PAIR SAMPLING IN HOUSEHOLD SURVEYS

James R. Chromy and Michael A. Penne
Statistics Research Division, RTI International, Research Triangle Park, NC 27709

KEY WORDS: Household Rosters, Person Selection, Unequal Probability Sampling, Pair Sampling

Computer-assisted dwelling unit screening provides the mechanism for targeted sampling of both individuals and pairs of individuals. Selection algorithms can be programmed into the screening instrument to implement unequal probability sample selection. This paper discusses a modification of Brewer's method for samples of size two as a means of selecting samples of 0, 1, or 2 persons from eligible dwelling units. A second adaptation which can be used to control the number of pairs selected is also discussed. Some empirical data on household roster composition, and response rates based on the number of persons selected, are presented from the National Survey on Drug Use and Health (NSDUH). Finally, the results of some simulation of the sample selection and response process are presented as a means of evaluating alternatives.

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

Allowing selection of up to 2 persons per dwelling unit can help reduce the number of dwelling units that must be surveyed and, as a result, reduce overall survey costs. If the selection procedure is designed so that any two persons in the same dwelling unit always have a positive probability of both being selected, then it is also possible to support an analysis of pair data and study the relationships among person pairs residing in the same dwelling unit. Possible negative impacts of sampling pairs of persons from the same dwelling unit include possible reductions in response rates, an increased clustering effect due to multiple observations from the same dwelling unit, and possible response biases.

The National Survey of Drug Use and Health is designed to produce data both about individuals and about person pairs residing in the same dwelling unit. About 200,000 dwelling units are screened for eligibility and to obtain a roster of eligible[1]persons aged 12 or older. The target sample size of about 70,000 persons is targeted by state and by five age groups within each state. The highest sampling rates are applied to persons aged 12 to 17 and persons aged 18 to 25. To achieve near equal probability sampling (*Epsem*[2]) within state-age groups, within dwelling unit samples of 0, 1, or 2 persons are allowed. The sample design requires that every eligible persons must have a positive probability of selection and every within-dwelling unit person-pair must also have a positive probability of selection. The screening interview is conducted using a hand-held computer. After the interviewer has enumerated and recorded all eligible persons at a dwelling unit, the computer is programmed to select the sample of 0, 1, or 2 persons.

The Basic Sampling Algorithm (Option 1)

The use of a computer-aided dwelling unit screening procedure permits the application of sampling procedures which can adapt to the household composition and more nearly meet the stated sample allocation goals for both persons and person-pairs. Brewer's (1963, 1975) sample selection method for samples of size 2 was adapted to the problem of selecting 0, 1, or 2 persons from each dwelling unit. Within each state, target sampling rates were set by five age groups. Target selection probabilities were set within dwelling unit to achieve an Epsem design within state-age groups. The target selection probability for person *i* in dwelling unit *h* was set defined as P(*hi*). To insure that all person-pairs would have a positive probability of selection, all individual person probabilities had to be strictly less than 1; arbitrarily, the maximum P(*hi*) was set at 0.99.

Since the design only allowed for selection of 0, 1, or 2 persons per dwelling unit, it was necessary to bound the sum of the final probabilities to be less than 2. If their sum, S, was more than 2, a multiplicative scaling factor, F=2/S, was applied to all the target selection probabilities so that they were scaled down to sum to exactly 2. If their sum was strictly less than 2, it was not possible to apply Brewer's method for sample size of 2. This problem was remedied by creating 3 dummy persons and distributing the remaining size measure, (2 - S), to them equally. Now by including the three dummy persons, the sum of the adjusted person probabilities was exactly 2. Brewer's method could then be applied to select a person pair using the pair selection formula

[1]The 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.

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

$$P(hi, hj) = \frac{P(hi)P(hj)}{K}\left[\frac{1}{1-P(hi)} + \frac{1}{1-P(hj)}\right]$$

where

$$K = 2 + \sum \frac{P(hi)}{1-P(hi)}.$$

If the selected pair consisted of two real persons (no dummy persons), then both persons were selected. If the selected pair consisted of one real person and one dummy person, then the one person was selected. If the selected pair consisted of two dummy persons, no one was selected from that dwelling unit. Person and person-pair probabilities [P(*hi*), P(*hj*), and P(*hi*,*hj*)] were retained from the values used in selecting the pair sample.

This basic approach, designated as Option 1, was utilized in the 1999, 2000, and 2001 surveys.

Sampling Algorithm Option 2

Based on a review of the 1999 and 2000 sample results, some ways of increasing the number of pairs were considered. Sampling algorithm Option 2 represents the maximum potential to increase the number of pairs and still maintain the Epsem person selection within state-age groups while utilizing a simple adaptation of Brewer's method.

Option 2 modified the Option 1 methodology only when the sum of person probabilities, S, was less than 2. The general approach was to select two persons whenever possible, but to maintain the person selection probabilities. To accomplish this, we maintained the relative sizes of the selection probabilities within all dwelling units, and scaled the individual person probabilities to sum to 2 or as close to as possible without violating some other condition required for probability sampling of persons and pairs. This reduced to computing a scaling factor (greater than 1) for dwelling units where the initial sum of target person selection probabilities was less than 2. Then, based on a preliminary random number, the computer did not select any persons from the dwelling unit or it applied the Option 1 algorithm with the scaled up probabilities of selection. Operationally, when the initial sum of person probabilities, S, was less than 2, the scaling factor was computed as

$$F_s = Min\left\{\frac{2}{S}, \frac{0.99}{Max\{P(hi)\}}\right\}.$$

This ensured that no person selection probability was adjusted to be greater than 0.99; if this was not a limitation, the sum of person selection probabilities was set to exactly 2. Based on a preliminary uniform (0,1) random number, R_1, the Option 1 algorithm was applied with the scaled up person selection probabilities if $R_1 \leq 1/F_S$. Otherwise, no persons were selected from the dwelling unit.

This had the effect of preserving the person selection probabilities and maintaining or increasing the pair selection probabilities. Pair selection probabilities were increased since the method resulted in fewer instances of selecting exactly one (real) person and more instances of selecting 0 or 2 persons. The 2000 survey dwelling unit roster and target selection probabilities were used to simulate the impact of applying the Option 2 algorithm. Option 1 and 2 results are compared in Table 1. The changes in persons selected, completed screeners required, and dwelling unit (DU) selections required are small and, probably, attributable to the random nature of the simulation. Option 2 was successful in increasing the number of pairs selected by over 40 percent.

Table 1. Selected Sample Characteristics using Two Selection Algorithms

Item	Option 1	Option 2	Change
Persons selected	92,942	92,339	-0.65%
Pairs selected	22,849	32,031	40.19%
Completed screeners	169,069	170,173	0.65%
DU selections	214,970	216,374	0.65%

One suspected adverse impact of selecting more pairs was that person response rates might decrease when more than one person in a dwelling unit is asked to participate in the study. The 2000 survey data on response rates by pair selection domains (all combinations of 5 age groups plus no other person selected) were used to compare projected response rates using the two algorithms. The results, summarized in Table 2, show a projected reduction in response rates of over 1 percent overall, with little difference in the younger age groups and quite a large reduction in response rates (over 6 percent) for persons 50 or older. The full detail cannot be shown in a small table, but most of the increase in selected pairs when moving from Option 1 to Option 2 occurs at the older age groups. The younger age groups are already being included in pair selections naturally with the application of the Option 1 algorithm.

Table 2. Projected Response Rates using Two Selection Algorithms

Age group	Option 1	Option 2	Difference
12 or older	77.53%	76.34%	-1.19%
12 to 17	81.85%	81.84%	-0.01%
18 to 25	76.99%	76.60%	-0.39%
26 to 34	75.26%	73.39%	-1.87%
35 to 49	73.74%	71.82%	-1.92%
50 or older	71.74%	65.46%	-6.28%

Sampling Algorithm Option 3

To overcome the large adverse impact on response rates while still increasing pair selections somewhat, an intermediate approach was proposed. Option 3, like Option 2, also only impacts dwelling units where the sum of person probabilities is initially less than 2. Instead of scaling the person probabilities up the maximum amount allowed by other sample design constraints, Option 3 allows a partial scaling adjustment. Operationally, this is achieved by defining an intermediate target for the sum of the size as $T(\lambda) = S + \lambda(2 - S)$. We then modify the computation of the scaling factor as $T(\lambda) = Min\left\{\frac{T(\lambda)}{S}, \frac{0.99}{Max\{P(hi)\}}\right\}$.

Note that with $\lambda = 0$, this reduces to Option 1 and that with $\lambda = 1$, this reduces to Option 2. Operational procedures for Option 3 are analogous to the Option 2 procedures

Table 3 shows the simulated number of pairs selected for intermediate values of λ. Table 4 shows interpolated response rates by age group.

Table 3. Simulated Pair Selections

λ	Projected Pair Selections
0.00 (Option 1)	22,849
0.25	25,145
0.50	27,440
0.75	29,736
1.00 (Option 2)	32,031

Table 4. Interpolated Response Rates

Age group	Option 3 with λ =				
	0.00	0.25	0.50	0.75	1.00
12 or older	0.775	0.772	0.769	0.766	0.763
12 to 17	0.819	0.818	0.818	0.818	0.818
18 to 25	0.770	0.769	0.768	0.767	0.766
26 to 34	0.753	0.748	0.743	0.739	0.734
35 to 49	0.737	0.733	0.728	0.723	0.718
50 or older	0.717	0.701	0.686	0.670	0.655

After further consideration, Option 3 with $\lambda = 0.50$ was implemented for the 2002 Survey. This option increased the number of pairs by about 20 percent with only moderate impact on response rates by age group. Examination of early results from the first quarter of work confirmed the projected yields of persons and pairs. A detailed analysis of the response rates by age group and by sample composition still needs to be completed.

REFERENCES

Brewer, R. K. 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.

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