	10	23	35
26-34, 35-49, 50+	0	1	15	32	43
26-34, 50+, 50+	0	3	21	28	30
35-49, 35-49, 35-49	0	1	5	13	23
35-49, 35-49, 50+	0	0	12	51	48
35-49, 50+, 50+	0	3	8	43	32
50+, 50+, 50+	0	2	9	28	58
All ages	5,782	6,073	6,319	6,549	6,740

In summary, based on a single simulation from the NSDUH 2002 data for each value of λ , the effects of the scaling factor as a function of λ are listed as follows:

- No effect on the number of persons selected.
- Almost no effect on the number of pairs or triads selected if the youngest selected person in a DU is 12 to 17 years old.
- Some effect on the number of pairs or triads selected if the youngest selected person in a DU is 18 to 25 years old.
- If the youngest selected person in a DU is at least 26 years old, then some choices of λ substantially increase the number of pairs selected, substantially reduce the number of DUs in which only one person is selected, and increase the number of triads from zero to double-digit numbers in all cases.

NSDUH 2002 Response Rates

Weighted response rates based on NSDUH 2002 data have been estimated for different combinations of age groups and roster information, as an aid to choosing an appropriate value of λ , since the number of pair and triad selections for different age group combinations may have an effect on specific and even overall response rates.

In Table 3a, weighted response rates of *single-person selections* for each age group are given

separately for two DU composition types: (1) DUs containing only one eligible person, and (2) DUs containing at least two eligible persons. In addition, the difference in response rates between the two DU types is also given.

Note that when all age groups are combined, there is little difference in response rates between the DU types. However, this is no longer true when differences in response rates are estimated within each age group. But before comparisons within age groups are carried out, first note that there are only 44 cases in which the only person eligible in a DU is 12 to 17 years old. This is not due to a sampling aberration, but rather to the fact that in the US relatively few DUs contain a single 12 to 17 year old, or a single 12 to 17 year old who heads a family of other persons younger than 12. This fact will have to be taken into account for any trend and comparison analyses.

Thus, by excluding the cell representing 12 to 17 year olds of the first DU type, there is clear evidence of a nearly linearly decreasing response rate by age group within each DU type. And, by excluding the youngest age group, there is also statistically significant evidence that the response rate is higher for the first DU type, for all other age groups, except the oldest. And even in this latter age group, the difference is still positive.

Table 3a. Weighted response rates (RRs) of single-person selections by number eligible per DU (1 versus 2 or more) and age group (NSDUH 2002).

Age Group	1 person eligible in DU			≥ 2 persons eligible in DU			Difference (1 person – ≥ 2 persons)			
	N	RR	SE	N	RR	SE	N	RR	SE	P-val
12-17	44	0.881	0.0575	8,504	0.902	0.0039	8,548	-0.021	0.0576	0.712
18-25	2,840	0.893	0.0077	7,063	0.833	0.0054	9,903	0.060	0.0092	<0.001
26-34	1,079	0.840	0.0133	2,598	0.791	0.0100	3,677	0.050	0.0167	0.003
35-49	1,445	0.798	0.0130	2,973	0.765	0.0095	4,418	0.034	0.0156	0.031
50+	1,693	0.734	0.0126	2,552	0.724	0.0108	4,245	0.010	0.0167	0.566
All ages	7,101	0.778	0.0079	23,690	0.774	0.0053	30,791	0.003	0.0094	0.728

Weighted response rates based on NSDUH 2002 data have also been estimated for *each person in a pair* by the age group of that person (age group 1) and the age group of the other in the pair (age group 2), and these are given in Table 3b. The response rate of one in a pair typically decreases as the age group of the other in the pair increases. The converse of this is also true: The response rate of one in a pair typically increases as the age group of the other in the pair decreases. By comparing Tables 3a and 3b and by limiting the comparison to DUs with 2 or more persons eligible, we also see that the response rate of one in a pair is always greater than that of a single-person selection of the same age

group, if the other in the pair is 12 to 17 years old. It is also always the case that the response rate of one in a pair is always less than that of a single-person selection of the same age group, if the other in the pair is 50 or older.

In summary, based on the NSDUH 2002 data, response rates appear to be affected more by the composition of the DU (i.e., containing one vs. at least two eligible persons), and by the age group of the respondent (and age group of other in pair, if a pair is selected), than by whether or not one or two persons are selected from DUs containing at least two eligible persons.

Table 3b. Weighted response rates of each person in selected pairs by the age group of that person (Age Group 1) and the age group of the other person in the pair (Age Group 2) (NSDUH 2002).

Age Group 1	Age Group 2	Number Selected	Response Rate	SE
12-17	12-17	9334	0.914	0.00498
	18-25	3245	0.886	0.00792
	26-34	826	0.904	0.01257
	35-49	3795	0.879	0.00642
	50+	482	0.837	0.02202
18-25	12-17	3245	0.869	0.00811
	18-25	11040	0.859	0.00522
	26-34	975	0.804	0.01551
	35-49	1449	0.828	0.01293
	50+	604	0.799	0.02100
26-34	12-17	826	0.844	0.01566
	18-25	975	0.781	0.01663
	26-34	1548	0.766	0.01807
	35-49	450	0.815	0.02083
	50+	196	0.662	0.04102
35-59	12-17	3795	0.840	0.00713
	18-25	1449	0.763	0.01505
	26-34	450	0.786	0.02303
	35-49	1614	0.749	0.01646
	50+	350	0.721	0.02882
50+	12-17	482	0.765	0.02496
	18-25	604	0.737	0.02341
	26-34	196	0.705	0.04085
	35-49	350	0.709	0.02983
	50+	1510	0.657	0.01862

Conclusions

This paper has demonstrated the feasibility of selecting samples of 0, 1, 2, or 3 persons from DUs containing at least one eligible person, while at the same time maintaining the target person probabilities by age group. A modification of the method to increase the number of pairs and triads is under the control of a tuning parameter that has also been demonstrated through simulation. Empirical data on response rates based on DU composition and the age of the selected persons show that household composition and age may have greater effects on response rates than the choice of tuning parameter.

Acknowledgements

This paper was developed with the support of the Substance Abuse and Mental Health Services Administration, Office of Applied Studies, under Contract no. 283-2004-00022.

References

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

Brewer, K. R. W. (1975), "A Simple Procedure for Sampling ppswor," *Australian Journal of Statistics* 17:166-172.

Chromy, J. R., and M. A. Penne (2002), "Pair Sampling in Household Surveys," *Proceedings of the Survey Research Methods Section, American Statistical Association*. New York, NY.

Kish, Leslie. (1965). *Survey Sampling*. New York: John Wiley & Sons, Inc.

Sampford, M. R. (1967). "On Sampling without Replacement with Unequal Probabilities of Selection." *Biometrika* 54:499-513.