ALTERNATIVE DUAL SYSTEM NETWORK ESTIMATORS

Robert J. Casady and Monroe G. Sirken, National Center for Health Statistics
Gad Nathan, Department of Statistics, Hebrew University, Jerusalem, Israel

INTRODUCTION

When there are two or more data systems and none of them enumerates the population at an acceptable completeness level, concern about the bias due to under-enumeration suggests a dual system estimator which makes joint use of data compiled by the combined imperfect data systems. In some countries, for example, adequate single data systems for enumerating the number of vital events do not exist, although incomplete counts are obtainable from vital registration systems and from household sample surveys. In these instances, the components of population change are sometimes estimated by dual system estimators which make joint use of the data compiled by a vital registration system and a household sample survey - Marks, Selzer and Krotki (1974). Dual system estimators of this type, require that the population be enumerated by two different data systems and the data files be matched to identify the persons enumerated by both systems. Henceforth, we will refer to them as conventional dual system estimators in contrast with the dual system network estimators that are presented in this paper.

The conventional dual system estimator of the number of vital events, say N, is given by

$$\hat{N} = X_1 X_2 / Z_{12}$$

where X_1 and X_2 are estimates of N based on data compiled by the first and second imperfect data systems, respectively. Z_{12}, which estimates the number of events that are enumerated by both systems, is obtained by matching the files of the two data systems and identifying the events that were enumerated in both. Even though X_1 and X_2 are biased, $\hat{N}$ is a consistent estimate of N if the Bernoulli variables representing the enumerations of the same event by each of the data systems are independent.

Conventional dual system estimation poses a number of design problems including the expense of establishing two different data collection systems and the difficulties of matching the events enumerated by both systems. The dual system network estimators presented in this paper are not subject to these particular problems because they do not involve two distinct data systems and do not require matching. It will be easier to describe the difference between dual system network and conventional estimation if the difference between single system network and conventional estimation is dealt with first.

The essential difference between conventional and network estimators is that the former are based on conventional counting rules and the latter are based on multiplicity counting rules. Conventional counting rules specify that none of the population elements is countable at more than one enumeration unit, while multiplicity rules specify that any of the elements may be countable at more than one unit. If, for example, a household sample survey adopts a conventional counting rule such as a de jure residence rule, individuals would be enumerable only at their de jure residences, and if the survey adopts a multiplicity rule, such as a kinship rule, individuals would be enumerated at the households of their relatives.

Dual system network estimators assume a main survey and a follow-up quality check survey. The main survey adopts a primary multiplicity counting rule that combines two partial counting rules. It is assumed that the primary rule gives complete coverage in the sense that all events are eligible to be enumerated by the primary rule. Furthermore, the two partial counting rules are assumed to be mutually exclusive in that any particular event is eligible for enumeration by at most one of the two partial rules at a given enumeration unit.

The events that are enumerated in the main survey are eligible for inclusion in the quality check survey if they are in the dual coverage set, that is, enumerable by both partial counting rules. The events selected for the quality check are re-enumerated at a different set of households than those at which they were originally enumerated in the main survey. They are re-enumerated in the quality check survey at households that are eligible to report them by the complement of the partial counting rule by which they were enumerated in the main survey. Assume, for example, a main survey that enumerates retrospectively the people who died in a prior reference period, by a multiplicity rule that includes the de jure residence rule and the next-of-kin counting rule. The deaths in the dual coverage set that had been enumerated at their former de jure residences in the main survey would be re-enumerated at the residences of their surviving next-of-kin in the quality check survey, and vice versa for the deaths that had been enumerated at the residences of their next-of-kin in the main survey. The household addresses for the quality check survey are obtained from the respondents who reported the deaths in the main survey. A quality check survey of this type conducted in Israel is described in Nathan, Schmelz and Kenvin (1977).

Three dual system network estimators are presented in this paper. One estimator was previously proposed by Sirken (1979) and is the natural analogue of the conventional dual system estimator. The two other dual system network estimators that are presented in this paper were proposed as potential improvements, although neither of them is natural analogue of the conventional dual system estimator. The three dual network estimators are evaluated and their design effects are compared analytically and empirically with one another, and with those of the single system conventional and network estimators.

ESTIMATION OF N

First we define the following counting rule related population parameters:

N_1 = number of events eligible to be reported by the first partial counting rule,

N_2 = number of events eligible to be reported by the second partial counting rule,

and N_{12} = number of events eligible to be reported by both partial rules.

As we have previously assumed that the primary counting rule gives complete coverage it follows immediately that

$$N = N_1 + N_2 - N_{12}. \qquad (1)$$

Secondly, we define the following set of statistics to estimate the components of the right hand side of (1):

X_k = "natural" multiplicity estimator (via partial rule k) of N_k based on event reports in the main survey

Y_k = "natural" multiplicity estimator (via partial rule k) of N_{12} based on event reports in the main survey

W_k = "natural" multiplicity estimator (via partial rule k) of N_k based only on the subsample of event reports in the quality check survey

Z_1 = "natural" multiplicity estimator (via partial rule 2) of N_{12} based on quality check sample of events reported in main survey by partial rule 1

Z_2 = analogue of Z_1 but roles of partial rules 1 and 2 are reversed

In addition, the statistics X_k^*, W_k^* and Z_k^* are defined as analogues of X_k, W_k and Z_k, respectively, with event reports weighted by the reciprocal of the event's total multiplicity (i.e., its multiplicity by the primary counting rule) instead of the reciprocal of its partial multiplicity by rule k.

Finally, using these estimators for the parametric components of N, the following statistics are proposed as estimators of N:

$$\hat{N}_0 = X_1 \qquad (2)$$

$$\hat{N}_A = X_1X_2/[\lambda Z_1 + (1 - \lambda)Z_2] \qquad (3)$$

$$\hat{N}_B = (X_1-\lambda Y_1)\, W_2/Z_2+[X_2-(1-\lambda)Y_2]\, W_1/Z_1 \qquad (4)$$

$$\hat{N}_C = X_1^*\, W_2^*/Z_2^* + X_2^*\, W_1^*/Z_1^* \qquad (5)$$

$$\hat{N}_D = X_1^* + X_2^* \qquad (6)$$

where λ is a constant such that $0 \leq \lambda \leq 1$. In practice λ will be chosen to minimize the variance.

$\hat{N}_0$ is easily seen to be the conventional single system estimator when partial rule 1 is assumed to be the conventional household counting rule and $\hat{N}_D$ is a single system network estimator without quality check. $\hat{N}_A$, $\hat{N}_B$ and $\hat{N}_C$ are variants of dual system network estimators. $\hat{N}_A$ has been previously proposed by Sirken (1979) and is the natural analogue of the conventional dual system estimator. It should be noted that if both partial rules give complete coverage then $X_k= Y_k (k=1,2)$ so that

$$\hat{N}_B= (1-\lambda)X_1W_2/Z_2+\lambda X_2W_1/Z_1.$$

ANALYTICAL RESULTS

The results presented in this section are based on a complex statistical model involving the counting rule coverage, a response error model and sampling specifications for both the main survey and the quality check survey. The principle conditions and assumptions of the statistical model for the dual system network survey are given below:

(a) the primary counting rule gives complete coverage,

(b) sampling in the main survey is simple random without replacement and finite population corrections are negligible,

(c) Bernoulli sampling is utilized for the selection of event reports for inclusion in the quality check survey,

(d) a Hansen, Hurwitz, Bershad (1961) response error model is appropriate for event reporting in both the main survey and quality check survey and event reporting in the two surveys is independent,

(e) at most one event is reported by the primary counting rule at any enumeration unit (only required for derivation of variances).

The following analytic results can be derived from the statistical model specified above:

Result 1 - The expections of the estimators $\hat{N}_A$, $\hat{N}_B$ and $\hat{N}_C$ do not depend on the magnitude of the reporting probabilities specified by the response error model. However, the variances of all three estimators increase as the reporting probabilities decrease.

Result 2 - The estimators $\hat{N}_B$ and $\hat{N}_C$ are consistant for N but $\hat{N}_A$ is an overestimate. $\hat{N}_A$ is also consistant if condition (a) above is strengthened so that at least one of the partial rules gives complete coverage.

Result 3 - If condition (a) is strengthened so that both partial counting rules give complete coverage then $\text{Var}(\hat{N}_B) \leq \text{Var}(\hat{N}_A)$ but $\text{Var}(\hat{N}_B)$ is not necessarily smaller than $\text{Var}(\hat{N}_C)$.

A complete description of the statistical model and the derivation of the results presented above may be found in Casady, Nathan and Sirken (1983).

EMPIRICAL FINDINGS

The empirical evaluation presented in this section is based on the data from pilot study carried out by the National Center for Health Statistics to study the error effects of different counting rules in reporting of deaths retrospectively in a single time sample survey.

A complete description of the study is given by Royston and Sirken (1978). The study involved three stages, based on a stratified sample (by age and by color) of about 1700 death records registered in North Carolina over a ten-month period. In the first stage, for each death, a list of names and addresses of surviving relatives in the United States (spouse, children, siblings and parents) and the names and addresses of the members of the key household (that of the decedent's de jure address at time of death) were obtained from the death record informant.

In the second stage, interviews were conducted at a sample of these addresses to see whether they would report the death. In the third stage, deaths reported by the second stage sample household were matched against the state file of death certificates to evaluate the completeness of death registration.

The results presented in the following relate only to the first and second stages of the pilot study with respect to deaths of persons aged 17 and above.

The maximal network counting rule considered was that linking the death to the key household (KH) and to the households of the decedent's surviving spouse (SP), sibling (SI) and children (CH), if living outside the key household in North Carolina. The primary counting rules considered included the key household and some or all of the households of the decedents' relatives. The first partial rule (k = 1) was the conventional rule (death linked to the key household only) and the second partial rule (K = 2) linked the death to the households of relatives not living in the key household.

Thus, for this empirical example it is reasonable to assume that conditions (a) through (e) specified in the previous section hold. In fact, as the conventional rule gives complete coverage, the stronger assumption specified in Result 2 holds so that N_1=N and N_2=N_{12} < N. Hence, all three of the dual system network estimates are consistant but Result 3 does not necessarily hold.

The dual system network estimators require visits to additional enumeration units (for evaluation in the quality check survey) compared to the conventional and to the single system network for the same initial sample size m. The ratio of the total expected sample size m* required for the dual system to the initial sample size is given by:

$$m^*/m = 1 + (N_{12}/M)\ (f_1 p_1 \bar{s}_1 + f_2 p_2 \bar{s}_2)$$

where

M = number of enumeration units in universe,

f_k = quality check survey sampling rate for events reported in main survey by rule k,

p_k = conditional probability of reporting an event under partial rule k, given it is eligible to be reported at a sampled enumeration unit,

$\bar{s}_k$ = average multiplicity by partial rule k for for events eligible to be reported by partial rule k.

For each of the counting rules considered in Tables 1, 2-a and 2-b the factor m*/m was estimated and the initial sample size was reduced accordingly in computing the variances of the dual system network estimators. This assures that for the comparisons presented in the tables the sample size for the single system estimators and the expected total sample size (i.e., the main survey sample size plus the expected quality check survey sample size) for the dual system network estimators are identical.

Table 1 gives the relative standard errors, the relative biases and the relative root mean square errors for the conventional estimator, $\hat{N}_0$, (first line) and for the single system and dual system network estimators for each of the four counting rules and the two sample sizes considered. The results in Table 1 are for $f_1 = f_2 = 1$ (i.e., all reported events are evaluated).

The table shows that the biases of the single system estimators (which are independent of the sample size) are large relative to the standard errors, especially for the larger sample size. The sampling errors of the single system network estimators are substantially smaller than those of the single system conventional estimator and the biases of the single system network estimators are generally less than that of the single system conventional estimator but for one counting rule (KH+SP+SI) the bias is larger.

The standard error of the single network estimators are less than those of the conventional estimator by over 20%, resulting in overall gains in the root M.S.E., except for the counting rule KH+SP+SI for sample size 40,000. All the dual system estimators show considerable gains in root M.S.E. over the single system estimators. The gains are due to the elimination of the biases since the standard errors for the dual system estimators are larger than the standard errors for the single system estimators. All three dual system estimators have similar errors for $f_1 = f_2 = 1$. There are only small differences between counting rules for the dual system estimators whereas for the single system estimator the rule KH+SP+SI performs considerably worse than the other three.

Table 2-a and 2-b present the design effects (DEFTS) of the single system and dual system network estimators relative to the conventional estimator for two counting rules - KH+SP+SI+CH and KH+SP+CH for all the combinations of subsampling fractions f_1 and f_2 and for the two sample sizes considered.

The marked superiority of $\hat{N}_B$ and $\hat{N}_C$ over $\hat{N}_A$ and to a lesser extent the superiority of $\hat{N}_B$ over $\hat{N}_C$ for small values of f_1 is seen from the table. The superiority decreases for higher values of f_1 and of f_2. While $\hat{N}_B$ and $\hat{N}_C$ have considerably smaller root mean square errors than both single system estimators for all combinations of the subsampling fractions, $\hat{N}_A$ only achieves this superiority for high subsampling fractions.

For all three estimators the highest efficiencies are attained for $f_1 = 1$ and for $f_2 = 1$. It should be noted however that the comparison is with respect to total expected sample size, whereas costs of the evaluation survey may be higher than those of the main survey.

ACKNOWLEDGEMENTS

The authors are grateful to Janice Melvin for her patience and her efficient typing of the manuscript.

REFERENCES

Casady, Robert J., Gad Nathan, and Monroe Sirken (1983). "Dual System Network Estimation", Manuscript submitted for publication to the International Statistical Review.

Hanson, M. H., W. M. Hurwitz, and M. S. Bershad (1961). "Measurement Errors in Censuses and Surveys", Bulletin of the International Statistical Institute 38, part 2, Tokyo: International Statistical Institute.

Marks, Eli S., William Seltzer and Karol J. Krotki (1974). Population Growth Estimation, The Population Council, New York.

Nathan, Gad, Usiel O. Schmelz, and Jay Kevin (1979). Multiplicity Study of Marriages and Births in Israel, DHEW Publication No. (HRA) 77-1344. Public Health Service. Washington, D.C., U.S. Government Printing Office.

Royston, Patricia N., Monroe G. Sirken and Jean Bergston (1978). "Bias and Sampling Errors and Mortality Counts Based on Network Surveys", American Statistical Association Proceedings of the Social Statistics Section, pp. 471-475.

Sirken, Monroe G. (1979). "A Dual System Network Estimator", American Statistical Association Proceedings of the Section on Survey Research Methods, pp. 340-342.

TABLE 1: Relative standard errors, relative biases and relative root mean square errors (RRMSE) of the conventional estimator and of the single system and dual system network estimators (deaths 17+; $f_1 = f_2 = 1$).

Estimator

Sample Size	Counting rule	Single System - $\hat{N}_0$ or $\hat{N}_D$			Dual System - RRMSE		
		Rel. S.E.	Rel. Bias	RRMSE	$\hat{N}_A$	$\hat{N}_B$	$\hat{N}_C$
2,500	KH (conventional)	0.124	-0.087	0.151	-	-	-
	KH+SP+SI+CH	0.079	-0.077	0.111	0.097	0.096	0.095
	KH+SP+SI	0.092	-0.102	0.138	0.112	0.110	0.111
	KH+SP+CH	0.098	-0.070	0.121	0.114	0.111	0.113
	KH+SI+CH	0.080	-0.077	0.111	0.097	0.096	0.095
40,000	KH (conventional)	0.031	-0.087	0.092	-	-	-
	KH+SP+SI+CH	0.020	-0.077	0.080	0.024	0.024	0.024
	KH+SP+SI	0.023	-0.102	0.105	0.028	0.028	0.028
	KH+SP+CH	0.025	-0.070	0.075	0.029	0.028	0.028
	KH+SI+CH	0.020	-0.077	0.079	0.024	0.024	0.024

Table 2-a: DEFTS of Single System and of Dual System Network Estimators (Ratio of Root M.S.E. to that of the Conventional Estimator ($\hat{N}_0$) With Same Expected Sample Size)

Network Rule: KH+SP+SI+CH

Sample Size:		2500			40000		
Single System		0.7316			0.8628		
Dual System f_1	 f_2	$\hat{N}_A$	$\hat{N}_B$	$\hat{N}_C$	$\hat{N}_A$	$\hat{N}_B$	$\hat{N}_C$
0.10	0.10	1.8910	0.7585	0.8442	0.7704	0.3106	0.3457
	0.25	1.3896	0.7362	0.8265	0.5691	0.3015	0.3385
	0.50	1.1061	0.7318	0.8256	0.4530	0.2997	0.3381
	1.00	0.9123	0.7376	0.8346	0.3736	0.3021	0.3418
0.25	0.10	1.5133	0.6851	0.7208	0.6198	0.2806	0.2952
	0.25	1.2207	0.6719	0.6986	0.4999	0.2752	0.2861
	0.50	1.0158	0.6705	0.6953	0.4160	0.2746	0.2840
	1.00	0.8606	0.6772	0.7015	0.3525	0.2773	0.2873
0.50	0.10	1.1657	0.6517	0.6756	0.4774	0.2669	0.2767
	0.25	1.0158	0.6419	0.6514	0.4160	0.2629	0.2668
	0.50	0.8893	0.6417	0.6470	0.3642	0.2628	0.2650
	1.00	0.7810	0.6486	0.6520	0.3198	0.2656	0.2670
1.00	0.10	0.7974	0.6340	0.6538	0.3266	0.2596	0.2677
	0.25	0.7447	0.6260	0.6282	0.3050	0.2564	0.2573
	0.50	0.6912	0.6263	0.6231	0.2831	0.2565	0.2552
	1.00	0.6385	0.6333	0.6275	0.2615	0.2594	0.2570

Table 2-b: DEFTS of Single System and of Dual System Network Estimators (Ratio of Root M.S.E. to that of the Conventional Estimator ($\hat{N}_0$) With Same Expected Sample Size)

Network Rule: KH+SP+CH

Sample Size:		2500			40000		
Single System		0.7982			0.8073		
Dual System f_1	 f_2	$\hat{N}_A$	$\hat{N}_B$	$\hat{N}_C$	$\hat{N}_A$	$\hat{N}_B$	$\hat{N}_C$
0.10	0.10	2.3898	0.7918	0.9286	0.9787	0.3243	0.3803
	0.25	1.7201	0.7750	0.8394	0.7045	0.3174	0.3438
	0.50	1.2790	0.7702	0.8087	0.5238	0.3155	0.3316
	1.00	0.9264	0.7718	0.7987	0.3794	0.3161	0.3271
0.25	0.10	1.9071	0.7600	0.8976	0.7811	0.3113	0.3676
	0.25	1.5116	0.7482	0.8047	0.6191	0.3064	0.3296
	0.50	1.1862	0.7454	0.7733	0.4858	0.3053	0.3167
	1.00	0.8894	0.7477	0.7611	0.3643	0.3062	0.3117
0.50	0.10	1.5072	0.7482	0.8881	0.6173	0.3064	0.3637
	0.25	1.2881	0.7302	0.7937	0.5275	0.3023	0.3251
	0.50	1.0689	0.7360	0.7617	0.4378	0.3014	0.3119
	1.00	0.8368	0.7387	0.7491	0.3427	0.3025	0.3068
1.00	0.10	1.1421	0.7436	0.8853	0.4677	0.3045	0.3626
	0.25	1.0377	0.7344	0.7899	0.4250	0.3008	0.3235
	0.50	0.9127	0.7325	0.7575	0.3738	0.3000	0.3102
	1.00	0.7555	0.7353	0.7446	0.3094	0.3012	0.3049